5. Recognizing Paralinguistic events

5.1 Introduction

Applications of Computational Paralinguistics have grown rapidly over the last decade and span both human-human as well as human-machine interactions. The ComPare Paralinguistics challenges have been playing a significant role in driving progress in the diverse use of paralinguistics. Besides the traditional task of affect recognition using suprasegmental non-verbal aspects of speech, novel tasks were introduced, such as, the detection of speaker traits, deception, conflict, eating and autism [**schuller2013interspeech**; **schuller2010interspeech**; **schuller2015interspeech**; **schuller2017interspeech**]. These challenges have shown that paralinguistic information can be used not only to identify affect but also clues that are helpful to detect abnormalities indicating disorders. Paralinguistic information also has applications in other domains of speech processing such as dialog systems, speech synthesis, voice conversion, assistance systems, and eHealth systems.

Typical approaches for classification and prediction of paralinguistic features include extraction of low level descriptive features followed by a machine learning model. Examples of low level descriptors are Mel-Frequency Cepstral Coefficients (MFCCs), log Mel-scale filter banks energies (FBANK) and several suprasegmental acoustic features that can be extracted using the openSMILE tool [**opensmile2010**]. These features act as general purpose feature set and are expected to achieve competitive results in a wide range of paralinguistic problems. However, derived neural representations using unsupervised learning have shown impressive results on many speech and image based tasks recently [**aytar2016soundnet**]. These features usually embed the task relevant information from the entire utterance in a compact form. Also end-to-end learning models have been employed in affect classification using Long Short-Term Memories (LSTMs) or Gated Recurrent Units (GRUs) [**trigeorgis2016adieu**; **Interspeech2018**].

5.2 Impementations

5.2.1 Data

Used self assessed affect dataset

5.2.2 Low level features

For acoustic feature extraction each utterance was divided into (length is 8) into 20 segments and no frame shift. For each frame we extract 60 dimensional MGC.

Implementing attention as in (Gor18)

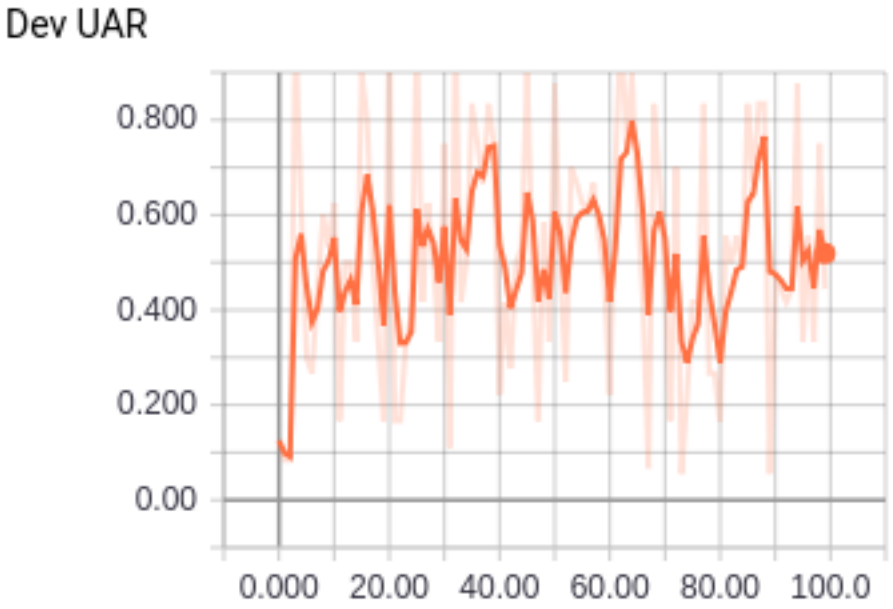

Figure 5.1: UAR plot per 100 epochs similar to figure 3 in [Gor18]