Word-Level Emotion Recognition
Using High-Level Features

Johanna D. Moore, Leimin Tian, and Catherine Lai

University of Edinburgh
School of Informatics
Informatics Forum, EH8 9AB Edinburgh, UK
J.Moore@ed.ac.uk, s1219694@sms.ed.ac.uk, clai@inf.ed.ac.uk

Abstract. In this paper, we investigate the use of high-level features for recognizing human emotions at the word-level in natural conversations with virtual agents. Experiments were carried out on the 2012 Audio/Visual Emotion Challenge (AVEC2012) database, where emotions are defined as vectors in the Arousal-Expectancy-Power-Valence emotional space. Our model using 6 novel disfluency features yields significant improvements compared to those using large number of low-level spectral and prosodic features, and the overall performance difference between it and the best model of the AVEC2012 Word-Level Sub-Challenge is not significant. Our visual model using the Active Shape Model visual features also yields significant improvements compared to models using the low-level Local Binary Patterns visual features. We built a bimodal model by combining our disfluency and visual feature sets and applying Correlation-based Feature-subset Selection. Considering overall performance on all emotion dimensions, our bimodal model outperforms the second best model of the challenge, and comes close to the best model. It also gives the best result when predicting Expectancy values.

1 Introduction

Affective Computing, the study of recognizing, understanding, and synthesising human emotions using computational technologies, has shown great potential both in academic studies of human behaviour as well as industrial applications. For example, by detecting affective states, such as boredom, an Intelligent Tutoring System can improve student learning and increase user satisfaction [1]. Multimodal emotion recognition has recently become a focus of affective computing. However, this task remains challenging, especially with respect to spontaneous spoken dialogue. Much of the early work on this topic was based on acted expressions of emotions [2], leading to models with good performance when the test and training data are similar, but which perform poorly when applied to a system working in a more natural environment. Moreover, differences in data collection and annotation style make it difficult to compare results across studies.

To address these issues, recent studies have focused on recognizing emotions in more realistic dialogues while shared tasks such as the annual Audio/Visual
Emotion Challenge (AVEC) have been held with the goal of comparing different approaches on common datasets of spontaneous speech. Despite these steps forward, the performance of existing multimodal emotion recognition models leaves much room for improvement. Predicted values from the top competitors in AVEC2012 [3], for example, exhibit relatively weak correlations for both the frame and word level subchallenges. As a regression task, the average correlation-coefficients over all test sessions for the best Fully-Continuous Sub-Challenge (FCSC) model [4] and the best Word-Level Sub-Challenge (WLSC) model [5] are 0.456 and 0.280 respectively, i.e., weak to moderate correlations. A possible reason for this poor performance is that the lexical, acoustic, and visual features often examined in these tasks are too low-level to predict emotions.

In this paper, we investigate the predictiveness of high-level features in the word-level emotion recognition task. These features include six disfluency features and locations of facial landmarks. We extracted our features from the AVEC2012 database and compared the predictiveness of our high-level features with that of the conventional lower-level audio and visual features used by the AVEC2012 WLSC baseline models. These include spectral and prosodic (SP) features and Local Binary Patterns (LBP) [6]. We compare our models to the corresponding AVEC2012 WLSC baseline models, as well as the three best performing models from the AVEC2012 WLSC. We find that our high-level features are more predictive than the low-level features, and the performance of our best bimodal model is competitive with the highest scoring models from the AVEC challenge, while using at most 22 features.

1.1 Background

Previous approaches to emotion prediction based on the AVEC data work with a high dimensional space of low-level features (1842 SP features and 5908 LBP features in the baseline model). However, the results from the top performing WLSC model [5] show that significant gains can be made by including lexical features. In this paper, we investigate whether other higher level features can be used to reduce feature space dimensionality and improve performance for this task.

Studies of both human cognition [7] and natural language processing [8] suggest that disfluencies are powerful clues for recognizing the emotional states of a speaker. Thus, disfluency features may have a stronger relationship with emotions than SP features or more general lexical features extracted from content words, and may contain less noise. Therefore, we conjecture that a unimodal emotion recognition model using disfluency features will outperform models using SP features or more general lexical features, and may contain less noise. Therefore, we conjecture that a unimodal emotion recognition model using disfluency features will outperform models using SP features or more general lexical features.

Both the best [4] and the second best [9] performing FCSC models of AVEC2012 chose high-level visual features that describe the facial expressions of the speaker using positions of facial landmark points, instead of the LBP features that describe