Demonstrating Laughter Detection in Natural Discourses

Stefan Scherer[1], Volker Fritzsch[1], Friedhelm Schwenker[1], and Nick Campbell[2]

[1] Institute of Neural Information Processing, Ulm University
[2] Center for Language and Communication Studies, Trinity College Dublin

1 Introduction

This work focuses on the demonstration of previously achieved results in the automatic detection of laughter from natural discourses. In the previous work features of two different modalities, namely audio and video from unobtrusive sources, were used to build a system of recurrent neural networks called Echo State networks to model the dynamics of laughter. This model was then again utilized to detect laughters from presented test data. The approach was used to confirm human labels of laughter and to detect laughter in previously unmarked data, which resulted in nice results in offline applications. As reported in a publication currently under revision accuracies of 90 % were achieved using the multi modal input data relying on modulation spectral features from one microphone and movement data of a 360 degree camera positioned in the middle of a conference table [1]. In this work however, we would like to show a proof of concept for the online and on the fly recognition of laughter performing close to real-time. The goal of this work is to use a previously trained model of laughter in a modular process engine environment, which is currently under development overcoming known difficulties of pattern recognition and information fusion tasks, to detect laughter from a continuous microphone input. The intended application may then be used in online applications such as robots interacting with humans or the evaluation of human to human communication. Furthermore, the detection of laughter is an integral part of the improvement of dialog systems towards an affect understanding machine, since laughter is an important part of a healthy and natural communication.

2 Recognition System

As mentioned before we utilize Echo State networks (ESN) [2], making use of the sequential characteristics of the modulation spectrum features that are scaled perceptually, in our approach. The features are extracted every 20 ms and comprise data of 200 ms in order to be able to give accurate on- and offset positions of laughter, but also comprise around a whole "laughter syllable" in one frame [3]. The biologically inspired features are extracted using standard methods like Mel filtering and Fast Fourier Transformations. In short they represent the rate

of change of frequency, since they are based on a two level Fourier transformation. These low dimensional ($dim = 8$) features are used as input to the ESN, consisting of a large number of sparsely interconnected neurons in a dynamic reservoir. Furthermore, the ESN is trained efficiently using the direct pseudo inverse function to adapt the weights from the reservoir towards the output layer [4]. ESNs are capable due to the characteristics of the dynamic reservoir to make use of the past few inputs to predict upcoming events, such as laughter in a sequential input.

3 Process Engine

The real time environment is realized by our own process engine, that is under constant development, providing rapid prototyping possibilities to the user. The goal is to be able to plug different black boxes, resembling various modules comprising filters, classifiers, and sources, together in an abstraction level above common programming languages. Furthermore, the engine will be able to deal with multiple sources in order to fulfill information fusion tasks at different steps, such as feature fusion or classifier fusion. Therefore, it is necessary to handle sources with different starting, and end points as well as different sampling rates as it is the case for audio and video sources. The integration of multiple modalities however, is a matter of future development. At the current state we will demonstrate a proof of concept using a processing pipeline composed by the following elements:

- input source node (microphone)
- feature extraction node (modulation feature extraction using matlab)
- classifier node (an ESN provided by matlab)
- output node for visual demonstration

Acknowledgements: This research would not have been possible without the generous funding of the German Academic Exchange Agency (DAAD), and within the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

References

1. Scherer, S., Schwenker, F., Campbell, N.: Multi modal laughter detection in natural discourses. In submission for International Journal Computer Vision and Image Understanding (special issue on Sensor fusion), Elsevier (2009)
2. Jaeger, H., Haas, H.: Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. Sience 304. pp, 78-80. (2004)
3. Knox, M., Mirghafori, N.: Automatic laughter detection using neural networks. Proceedings of Interspeech. pp. 2973-2976. ISCA, (2007)
4. Jaeger, H.: Tutorial on training recurrent neural networks, covering bppt, rtrl, ekf and the echo state network approach. Technical Report 159, Fraunhofer-Gesellschaft, St. Augustin Germany. (2002)