COMPARING FEATURE SETS FOR ACTED AND SPONTANEOUS SPEECH IN VIEW OF AUTOMATIC EMOTION RECOGNITION

Thurid Vogt[*,†], *Elisabeth André*[*]

[*]Augsburg University, Germany
Multimedia concepts and applications

[†]Bielefeld University, Germany
Applied Computer Science

{andre,vogt}@informatik.uni-augsburg.de

ABSTRACT

We present a data-mining experiment on feature selection for automatic emotion recognition. Starting from more than 1000 features derived from pitch, energy and MFCC time series, the most relevant features in respect to the data are selected from this set by removing correlated features. The features selected for acted and realistic emotions are analysed and show significant differences. All features are computed automatically and we also contrast automatically with manually units of analysis. A higher degree of automation did not prove to be a disadvantage in terms of recognition accuracy.

1. INTRODUCTION

Many features for emotion recognition from speech have been explored. However, there is still no agreement on a fixed set of features. We present a data-mining experiment where we computed a large set of acoustic features providing different views on pitch, energy and MFCC time series of the data. Then, we automatically selected from these the best subsets for given data sets. This approach is quite common in speech emotion recognition (e. g. see [1],[2],[3]), but unlike previous work we start with more than 1000 features as opposed to just a few hundred.

In view of a future online emotion recognition system we want to investigate the following questions: Does a large number of features provided to the selection algorithm enable the selection of a better feature set? What degree of automation is feasible, i. e. which analysis units and features can be calculated automatically in an online system and still yield good results? Acted and realistic emotions have been compared before (e. g. [2],[3]) but not in regard to feature sets. Therefore, the question arises of how do optimal feature sets for both types of data differ?

The steps of the feature extraction from the speech signals are described in the next section. Then we present the databases on which we performed the experiments and give our evaluation results.

This work was partially funded by a grant from the DFG in the graduate program 256 and by the EU Network of Excellence Humaine.

2. FEATURE EXTRACTION

Prosodic features that are commonly used in the literature for speech emotion recognition are based on pitch, energy, MFCCs (Mel Frequency Cepstral Coefficients), pauses, duration and speaking rate, formants and voice quality features (e. g. [1], [2],[3]). Features are derived from these measurements over a given time segment. In our approach, we compute a multitude of features and then select the most relevant ones for the given application. While this concept is followed also by others, this work is intended to be more exhaustive: instead of reducing from 100–200 features, we start with almost 1300 features.

The process of feature extraction can be divided into 3 steps: choosing a segment length, calculating features over that segment and then reducing the feature set to the most relevant ones. These steps are now explained in more detail.

2.1. Segment length

Single pitch or energy values are not meaningful for emotions, but rather their behavior over time. Therefore, normally statistics such as mean, minimum or maximum from time series of these measures are computed. Thus, the time series of values have to be segmented into chunks from which to compute the statistics. These time segments have to be chosen very carefully as they have to fulfil two conflicting conditions: 1) emotion changes can occur very quickly, but the segment length sets the temporal resolution of recognizable changes, 2) reliable statistical features can often only be computed over longer segments. To find the best trade-off we experimented with several kinds of segments.

One possibility is to use a fix segment length, e. g. 500 ms. Other units can be linguistically motivated such as words, words with context, segments delimited by pauses or whole utterances. While whole utterances usually exhibit very distinctive contours for emotional states, they are