INTELLIGENT TUTOR FOR FIRST GRADE CHILDREN’S HANDWRITING APPLICATION

Felix Albu[1], Daniela Hagiescu[2], Mihaela-Alexandra Puică[2], Liviu Vladutu[3]

[1]*Valahia University of Targoviste, Romania*
[2]*Danimated SRL, Romania*
[3]*University “Politehnica” of Bucharest, Romania*

Abstract

Although the teaching of handwriting is not compulsory in all countries, it is widely accepted that this activity highly improves young children’s personality, basic coordination abilities and communication skills. In this context, we present an intelligent tutor that evaluates not only the quality of the written symbol, but also the child’s personality and emotional state in order to adapt its teaching strategy.

In the first part of the paper, we propose a tool designed for automatic quality evaluation of handwritten symbols. We acquire the letters using a digital pen that transmits the space and time coordinates. We transform these coordinates in a binary image representation of the letter which is compared with the prototype letter. The evaluation module computes several parameters related to the legibility, size and space. An overall quality evaluation of the handwritten symbol is made.

Human communication is a combination of verbal and non-verbal interactions. Our intelligent tutor tries to follow the behavior of a teacher by assessing child’s facial expression and voice pattern. The pedagogical strategy is modified depending on the child’s interest in the handwritten application. Therefore, an important aspect of our intelligent tutoring system is the recognition of the child personality traits (based on a simple and intuitive personality test) and affective state (based on recorded speech and face information). The microphone and the camera of the system are used to collect speech and image signals.

The second part of the paper describes the approaches used to recognize child’s emotions using two modalities: recorded speech signals and face detected images. The first modality presents the features computed for the speech signal. The second modality describes the face landmarks considered for expression classification, including space coordinates for eyes, eyebrows and lips. Both proposed modalities use similar training methods, namely the multilayer perceptron neural network and the radial basis function network. Each modality identifies three types of emotions - positive, negative and neutral – and the final result is computed through decision fusion. Depending on the identified emotion and long term child’s attitude, a recommendation for a specific strategy is made. It is expected that the correct interpretation of the child’s affective state and the empathy with the animated tutor will encourage the child's interest in the calligraphy application.

Keywords: Intelligent tutor, handwriting, emotion recognition

1 INTRODUCTION

The building of an intelligent tutor that evaluates the school-aged children’s handwritten symbols and takes into account the emotions is a difficult problem. There is a worldwide debate if the typing should replace handwriting in schools, because of the widespread use of tablets, smartphones, computers etc. In many American states the teaching of handwriting skills for young pupils is not compulsory [1]. Recently, a decision has been taken in Finland to stop teaching handwriting in schools from 2016 [2]. However, it is known that the handwriting skill is useful for children in their basic coordination abilities and communication skills etc. [3]. Therefore, an automatic quality evaluation of handwritten symbols is a useful tool for assessing the calligraphy skill progress of the children [4]. It is essential that the intelligent tutor mimic the behavior of a teacher. Usually, the teacher is assessing child’s facial expression and voice pattern. There are multiple methods to assess the child interest starting from recorded voice and face detected images [5]. We propose to identify three types of emotions - positive, negative and neutral. Although most papers (e.g. [6]-[8] etc.) identify more emotions, we found that the mentioned ones are useful for our application. Depending on the identified emotion and long term child's attitude, a recommendation for a specific strategy is made. It is also known that the child personality traits are important in establishing the pedagogical strategy.

The paper is organized as follows. Section 2 presents few details about the handwritten evaluation software. In Section 3, details about the speech emotion detection method are provided. Section 4 describes the expression classification technique using face landmarks such as space coordinates for eyes, eyebrows and lips. Section 5 presents succinctly the strategy of the tutor by using a fusion of the detected emotion from speech and face detected images and the results of the handwritten evaluation module. Finally, the conclusions regarding the emotion based strategy and handwritten evaluation score are given and ideas for further improvements of the intelligent tutor are proposed.

2 HANDWRITEN EVALUATION TOOL

The handwritten evaluation software is an improved version of the method described in [9]. It relies on the two-dimensional space coordinates acquired using a tablet and a digital pen and described in [9]. We compare the handwritten symbol with the prototype symbol. It is known that the Euclidean distance between vectors is not suitable in case of handwritten letters [10]. The Dynamic Time Warping (DTW) method is intuitively connected with the way people evaluate the handwritten text [11]. Several parameters of the written character are investigated (e.g. the centre of mass of both prototype and test letters, the height over width ratio etc [9]. Usually, high DTW distances between the written letter parameter and the prototype parameters indicates alignment or written errors. Also, we determine handwritten errors by using the projection on horizontal and vertical axis of bitmapped letters. These projection vectors are compared in order to identify differences from the prototype letters. The horizontal, vertical, angle and rotational distances are computed by using the corresponding integral projection vectors [9], [12]. An example of such computed distances for 14 written "B" letters is shown in Fig. 1. The rotational distances are based on integral projection vectors of rotated binary images and are useful to identify wrongly tilted symbols. The global distance is a weighted combination of these four distances, with the highest weight being allocated to both horizontal and vertical distances. Therefore a sorting of the overall distances for a specific handwritten symbol is possible. An example is shown in Fig. 2 and it can be noticed that the left image is visually closer to the prototype image than the image form the right side.

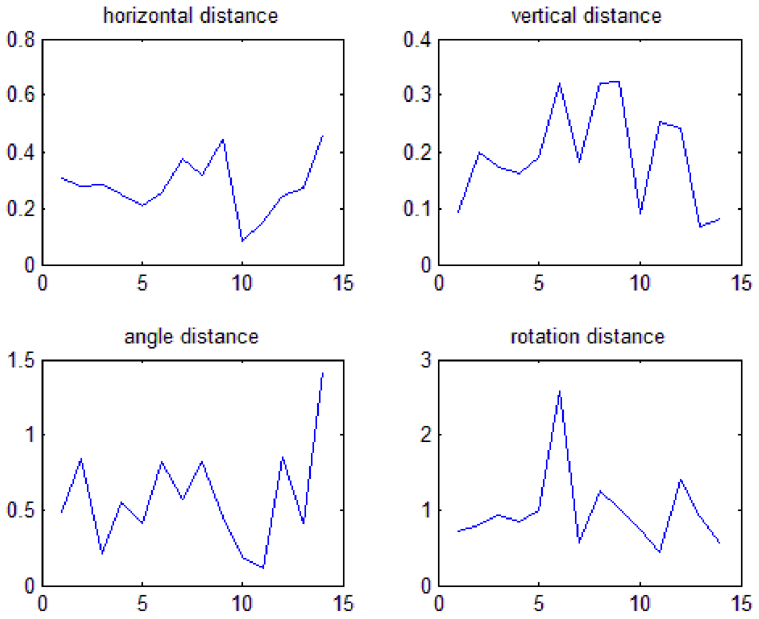


Fig. 1. DTW distance for 14 letters "B"; a) horizontal distances; b) vertical distances; c) angle distances; d) rotational distances.

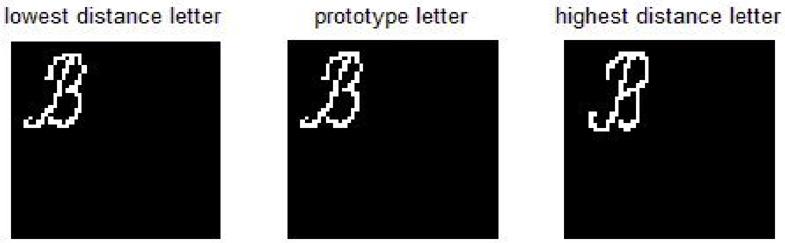


Fig. 2 The closest binary image of “B” (left); (center) prototype letter “B”; c) (right) farthest letter binary image of “B”

The total accumulated errors are calculated and compared to a threshold. If the error goes beyond a threshold, a message is displayed. Also, the mean and variance of these distances provide a measure of the child’s proficiency in replicating the prototype letter. A histogram of the DTW distances can show the specific letters where the child encounters writing problems. The sum of the division between all the DTW distances and the number of letters of the corresponding bin is computed. In the initial phases of the learning process, this computed parameter has high values; afterwards, when the child turns to a proficient writer, it becomes much smaller. An example of average and standard deviation values of the global distances for handwritten small letters is shown in Fig.3.

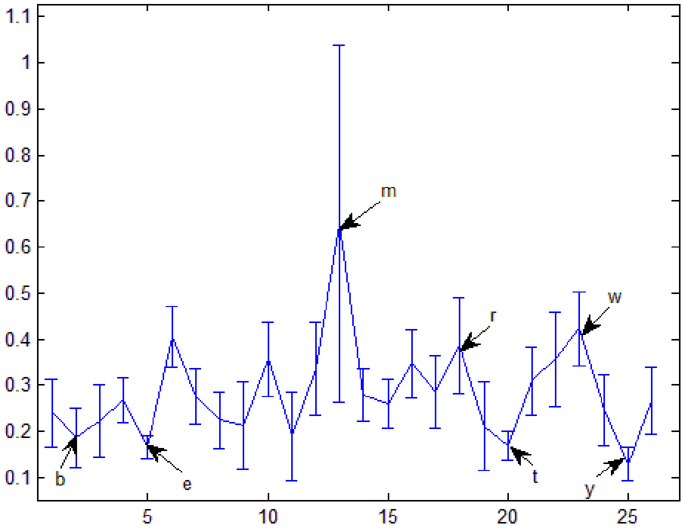


Fig. 3 Average and standard deviation values of the computed global DTW distances for 26 handwritten small letters

A small average values for one letter indicates proficiency while a small standard deviation values indicates consistency of the child in handwriting the letter. The standard deviation is shown by vertical segments centered on the DTW average value of the letter. For example, we can deduce from Fig. 3 that the letters with the smallest average values are “y”, “e”, “t” and “b”, while the letters with increasing value of the standard deviation are “e”, “t”, “y”, “b” etc. It can be easily seen that the letters where the child encounter problems are “m” followed by “w”. Our results were consistent with previous works [3], [4].

3 SPEECH EVALUATION APPROACH

3.1 Voice activity detection

The speech from the recorded signal is firstly detected. It is known that up to 40% of the speech signal can be represented by silence [11]. The silence does not bring important information regarding the emotion. The first block is the voice activity detection. The method from [13] is used and the output of such VAD system is shown in Fig. 4.

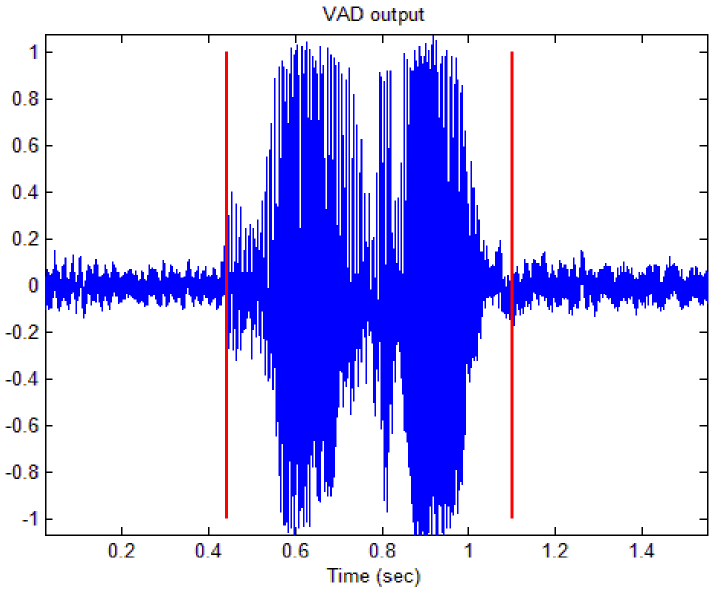


Fig. 4: VAD output

3.2 Feature selection

The mel frequency cepstral coefficients (MFCC) are frequently used for speech applications for the following reasons: they have a good accuracy for noise-less speech signals, and robustness [11].

The preprocessing step involves the use of pre-emphasis filter and background noise reduction. The pre-emphasis is made with a fixed-pole (close to one) filter. The high frequency content is improved. The signal is analyzed on 10-30 ms frames with a superposition of frames on half. The Hamming window is also applied on these frames. The scheme for computation of MFCC coefficients is shown in Fig. 5a. More details can be found in [11]. Also in Fig. 5b example of MFCC coefficients for a sentence with a neutral emotion is shown.

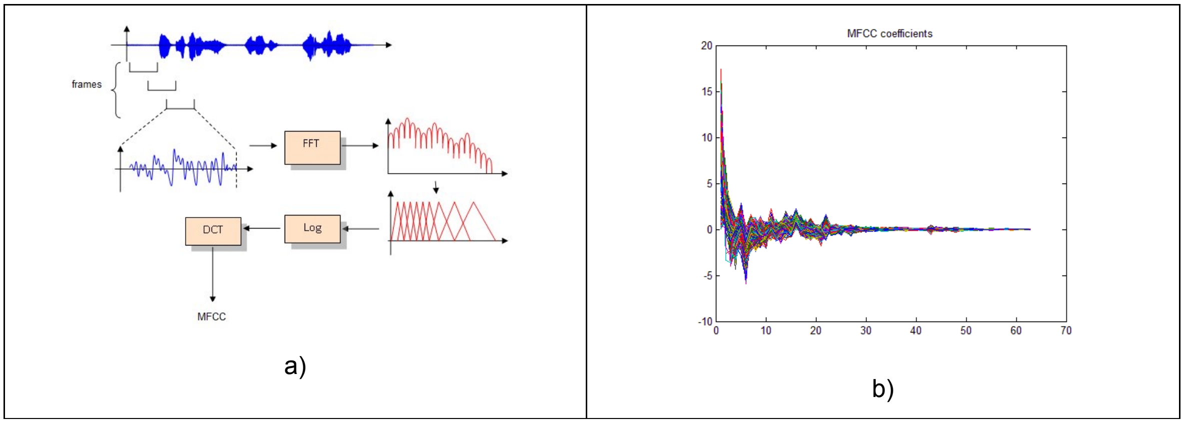
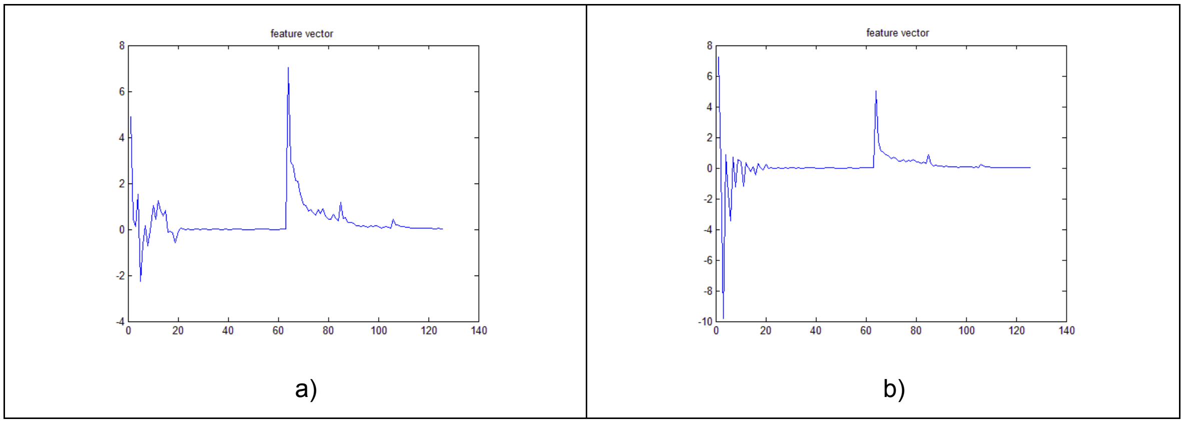


Fig. 5: a) MFCC computation scheme; b) an example of MFCC coefficients for a “neutral sentence”

The average and standard deviation is computed from each sentence and these values represent the acoustic vectors used for training and classification. Next figure shows these acoustic vectors for neutral sentence (Fig. 6a) and positive sentence (Fig. 6b) respectively.



a)

b)

Fig. 6: Example of feature vectors for different sentences. a) Neutral emotion sentence; b) Positive emotion sentence

3.3 Classification with Neural Networks

A MLP neural network with one hidden layer [14] is used to classify the acoustic vectors. The number of neurons in the hidden layer is 50, the number of output neurons is 3, each one corresponding to one of the investigated emotions (neutral, positive and negative). The step size is 0.005. The training of the network is made on the training set after a specified number of epochs. The neuron with the highest output node gives the identified emotions from the speech sentence. The classification rate a testing subset from the database of [7] is around 65%. Better results are obtained by using RBF with five centers for classification. The classification rate for RBF is about 76 % with a standard deviation of about 3%. This indicate that the RBF networks are much better than MLP for this particular classification task. Similar results were obtained using another speech emotion database [15].

4 EMOTION EVALUATION FROM FACE DETECTED IMAGES

The recorded images are from a database [16] of children's faces. From the images corresponding to 20 children (10 boys and 10 girls) we extracted the coordinates of five bounding boxes (corresponding to the eyebrows, eyes and mouth) from which 11 distances were computed. For each child, up to two images with face expressions were evaluated.

The emotional expressions to be classified are of 3 kinds:

a) interested – which corresponds to a positive expression (the images from this class correspond to 3 emotions from the above mentioned database, I.e: “happy”, “surprise” and “pleased”);

b) not-interested - corresponding to negative expression (two emotions correspond to this negative class: “disgust” and “sadness”)

c) neutral (nor boredom, sadness, happiness or surprise was detected on subject's face).

The database with 11 computed distances was used to buílld a training set and a testing set for Radial Basis Functions (RBF) Networks classifiers. The RBF networks were introduced back in 1988 by Broomhead and Lowe [17]. The distances were computed using the coordinates of the five rectangles obtained with OpenCV for the facial landmarks, in a slightly different way than in publication [18] (lower-eyebrows distances and inner-lips distances were excluded).

In our implementation of RBF networks we use two-layer feed-forward neural networks: an input layer with 11 neurons and a hidden layer with a variable number of sigmoidal units, that can learn to approximate functions. This number of units is to be determined dinamically, in order to fulfill a constant restriction (the maximum sum-squared error goal). It is a modern multi-stage implementation of RBF ([19]), which has proven in many test cases to be more intuitive and with better performances than MLP (multi-layer perceptron), the 1st type of feedforward ANN introduced. There are 3 output neurons (one for each category of emotional expression to be evaluated). The one with the maximum score (for a certain pattern at input) prevails

There are three main steps of the training process:

1. Prototype selection through k-means clustering;
2. Evaluation of the beta coefficient (the one which controls the width of the RBF neuron activation function) for each RBF neuron;
3. Training of output weights (see Figure 7) for each category using gradient descent.

Having defined as in [17], $(x^{\mu}, y^{\mu}), \mu = 1,\ldots, M$ the set of training examples, $H_{\mu j} = h_j(x^{\mu})$ the outcome of the j-th basis function, and $\mathbf{Y} = (Y_{\mu j})$ the vector of the output layer weights **W** is the result of the minimization of the error function:

$E(W) = (\|HW - Y\|)^2$. The norm in the figure 7 below is the Euclidean one.

Note: For the neurons in the hidden layer, the activation functions have a slightly modified expression:

$\Phi(x) = \exp\left(-\beta \|(x-\mu)^2\|\right)$ where β dictates the width of the RBF activation curve.

Also, table 1 below presents the best classification performances obtained using different types of feed-forward, “RBF-like” neural networks.

Table 1. Best classification results and corresponding classifiers

Crt. No.	Topology of the Neural Network	Classification Result (test set)
1.	RBF-Multi stage Neural Network [19]	94%
2.	Classic RBF-Net Matlab Toolbox	84%
3.	Probabilistic Neural Network -PNN [20]	94%

Observations:

1. Although we've got the same good classification results with PNN and Multi-stage RBF networks, the second one runs significantly faster.
2. Results obtained with MLP and RBF nets using multi-quadrics (not exponential like in the expression of Φ above), were far below the results from the table above.
3. Best results with the Matlab ANN-Toolbox (line 2 in table 1) were obtained for a sum-squared-error of 0.01, a spread-constant of 0.5 and MN = 25 (number of neurons in the hidden layer).

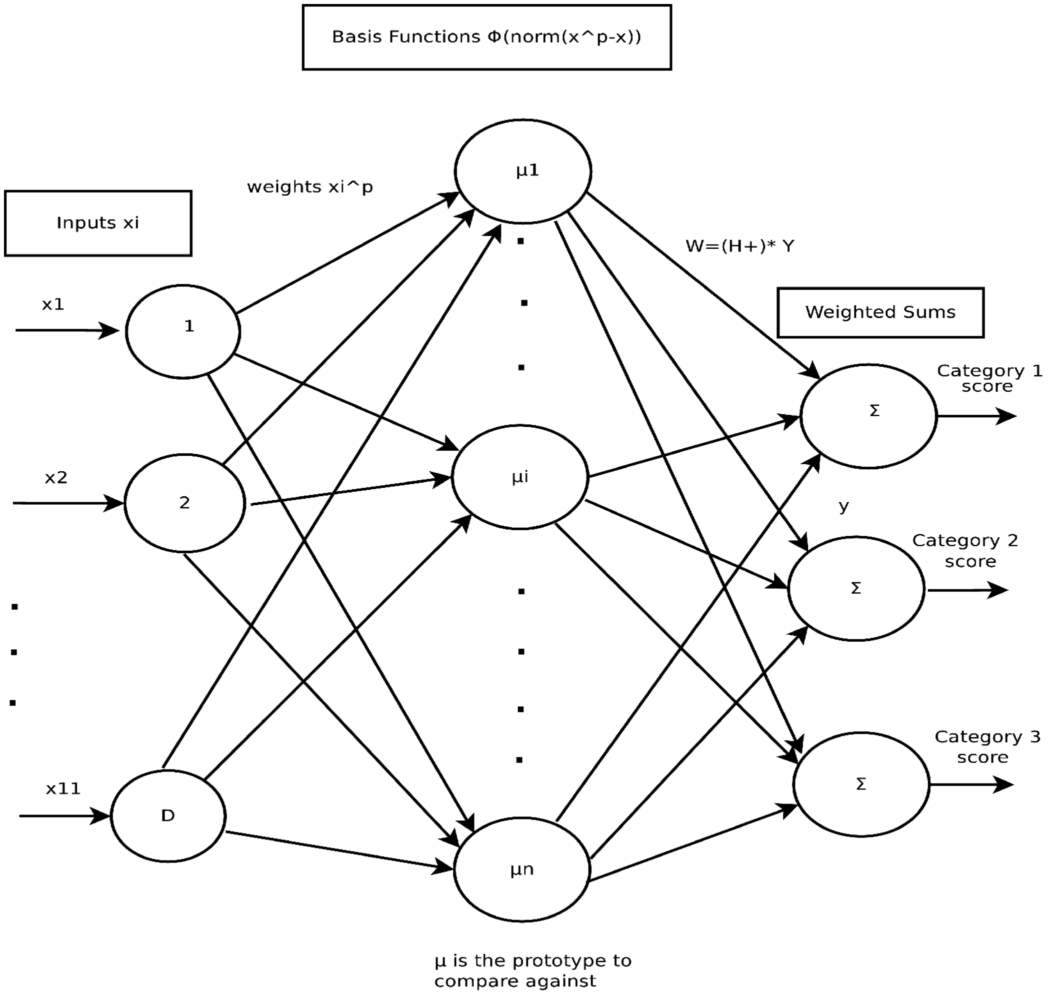


Fig7.Architecture of a RBF- Fig Neural NETWORK

5 INTELLIGENT TUTOR

The aim of building this intelligent tutoring system is to help children learn to write the Latin alphabet letters in a fun and efficient way. Therefore, we are not concerned only with the correctness of the written letters, but also with the personality and emotional state of the child.

Fig. 8 shows the proposed architecture of the intelligent tutoring system. Section 2 presented the methodology for the cognitive assessment of the lesson, while Section 3 and 4 presented the algorithms for its affective assessment. These results, together with the information about the student regarding its skills, evolution and personality are used by the tutor to make its decisions, according to a specified set of pedagogical strategies.

The personality is devised using a short initial questionnaire in which the child has to choose, between two images representing two opposing situations, the one that (s)he prefers most. The answers position the child on two axes (extrovert-introvert and psychologically stable-unstable), giving the four temperament types: choleric, melancholic, phlegmatic, sanguine.

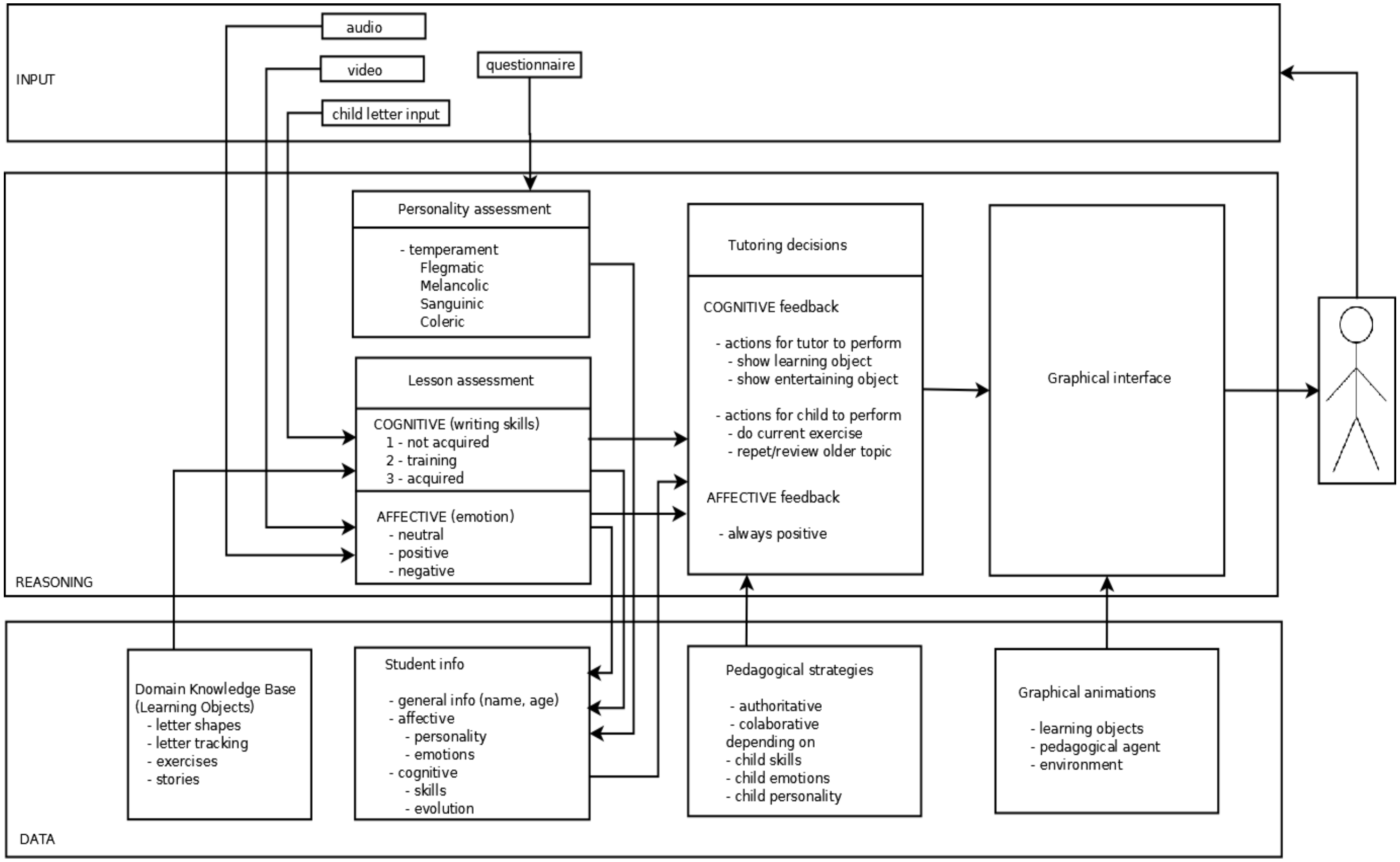


Fig. 8. Proposed architecture of the Intelligent Tutoring System

The affective state is assessed from recorded speech signals and facial expression images, as discussed in Section 3 and Section 4. As stated previously, we consider only three emotion types: neutral, positive and negative. A positive facial expression implies that the child is interested in the lesson being taught and so the tutoring session can continue with the default pedagogical strategy. A negative facial expression implies that the child does not like or does not understand the lesson, which suggests that the tutor must take into account presenting the lesson using a different approach. A neutral expression indicates that the pupil is not interested in the lesson and therefore the strategy needs to be changed in order to attract child attention.

As Sintija Petrovica notices in her work [21], there is little attention paid to the student emotional state when it comes to defining pedagogical strategies. The usual approaches for selecting the next teaching action are rule-based methods [22] and case-based reasoning [23]. In his PhD thesis [24], Samuel Alexander argues that case-based reasoning is more appropriate, since rule generation require apriori knowledge regarding the correct course of action to be taken in a given context and therefore their soundness is arguable. Moreover, the 15 production rules defined for AutoTutor [22] do not take into consideration student affect. The case-based reasoning approach uses a database of action sequences taken in turns by the tutor and the student, together with the most appropriate action to be selected when a similar sequence is encountered. Thus, this method accounts for the interaction history, not just the current situation.

The result of the decision making process is presented to the student through the graphical interface using the animations for the learning objects (e.g. show how to draw a letter) and the pedagogical agent (the embodiment that the tutor takes). An important aspect is for the tutor to always have a positive attitude, such that the child is encouraged to exercise further.

6 CONCLUSIONS

In this paper we propose an affective intelligent tutoring system that evaluates the handwritten symbols and takes into account the pupil’s emotion and personality when selecting the next teaching action. The emotion recognition is based on recorded speech signals and face detected images and the system can recognize three emotional expressions: positive, negative and neutral. These expressions are sufficient for our purpose, namely to select an appropriate teaching activity in order for the maintain child's interest for the lesson. The preliminary results on emotion detection are reasonably accurate. Our future work will be focused on improving the emotion recognition performance for various noise levels and luminance variation. More classification methods and feature vectors selection will be evaluated.

ACKNOWLEDGMENTS

The results presented in this work concerns the research carried out for the "ATIBo" research project, co-financed through the European Regional Development Fund, POSCCE-A2-O2.3.3-2011 grant, Competitiveness through Research and Technological Development and Innovation.

REFERENCES

[1] Appleby, M. (2013). An Introduction to Comparative Education*.* Education in the United States of America 177*,* pp. 1–11.

[2] Online resource: http://www.bbc.com/news/blogs-news-from-elsewhere-30146160 Last time accessed: 18-12-2014

[3] Stefansson, T., and Karlsdottir, R. (2002) “Formative evaluation of handwriting quality”, Perceptual and Motor Skills, 97, pp. 1231-1264.

[4] Falk T.H., Tam C., Schellnus H., Chau T., (2011). On the development of a computer-based handwriting assessment tool to objectively quantify handwriting proficiency in children”, Comput Methods Programs Biomed.104, (3):e102-11

[5] Haq, S. and Jackson, P.J.B. (2010). Multimodal Emotion Recognition, In W. Wang (ed), Machine Audition: Principles, Algorithms and Systems, IGI Global Press, ISBN 978-1615209194, chapter 17, pp. 398-423, 2010

[6] Slot, K., Cichosz J., Bronakowski L. (2009) Application of Voiced-Speech Variability Descriptors to Emotion Recognition. CISDA 2009: 1-5

[7] Espinosa H. P., García C. A. R., Pineda L. V. (2011). EmoWisconsin: An Emotional Children Speech Database in Mexican Spanish. ACII (2) 2011: 62-71

[8] Krishna K.K.V., Satish P.K. (2013). Emotion Recognition in Speech Using MFCC and Wavelet Features. 3rd IEEE International Advance Computing Conference.

[9] F. Albu, D. Hagiescu, and M. Puica, "Quality evaluation approaches of the first grade children's handwriting", The 10th International Scientific Conference "eLearning and software for Education" Bucharest, April 24-2, 2014, pp. 17-23.

[10] Kruskall, J. and M. Liberman, (1983) The Symmetric Time Warping Problem: From Continuous to Discrete", In Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, pp. 125-161, Addison-Wesley Publishing Co., Reading, Massachusetts.

[11] Deller J.R., Proakis J. G. and Hansen J. H.L., (1993) Discrete Time Processing of Speech Signals, Macmillan, New York.

[12] Albu F., Florea C., Zamfir A., Drimbarean A. (2008).Low Complexity Global Motion Estimation Techniques for Image Stabilization", in Proc. of ICCE 2008: 465-467

[13] Giannakopoulos, T., & Pikrakis, A. (2014). Introduction to Audio Analysis: A MATLAB® Approach. Academic Press.

[14] Haykin, S. (1998). Neural Networks: A Comprehensive Foundation (2 ed.). Prentice Hall.

[15] Database of Polish Emotional Speech, available: http://www.eletel.p.lodz.pl/bronakowski/med_catalog/ (Accessed 21.01.2015).

[16] Dalrymple, K.A., Gomez, J., & Duchaine, B. (2013). The Dartmouth Database of Children's Faces: Acquisition and validation of a new face stimulus set. PLoS ONE, 8(11): e79131.

[17] Broomhead, D., & Lowe, D. (1988). Multivariable functional interpolation and adaptive networks. Complex Systems, 2, 321-355

[18] Puică, M.A., Mocanu,I., Florea A.M., Agent-based system for affective intelligent environment. In Filip Zavoral, Jason J. Jung, and Costin Badica, editors, Proceedings of the 7th International Symposium on Intelligent Distributed Computing (IDC 2013), volume 511 of Studies in Computational Intelligence, pp. 335–342. Springer, 2013.

[19] Schwenker, F., Kestler, H. A., Palm, G., Three learning phases for radial-basis-function networks, Neural Networks, 14 (2001) 439-458

[20] Specht, D.F., Probabilistic Neural Networks, Neural Networks, Vol. 3, 109-118, 1990.

[21] Petrovica, S. Design of the Pedagogica Module for an Emotionally Intelligent Tutoring System, Science – Future of Lithuania, Vilnius Gediminas Technica University, 2014 6(3): pp.138-146.

[22] Graesser, A.C., Person, N.K., Harter, D., & The Tutoring Research Group. Teaching Tactics and Dialog in AutoTutor, Artificial Inteligence in Education, 2001 12(3): pp. 257-279.

[23] Sarrafzadeh, A., Alexander, S., Dadgostar, F., Fan, C., Bigdei, A. (2008). How do you know that I don't understand? A look at the future of intelligent tutoring systems, Computers in Human Behavior, 2008 24(4): pp.1342-1363.

[24] Alexander, S. An affect-sensitive intelligent tutoring system with an animated pedagogical agent that adapts to student emotion like a human tutor, PhD thesis, Massey University, Albany, New Zealand, 2007.