English
Research Express@NCKU Volume 29 Issue 2 - May 8, 2015
[http://research.ncku.edu.tw/re/articles/e/20150508/4.html]

Two-Level Hierarchical Alignment for Semi-Coupled HMM-Based Audiovisual Emotion Recognition with Temporal Course

Chung-Hsien Wu*, Jen-Chun Lin, and Wen-Li Wei
Department of Computer Science and Information Engineering, National Cheng Kung University
chunghsienwu@gmail.com

IEEE Trans. Multimedia, DOI (identifier) 10.1109/TMM.2013.2269314, VOL. 15, NO. 8, December 2013, pp.1880-1895.

A complete emotional expression typically contains a complex temporal course[1] in face-to-face natural conversation. In this study, we focused on exploring the temporal evolution of an emotional expression for audio-visual emotion recognition. A novel data fusion method with respect to the temporal course modeling scheme named Two-Level Hierarchical Alignment-Based Semi-Coupled Hidden Markov model (2H-SC-HMM)[1] is proposed to effectively solve the problem of complex temporal structures of an emotional expression and consider the temporal relationship between audio and visual streams for increasing the performance of audio-visual emotion recognition in a conversational utterance.

Temporal Course of Emotional Expression

Previous psychologist research[2][3][4] showed that a complete emotional expression can be characterized in three sequential temporal phases: Onset (application), Apex (release), and Offset (relaxation), when considering the manner and intensity of expression. Although the temporal course of emotional expression was demonstrated sequentially in time, a complete emotional expression is expressed by more than one utterance in natural conversation, and in more detail, each utterance may contain several temporal phases of emotional expression as shown in Fig. 1.

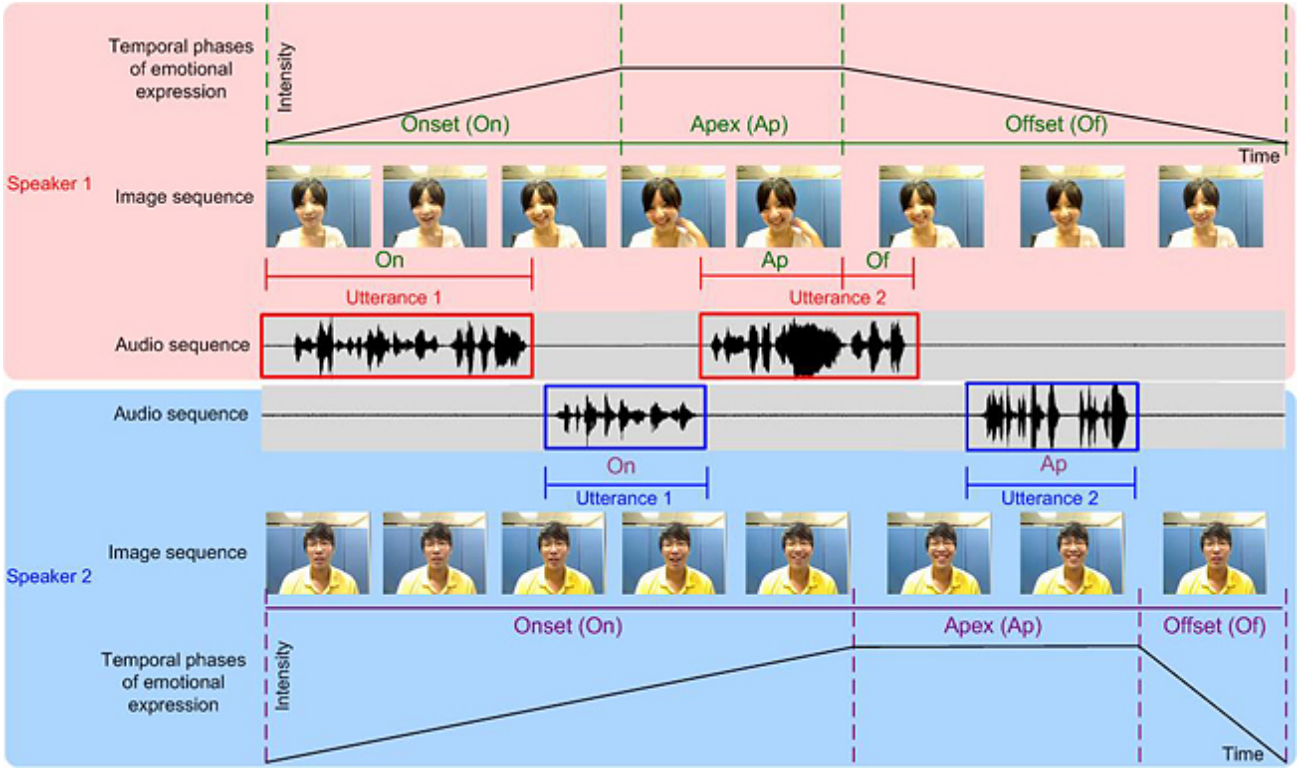

Figure 1. An example of various temporal phases of happy emotional expression occurred to different utterances in a real conversational environment.

Modeling the Temporal Course of Emotional Expression

Based on aforementioned analysis, to model the complex temporal course of emotional expression, a temporal course modeling scheme is proposed in this study to characterize the temporal evolution involved in an emotional state that occurs in an isolated utterance. An isolated utterance in a conversation can express one or several emotional sub-states, which are defined to represent the temporal phases (i.e., onset, apex, or offset with low or high intensity) of an emotional expression, and an HMM is used to characterize single emotional sub-state, rather than the entire emotional state as shown in Fig. 2.

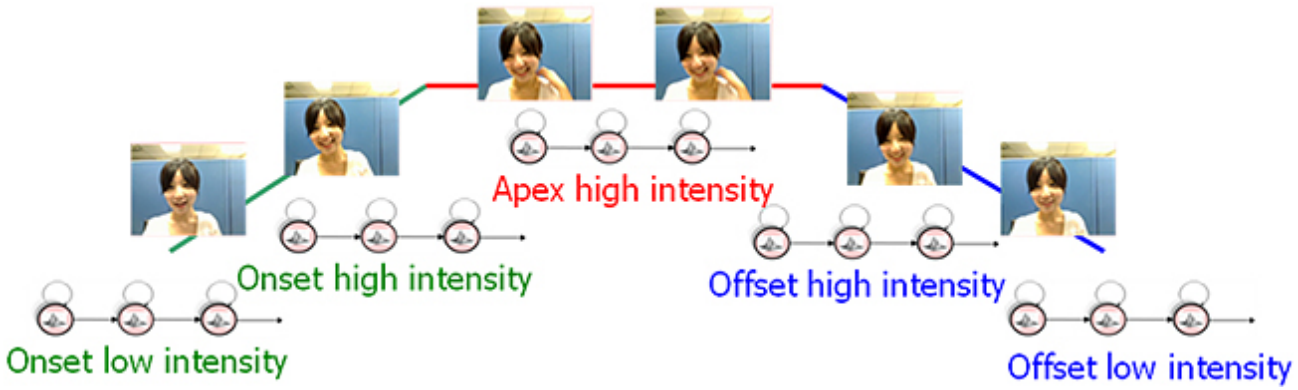

Figure 2.Emotional Temporal phases modeling based on various HMMs

Two-Level Hierarchical Alignment-Based Semi-Coupled Hidden Markov Model (2H-SC-HMM)

For effective emotion recognition, a two-level hierarchical alignment mechanism is further proposed and applied to the SC-HMM to align the relationship within and between the temporal phases in the audio and visual HMM

sequences at the model and state levels as shown in Fig. 3. By integrating the emotional sub-state language model, which model the temporal transition between emotional sub-states expressed in an isolated utterance, the proposed two-level hierarchical alignment-based SC-HMM (2H-SC-HMM) can further provide a constraint on allowable temporal structures to obtain an optimal recognition result of emotional state in each utterance. The formula of the proposed 2H-SC-HMM is shown as

$$\hat{E} = \arg\max_{E} \left\{ \max_{\Lambda^a,\Lambda^v} \left[P(\Lambda^v \mid \Lambda^a, E) P(\Lambda^a \mid E) \max_{S^a,S^v} \left(P(O^a, S^a \mid \Lambda^a, E) P(S^v \mid S^a, \Lambda^a, E) \right.\right.\right.$$
$$\left.\left.\left. P(O^v, S^v \mid \Lambda^v, E) P(S^a \mid S^v, \Lambda^v, E) \right) P(\Lambda^a \mid \Lambda^v, E) P(\Lambda^v \mid E) \right] P(E) \right\}$$

隱藏式馬可夫模型序列辨識機率 | 隱藏式馬可夫模型序列狀態序列校準機率 | 隱藏式馬可夫模型序列校準機率 | 情緒子狀態語言模型 | 模型序列機率

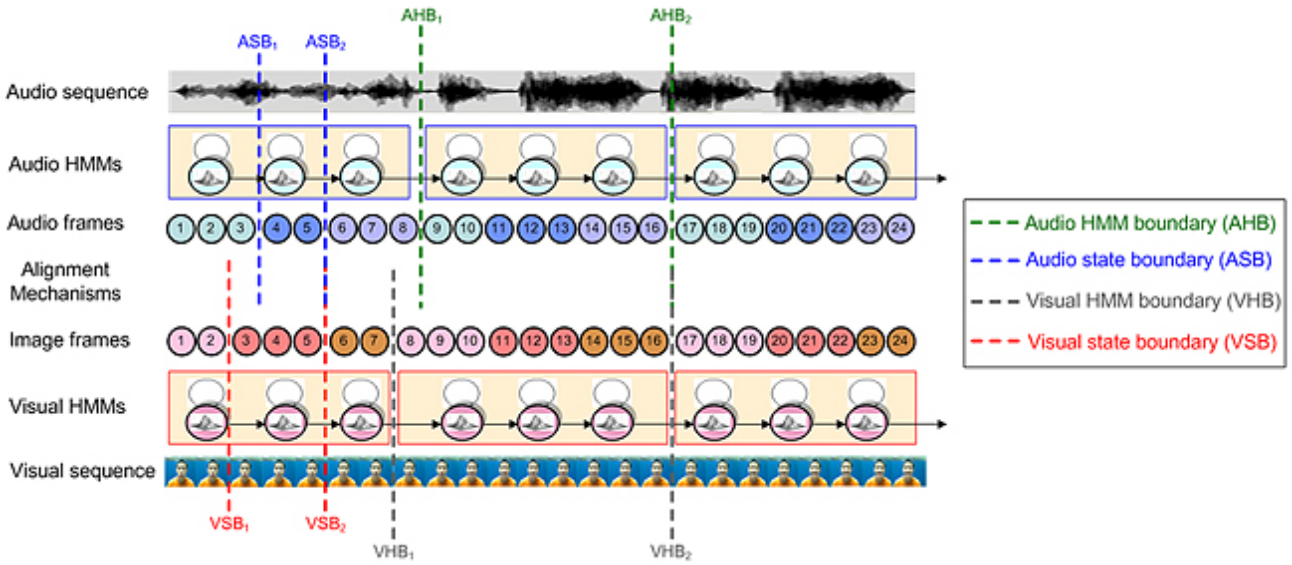

Figure 3. An example illustrating model- and state-level alignment between audio and visual HMM sequences in the happy emotional state. The green and gray dotted lines represent the audio and visual HMM boundaries respectively and are used for model-level alignment estimation; the blue and red dotted lines represent the state boundaries under audio and visual HMMs respectively and are used for state-level alignment estimation. The audio and image frames are represented by the numbered circles.

For performance evaluation, two databases[5][6][7] are considered: the posed MHMC database and the spontaneous SEMAINE database. The recognition accuracy achieved 91.55% and 87.5% for posed MHMC database and the spontaneous SEMAINE database, respectively. Experimental results show that the proposed method not only outperforms other fusion-based bimodal emotion recognition methods for posed expressions but also provides acceptable results for spontaneous expressions.

References

1. C. H. Wu, J. C. Lin, and W. L. Wei, “Two-level hierarchical alignment for semi-coupled HMM-based audiovisual emotion recognition with temporal course,” IEEE Trans. Multimedia, vol.15, no.8, pp. 1880–1895, 2013.
2. P. Ekman, *Handbook of Cognition and Emotion.* Wiley, 1999.
3. M. F. Valstar and M. Pantic, “Fully automatic recognition of the temporal phases of facial actions,” *IEEE*

Trans. Systems, Man and Cybernetics–Part B, vol. 42, no.1, pp. 28–43, 2012.

4. M. F. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," *Int'l Conf. on Computer Vision and Pattern Recognition*, vol. 3, 2006.
5. J. C. Lin, C. H. Wu, and W. L. Wei, "Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition," *IEEE Trans. Multimedia*, vol. 14, no.1, pp. 142–156, 2012.
6. G. Mckeown, M. F. Valstar, R. Cowie, M. Pantic, and M. Schroe, "The SEMAINE database: annotated multimodal records of emotionally coloured conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no.1, pp. 5–17, 2012.
7. G. Mckeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpous of emotionally coloured character interactions," *IEEE Int'l Conf. on Multimedia and Expo*, pp. 1079–1084, 2010.