Automatic Recognition of Emotions from the Acoustic Speech Signal

1. Research Team

Project Leader: Prof. Shrikanth Narayanan, *Electrical Engineering*

Graduate Students: Chul Min Lee

Industrial Partner(s): Speechworks International

2. Statement of Project Goals

This research aims at investigating several feature sets such as acoustic, lexical, and discourse features, and classification algorithms for classifying spoken utterances based on the emotional state of the speaker [1]. Besides applications in enabling natural human machine interfaces, the problem motivates development of novel classification algorithms that can operate with sparse data. Another aim is to find out the way to seamlessly combine the emotion recognition system with state-of-the-art speech recognition system, and the satisfactory dialog management of the automated call center system to better support human-computer interaction in that application.

3. Project Role in Support of IMSC Strategic Plan

The proposed work contributes to enabling natural and customizable interactions, a key element of IMSC's strategic plan.

4. Discussion of Methodology Used

Combination of other information: there are two kinds of information useful for emotion recognition in a given application. Those are lexical information and discourse information. When properly combined with acoustic information, both lexical and discourse information could improve the performance of the emotion recognizer. Especially, in a domain-specific application, content-based information should be incorporated in the proper way, and combining several kinds of information for an emotion recognition system is explored.

5. Short Description of Achievements in Previous Years

Over all, results show that combining all the information improves emotion classification by 40.7% for male and 36.4% for female (linear discriminant classifier used for acoustic information) over using only acoustic information. The results were submitted to ICSLP '02 and IEEE Transaction on Speech and Audio Processing.

5a. Detail of Accomplishments During the Past Year

We investigated the combination of the other sources of information with acoustic information in order to improve the performance with respect to overall classification error**.** The scheme for combining various sources of information was decision level fusion. We can generate multiple

classifiers to manipulate multiple sources of information available to the learning algorithm. In our case, we have three independent classifiers, one each for, acoustic, language, and discourse information, and then a final decision is made by combining the output results from these classifiers. Overall, combining other sources of information with acoustic information leads to performance improvements in motion recognition. Note the combination of acoustic and language information showed the best performance in almost all the settings. The inclusion of discourse information for this data does not seem to provide any significant improvements when used in conjunction with acoustic and lexical information. This may be due to the fact that lexical information is highly correlated to discourse information.

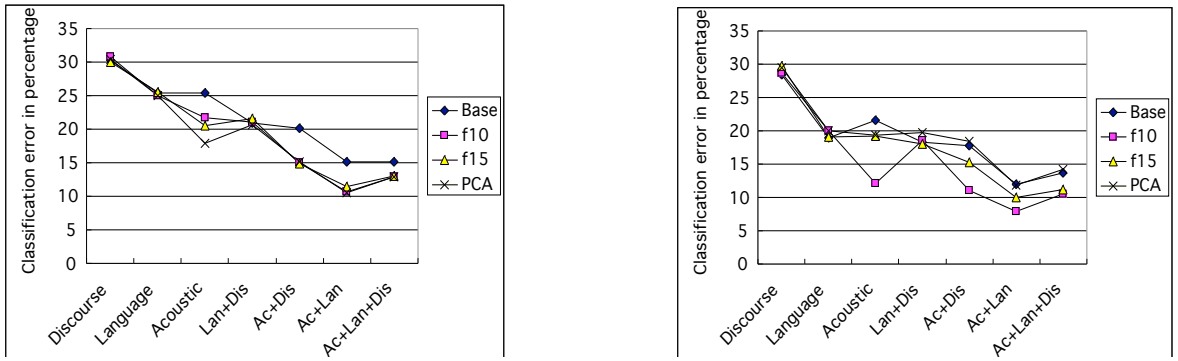

Figure 1. Comparison of classification errors for several schemes of combination of information. Classification methods used were linear discriminant classifiers for each information. (a) male data (b) female data.

6. Other Relevant Work Being Conducted and How this Project is Different

Various pattern recognition approaches have been explored for automatic emotion recognition [5,6]. Dellaert et al., for instance, used maximum likelihood Bayes classification, Kernel regression, and k-nearest neighbor methods [7], whereas Roy and Pentland used Fisher linear discrimination method [8]. In this paper we used linear discrimination (LDC), and k-nearest neighborhood (k-NN) classifiers as the classification methods. Petrushin developed a real-time emotion recognizer using neural networks for call center applications, and achieved ~77% classification accuracy in two emotion states, 'agitation' and 'calm' for 8 features chosen by a feature selection [5].

In this study, we used a corpus of sentences from a human-machine spoken dialog application deployed by SpeechWorks used by real customers [9]. So far, most of the reported studies have used speech recorded from actors that were asked to express pre-defined emotions. A notable exception is the study by Batliner et al [6]. They adopted a 'Wizard-of-Oz' scenario to collect data, which assumed that naive subjects were asked to communicate with a real computer, in two emotion categories such as 'emotional' and 'neutral'.

As for the combination of various sources of information, we used 'emotionally salient' words as lexical information, and discourse information; i.e., rejection, repetition, and other discourse

tags. Combining context-related information with acoustic information to improve the performance of the emotion recognizers has been studied by Batliner et al [6], and Ang et al [12].

7. Plan for the Next Year

After finishing analysis of the contribution of each source of information to emotional states, we will develop and implement an emotion recognizer mounted on a state-of-the-art speech recognition system.

8. Expected Milestones and Deliverables

Data, tagged by subjective judgments.
Algorithms for emotion categorization, published.
Emotion recognition system mounted on state-of-the-art speech recognition system.

9. Member Company Benefits

N/A

10. References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J.G. Taylor, "Emotion Recognition in Human-Computer Interaction," *IEEE Signal Proc. Mag.*, 18(1), pp. 32-80, Jan 2.
[2] C.M. Lee, S. Narayanan, R. Pieraccini, "Recognition of Negative Emotions from the Speech Signal," *Proc. Automatic Speech Recognition and Understanding*, (Trento, Italy), Dec. 2001.
[3] A. Ortony, G.L. Clore, and A. Collins, *The Cognitive Structure of Emotions*, Cambridge Univ. Press, Cambridge, UK, 1988.
[4] K. Scherer, "A Cross-Cultural Investigation of Emotion Inferences from Voice and Speech: Implications for Speech Technology," *ICSLP 2000*, Beijing, China, Oct. 2000.
[5] V. Petrushin, "Emotion in Speech: Recognition and Application to Call Centers," *Artificial Neu. Net. In Engr. (ANNIE '99)*, pp. 7-10, Nov. 1999.
[6] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Noth, "Desperately Seeking Emotions: Actors, Wizards, and Human Beings," *Proceedings of the ISCA Workshop on Speech and Emotion*, (to appear).
[7] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing Emotion in Speech," *ICSLP'96 Conference Proceedings*, Philadelphia, PA., pp. 1970 -1973, 1996.
[8] D. Roy and A. Pentland, "Automatic Spoken Affect Analysis and Classification". *In the Proceedings of the International Conference on Automatic Face and Gesture Recognition*, Killington, VT. 1996.
[9] http://www.speechworks.com/indexFlash.cfm
[10] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk and S. Stroeve, "Approaching Automatic Recognition of Emotion from Voice: A Rough Benchmark," *ISCA Workshop on Speech and Emotion*, Belfast 2000.
[11] Sudha Arunachalam, Dylan Gould, Elaine Andersen, Dani Byrd and Shrikanth S. Narayanan, ``Politeness and frustration language in child-machine interactions", in *Proc. Eurospeech*, (Aalborg, Denmark), pp. 2675-2678, 2001.
[12] J. Ang and R. Dhillon and A. Krupski and E. Shriberg and A. Stolcke, "Prosody-Based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog," ICSLP '02, Denver, CO.