

Isolated face region analysis for emotional speech synthesis

N.N. Nadtoka, J.D. Edge, P.J.B. Jackson, A. Hilton

University of Surrey, UK, {N.Nadtoka, J.Edge, P.Jackson, A.Hilton}@surrey.ac.uk

Abstract

This work aims to improve the quality of visual speech synthesis by modelling its emotional characteristics. The emotion specific speech content is analysed based on the 3D video dataset of expressive speech. Preliminary results indicate a promising relation between the chosen features of visual speech and emotional content.

Non-verbal communicational cues are important for natural speech synthesis. Existing state of the art animation models are capable of automated synthesis of plausible novel animated speech ([1], [2]), however, their results are far from the real performance due to the lack of correct model of non-verbal cues, e.g., emotion. Our work aims to learn and reproduce emotional characteristics for expressive content synthesis.

A commercial 3DMD dynamic surface capture system is used to recover 3D facial data of a native English speaker. The data consists of sentences performed both in a specified emotion, and in a neutral manner. The dataset presents all fundamental cross-culturally recognised emotions: Anger, Surprise, Sadness, Happiness, Fear and Disgust. A total of 110 sentences are selected on the basis of having strong expressive content (based on the dictionary of affect [3]) and good phonetic sampling. All sentences are also recorded without emotion (neutral). The synchronised audio is captured at 44.1 kHz. The visual data consists of a sequence of 3D scans (shape and texture) and associated normal maps recorded at 60 frames per second. The painted blue visual markers and lipstick are used in order to perform a temporal registration of the mesh geometry and texture.

Analysis of the emotional component of visual speech is performed via partitioning of the facial data into overlapping upper and lower face regions. Our analysis concentrates upon the isolated upper region as the expressive emotional gestures in this part of face are less contaminated by the articulatory face motion.

The upper facial region is subdivided into sub-regions using tracked marker points (see Fig. 1). Principal component analysis is used to reduce the dimensionality of the data. The first principal component (pc 1) of the region shown on the Fig. 1 reflects the position of the eyebrows with high values corresponding to the raised and low to lowered state. The fundamental frequency is extracted from the audio signal with Speech Filing System. Visual (pc 1) and auditory (f0)

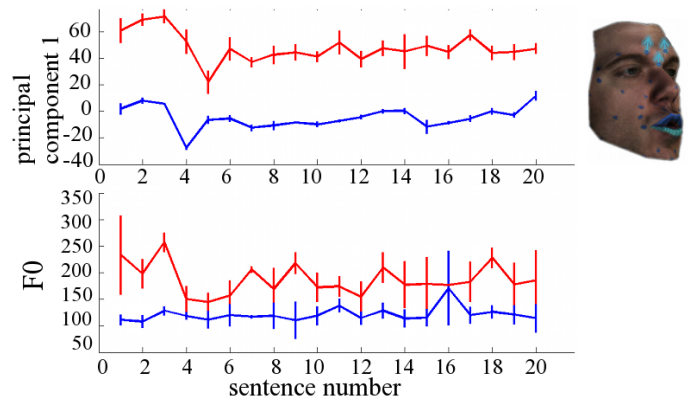


Figure 1: From left to right, top to bottom: Mean value and standard deviation (vertical bars) of the pc 1 of the upper face subregion for each sentence; red corresponds to fear and blue to neutral emotion. The main direction of variation of pc 1. Mean and standard deviation of f0 for corresponding audio signals for each sentence.

signals show similar trend in the differences of their mean value between neutral and fear. Therefore, the fundamental frequency of the audio signal for this emotion can be used to drive the visual animation. For other emotions pc 1 demonstrates more differences in the upper facial region, whereas f0 doesn't show distinctive behaviour patterns. E.g., in examples of sadness the eyebrows are constantly lower than for its neutral counterparts, whereas average f0 signal levels don't show much of the difference between sadness and neutral.

Observed relationships between the principal component features of visual signal in the upper face and the fundamental frequency of corresponding audio signal will be further used to drive the emotional component of synthetic visual speech.

References

- [1] M. Brand. Voice puppetry. *SIGGRAPH'99*, 1999.
- [2] Z. Deng, U. Neumann, J.P. Lewis, T. Kim, M. Bulut, and S. Narayanan. Expressive facial animation synthesis by learning speech coarticulation and expression spaces. *IEEE Transactions on visualization and computer graphics*, 2006.
- [3] C.M. Whissell. The dictionary of affect in language. In *Emotion: Theory, Research, and Experience*, pages 113–131, 1989.