Speaker State Recognition with Neural Network-based Classification and Self-adaptive Heuristic Feature Selection

Maxim Sidorov[1], Christina Brester[2], Eugene Semenkin[2] and Wolfgang Minker[1]
[1]*Institute of Communication Engineering, Ulm University, Ulm, Germany*
[2]*Institute of System Analysis, Siberian State Aerospace University, Krasnoyarsk, Russia*
{*maxim.sidorov, wolfgang.minker*}*@uni-ulm.de, abahachy@mail.ru, eugenesemenkin@yandex.ru*

Keywords: Speech-based emotion recognition, speech-based speaker and gender identification, neural network, genetic algorithm-based feature selection, speech corpora analysis.

Abstract: While the implementation of existing feature sets and methods for automatic speaker state analysis has already achieved reasonable results, there is still much to be done for further improvement. In our research, we tried to carry out speech analysis with the self-adaptive multi-objective genetic algorithm as a feature selection technique and with a neural network as a classifier. The proposed approach was evaluated using a number of multi-language speech databases (English, German and Japanese). According to the obtained results, the developed technique allows an increase in emotion recognition performance by up to 6.2% relative improvement in average F-measure, up to 112.0% for the speaker identification task and up to 6.4% for the speech-based gender recognition, having approximately half as many features.

1 INTRODUCTION

Machines are still quite poor at analysing human-human dialogues, meanwhile such a possibility as the automatic understanding of emotional state, gender and speakers itself might be useful in various applications, including the improvement in performance of spoken dialogue systems (SDSs) or the quality monitoring of call-centres. Such tasks as emotion recognition, gender and speaker identification (or recognition) could be conditionally named as *speaker state recognition* or as a part of a *speech analysis* procedure.

Our proposal is a usage of the multi-objective genetic algorithm (MOGA), which is a baseline heuristic method of pseudo-boolean optimization, in order to select informative features based on two metrics characterising the relevancy of the reduced data sets. In fact, the efficiency of GA applications dramatically depends on its parameters and setting reasonable parameters requires the expert knowledge of an end-user. Therefore, we also proposed here a self-adaptive scheme of MOGA, which removes the necessity of choosing the algorithm's parameters.

It turns out that by using the proposed techniques we could significantly improve the performance of such procedures as emotion recognition, and speaker and gender recognition based on speech signals.

The rest of the paper is organized as follows: Significant related work is presented in Section 2, and Section 3 describes the applied corpora and renders their differences. Our approach to improving the automated dialogue analysis is proposed in Section 4, and the results of numerical evaluations are in Section 5. The conclusion and future work are described in the Section 6.

2 SIGNIFICANT RELATED WORK

One of the pilot experiments which dealt with speech-based emotion recognition has been presented by Kwon et al. (Kwon et al., 2003). The authors compared the emotion recognition performance of various classifiers: support vector machine, linear discriminant analysis, quadratic discriminant analysis and hidden Markov model on SUSAS (Hansen et al., 1997) and AIBO (Batliner et al., 2004) databases of emotional speech. The following set of speech signal features has 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

Database	Language	Full length (min.)	File level duration		Num. of Emotions	Num. of Speakers	Num. of Gender
			Mean(sec.)	Std. (sec.)			
Berlin	German	24.7	2.7	1.02	7	10	2
SAVEE	English	30.7	3.8	1.07	7	4	1
VAM	German	47.8	3.02	2.1	4	47	2
UUDB	Japanese	113.4	1.4	1.7	4	10	2
LEGO	English	118.2	1.6	1.4	5	291	2

Table 1: Databases description.

Gaussian support vector machine.

The authors in (Gharavian et al., 2012) highlighting the importance of feature selection for the ER used the fast correlation-based filter feature selection method. A fuzzy ARTMAP neural network (Carpenter et al., 1992) 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 (Bijankhan et al., 1994).

While our research of identifying cross-lingual salient features includes more different emotions, Polzehl et al. (Polzehl et al., 2011) only focused on the emotion anger. The authors analysed two different anger corpora of German and American English to determine the optimal feature set for anger recognition. The German database contains 21 hours of records from a German Interactive Voice Response (IVR) portal offering assistance troubleshooting. For each utterance, three annotators assigned one of the following labels: not angry, not sure, slightly angry, clear anger, clear rage, and garbage. Garbage-marked utterances are non-applicable, e.g., contain silence or critical noise. The English corpus originates from an US-American IVR portal which is capable of fixing Internet-related problems. Three labelers divided the corpus into angry, annoyed, and non-angry turns. A total of 1,450 acoustic features and their statistical description (e.g, means, moments of first to fourth order, the standard deviation) have been extracted from the speech signal. The features are divided into seven general groups: pitch, loudness, MFCC, spectrals, formants, intensity, and other (e.g., harmonics-to-noise). Analysing each feature group separately, the authors achieved in a baseline approach without further feature selection a maximal $f1$ score of 68.6 with 612 MFCC-based features of the German corpus and a maximal $f1$ score of 73.5 with 171 intensity-based features of the English corpus.

3 DATABASES

For the study, a number of speech databases have been applied for the dialogue analysis. In this Section, a brief description of each corpus is provided.

The *Berlin* emotional database (Burkhardt et al., 2005) was recorded at the Technical University of Berlin and consists of labeled emotional German utterances which were spoken by 10 actors (5 female). Each utterance has one of the following emotional labels: neutral, anger, fear, joy, sadness, boredom, and disgust.

Haq and Jackson (Haq and Jackson, 2010) recorded the *SAVEE* (Surrey Audio-Visual Expressed Emotion) corpus for research on audio-visual emotion classification from four native English male speakers. The emotional label for each acted utterance is one of the standard emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral).

The *RadioS* database consists of recordings from a popular German radio talk-show. Within this corpus, 69 native German speakers talked about their personal troubles. Labelling has been done by only one evaluator at Ulm University, Germany. One of the following emotional primitives has been set as a label for each utterance: *happiness*, *anger*, *sadness* and *neutral*.

The *UUDB* (The Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies) database (Mori et al., 2011) consists of spontaneous Japanese human-human speech. The task-oriented dialogue produced by seven pairs of speakers (12 female) resulted in 4,737 utterances in total. Emotional labels for each utterance were created by three annotators on a five-dimensional emotional basis (interest, credibility, dominance, arousal, and pleasantness). For this work, only pleasantness (or evaluation) and the arousal axes are used.

Based on the popular German TV talk-show "Vera am Mittag" (Vera in the afternoon), the *VAM* database (Grimm et al., 2008) has been created at the Karlsruhe Institute of Technology. The emotional labels of the first part of the corpus (speakers 1–19) were given by 17 human evaluators and the rest of the utterances (speakers 20–47) were labelled by six annotators.

The *LEGO* emotion database (Schmitt et al., 2012) comprises non-acted American English utterances extracted from an automated bus information