Procesamiento del lenguaje Natural, nº 41 (2008), pp. 307-308 recibido 7-05-2008; aceptado 16-06-2008

ESEDA: Tool for enhanced speech emotion detection and analysis

ESEDA: una herramienta para el reconocimiento de emociones en el habla

Julia Sidorova
Universitat Pompeu Fabra
Ocata, 01 Barcelona
julia.sidorova@upf.edu

Toni Badia Cardús
Universitat Pompeu Fabra
Ocata, 01 Barcelona
toni.badia@upf.edu

Resumen: Esta demostración presenta una herramienta para el reconocimiento de emociones en el habla. Se basa en métodos estándar de aprendizadaje automático supervisado, ampliado con un módulo de análisis y mejora de errores en la clasificación. Los resultados experimetales muestran la validez de esta ampliación.
Palabras clave: reconocimiento de emociones en el habla, paralingüística

Abstract: This demo paper presents a speech emotion recognition tool, based on standard supervised machine learning methods and enhanced with an additional block of classification error analysis and fixing. Experimental results demonstrate validity of this enhancement.
Keywords: emotion recognition, paralinguistics

1 Introduction

In a number of applications such as man machine interfaces it is important to be able to recognise people's emotional state. An aim of a speech emotion recognition (SER) engine is to produce an estimate of the emotional state of the speaker given a speech fragment as an input. The standard way to do SER is through a supervised machine learning procedure. Recently a number of alternative classification strategies has been offered, which are preferable under certain conditions, e.g. unsupervised learning or numeric regression. The SER tool presented allows for these alternative classification strategies. We propose the ESEDA classification strategy, based on standard supervised learning techniques and enhanced with an additional block of classification error analysis and fixing. The achieved improvement is 12.7% of recognition accuracy averaged over all classes, and 32.1% of accuracy for the anger class.

2 System Architecture

The standard part of the system is comprised of 3 modules: Feature Selection (FS), Feature Extraction (FE) and Classification. Their performance will serve as a baseline to validate the enhancement proposed.

FE and FS In the literature there is a consensus that global statistics features lead to higher accuracies compared to the dynamic classification of multivariate time-series. The FE module extracts 116 statistical features. The FS implements the wrapper approach with forward selection. The resulting vector depends on the language, for example for the French data set in this study it had 8 features: intensity mean, harmonicity mean, long-term average spectrum value at 1500 Hz as a function of frequency, max of long-term average spectrum, frequency of minimum of the power spectral density, min of pitch, std of pitch, and mean absolute slope of pitch.

Classification The classification module takes as input the feature vector created by the feature selector, and applies the Multilayer Perceptron classifier, in order to assign a class label to it, the labels are the emotional states to discriminate among. Multilingual classifier is constructed by merging the data of several languages and further training and testing on this merged data set.

Error analysis and fixing We propose *improved classification settings* ESEDA, which is the standard classification step as described above, but enhanced with the following procedure:

I. We identify the class of special interest (denote it class I), for which the recognition rates are to be improved. It can be the worst recognised class or a class of spe-

ISSN: 1135-5948

cial interest for some application. From the confusion matrix of the standard classification step it is deduced with which other class the class of interest is most frequently confused (denote it class J). Then we classify in two steps: among the new classes (the new labels are the old labels, except that we have a joint label for the class I and class J), and then between classes I and J.

II. If the minority class problem is present and hampers the classification, we employ cost-sensitive training (more specifically, we duplicate every minority class sample in the database).

3 Experimental work

Validation of baseline performance. We did the validation on acted emotional speech from the Interface databases. Although acted material has a number of drawbacks, it was used to establish a proof of concept for the methodology proposed; for future work it is planned to test ESEDA on real emotions. There are six emotions (anger, disgust, fear, joy, surprise, sadness and neutral) from two male and female speakers. The database contains isolated words and sentences (both affirmative and interrogative) of various lengths: short (five to eight words), medium (13 w.) and long (14–18 w.). The recordings were made in a studio environment with a sampling frequency of 48 kHz and quantisation of 16 bits. A randomly chosen subset of the Interface databases was used (3711, 3805, and 4030 utterances for English, Slovenian and French respectively). The proportion of classes in the validation subset is preserved as in the whole databases.

For the testing protocol, 10-fold cross-validation was used. (We also considered disjoint sets for training (50%), validation (25%) and testing (25%) sets. We found that the accuracies in the two modes differed in 1%, which is due to the homogeneity of the Interface databases, i.e. distributions are the same in different chunks of the database. Therefore cross validation can be used without loss of generality.) In case of *monolingual validation*, the obtained accuracy is 73%. Accuracies for individual classes are as follows: 76% for neutral, 70% for angry, 94% for disgusted, 53% for fear, 83% for joy, 63% for surprise, and 72% for sad. As follows from these numbers, on average the accuracies are good, with the exception of fear (is often confused with surprise and sad) and surprise (is often confused with fear). *As for multilingual validation*, the accuracy is 69.5%.

	An	**Ne**	**Total**
Baseline	70%	76%	73.3%
+ class. step	84%	95%	76.8%
c-s. train.	99.5%	93%	86%

Table 1: The consecutive improvements in accuracies: baseline, adding the classification step between anger and neutral, adding cost-sensitive training.

Validation for the enhanced architecture. Due to improved classification settings, the system performance improved by 12% (averaging over the three languages). For example for the French database, anger was taken as a class of special interest as required in a number of applications. For example, in call centres anger detection is needed for the off-line control of how well conflict dialogues are resolved, etc. From the confusion matrix obtained with the baseline classification it was deduced that anger is mostly confused with neutral. Therefore the classification was done in two steps: among the new classes (the new labels are the old labels, except that there is a joint label for anger and neutral), and then add an extra classification step to classify between anger and neutral. The minority class problem was detected, therefore every angry sample was duplicated in the database.

4 Discussion and conclusions

Table 1. sums up the consecutive increase of classification rates. Adding an extra classification step brought the overall accuracy improvement of 3.5% (the accuracy for anger and neutral improved by 14% and 19% respectively). The cost-sensitive training brought 15.5% and 9.2% more for anger and neutral respectively. As the recognition rates improve, the false alarm rate increases only by 2% (i.e. the accuracy for the neutral class drops from 95% to 93%).

We presented a SER tool based on the ESEDA method, which is the standard supervised machine learning methods enhanced with an additional block of classification error analysis and fixing. Although this enhancement is simple from the theoretical point of view, it is of practical use.