Design and Implementation of an innovative framework for Speech Emotion Recognition

Dissertation in Human-Computer Interaction

Supervisors:
Prof. Vitoantonio BEVILACQUA
Prof. Pietro GUCCIONE
Eng. Luigi MASCOLO

Student:
Angelo Antonio SALATINO

Abstract

Automatic Speech Emotion Recognition is a very active research topic, having a wide range of applications. It can be used to detect the customers dissatisfaction in automatic remote call centres, to monitor the mental level of attention of a pilot in an aircraft cockpit, the trend of depressive symptoms in patients with mood disorders, the level of captivation skill of a teacher during a lesson in order to enhance the quality of the lesson, and in many other contexts.

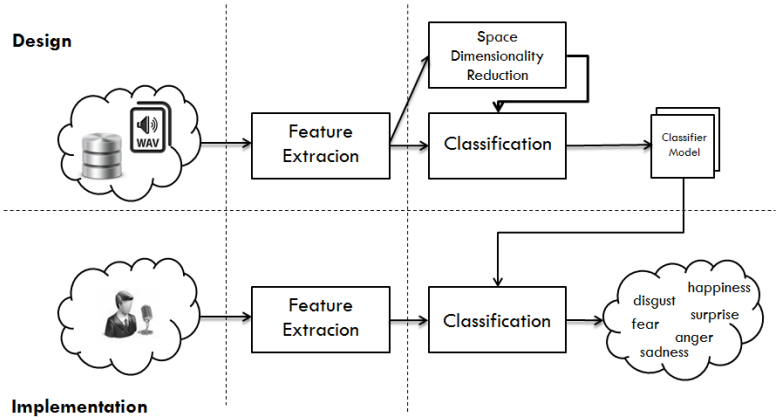

Figure 1: System Workflow.

The aim of my thesis work was a novel organization of the existing material (such as fundamental algorithms, output classes, software tools and databases) in order to provide an innovative and integrate framework for a decision support system. In particular, the building of a system able to perform a supervised learning from data and then infer emotional status.

Firstly, some emotional databases have been sought as example cases for the learning process of the system. EmoDB and DaFEx have been found that are German and Italian tongue emotional databases, respectively. They consist of several short tracks labelled by a subject-matter expert with the fundamental emotions (anger, happiness, boredom, sadness, fear, disgust, surprise and neutral).
Secondly, from those databases some prosodic and spectral features have been extracted, these include pitch, mfcc, formants, linear predictive coefficients, armonicity and zero crossing rate. All these features have been extracted frame-by-frame using Praat by means of its scripting language that allows running actions automatically and iteratively. About these features only those which are extracted from voiced frames are kept and saved inside text files. Voiced frames are those frames in which the amount of energy exceeds a dynamic threshold calculated on the input signal.
At this stage, Matlab has been used in order to compute some statistical indexes like mean, standard deviation, minimum value, maximum and range for each feature along the entire track. The calculus of statistical features has been motivated due to the high variability of frame-by-frame feature that will make the system unstable. The output of this step were two Weka compliant file, one for each database, that contain as many rows as the number of database tracks and as many columns as the number of statistical features and the emotional label.

Afterwards, many machine learning techniques have been investigated for the classification and space dimensionality reduction step. In particular, Information Gain (IG) and Svm-Recursive Feature Elimination (SVM-RFE) have been exploited to create feature rank so to retain the most significant ones and to reduce overfitting. Therefore, Artificial Neural Network (ANN) and Support Vector Machine (SVM) have been tested in order to achieve best performances. Several topologies have been tested for ANN, changing either the number of hidden layer or the number of neurons for layer. For SVM, instead, a grid search has been performed in order to find the best C and γ that are penalty and RBF kernel parameter, respectively.

After numerous tests, the performances of both ANN and SVM were quite similar. In fact, using only EmoDB the performance achieved were 86% for ANN and 88% for SVM, respectively. These performances decreased in about 10% with DaFEx due to noise in tracks because they were not recorded in an anechoic chamber as EmoDB. Another step was the merging of the two datasets where the performances stated at 80%. All the performances have been evaluated in terms of F-Measure.

Considering the obtained results, to build the final classifier and then the classifier model, SVM and 44 features have been considered.

About the implementation, C++ language with Qt Framework has been chosen. This choice has been dictated by the available libraries. In fact, with FFMPEG, every multimedia format could be read. For feature extraction Praat in combination with Gnu Scientific Library (GSL) has been used and LibSVM for the classification step.

In conclusion, both design and implementation gave satisfactory results, because the final system is able to extract emotive information from speech even if there is a room of improvements.

Acknowledgements

This study was supported by the Italian PON FIT Project called “Sviluppo di un sistema di rilevazione della risonanza (SS-RR) N° B01/0660/01-02/X17” - Politecnico di Bari and AMT Services s.r.l. – Italy.