Robot Emotional State through Bayesian Visuo-Auditory Perception

José Augusto Prado[1,] , Carlos Simplício[1,2], Jorge Dias[1]

[1] Instituto de Sistemas e Robotica ISR, FCT-UC, Universidade de Coimbra, Portugal
[2] Instituto Politécnico de Leiria, Portugal
{jaugusro,jorge}@isr.uc.pt, carlos.simplicio@ipleiria.pt

Abstract. In this paper we focus on auditory analysis as the sensory stimulus, and on vocalization synthesis as the output signal. Our scenario is to have one robot interacting with one human through vocalization channel. Notice that vocalization is far beyond speech; while speech analysis would give us what was said, vocalization analysis gives us how was said. A social robot shall be able to perform actions in different manners according to its emotional state. Thus we propose a novel Bayesian approach to determine the emotional state the robot shall assume according to how the interlocutor is talking to it. Results shows that the classification happens as expected converging to the correct decision after two iterations.

Keywords: Bayesian Approach, Auditory Perception, Robot Emotional State, Vocalization.

1 Introduction

In the context of human robot interaction, a core problem is how to reduce the estrangement between humans and machines. In order to do this, recently researchers are investigating how to endow the robots an emotional feedback. There has never been any doubt about the importance of emotions in human behavior, especially in human relationships. The past decade, however, has seen a great deal of progress in developing computational theories of emotion that can be applied to building robots and avatars that interact emotionally with humans. According to the main stream of such theories [1], emotions are much intertwined with other cognitive processing, both as antecedents (emotions affect cognition) and consequences (cognition affects emotions). In our scenario, a pre-defined story board exists, which the human and the robot shall follow, though removing the importance of what is said and focusing the experiments on the detection of emotion. In the simplest case, robot will mimic the detected emotion.

2 Contribution to Sustainability

Schroder et. al. [2] presented the SEMAINE API as a framework for enabling the creation of simple or complex emotion oriented systems. Their framework is rooted in the understanding that the use of standard formats is beneficial for interoperability and reuse of components. They show how system integration and reuse of components can work in practice. An implementation of a dialogue system was done using a 2D displayed avatar and speech interface. More work is needed in order to make the SEMAINE API fully suitable for a broad range of applications in the area of emotion-aware systems [2]. Classifying emotions in human dialogs was studied by Min [3] presenting a comparison between various acoustic feature sets and classification algorithms for classifying spoken utterances based on the emotional state of the speaker. Later, Wang [4] presented an emotion recognition system to classify human emotional state from audiovisual signals. The strategy was to extract prosodic, mel-frequency Cepstral coefficient, and formant frequency features to represent the audio characteristics of the emotional speech. A face detection scheme based on HSV color model was used to detect the face from the background. The facial expressions were represented by Gabor wavelet features. This proposed emotional recognition system was tested and had an overall recognition accuracy of 82.14% of true positives. Recently, Cowie [5] it was described a multi-cue, dynamic approach to detect emotion in video sequences. Recognition was performed via a recurrent neural network.

Our approach presents a novel probabilistic model for emotion classification based on vocalization analysis and Bayesian Networks applied for Human Robot Interaction. Our prototype robot can be seen in figure 1. Furthermore, we propose a model for integration of two modalities (visual and aural), more specifically facial expression analysis following Ekman [6] and vocalization analysis.

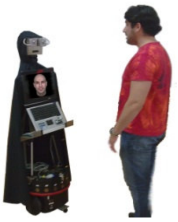

Fig. 1. Our prototype robot.

3 Emotional States

Spinozza [7], during the seventeenth century, proposed a definition of how human emotions behave. His work was recently continued and extended by Damasio [8] [9] who proposed an approach with the joint behavior of four groups of emotional states: three of them are related to the lost of some capability of communication. A fourth group, associated to success, was also considered. Each group contains the social emotion and the Emotional Competent Stimulus (ECS) for that emotion. Damasio did not define ECS for the neutral state. Here we propose the addition of a fifth group where the neutral state is. The four groups of emotional states proposed by Damasio [9], and plus the neutral state added by us, can be summarized as follow: *fear, anger, sad, happy and neutral*. According to Damasio [9], the emotional state can be influenced by what is happening with the individual's, and also to interlocutor emotional state. Taking this into account, our system is composed by analysis and synthesis (see figure 2). In the analysis part, we are determining what are the vocal expressions produced by the human. Later in the synthesis part, the emotional state is established and the reaction is synthesized. A combination with an input from human's emotional state, which is given by facial expression analysis following Ekman [6], is also proposed on the synthesis part.

4 Bayesian Modeling

The Bayesian approach is characterized by assigning probabilities to characterize the degree of belief associated with the state of the world. Bayesian approach defines how new information should be combined with prior beliefs and how information from several modalities shall be integrated. Bayesian decision theory defines how our beliefs should be combined with our objectives to make optimal decisions.

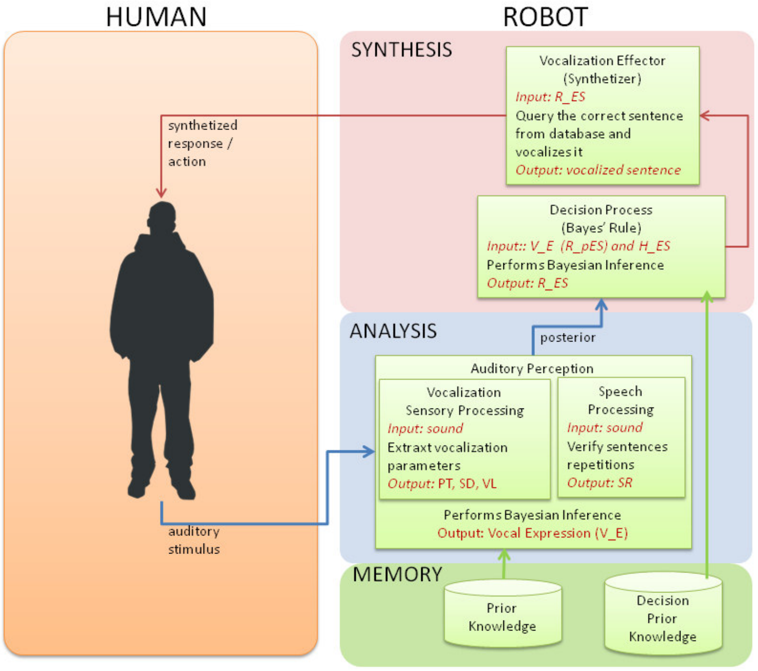
HUMAN
ROBOT
SYNTHESIS
Vocalization Effector (Synthetizer)
Input: R_ES
Query the correct sentence from database and vocalizes it
Output: vocalized sentence
synthetized response / action
Decision Process (Bayes' Rule)
Input:: V_E (R_pES) and H_ES
Performs Bayesian Inference
Output: R_ES
ANALYSIS
posterior
Auditory Perception
Vocalization Sensory Processing
Input: sound
Extraxt vocalization parameters
Output: PT, SD, VL
Speech Processing
Input: sound
Verify sentences repetitions
Output: SR
Performs Bayesian Inference
Output: Vocal Expression (V_E)
auditory stimulus
MEMORY
Prior Knowledge
Decision Prior Knowledge

Fig. 2. Analysis and synthesis' system schema.

Figure 2 shows the system implementation modules. Auditory perception feature extraction and the Bayesian model are described on section 4.1. Synthesis will be briefly presented in section 4.2.

4.1 Human Vocalization Analysis

In our model of auditory perception, the vocalization analysis classifies a vocal expression. The robot needs to be capable of classifying among the possible vocal expressions, which are in the same scope as the facial expressions as defined by Ekman [6]: {*anger, fear, happy, sad, neutral*}.

4.1.1 Vocalization Analysis

All waves are effectively combinations of sinusoids. The Fourier transform takes a waveform and turns it into a function describing which sinusoids are present in the waveform. So, as one can see in figure 3, a digitalized sound wave comes with positive and negative values. However, it is only in the frequency of oscillation of the signal that sound can exist. Obviously, a single sample cannot represent any oscillation.

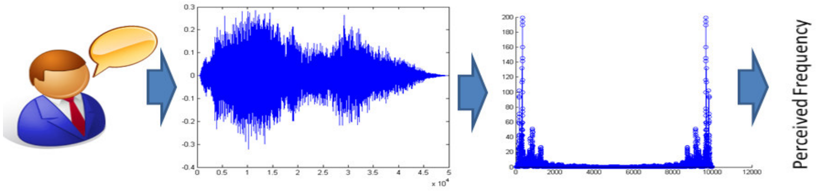

Fig. 3. Sample recorded speech waveform of an utterance, sad. The x axis is the number of samples while the y axis is the amplitude in dB. A sampling frequency of 16000Hz was used; it has 49679 samples and 3.1049 seconds of duration. At right a FFT of the interval from 1 to 2 seconds; after applying the correlation method presented by Sondhi [10] it is possible to get the perceived frequency.

In order to classify emotions from a waveform, first it is necessary to extract features from it. Here we define which features we are going to extract and also the Bayesian network to structure the relationship among them (figure 4).

There are several methods to extract pitch [10][14][15][16]: zero-crossing, autocorrelation function, cepstrum, average magnitude differential function, comb transformation, FIR filter method of periodic prediction. Lopes [11] [12] [13] extensively studied vocal tract's length normalization using pitch's features for it.

4.1.1 Auditory Perception Bayesian network

To classify the vocal expressions performed by the human, a Bayesian network was developed. The structure of this network of two levels is illustrated in figure 4. A vocal expression will be classified after a sentence finish. In other words, for the Bayesian network, the time 1 is just after sentence 1 is completed, time 2 is just after sentence 2 is completed; and so one. This is independent of each sentence's (real time) duration. Now it is implemented over Matlab and off line, so we don't need a state switching since the sentences are recorded separated. In future, to determine the state switching, we expect to use a silence detector as presented by Hoelper [17]. If the silence period is bigger than a threshold (3 seconds), this event may trigger the state switching.

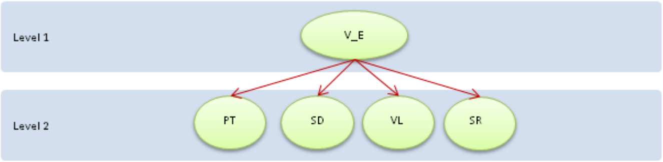

Fig. 4. Bayesian network Auditory Perception.

In the Bayesian network's first level there is only one node. The global classification result obtained is provided by the belief variable associated with this node: V_E$\in$ {*angry, fear, happy, sad, neutral*}; where the variable name stands from Vocal Expression. Considering the structure of the Bayesian network, the variables in the second level have as parent this one in the first level: V_E.

In the second level there are four belief variables,

- PT$\in$ {*short, normal, long*} is variable which a belief is related with Pitch. Pitch represents the perceived fundamental frequency of a sound. We are using the pitch extraction by autocorrelation method proposed by Sondhi [10].

The voice pitch changes significantly along a sentence and an important part of our voices are un-pitched, however, since the conversation follows a pre-defined story board, the mean pitch of the sentence will help to distinguish the emotional state that was there when this very sentence was spoken.

- SD$\in$ {*short, normal, long*} is variable which a belief is related with Sentence Duration. Since we know the sampling frequency (*sfreq*) of the acquired sound, and we also know the beginning and the end of each sentence, consequently the number of samples (*nsam*) then it is trivial to determine the duration in seconds by $SD = nsam/sfreq$. This variable contributes to the classification. By example, when a person speaks the same sentence with a happy emotion it usually speaks faster than with a sad emotion. For some emotional states the duration might be exactly the same, but then the other variables will contribute for the disambiguation.
- VL$\in$ {*low, medium, high*} is a belief variable which stands for Volume Level. This variable is actually the energy y of the signal, which for a continuous-time signal *x(t)* is given by $VL = \int |x(t)^2|dt$.
- SR$\in$ {*zero, one, two, three_or_more*} is the belief variable which stands for Sentence Repetition. It is associated with the number of sentences repetitions that the interlocutor may perform. The value of this is given by the comparison of the previous three variables along four previous times.

The following equations illustrate the joint distribution associated to the Bayesian Vocal Expressions Classifier:

$$P(V\_E, PT, SD, VL, SR) = \\ P(PT, SD, VL, SR|V\_E).P(V\_E) = \\ P(PT|V\_E).P(SD|V\_E).P(VL|V\_E).P(SR|V\_E).P(V\_E) \qquad \textbf{(1)}$$

The last equality can be done only if it is assumed that belief variables PT, SD, VL and SR are independent.

From the joint distribution, the posterior can be obtained by the application of the Bayes' Formula as follow:

$$P(V\_E|PT,SD,VL,SR) = \frac{P(PT|V\_E).P(SD|V\_E).P(VL|V\_E).P(SR|V\_E)}{P(PT,SD,VL,SR)} \quad \textbf{(2)}$$

4.2 Robot Vocalization Synthesis

According to figure 2, the Decision Process receives as input H_ES (the Human Emotional State), which is given by the external facial expression classifier and V_E (the inferred Vocal Expression). It will take a decision according to these inputs and to the MEMORY contents.

It is assumed that the robot will initially internalize the vocal expression of the person, thus V_E implies on a robot pre-emotional state (R_pES). The Decision Process will then combines R_pES with the H_ES in order to determine R_ES (the final emotional state that the robot will assumes).

This fusion is proposed in order to determine the robotic emotional state in similar way of that established by Damasio [8] for human beings. As a consequence of the Bayesian framework, the prior knowledge will determine the balance of the fusion. This brings up to which side the decision will fall: in one side the robot is more confident in its own emotional state (from vocalization analysis); in another extreme the robot is less confident and uses the human emotional state (from facial expression analysis).

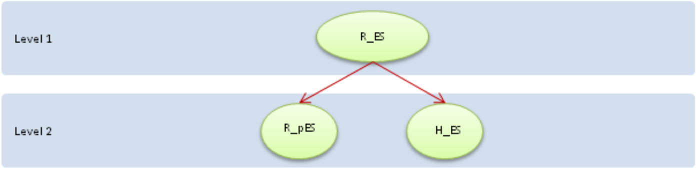

Fig. 5. Synthesis Bayesian Network.

5 Results

Results of learned likelihoods for Auditory Perception.

After teaching the system, by pointing which is the correct Vocal Expression for a given input, is obtained a histogram table exactly as shown at Table 1. This histogram is the likelihood knowledge for the Bayesian algorithm to perform the inferences later: the joint distribution (see eq. 1) contains all the information needed.

Table. 1. Learning for Analysis of Vocal Expressions.

V_E	PT			SD			VL			SR			
	low	med	high	short	norm	long	low	med	high	0	1	2	3+
Ang	0.8	0.10	0.10	0.10	0.10	0.80	0.10	0.10	0.80	0.97	0.01	0.01	0.01
Neu	0.1	0.80	0.10	0.80	0.10	0.10	0.10	0.80	0.10	0.97	0.01	0.01	0.01
Sad	0.1	0.10	0.80	0.80	0.10	0.10	0.80	0.10	0.10	0.97	0.01	0.01	0.01
Hap	0.1	0.80	0.10	0.25	0.50	0.25	0.10	0.80	0.10	0.97	0.01	0.01	0.01
Fear	0.34	0.33	0.33	0.33	0.34	0.33	0.33	0.33	0.34	0.01	0.23	0.33	0.43

Results of Bayesian Network for Auditory Perception.

The robot is able to infer over the likelihoods (see eq. 2) when interacting to the user. The expected results for the ANALYSIS part are correct classifications of vocal expressions according to what is expected. Convergence is also expected to appear among the time, since both are Dynamic Bayesian Networks. Figure 6 shows results of the Bayesian inference during five iterations with the following constant evidences:

Pitch=long, SentenceDuration=short,
VolumeLevel=low, SentenceRepetition=zero.

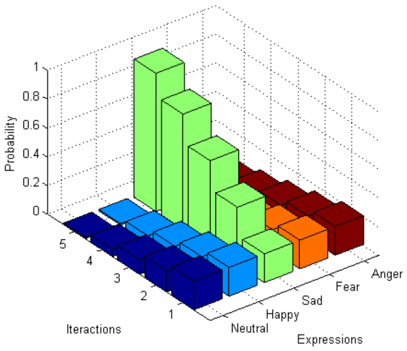

Fig. 6. Results for Classification of a Vocal Expressions (Sad) - The convergence happens after the second iteration.

6 Conclusions and Future Work

This work presented a novel approach to determine robot emotional state through vocal expressions, according to philosophic references on how humans do it for themselves. The results show that correct classifications are done: the inferred emotional state is correct during an interaction between a robot and a human. This approach turns interaction more inclusive and reduces the estrangement between humans and machines. Our approach endows the robot to say the same sentence with different characteristics, according to its emotional state. We just presented one example of an utterance in sad; however, we are preparing a dataset with different sentences uttered in all the five emotional states here considered.

The current implementation of our Bayesian network is with limited values and it shows a proof of concept; however, we expect to experiment it with a larger scope of possibilities for each variable.

As the proposed model is simple, it is advantageous in particular contexts; specially multimodal fusion; where a quick (less complex) form of predicting an emotional state is better than a large model of human emotional processing.

References

1. J. Gratch, S. Marsella, and P. Petta, “Modeling the cognitive antecedents and consequents of emotion”, Cognitive systems, vol. 10(1), pp. 1–5, 2008.
2. M. Schroder, “The semaine api: Towards a standards-based framework for building emotion oriented systems”, Advances in human-computer interaction, article ID 319406, 21 pages. doi:10.1155/2010/319406, vol. 2010, 2010.
3. C. M. Lee, S. S. Narayanan, and R. Pieraccini. “Classifying emotions in human-machine spoken dialogs”,ICME, 2002.
4. Y. Wang and L. Guan. “Recognizing human emotion from audiovisual information”, ICASSP IEEE, 2005.
5. R. Cowie, E. Douglas-Cowie, K. Karpouszis, G. Caridakis, M.Wallace, and S. Kollias. “Recognition of emotional states in natural human-computer interaction”, School of Psychology, Queen's University, 2007.
6. P. Ekman and E. L. Rosenberg. “What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS)”, Oxford University press. Second expanded edition, 2004.
7. Spinoza, Ethics, 1677.
8. A. Damasio , “Looking for Spinoza”, Harcourt, Inc ISBN 978-0-15-100557-4, 2003.
9. A. Damasio , “The feeling of what happens”, Harcourt, Ed. Harcourt, Inc - ISBN 978-0-15-601075-7, 2000.
10. M. M. Sondhi, "New methods of pitch extraction',' IEEE Trans. on audio and electroacoustics, vol. 16, no. 2, pp. 262 266, 1968.
11. C. Lopes and F. . "VTLN through frequency warping based on pitch", *Revista da Sociedade Brasileira de Telecomunicações*, Vol. 18, No. 1, pp. 86 - 95, June, 2003.
12. C. Lopes and F. Perdigão. "VTLN through frequency warping based on pitch", Proc*IEEE International Telecommunications Symp.*, Natal, Brazil, September, 2002.
13. C. Lopes and F. Perdigão. "On the use of pitch to perform speaker normalization", *Proc. International Conf. on Telecommunications, Electronics and Control*, Santiago de Cuba, Cuba, July, 2002.
14. S. Zieliński, Papers from work on comb transformation method of pitch detection ("Description of assumptions of comb transformation", "Comb transformation - implementation and comparison with another pitch detection methods"), Technical University of Gdansk, 1997.
15. P. R. Cook, D. Morill and J.O. Smith, "An automatic pitch detection and MIDI control system for brass instruments", In Proc. of special session on automatic pitch detection, 1992.
16. W. Hess, "Pitch determination of speech signals," Berlin: Springer Verlag, 1983.
17. C. Hoelper, A. Frankort, and C. Erdmann, “Voiced/unvoiced/silence classification for offline speech coding”, in Proceedings of international student conference on electrical engineering (Prague), 2003.