MULTIMEDIA DATA ANALISYS FOR EMOTION RECOGNITION

Author: Fernando García Novo
Thesis Advisor: Carmen García Mateo
Departamento de Teoría do Sinal e Comunicación

MOTIVATION OF THE WORK

The motivation that led to the choice of this topic for the Doctoral Thesis is four-fold:

•The Major Depressive Disorder (MDD), is a mental disorder affecting approximately 3% of the population. Fortunately, medical studies show that the depression is curable, and early detection of depression is key for a successful treatment. Traditional approaches of depression analysis are highly dependent on the verbal reports of patients, and the mental status examination such as SANS, HRSD, BDI-II, PHQ-8, etc. Besides, they commonly require extensive human expertise and are time consuming, and therefore, very expensive. Thus, if mass detection campaigns for the detection of this disorder are to be carried out, a focus on Automatic Depression Detection (ADD) is needed.

• The MDD is a pathological emotion characterized by a pervasive and persistent low mood, and as stated above, the focus of our research.

• It is a field of research much less developed than automatic speech recognition, ADD has not been investigated until 2009.

•The rapid development in machine learning, especially ragrding Deep Neural Networks. The possibility of applying these new techniques in the field of automatic depression classification will open many lines of research with promising results.

•Increasingly, there are better available databases to study this problem.

THESIS OBJECTIVES

To develop a methodology that allows to detect the depression through multimedia data.

SECONDARY OBJECTIVES

• To detect cases of use in which depression detection has an application of interest.

• To select multimedia records that they are interesting for the resolution of the problem: voice, image, movement records, etc.

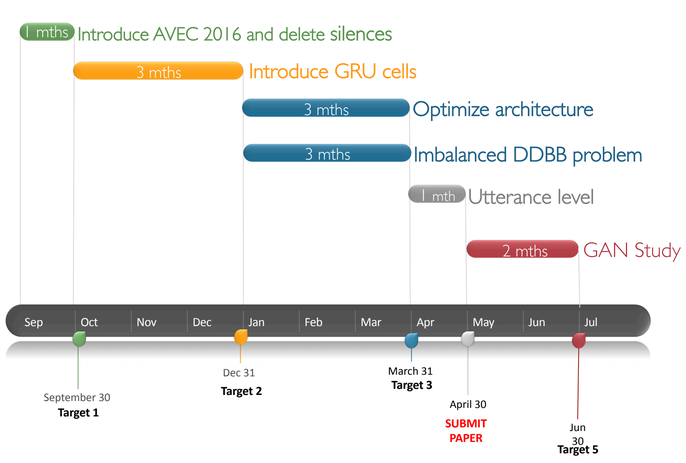
• If necessary, we will acquire a new database enabling us to get new results.

• To detect those algorithms and/or methodologies that have the best behaviour to solve the problem.

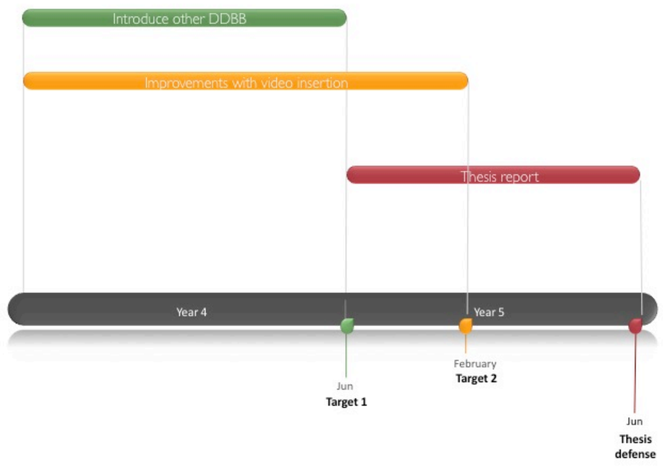
• To improve, if it is possible, the behaviour of the selected algorithms.

• To define an architecture that enables to solve these problems in real-time.

NEXT YEAR PLANNING:



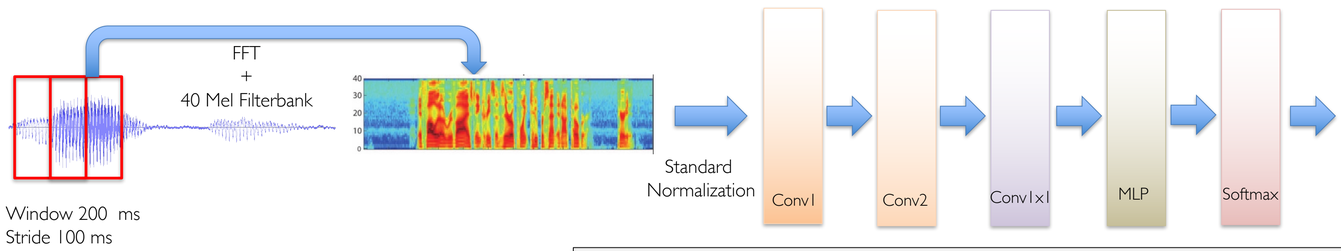
RESEARCH PLANNING:



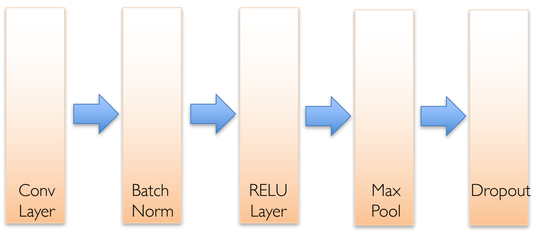
FUTURE:

• **Work with AVEC 2016:** The train and test sets don't share speakers in AVEC 2016 DDBB.
• **Delete long silences in preprocessing phase:** To improve the training process, they do not provide information.
• **Introduce GRU cells in the architecture:** To detect relationships across time in a window.
• **Optimize architecture:** To search the best architecture with CNN+RNN, we will study if there are improvements with inception or rest techniques.
• **Study techniques to compensate for the imbalanced of the DDBB:** To eliminate bias.
• **Work in utterance level not in window level:** The objective is to detect the depression of a speaker not just it in a fragment of his speech.
• **Study the use of Generative adversarial networks in semisupervised trainning:** To solve the problem of low data volume. It's very expensive to label the data.

PROPOSED ARQUITECTURE BASED IN DEEP NEURONAL NETWORKS



Conv1 and Conv2 structure:



ARCHITECTURE:

DDBB:
•80% Train, 10% Validation, 10% Test
•Unbalanced, SMOTE doesn´t work well

Conv1:
• 256 kernels, dimensions 3x3
•Max pool kernels 2x2 with stride 2x2

Conv2:
• 600 kernels, dimensions 20x5
•Max pool kernels 1x2 with stride 1x2

Conv1x1:
• dimension reduction from 600 to 256

MLP:
• RELU activation
• layer 1 9024 neurons
• layer 2 5024 neurons
• layer 3 512 neurons

Softmax:
• 46 outputs
(Number of lebels used in DDBB)

Other techniques:
•Early Stop
• Adam optimization
• Dropout: 0.7

REFERENCES

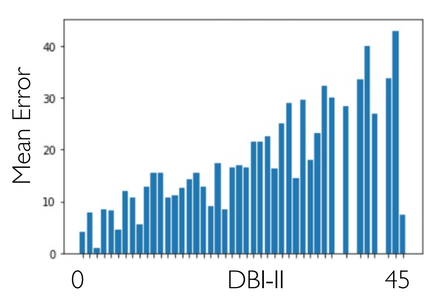
[1] Ioffe, Sergey and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." ICML (2015).
[2] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich;"Going Deeper with Convolutions", The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1-9
[3]Deep Neuronal Network and Extreme Learning Machine", INTERSPEECH (2014).
[4]. Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. 2014. AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge. In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge(AVEC '14).
[5] W. Q. Zheng, J. S. Yu and Y. X. Zou, "An experimental study of speech emotion recognition based on deep convolutional neural networks," Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on, Xi'an, 2015, pp. 827-831.doi: 10.1109/ACII.2015.7344669.
[6] Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. 2016. DepAudioNet: An Efficient Deep Model for Audio based Depression Classification. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (AVEC '16).

PRELIMINARY RESULTS

BEST INDICATORS:

MAE: 6.71
RSE: 11.61
Pearson Correlation: [0.553,2.661]
Test accuracy: 45.3%
F1: 0.45

Bias in the DDBB -> Bias in error



AVEC 2013:
• **340 videos:** Length between (50-20 min). Total duration 240 hours.
• **292 subjects:** 5 subjects appears in 4 recording, 93 in 3, 66 in 2 and 128 in 1.
• The subjects speech in **German**: Reading a book or telling a story.
•The level of depression is labelled with a single value per recording using the **BDI-II** questionnanaire.
• BDI-II contains 21 questions. The final score ranges form 0-63: 0-13 minimal depresssion, 14-19 mild, 20-28 moderate, 29-63 severe. In AVEC 2013, **the depression level of the subjects ranges from 0 to 45.**

fgnovo@gmail.com