Synthesizing Emotions (from Machines to Humans) *Seminar Paper* *

Adel Rizaev
University of Fribourg
Departement of Computer Science
DIVA Group
CH-1700 Fribourg, Switzerland
adel.rizaev@unifr.ch

ABSTRACT

The paper represents a desk research, which is done for the seminar *Emotion Recognition* in the University of Fribourg, Switzerland. It is also the first steps of the author in the "world" of emotion synthesis. In the paper the concepts like: modalities and the cues from emotion synthesis point of view, rules based and data-driven implementations of emotion synthesis were reviewed. The actual achievements are also proposed.

Categories and Subject Descriptors

I.6 [**Simulation and Modeling**]: Miscellaneous, General; J.4 [**Computer Applications**]: Social and Behavioral Sciences : Psychology; H.1.2 [**Information Systems**]: User / Machine Systems: Human Information Processing

General Terms

Human Factors, Experimentation

1. INTRODUCTION

In recent years the interest for the methods of embedding emotions in the machines has been increased. We can consider different applications, which implies use of emotions in machines. For example, animated characters in e-learning systems, avatars in virtual environments or computer games, automatic speech services, animated agents, which could interact with a user in a natural way, using gesture, facial and speech expression. Using emotions in machines can bring the interaction between human and computer on a new, more natural way [4, pp. 1-15]. This new interaction would be more efficient, robust and intuitive. Moreover, emotion synthesis as a fundamental research is not less interesting. Because, during the creation or synthesizing them, we can better and more deeply understand the structure of a human in terms of non-verbal communication. In [1] R. Picard writes, that psychologists of the nineteenth century "argued for the importance of understanding emotion in human behavior." But in the last century the constructed model of human underestimates the role of emotions, that is, obviously, describing the nature only in a part: superficially and crudely. Then the author hopes, that in new century we will use computing not only for producing a high quality images, audio, video, etc., but computers would also be able to understand emotional expressions and communicate using emotional expressions, since they have direct relationship to human's soul.

Emotion synthesis is a very complicated task for researchers, because in order to solve it they are have to meet with many practical problems. The solutions for that practical problems ask for compromises between many parameters of the emotion synthesis system. The first practical problem is associated with the choice of the appropriate theory of emotions. Currently, a great number of theoretical models of emotions are available, each of them has so much advantages as disadvantages (discussing the emotional theories is out of the scope of that paper). Accordingly, there are many possible ways, how the parameters of emotions could be registered, interpreted and implemented in machines. Another problem is related to the defining of the emotional cues, because expressing of emotions differs from culture to culture, from language to language. Many emotional cues are individual. Technology is another important factor making an influence to the performance of the system, which requires much computational power. And, finally, perception of expressed emotions is very individual: some emotional cues may appear in expression of distinct emotions for different peoples.

From one point of view, an emotion, expressed in some interaction, has objective characteristics, which could be measured, like pitch of the voice, heart rate, skin conductivity, some muscle activity of the face or body and so on. After measuring of parameters some correlations between measured or observed signals and emotional state of the object could be found, which could be used in modeling of emotional behavior. Another idea is to collect labeled emotional data, select relevant signals for expression of emotion, take the values from that signals and assign them in order to obtain desired emotional behavior. In that case we say,

*The Paper is a delivrabled of the Msc Research Seminar *Emotion Recognition* in 2011, DIVA Group University of Fribourg, supervised by Dr. D. Lalanne, Dr. F. Ringeval

that emotional synthesis system "imitates" or "duplicates" the emotional behavior.

This paper is organized as following: first, we consider, which modalities and cues are relevant for synthesizing emotions. Then we will discuss the two dominant approaches such as, rule-based and data-driven synthesis and make a conclusion.

2. MODALITIES

Let's try to make clear, where a source of emotions is, i.e where a birth to emotions is given. By the nature, emotions are the product of interaction between human or animal with some objects from "material world", "not-material world" or composition of both. There are different behavioral cues and signals, which could tell us about human's emotional state. Most of that signals and cues can be read from different modalities: via acoustic, visual and tactual expressions. Then affective states of a human could be recognized from different kind of visible signals (facial expressions, body gestures, head movements), speech (pitch, energy, frequency and duration), physiological signals (heart rate, skin conductivity), brain and scalp signals, and thermal infrared sensors [3].

Usually, speech and visual signals are widely accepted for synthesizing emotions in machines. In [5] the authors uses speech in emotion synthesis. Another work [6] is dedicated to the generation of facial expressions of emotions. Niewiadomski et al. used a model, which enables to synthesize multimodal expressions of emotions in [14]. The model generates nonverbal behaviors using different modalities: facial expressions, head and gaze movements, gestures, torso movements and posture.

3. RULES BASED SYSTEMS

In 1988 Ortony et al. proposed a cognitive model [7]. This model, also known as OCC (Ortony, Clore, Collins, 1988) model, represents the emotions not by sets of basics emotions, but it groups all emotions according to "positive" or "negative" experiences, which appears from different evaluations of situations. Emotions are: *"valenced reactions to events, agents, or objects, with their particular nature being determined by the way in which the eliciting situation is constructed"* [7, p.13]. Ortony et al. also defined 22 emotion types using OCC model. The model was not fully implemented on any system, but it famous with *framing rules, which are easy to implement in computers*[8]. R. Picard in her book "Affective Computing" wrote, that OCC model *becomes the default model for synthesizing emotions in computers.* Let's consider the generating of a joy emotion using the method, introduced in [7]:

$$IF desire(p,e,t) > 0$$

$$THEN$$

$$SET joy_{potential}(p,e,t) = f_j[desire(p,e,t), I_g(p,e,t)];$$

Here the f_j is the function specific to joy, *desire(p,e,t)* is the value of a function, that assigns desirability level to event e by a person p at the time t. The function $I_g(p,e,t)$ returns the value of the combined effects of the global intensity variables.

So, if desired level of event is positive, then the algotithm sets the $joy_{potential}(p,e,t)$. The second part of the algorithm updates an intensity level and changes the state of joy feeling by the person p:

$$IF joy_{potential}(p,e,t) > joy_{threshold}(p,t)$$

$$THEN$$

$$SET joy_{intensity}(p,e,t) = joy_{potential}(p,e,t) - joy_{threshold}(p,t)$$

$$ELSE\ SET joy_{intensity}(p,e,t) = 0$$

The work introduced by E. Zovato et al. uses a rule based approach to simulate three basic emotional styles [9]. Doing that, they recorded some sentences about 10 words. Each sentence has been recorded in four emotional styles: neutral, angry, happy and sad. After recording of these sentences they evaluated by volunteers and verified if the corpus could be used to extract style dependent rules. Then they derived a simple rules to obtain desired pitch profiles together with duration and energy constraints. For this task, the syllable was chosen as the reference acoustic unit. For each syllable in the database they calculated prosodic parameters such as minimum F0, maximum F0, F0 mean, F0 range and RMS energy. Then syllabic segmentation and labeling has been applied. The rules manipulate with waveforms and change above listed acoustic parameters of the Text-to-speech audio output in respect to prosodic neutral situation to obtain desired emotional style.

In the emotional speech synthesis system, considered by Butut et al. in [10], time domain prosody modifications have been utilized. For modification of the input speech TD-PSOLA algorithm with different scaling factors for duration, energy, F0 median and F0 range has been applied. In order to evaluate synthesized emotional speech with selected scaling factors automatic emotion recognition module was implemented. Another module of the system sequentially tuned scaling factors and applied them to the input utterances in order to obtain better recognition rates by the recognizer. The first five of successful modifications for each emotional state were selected and applied to the input utterances. Then resynthesized utterances were presented to human raters for the listening test. The experimental results show that the parameters automatically selected by the system can be successfully used in resynthesis of an input neutral speech as an angry speech (recognition rate by human raters is about 97 %).

Catherine Pelachaud proposed rule-based approach to synthesize facial emotions in her PhD. thesis [11]. In common words, the idea was to make a clusterization of the phonemes (the smallest units of speech) into visemes (smallest units of visual information, i.e. visemes represent phonemes in the visual domain). Visemes are classified with different rank of deformation. She used results, which are empirically observed by the number of psychologists. To implement these cues several algorithms have been mentioned. The main algorithm sequentially calls sub-algorithms, that are responsible for different parts of the face. The sub-algorithms are: compute lips shape, compute brows action on, compute blink, compute head motion, compute eyes movements.

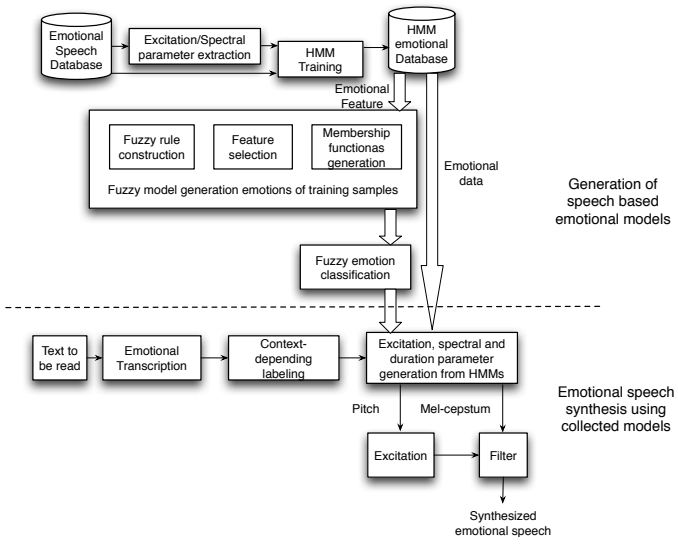

Figure 1: Architecture of HMM-based Fuzzy Emotion Synthesis Model.

Each sub-algorithm implements the corresponding rules. One of the rules for a sad person, for instance, is the following: *"eyebrow movements occur only on few pitch accents with low intensity, the head moves slowly, the speech rate is slow with a lot of long hesitation pauses"*. The numerical parameters are delivered from the practical measurements. Finally, the computing of the facial expression for each phoneme is done. The last step of the main algorithm is post-processing over the animation to restrict the abrupt movements in some facial regions. The main advantages of the rule-based method is flexibility, i.e. we could change the parameters of the synthesis in wide range. Another advantage is, that no database is required. But for believable emotion synthesis we have to find relative complex functions which depending on large (much) number of parameters.

4. DATA DRIVEN SYSTEMS

The principal difference between rule-based and data-driven solutions is, that in data-driven approach the expression of emotion depends on the collected data, while in the rule-based systems the emotional behavior is defined by the functions. Therefore, for successful implementation of the data-driven emotion synthesis huge database is needed. According to the Plutchik's theory [13] humans do not feel only one of the basic emotions, but they are affected by the complex emotional states at the same time. These emotional states vary in terms of intensity. Representing this statement using rules-based approach could be very difficult task, because, as we said before, we have to consider a complex functions with huge amount of parameters. Therefore using of the data-driven solution seems to be better.

The paper [5], proposed by Yuqiang Qin et al. describes fuzzy affective model for emotional speech synthesis. They suggest to use fuzzy emotion hypercubes to classify the emotional speech. Each axis of the hypercube corresponds to one basic emotion. Then some emotional state E_j could be represented as a linear combination of the basis vectors in hypercube, with corresponding basis emotions. Suppose, we have 3 basis emotions: sadness, happiness and anger. Then the emotional state E_1 with coordinates $E_1 = (0.2, 1.0, 0.9)$ corresponds to

$$FE_1 = \{(happiness, 0.2), (anger, 1.0), (sadness, 0.9)\},$$

which will represent the emotional state derived from anger and sadness [5, p.2]. The point $(0, 0, 0)$ is affective neutral state. We could also devide our cube with 3 basic emotions into 8 sub-cubes, and classify the members of that sub-cube as emotional states with sum of emotions, for example "Anger+Sadness". This approach enables the representation and synthesis of not only fixed set of basic emotions, but it makes possible the handling of derived emotions [5]. They considered three main phases according the speech synthesis schema: training using Hidden Markov Models (HMM), classification and emotional speech generation phases (*Fig. 1*). For the training phase the phonetic transcription of the emotional speech corpus is needed, therefore the precise boundaries in the waveform of each phoneme from the corpus have been defined. From the waveform of each phoneme the mel cepstrum (with first and second derivatives) and the pitch (with its first and second derivatives) has been extracted. In fact, in the spoken language the pronunciation of a particular phoneme varies depending upon the previous and the following phoneme. Therefore, phonetic transcription has been extended with context-dependent labeling. After initial preparation of the data they have trained the HMMs by learning the acoustic features according to the context-dependent labels. For generating of emotional speech after training and classification phases, desired emotional transcription, the phonetic transcription and the context-dependent dependent labeling of the text to be read must be done. Then phones duration and acoustical parameters of the synthesized speech will be taken from generated models.

The MSE (multimodal sequential expressions) - language proposed by Niewiadomski et al. in [14] might be used in multimodal emotion synthesis systems, which could be not classified as a pure rule-based or pure data-driven systems. From one point of view each emotional state could be defined with behavior set, consisting of signals of different modalities. The signals of different modalities, like smile, gesture, head and torso movements could be labeled and their parameters could be collected in the database. Each behavior set contains possible signals in the expression of one emotion. These behavior sets are developed according to the results, obtained by different psychology studies. For example, in emotional expression of embarrassment, the behavior set will be the following: *"two head movements (head down and head left), three gaze directions (look down, look right, look left), three facial expressions (smile, tensed smile and neutral expression), open flat hand on mouth gesture and bow torso movement"* [14, p.4]. Additionally, for each signal in the behavior set they define five main parameters: probabilities of occurrence of the signal at the beginning and at the end of a multimodal emotion expression, minimum and maximum possible duration of the signal, possibility if the signal might be repeated. From another side, each emotional state could be characterized with *constraint set*, another object defined in the MSE-language. Constraint set is a set of rules, which describes reliable configurations of signals. For example, *"some signal s can not occur at the end of the some*

emotional expression E" or "signals s and k occur simultaneously". At each step of synthesis the signal-candidate will be selected from the behavior set, then its parameters will be checked by the constraint-set for consistency with the previous steps of synthesis. Thus, they propose the idea, that each emotional state can be characterized by a behavior set, which describes the signals and orders them using the rules.

5. CONCLUSION

Despite of that two systems have the same aim - generating natural emotions - they do it using different paradigms. Rule-based systems synthesize the emotional behavior by deriving the rules, which are derived from psychology studies. These studies are based on the observations of affective behavior of humans. Rule based systems could be too straightforward and insensitive to the content, that has to be synthesized. In opposite to that stay data-driven systems. During synthesis of an emotion, the data-driven system turns to its emotional database for retrieving the corresponding to requested emotional state relevant and preferable (the most probable) parameters of the generating cue. Humans, in some sense, are doing that in a same way: expressing an emotion they appeal to their collected life experience: to the books, which the human had read, to the watched movies, to the memorized movements and so on. Data-driven systems could be difficult to generalize.

To avoid drawbacks, associated with creating complex functions of many variables in the rule-based systems and dependency from the database in the data-driven systems, could be mixed-mode systems utilized [2], [15].

A few words about perception of the synthesized emotions. In the emotions synthesized in speech the naturalness of the sound plays an important role, in the non-verbal emotional behavior synthesis naturalness and mulitimodality ([14]) makes an influence to the perception of synthesized emotional states. Of course, context (meaning of text or situation) is also very important to the perception of synthesized emotional states [16].

Very interesting concept of "emotional feedback" proposed R. Picard in [17]. May be machines don't need to have an emotions in the same sense as humans have, but machines could interact with users, giving them emotional feedback with their, "computer's", emotional cues. These machines could detect emotional behavior and expressions of the users, analyze, what the system doing at the time, and interact with the users by giving them some emotional feedback (which is also can be called a "synthesized emotion"), which will be perceived by the humans as computer have an emotions. May be the researches don't need to concentrate only on "human-like" or "animal-like" emotion synthesis, but try to find some new possibilities to express an emotion. In that case we can think about emotion synthesis as a creative process (create something, which does not exist in the nature) and not imitation or modeling.

6. REFERENCES

1. Rosalind W. Picard (2000), "Synthetic Emotion," IEEE Computer Graphics and Applications, Volume 20, No. 1, pp. 52-53, January/February 2000.

2. Rolf Carlson, Björn Granström (2005): "Data-driven multimodal synthesis".

3. Hatice Gunes, Massimo Piccardi, Maja Pantic (2008), "From the Lab to the real world: affect recognition using multiple cues and modalities. In: Affective computing: focus on emotion expression, synthesis, and recognition". In:Tech Education and Publishing, Vienna, Austria, pp. 185-218. ISBN 9783902613233

4. Toyoaki Nishida, Lakhmi Jain, Colette Faucher (eds.) "Modelling Machine Emotions for Realizing Intelligence: Foundations and Applications", Smart Innovation, Systems and Technologies Series, Springer, 2010.

5. Yuqiang Qin, Xueying Zhang, Hui Ying (2010) "A HMM-based fuzzy affective model for emotional speech synthesis".

6. Joshua M. Susskind, Geoffrey E. Hinton, Javier R. Movellan, Adam K. Anderson (2008) "Generating Facial Expressions with Deep Belief Nets".

7. Andrew Ortony, Gerald L. Clore, Allan Collins (1988) "The Cognitive structure of emotions".

8. Rosalind W. Pickard (1997) "Affectve computing", The MIT Press, MA, USA.

9. Enrico Zovato, Alberto Pacchiotti, Silvia Quazza, Stefano Sandri (2004). "Towards emotional speech synthesis: a rule based approach. Proc". 5th ISCA Speech Synthesis Workshop (pp. 219-220). Pittsburgh, PA, USA.

10. Murtaza Bulut, Sungbok Lee, Shrikanth Narayanan (2008), "Recognition for synthesis: automatic parameter selection for resynthesis of emotional speech from neutral speech".

11. Catherine Pelachaud (1992), "Communication and coarticulation in facial animation", doctoral dissertation.

12. Ya Li, Shifeng Pan, Jianhua Tao (2010) "HMM-based speech synthesis with a flexible mandarin stress adaptation model".

13. R. Plutchik (1990), "The Emotions", University Press of America, Inc., revised edition.

14. Niewiadomski R, Hyniewska S. J, Pelachaud C. (2011), "Constraint-Based Model for Synthesis of Multimodal Sequential Expressions of Emotions".

15. Susan R. Herz (2002): "Integration of Rule-Based Formant Synthesis and Waveform Concatenation: a Hybrid Approach to Text-to-Speech Synthesis".

16. Marc Schröder (2001): "Emotional Speech Synthesis: A Review".

17. Rosalind W. Picard (2001): "What does it mean for a Computer to "have" Emotions?", Chapter in "Emotions in Humans and Artifacts," Ed. By R. Trappl, P. Petta and S. Payr., MIT Press, 2003.