Salient Cross-lingual Acoustic and Prosodic Features for English and German Emotion Recognition

Maxim Sidorov, Christina Brester, Stefan Ultes, and Alexander Schmitt

Ulm University, Ulm, Germany,
{maxim.sidorov,stefan.ultes,alexander.schmitt}@uni-ulm.de
Siberian State Aerospace University, Krasnoyarsk, Russia
christina.brester@sibsau.ru

Abstract. While approaches on automatic recognition of human emotion from speech have already achieved reasonable results, still there remains a lot of room for improvement. In our research, we select the most essential features by applying a self-adaptive multi-objective genetic algorithm. The proposed approach is evaluated using data from different languages (English and German) with two different feature sets consisting of 37 and 384 dimensions, respectively. The obtained results of the developed technique have increased the emotion recognition performance by up to 49.8% relative improvement in accuracy. Furthermore, in order to identify salient features across speech data from different languages, we analysed the selection count of the features to generate a feature ranking. Based on this, a feature set for speech-based emotion recognition based on the most salient features has been created. By applying this feature set, we achieve a relative improvement of up to 37.3% without the need of time-consuming feature selection using a genetic algorithm.

Keywords: emotion recognition, speech analysis, dimensionality reduction, salient features

1 Introduction

Automatic recognition of human emotions based on the speech signal is in the focus of research groups all over the world. While enabling machines to recognize human emotions may be useful in various applications, e.g., improvement of Spoken Dialogue Systems (SDSs) or monitoring agents in call-centers, the recognition performance is still not satisfying.

Emotion recognition (ER) rendered as a classification problem may be solved with supervised learning approaches by extracting a huge amount of numerical features out of the speech signal. However, the curse of dimensionality [1], i.e., having more features results in worse performance, poses a critical issue. Some of the features may be highly-correlated or their level of variability may be dramatically low. Therefore, some attributes might not bring a beneficial impact to the

system—or even decrease its performance. Hence, feature selection techniques are applied which not simply results in a trade-off between time-consuming feature extraction and the accuracy of the model. Moreover, selecting an optimal feature set potentially increases the overall performance of the system. Furthermore, as feature selection is a time-consuming procedure, it is highly desirable to have a set of salient features which is known to result in good emotion recognition performance for different settings, e.g., for different languages.

However, a suitable feature set should be both, representative and compact, i.e., result in good performance while being as small as possible. Hence, in this contribution, we propose the usage of a multi-objective genetic algorithm (MOGA), which is a heuristic algorithm of pseudo-boolean optimization, in order to maximize ER accuracy and minimize the size of the feature set simultaneously. Furthermore, we proposed a self-adaptive scheme of MOGA which exempts from the necessity of choosing the parameters of the algorithm. Eventually, our main focus lies on identifying salient cross-lingual features for emotion recognition. Hence, we analysed the distributions of selected features of three different databases (English and German) in order to identify the most salient features across different databases and languages. Finally, the resulting feature set is evaluated for each of the databases used for creating the feature set.

The rest of the paper is organized as follows: Significant related work is presented in Section 2. Section 3 describes the applied corpora and renders their differences. Our approach on automated emotion recognition using MOGA-based feature selection is presented in Section 4 having its results of numerical evaluations in Section 5. In Section 6, we analyze the cross-lingual feature set and present the list of salient features. Conclusion and future work are described in Section 7.

2 Significant Related Work

One of the pilot experiments which deals with speech-based emotion recognition has been presented by Kwon et al. [2]. The authors compared emotion recognition performance of various classifiers: support vector machine, linear discriminant analysis, quadratic discriminant analysis and hidden Markov model on SUSAS [3] and AIBO [4] databases of emotional speech. The following set of speech signal features have been used in the study: pitch, log energy, formant, mel-band energies, and mel frequency cepstral coefficients (MFCCs). The authors have managed to achieve the highest value of accuracy (70.1% and 42.3% on the databases, correspondingly) using Gaussian support vector machine.

The authors in [5] highlighted the importance of feature selection for the ER manifested by determining an efficient feature subset was using the fast correlation-based filter feature selection method. A fuzzy ARTMAP neural network [6] was used as an algorithm for emotion modelling. The authors have achieved an accuracy of over 87.52% for emotion recognition on the FARSDAT speech corpus [7].