Extending Visual Speech Synthesis for Polish with basic emotion model

Jakub Bloch

Division of Teletransmission Systems, Institute of Telecommunications, The Faculty of Electronics and Information Technology, Warsaw Universisty of Technology, e-mail: jakub.bloch@gmail.com

Expressing emotions is a very important feature of Visual Speech Synthesis systems. In 1972 the first "basic emotions" list was introduced, by Paul Ekman. Since then few different classifications were published. Most famous "basic emotion" models are briefly described in this paper. In previous publication new Visual Speech Synthesis system for Polish was presented. The system was based on Xface toolkit and "Karol" face model. The aim of this paper is to add "basic emotion" model, according to Paul Ekman's classification, into "Karol" face model. To achieve this goal new emotional keyframes were proposed. This new functionality of "Karol" face model, allows to generate talking human face animations, which express emotions. The subjective test of new functionality are also included in the paper. The results showed that more information about speakers emotions is read from human face expression than form human speech signal. People can more easily recognize speakers emotion when they see his face expression.

Key words and phrases: Visual Speech Synthesis, emotion, Xface, Ekman

Introduction

In previous papers [HYPERLINK \l „Art101" 1] a new Visual Speech Synthesis (VSS) system for Polish was presented. The aim of this system was to generate human talking face animations according to phoneme transcription of speech signal. The FaceGen 2] application helped to create a new 3D face model called "Karol".. The face mesh itself does not allow any control of the movement. Therefore muscles, on a mesh for the MPEG-4 FAP [HYPERLINK \l „Pan02" 3] (Face Animation Points) movements, were implemented. "Karol" face supports also another type of animations – keyframe based animations. The created VSS system uses Xface framework 4] as a morphing engine. In this system visemes were used as the keyframes of the animations. The viseme is a face expression, which corresponds to specific phonemes (it is a graphic representation of phoneme). Since the phonemes and visemes inventories are language-specific, new set of visemes for Polish was proposed. Moreover, to achieve the better quality of animations, this set of face expressions includes also "half visemes" – less expressive visemes. The tests of the VSS system showed quite high quality of generated animation. In 5 grade MOS scale (Mean Opinion Score) naturalness gets 3,9 and synchronization - 4,2. . However, the emotions of the speaker were not expressed. "Karol" face model had the same expression during all animations.

The emotion and the emotion's classifications

What is an emotion? The answer is not so simple. There is more than 90 definitions of emotion [HYPERLINK \l „Rob" 5]. One of them describes the emotion as a feeling which corresponds to positive or negative affection. James Russell defined the emotion as an assessment process, which produces feelings 6]. On the other hand, John Broadus Watson defined the emotion as a visceral pattern of reactions [HYPERLINK \l „Hen02" 7]. So many different definitions of emotion make their classification very confusing. In 1972 Paul Ekman, based on gestures and face expressions, discovered that there is a set of emotions, which are universal between cultures and human races. This group of emotion is classified as basics. His research was focused on cultures from U.S., Japan, Chile, Argentina and Brasil. According to Ekman's classification, basic emotions are: anger, disgust, fear, happiness, sadness and surprise 8].

The other scientists Robert Plutchik classified eight emotions as the basic ones. His classification consists of joy, trust, fear, surprise, sadness, disgust, anger, anticipation [HYPERLINK \l „Rob" 5]. Those emotions creates Plutchik's "wheel of emotions". In accordance to his work, people are born with basic emotions that help survive and adapt to the world around. The complex "feelings" come

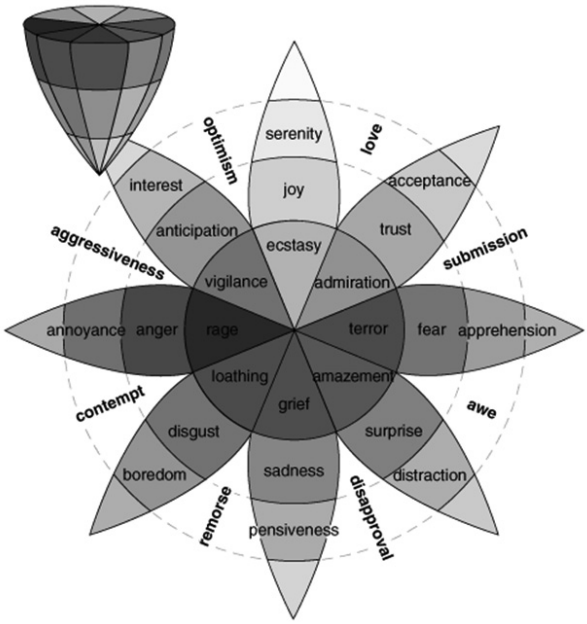

Fig. 1. The wheel of emotions [5]

into being as a mix of basic emotions. The Fig. 1 presents the "wheel of emotions". The basic emotions have their own color. The vertical dimension on the "wheel" corresponds to the intensity of sensation. The similarity of colors presents the similarity of emotions.

Applying emotions to "Karol" face model

The Xface framework allows for morphing two classes of keyframes. One class corresponds to visemes second to face expressions – emotions. Paul Ekman in [9] gave detailed description of face expressions, which matches to specific basic emotions. Fig. 2 shows the difference in human's eyebrows between "surprise" emotion and "neutral" emotion. Ekman's basic emotion classification was chosen to be implemented in "Karol" face model. In order to apply the emotions into the face model, the new keyframes were created in FaceGen application. The frames presents the face expressions, according to Ekman's work. The Fig. 3. illustrates one of the created emotional keyframes in comparison with "fear" face expression. The frames were exported as VRML (Virtual Reality Modeling Language) files and applied to face model.

Test of "Karol's" emotions

As a test of the new features of "Karol" face model, a two-part survey was proposed. The group of 30 people participated in it. The aim of this research was to observe the influence of face expressions on the emotion recognition process. Therefore ten short sentences (out of context), were taken from the J. K. Rowling's book – "Harry Potter and the Philosopher's Stone". The text was read by famous Polish actor – Piotr Fronczewski. The experiment was divided into two parts. The first assignment was to choose the most suitable emotion based only on human speech signal. In the second part, the same group of people were asked to point the most suitable emotion according to emotional face animation combined with speech signal. The given set of emotions included the Ekman's basic emotions and the neutral emotion (means there is no emotion). The emotion pointed by the most of interviewees, in first part of the research, is thought to be the correct emotion for the sentence. This emotion was used to generate the "emotional" animation for second part of the experiment. The second survey took place two weeks after the first one. In this part of examination the same group of people assessed the speakers emotion based on human-talking-face animation. The animations were generated by VSS system

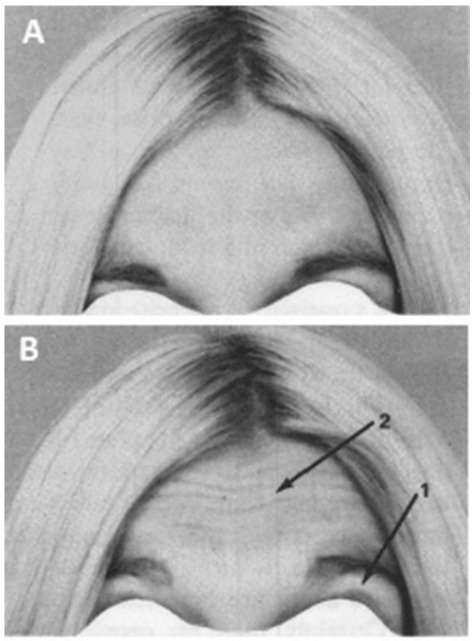

Fig. 2. The difference between humans brows during neutral (A) and surprise (B) emotion [9].

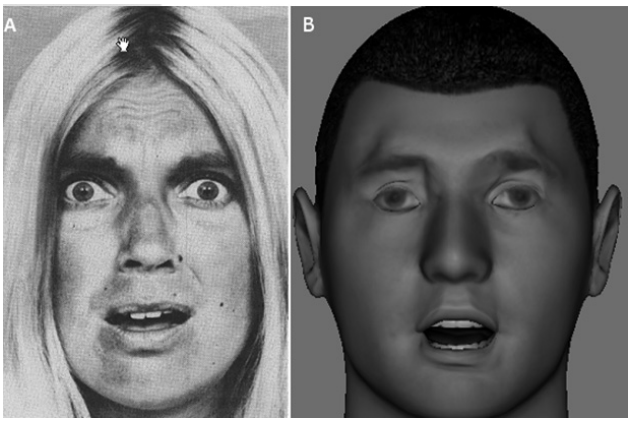

Fig. 3. (A) Ekman's face expression of fear emotion [9]; (B) created "Karol" face expression corresponding to fear emotion

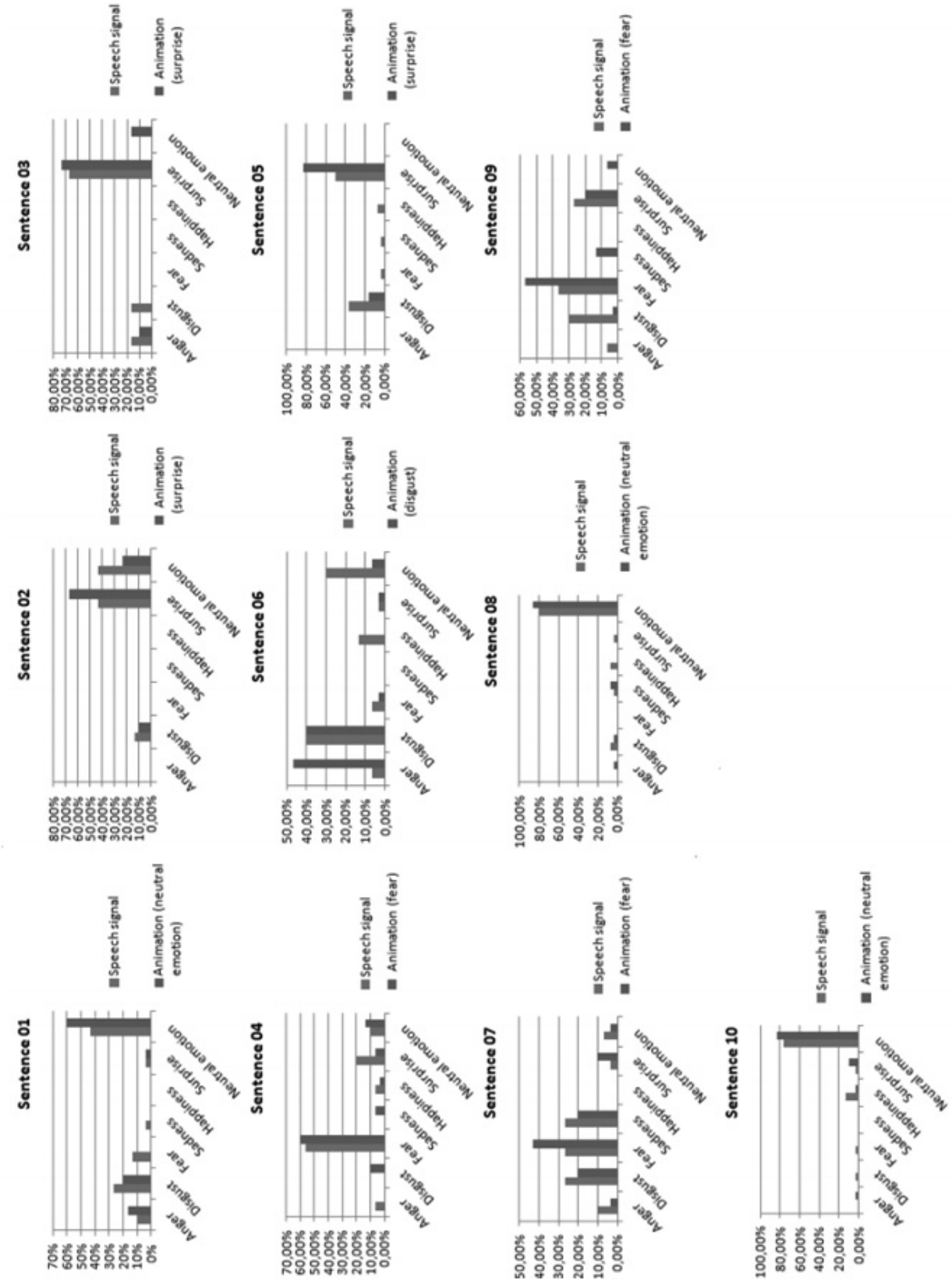

Fig. 4. The emotion recognition results, based on speech signal and human talking face animation which express the emotion given in brackets

using extended "Karol" face model. The phoneme transcriptions of the sentences and emotion identifier were used as an input data to the VSS system.

The study showed that assessing the emotions, based only on the information contained in short speech signal, is very complex and difficult. Only for five out of ten sentences the difference between most scored and the second-scored emotion was bigger than 15 %. Moreover for one of the sentences even three emotions received the highest score 26,67%. Those emotions were: disgust, fear and sadness. In case, when two or more emotions gained the same score, one emotion from them was chosen randomly to generate emotional animation.

During the second part of the research, new features of "Karol" face model gave interviewees additional information about speakers emotional state. For nine out of ten animations, the emotions, used in generation process, gained the highest score. In one case, where emotion was not recognized correctly the difference between highest and second scored emotion was less than 15 %. The second score belonged to emotion used in the animation generation process. The average of highest scores for emotion recognition based on the speech signal and on the animations were respectively 52% and 65%. Comparing the results of two surveys, it was much easier to match the emotion with the face animation than with speech signal. The face expression supports the emotion recognition process. The results of the two surveys are illustrated on Fig. 4.

Conclusion

Expressing the emotions is a very important feature of Visual Speech Synthesis systems. In this research, by successfully applying the Ekman's basic emotion model, a new Emotional VSS was developed. This functionality allows to read more information about speakers emotional state. The new feature of VSS system can have plenty of adaptations. It can be used not only to improve the human-computer interaction but also in area of modern psychology. The emotional human talking face animation can be used to study people emotional intelligence. Since the speech signal was the only input data used in this research, the future work should focus on automatic recognition of human emotions in it.

References

[1] Artur Janicki, Jakub Bloch, Karol Taylor, *Visual Speech Synthesis for Polish Using Keyframe Based Animation*, ICSES 2010 Gliwice.

[2] Singular Inversions Inc, *FaceGen – 3D Human Faces,* http://www.facegen.com.

[3] I.S. Pandzic and R. Forchheimer, *MPEG-4 Facial Animation: The Standard, Implementation and Applications*, The Atrium, Southern Gate, Chichester, West SussexPO19 1UD, England: John Wiley & Sons Ltd, 2002.

[4] K. Balci, *Xface: MPEG-4 based open source toolkit for 3D facial animation*, Proc. Advance Visual Interface 2004, pp. 399-402, 2004.

[5] Robert Plutchik, *The Nature of Emotion*, American Scientist, volume 89 No 4, 2001.

[6] Jeannette M. Haviland-Jones Michael Lewis, *Psychologia emocji*. Gdańsk, 2005. [7] Henryk Gasiul, *Teorie emocji i motywacji*. Warszawa, 2002. [8] Paul Ekman, Universals and Cultural Differences in Facial Expressions of Emotion, in Nebraska Symposium on Motivation, University of Nebraska Press, 1972. [9] Paul Ekman, *Unmasking the Face*, Englewood Cliff, New Jersey: PRENTICE-HALL INC., 1975.