中文

成大研發快訊　第二十九卷 第二期 - 2015年五月八日
[http://research.ncku.edu.tw/re/articles/c/20150508/4.html]

基於兩階層校準半耦合隱藏式馬可夫模型為基礎之語音視覺情緒辨識

吳宗憲*, 林仁俊, 魏文麗
國立成功大學資訊工程學系

chunghsienwu@gmail.com

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

在面對面的溝通過程中，一個完整的情緒展現往往包含著複雜的時間歷程[1]。因此，本論文主要針對情緒展現在時間歷程上之演化模式做探討，提出一個以時間過程建模方案為基礎之兩階層校準半耦合隱藏式馬可夫模型[1]，來有效地模組化一個情緒展現在一個句子當中時間的演化模式，並進一步考慮在情緒展現的過程中語音與視覺串流在時間上的關聯性來提升語音視覺情緒辨識之準確性。

情緒展現的時間歷程

根據過去心理學家的研究指出[2][3][4]，一個完整的情緒展現可被歸納為三個接續的時間階段：情緒被喚起時的初始階段(Onset)、情緒達到頂峰的階段(Apex)以及情緒慢慢緩和到中性情緒的階段(Offest)。然而，在自然的對話中，一個完整的情緒展現往往會被拆散到多個句子(utterances)之中，即一個句子往往包含著一到多個的時間階段，如圖1所示。

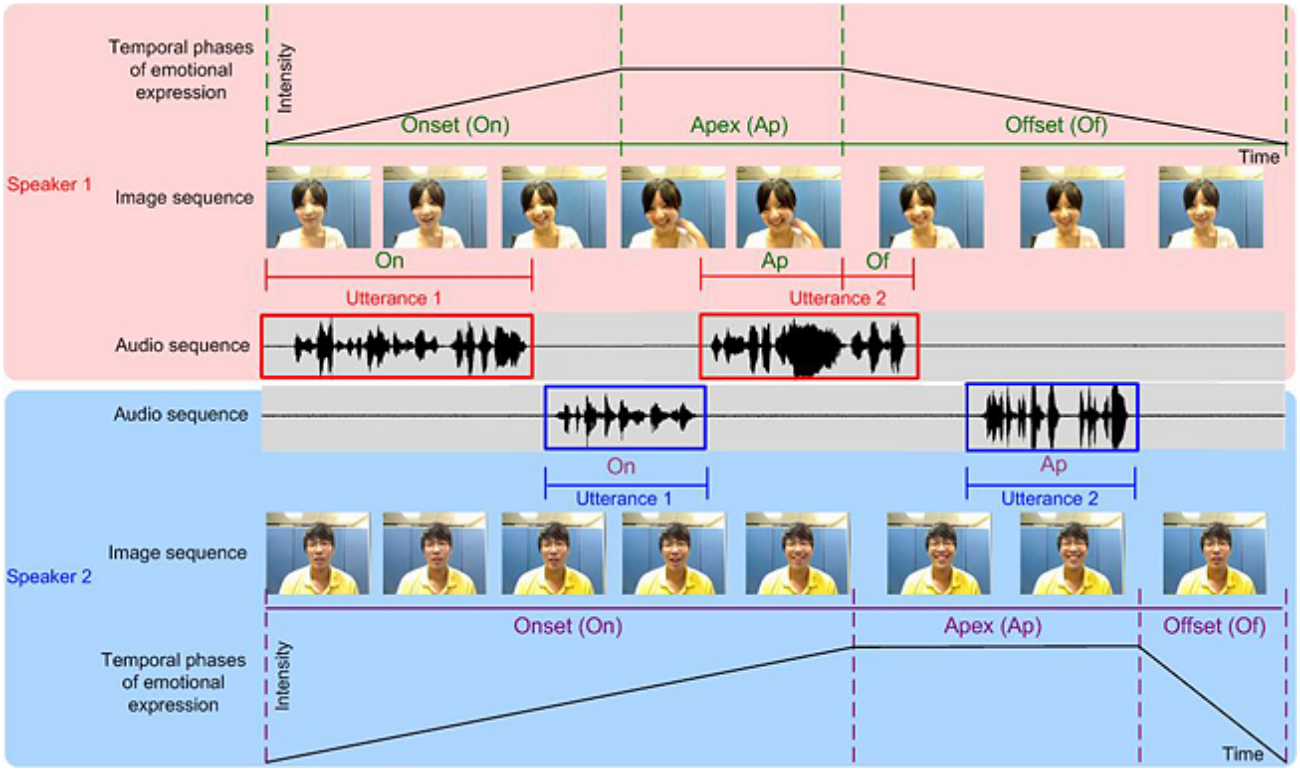

圖1 真實對話環境中，情緒展現“快樂”的各種時間階段發生在不同句子之中的實例。

模組化情緒展現的時間歷程

根據上述分析結果，欲模組化此複雜的時間結構，本論文首先以多個隱藏式馬可夫模型分別模組化情緒展現中的各種時間階段，如圖2所示。

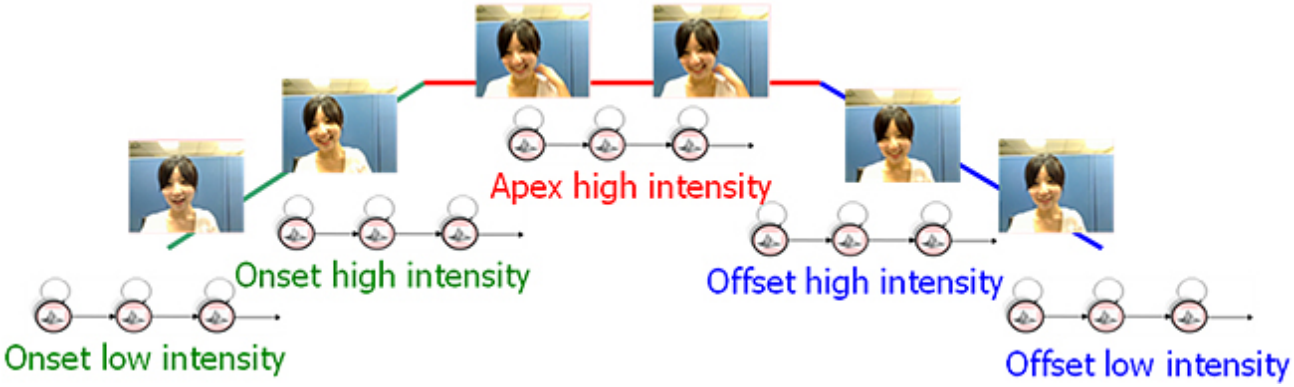

圖2 透過多個隱藏是馬可夫模型分別模組化情緒展現中的不同時間階段。

兩階層校準半耦合隱藏式馬可夫模型

辨識方面，本論文透過所提出之兩階層校準半耦合隱藏式馬可夫模型，串接各種模組化(情緒時間階段信息)之隱藏式馬可夫模型，來有效地模組化一個情緒展現在一個句子當中時間的演化模式，並進一步考慮在情緒展現的過程中語音與視覺串流在時間上的關聯性，達到情緒辨識之目的，如圖3所示。其模型公式如下；

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

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

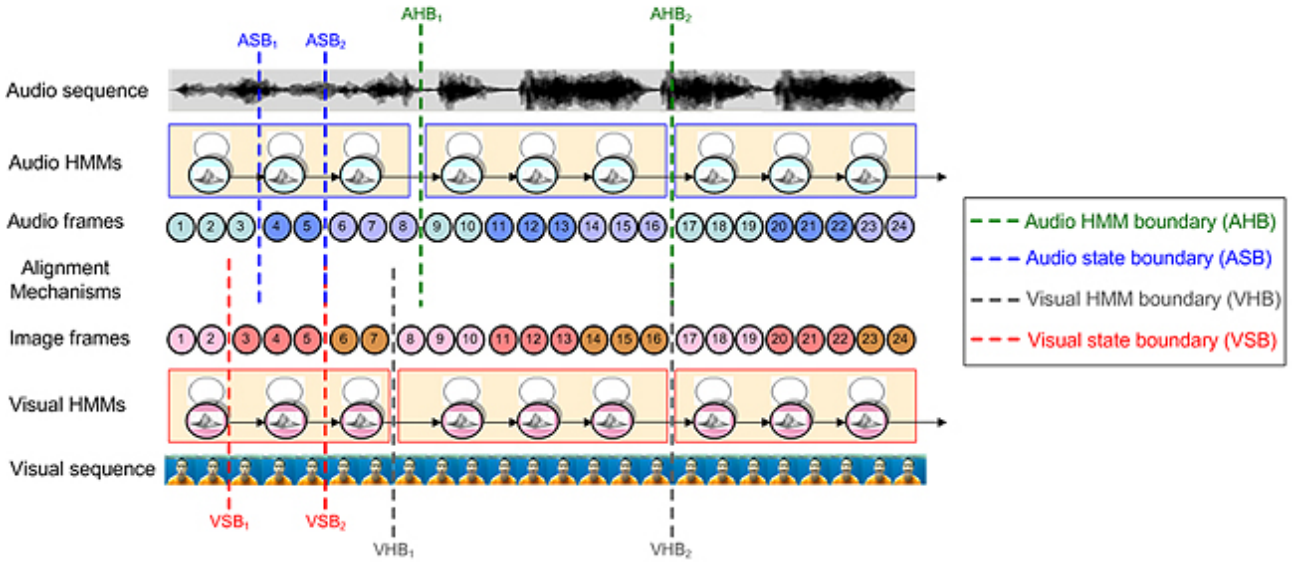

圖3兩階層(模型階層以及狀態階層)校準示意圖；其中，綠色和灰色虛線分別代表語音及視覺隱藏式馬可夫模型序列的邊界並進一步被用來估算語音及視覺馬可夫模型之間的關聯性；而藍色以及紅色虛線則分別表

是語音及視覺隱藏是馬可夫模型狀態序列的邊界並進一步被用來估算語音及視覺馬可夫模型狀態之間的關聯性；透過此關聯性來有效描述語音跟視覺串流在時間上之關聯。

最後，本論文主要採用兩個情緒資料庫[5][6][7]來對於所提出之兩階層校準半耦合隱藏式馬可夫模型進行實驗與檢定驗證顯示，本論文所提出之方法，在語音視覺情緒辨識的辨識準確率上分別可達到91.55%**以及**87.5%**，具有相當程度的改進。**

參考文獻：

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.