BELIEF NETWORKS IN NATURAL LANGUAGE PROCESSING FOR IMPROVED SPEECH EMOTION RECOGNITION

Ronald Müller, Björn Schuller, and Gerhard Rigoll

Technische Universität München
Institute for Human-Machine Communication
Arcisstr. 21, D-80333 München, Germany
{mueller, schuller, rigoll}@mmk.ei.tum.de

ABSTRACT

Within the history of research in Emotion Recognition from speech, focus has been put on investigations concerning vocal sound properties of utterances for the purpose of inferring to a potentially expressed emotion with sufficient success [1]. However, in many cases the spoken utterances also contain emotional information lying in the choice of words, which, as shown in the following, should not be ignored and condemned in order to improve recognition performance significantly. Hereby the speaker's expression of his emotion consists in usage of certain phrases that are likely to be mixed with other meaningful statements [2][3]. Therefore an approach with abilities in spotting for emotional relevant information is needed. Furthermore as an Automatic Speech Recognition (ASR) unit provides probably incomplete and uncertain data to work on, the processing interpretation algorithm should not only be able to deal with but use such knowledge. Hence, as mathematical background for the semantic analysis of spoken utterances we chose Belief Networks (BN) for their capabilities in spotting and handling uncertain and incomplete information on the input level as well as providing real recognition confidences at the output, which is valuable in regard to a subsequent late fusion with results from prosodic analysis [4]. The aim here is to make the Belief Network maximize the probability of the root node modeling the specific emotion expressed by a speaker via his choice of words and phrases (Fig. 1).

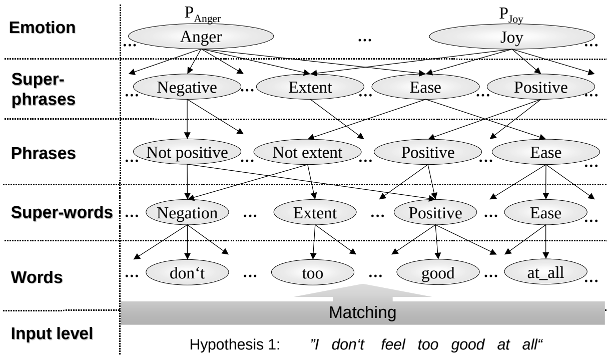

Fig. 1: Cutout of interpretation model for emotions

The approach presented here is to be based on integration and abstraction of semantically similar units to higher leveled units in several layers. This method proved to be applicable for Natural Language Interpretation in several restricted domains, like natural language man-machine dialogues for intuitive car-application controlling, with remarkable results.

Training and evaluation procedures ran on a large hand-labeled database comprising an equally distributed number of utterances for each emotion. Thereby 12% of items are free from any emotional content and are therefore assigned to "neutral". The contained emotion of a number of utterances recorded and transcribed in the database can unambiguously be identified only via their prosodic properties as from the spoken content alone a distinctive assignment is impossible even for a human mind. Hence, the average error rate at 40.4% of this semantic interpretation approach left on its own seems to be rather poor compared to methods to be based on prosody, which perform at error rates of 25.8%. Nevertheless this additional information appears quite valuable, as under application of an adequate Neural Network for the late fusion of recognition results from prosody and semantics a reduction of error rate by almost one third to overall 8.0% was achieved.

REFERENCES

[1] R. Cowie, et al.: "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[2] B. Schuller, G. Rigoll, M. Lang: "Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine - Belief Network Architecture", *to appear at ICASSP 2004*, Montreal, Canada, 17.-21.05.2004

[3] C. M. Lee, R. Pieraccini: "Combining acoustic and language information for emotion recognition," *Proceedings of the ICSLP 2002*, Denver, CO, USA, 2002

[4] F. V. Jensen, "*An Introduction to Bayesian Networks*", UCL Press, 1996.

From: MLMI 2005, Edinburgh, UK