English

Research Express@NCKU Volume 31 Issue 4 - March 3, 2017
[http://research.ncku.edu.tw/re/articles/e/20170303/1.html]

Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels

Chung-Hsien Wu*, Wei-Bin Liang
Department of Computer Science and Information Engineering, National Cheng Kung University
chunghsienwu@gmail.com

IEEE Trans. Affective Computing, VOL. 2, NO. 1, JANUARY-MARCH 2011, pp. 10~21.

【105 Outstanding Research Award】Special Issue

Speech is one of the most fundamental and natural communication means of human beings. With the exponential growth in available computing power and significant progress in speech technologies, spoken dialogue systems (SDS) have been successfully applied to several domains. However, the applications of SDSs are still limited to simple informational dialog systems, such as navigation systems, air travel information system, etc. [1][2]. To enable more complex applications (e.g. home nursing [3] educational/tutoring, and chatting [4]), new capabilities, such as affective interaction, are needed. However, to achieve the goal of affective interaction via speech, several problems in speech technologies, including low accuracy in recognition of highly affective speech and lack of affect-related common sense and basic knowledge, still exist. This work presents an approach to emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information (AP) and semantic labels (SLs). For AP-based recognition, acoustic and prosodic features including spectrum-, formant-, and pitch-related features are extracted from the detected emotional salient segments of the input speech. Three types of models Gaussian Mixture Models (GMMs), Support Vector Models (SVMs), and Multilayer Perceptrons (MLPs) are adopted as the base-level classifiers. A Meta Decision Tree (MDT) is then employed for classifier fusion to obtain the AP-based emotion recognition confidence. For SL-based recognition, semantic labels derived from an existing Chinese knowledge base called HowNet are used to automatically extract Emotion Association Rules (EARs) from the recognized word sequence of the affective speech. The maximum entropy model (MaxEnt) is thereafter utilized to characterize the relationship between emotional states and EARs for emotion recognition. Finally, a weighted product fusion method is used to integrate the AP-based and SL-based recognition results for final emotion decision.

Figure 1 illustrates the block diagram of the training and testing procedures for AP- and, SL-based emotion recognition. For AP-based approach, emotional salient segments (ESS) are firstly detected from the input speech. Acoustic and prosodic features including spectrum-, formant-, and pitch-related features are extracted from the detected emotional salient segments and used to construct the GMM-based, SVM-based, and MLP-based base-level classifiers. The MDT is then employed to combine the three classifiers by selecting the most promising classifier for AP-based emotion recognition. On the other hand, the word sequence recognized by a speech recognizer is used in SL-based emotion recognition. The semantic labels of the word sequence derived from an existing Chinese knowledge base called the HowNet [5] are extracted and then a text-based mining approach is employed to mine the Emotion Association Rules (EARs) of the word sequence. Next, the MaxEnt model [6] is employed to characterize the relation between emotional states and EARs and output the emotion recognition result. Finally, the outputs from the above two recognizers are integrated using a weighted product fusion method to determine the final emotional state. Furthermore, in order to investigate the effect of individual personality characteristic, the personality trait obtained from Eysenck Personality Questionnaire (EPQ) for a specific speaker

is considered for personalized emotion recognition.

For evaluation, 2,033 utterances for four emotional states (Neutral, Happy, Angry, and Sad) are collected. The evaluation results are shown in Table 1. According to the result based on EPQ, speaker A is an extrovert and the recognition performance of the corresponding emotions - happy and angry emotion which have stronger expression was improved. For speaker B who is neither extrovert nor introvert, the difference of the evaluation results is small. Besides this evaluation, the subjects were satisfied with the fine-tuned system after they tested this system again. The evaluation of the proposed approach proved that the proposed approach can work well on the emotion recognition task. In summary, the average recognition accuracy of this system can achieve 85.79% considering personality trait. The results confirm the effectiveness of the proposed approach.

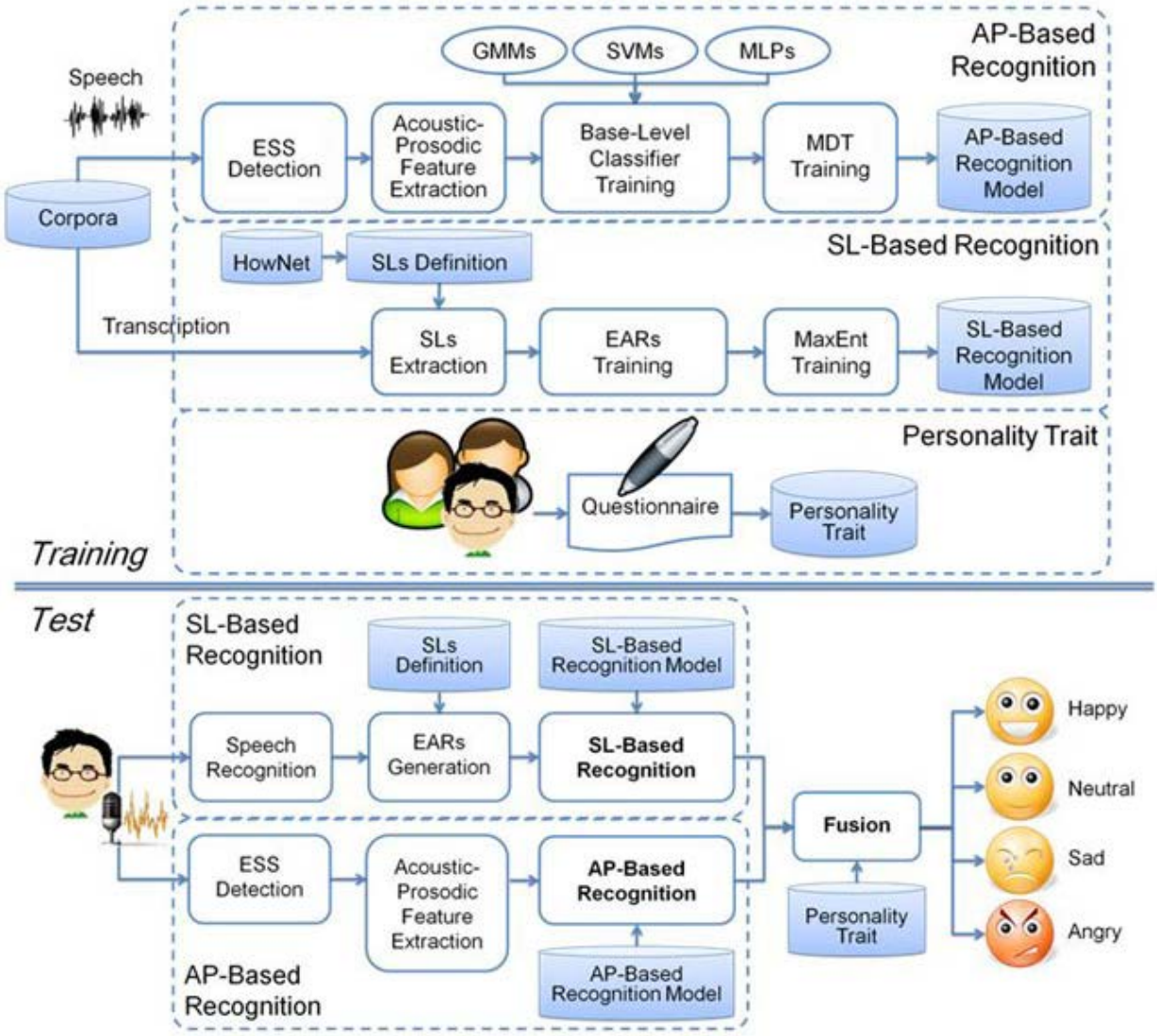

Figure 1: An overview of training and testing flowchart of the acoustic-prosodic information-based recognition, the semantic label-based recognition and the personality trait

Table 1: Evaluation results of AP-based and SL-based emotion recognition with personality trait

	MDT+MaxEnt (λ_{AP} = 0.4) (Accuracy %)				
	Neutral	Happy	Sad	Angry	Average
Speaker A	75.80%	85.97%	83.35%	87.81%	83.23%
Speaker B	80.40%	81.61%	88.49%	84.93%	83.86%
Coupus B	78.10%	83.79%	85.92%	86.37%	83.55%
	Proposed (MDT+MaxEnt+PT) (Accuracy %)				
	Neutral	Happy	Sad	Angry	Average
Speaker A	76.30%	88.79%	83.17%	89.91%	84.54%
Speaker B	86.48%	84.99%	91.49%	85.17%	87.03%
Coupus B	81.39%	86.89%	87.33%	87.54%	85.79%

References:

1. J. Liu, Y. Xu, S. Senef, and V. Zue, "CityBrowser II: A Multimodal Restaurant Guide in Mandarin," in Proc. *International Symposium Chinese Spoken Language Processing (ISCSLP)*, pp. 1-4, 2008.
2. C.-H. Wu and G.-L. Yan, “Speech Act Modeling and Verification of Spontaneous Speech with Disfluency in a Spoken Dialogue System,” *IEEE Trans. on Speech and Audio Processing*, Vol.13, pp.330-344, May 2005.
3. N. Roy, J. Pineau, and S. Thrun, "Spoken Dialogue Management Using Probabilistic Reasoning," in Proc. *Annual Meeting on Association for Computational* Linguistics (AM-ACL), pp. 93-100, 2000.
4. D. Jurafsky, R. Ranganath, D. McFarland, “Extracting Social Meaning: Identifying Interactional Style in Spoken Conversation,” in Proc. *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT)* , pp. 638-646, 2009.
5. Z. Dong, and Q. Dong, HowNet [Online] Available: http://www.keenage.com/
6. A. Berger, S. Della Pietra, and V. Della Pietra, “A Maximum Entropy Approach to Natural Language Processing,” *Computational Linguistics*, Vol.22, No. 1, pp. 39-71, 1996.

Copyright 2017 National Cheng Kung University