ELECTRONICS AND ELECTRICAL ENGINEERING
ISSN 1392 – 1215 *2011. No. 3(109)*
ELEKTRONIKA IR ELEKTROTECHNIKA

SYSTEM ENGINEERING, COMPUTER TECHNOLOGY
T 120
SISTEMŲ INŽINERIJA, KOMPIUTERINĖS TECHNOLOGIJOS

The Utilization of Feedback and Emotion Recognition in Computer based Speech Therapy System

O. A. Schipor, S. G. Pentiuc, M. D. Schipor
"Stefan cel Mare" University of Suceava,
str. Universitatii nr 13, RO-720229 Suceava, Romania, e-mail: schipor@eed.usv.ro

Introduction

A Computer Based Speech Therapy System (CBST) is a complex with both hardware and software architecture, which main role is to assist human speech and language pathologist (SLP). On the European level, there is a long-standing interest in develop and improve this kind of system due to the high number of potential beneficiaries [1].

Dyslalia is one of the most frequent speech disorders SLP work with and it affects the pronunciation of one or many sounds. As a result, a child suffering of dyslalia has difficulty in both of the communication and of the social integration [2].

Since 2006 our team has been developing a CBST adapted for dyslalia assessment in Romanian language - Logomon. Our first challenge was to implement and put together all classical required modules like: *Children Manager*, *3D Articulator Model*, and *Homework Manager* (installed on the child's PC or PDA) [3]. After the system was finished and tested we had tried to improve this classical architecture in order to archive a better similarity between human and artificial SLP.

The automatic generation of the exercises to be done by the child using a *fuzzy expert system* was a big step forward. The role of this new module was to suggest *optimal therapeutic actions* for each child (number, length and content of training sessions), based on specific information (tests' scores and social, cognitive and affective parameters) [4].

In this paper we present two new strategies that could make CBST's more "human like". First of these refers to a real time and a real environment feedback. The second involves the identification of emotional state base on the specific parameters of the audio stream.

We are not going to prove the possibility of replacing the SLP with the CBST. Instead of that, we believe that CBST's could become a better assistance in speech therapy process [5].

Real time feedback and emotion recognition in speech therapy

Feedback provided automatically by intelligent interface involves important advantages such as *objectivity* (e.g. SLP alone is not able to identify child's real progress because he tends to get used to the child) and *portability* (e.g. the development of a home training program).

Verbal behavior of the subject with pronunciation problems depends on the context: *controlled* environment (e.g. interaction with SLP or CBST) and *free* environment (e.g. family discussions, interaction with friends group, etc.) [6]. The controlled environment implies higher speech self-control than the free one. Therefore, the child verbal behavior during speech therapy sessions is better than the child verbal behavior manifested in social groups.

That is why in this paper we present opportunities for extension of traditional CBST's as providing feedback in free environments. Main challenges of this extension are related with the recording quality (noise, dialogue, several simultaneous voices) [7, 8].

Speech quality depends also of the subject emotional condition. Some paraverbal parameters such as speed, intensity, timbre and fluctuation of the speech provide evidence concerning the subject's affective state and they also can be used in order to identify emotional disorders that increase speech problems. These parameters can be transformed in bio-feedback and would be integrated in speech therapy process [9].

Intelligent interface can take two types of actions when some undesirable emotional state is identified:

- Reactive approach - The self adaptation behavior, in order to avoid some future situations that generate unwanted emotions;
- Proactive approach - The counseling of the child, in order to understand and solve his own affective state.

Free environments provide a wider range of emotional expressions and that is why we are concerned about the emotion recognition in this type of contexts [10].

Speech confusability - measuring hidden markov model similarity

A wide range of possibilities exist for parametrically representing the speech signal for the speech recognition task, such as: Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), and others [11].

Since MFCC simulates best the behaviour of human auditory apparatus, it became the most popular approach in speech recognition. MFCC is based on known variation of the human ear's critical bandwidth with frequency [12]. A block diagram of the structure of MFCC algorithm is illustrated in Fig. 1.

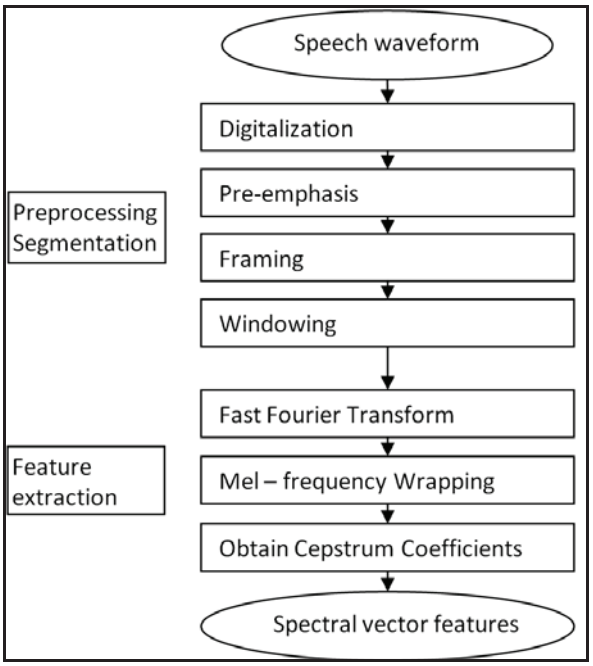

Fig. 1. Speech waveform feature extraction

The pre-emphasis block receives the speech signal $s_1(n)$ and sends it to a high-pass filter (1). The output signal of this block – $s_2(n)$ – is obtained using the following formula

$$s_2(n) = s_1(n) - a * s_1(n-1). \quad (1)$$

The coefficient – a – is the parameter to control the degree of pre-emphasis filtering and takes values between 0.9 and 1.0.

Speech is a highly non-stationary signal. That is why, speech analysis, must be carried out on short segments across which the speech signal is assumed to be stationary. Typically, the feature extraction is performed on 20 to 30 ms windows with 10 to 15 ms shift between two consecutive windows [13].

Each frame will be next multiplied with a hamming window (2-3). The aim of this step is to minimize the signal discontinuities at the beginning and end of each frame:

$$s_3(n) = s_2(n) * w(n), \quad (2)$$

$$w(n) = (1-\alpha) - \alpha\cos(2\pi n/(N-1)). \quad (3)$$

The parameter α usual value is between 0 and 0.5 and determines the attenuation degree performed by the window. The samples total number in each frame is denoted by N.

In order to convert each frame of N samples from time domain to frequency domain, is used The Fast Fourier Transformation. The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. Therefore, the powers of the spectrum have to be mapped onto the mel-scale, using triangular overlapping window (4). The mel scale is a perceptual one because the mel frequency is proportional to the logarithm of the linear frequency

$$Mel(f) = 2595 * \log_{10}(1 + f/700). \quad (4)$$

The transformation of mel spectrum into time domain is performed using Discrete Cosine Transform - DCT (5)

$$X_k = \sum_{n=0}^{N-1} x_n * \cos[\pi/N * (n+1/2) * k]. \quad (5)$$

The acoustic vectors (i.e. MFCC) can be used in three different scenarios, illustrated in Fig. 2.

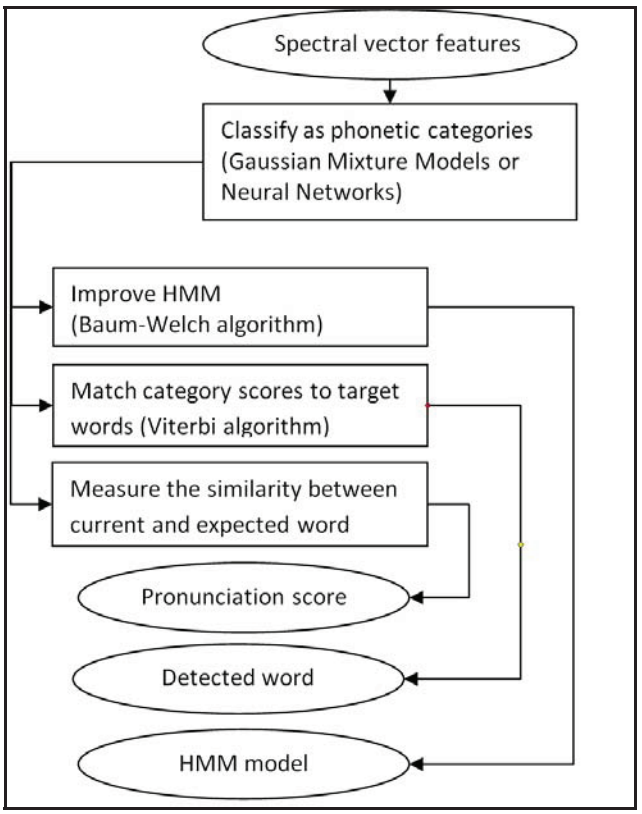

Fig. 2. Acoustic vectors utilization

A Hidden Markov Model – HMM is specified by a five-tuple (6):

- S – set of hidden states;
- O – set of observation symbols;
- Π – the initial state distribution;
- A – the state transition probability distribution;
- B – the observation probability distribution.

$$\begin{cases} \lambda = \{S, O, \prod, A, B\}, \\ S = \{1,2,\ldots,N\}, \\ O = \{o_1, o_2, \ldots, o_M\}, \\ \prod = \{\pi_i\}, \pi_i = P(s_0 = i), 1 \le i \le N, \\ A = \{a_{ij}\}, a_{ij} = P(s_t = j \,|\, s_{t-1} = i), 1 \le i, j \le N, \\ B = \{b_j(k)\}, \\ b_j(k) = P(o_k \,|\, s_t = j), 1 \le j \le N, 1 \le k \le M. \end{cases} \quad (6)$$

The first scenario refers to the *HMM model training* using the Baum-Welch algorithm. In this approach, we start with an approximate model λ_1 and a set of observations O and then we obtain an improved model λ_2. We generally want to choose parameters that will maximize the likelihood of data. For each HMM model, can be calculated the following score

$$Score(\lambda_i) = \sum_{k=1}^{T} P(O_k \,|\, \lambda_i). \quad (7)$$

The second scenario is used for the *best HMM model recognition* (i.e. phonemes or words) based on observations (i.e. acoustic vectors). The Viterbi algorithm picks and remembers the best path for a given HMM model (8). Subsequently, it is chosen the model with highest probability

$$V_t(i) = P(X_1^t, S_1^{t-1}, s_t = i \,|\, \lambda). \quad (8)$$

The third scenario assumes that HMM model and acoustic vectors are already known. We first establish what the subject has to say so we focus on a specific HMM model. Then the subject pronounces the indicated word and the system generate the correspondent acoustic vectors. Finally, the system compute the probability for a certain model in order to generate those observations.

There are two fundamental differences between speech recognition (i.e. second scenario) and pronunciation evaluation (i.e. third scenario) [14]:

- In the first case, the system does not know what the subject will say while in the second case, the next pronunciation word is established by the system;
- In the first case, a correct pronunciation provides a good recognition rate while in the second case, provides a good score.

Towards a new CBST model

In Fig. 3 there is a scheme presenting the main modules of our CBST architecture [4] and the relation between them and *Speech and Emotion Recognition Engine*.

It can be seen that almost all components (i.e. Child, Logopaed, Home Monitor, Lab Monitor and Expert System) can be improved by speech and emotion recognition capabilities.

The verbal stream is taken from the child using Lab Monitor Application or Home Monitor. The results are transformed into bio-feedback for the subject or/and they are recorded and transmitted afterwards to the logopaed.

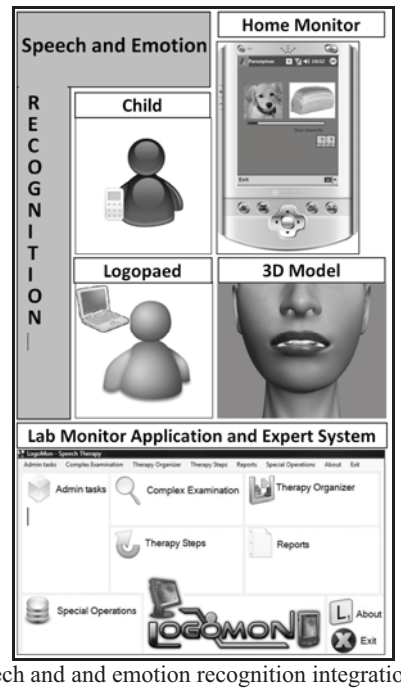

Fig. 3. Speech and and emotion recognition integration

Expert system might be extended to deal with personalized emotions profile because each child has his own personality [15]. Using this extended approach a CBST system could become a better SLP assistant. It cannot replace SLPs, but it could facilitate their assessment of speech by helping them to better target therapeutic intervention.

Conclusions

In this paper we indicate the two most appropriate ways of extending a CBST system. The first, relates to providing real-time quality feedback in free environments. Due to this extension, the system could evaluate child's pronunciation in most varied context such as family or friends group.

The second possible improvement concerns in the emotions recognition. There is a strong relation between child's affective state and the pronunciation quality. Therefore, we consider our new CBST architecture being a step forward.

Acknowledgements

This paper was supported by the project „Progress and development through post-doctoral research and innovation in engineering and applied sciences– PRiDE - Contract no. POSDRU/89/1.5/S/57083", project co-funded from European Social Fund through Sectorial Operational Program Human Resources 2007-2013.

References

1. **Tobolcea I.** Modern Audio-visual Techniques Used in the Treatment of Logoneurosis (in Romanian). – Romania: Ed. Spanda, Ia□i, 2001. – 206 p.
2. **Pentiuc S. G., Tobolcea I., Schipor O. A., Danubianu M., Schipor M.D.** Speech Therapy Programs for a Computer Aided Therapy System // Electronics and Electrical

Engineering. – Kaunas: Technologija, 2010. – No. 7(103). – P. 87–90.
3. **Zaharia M. H., Leon F.** Speech Therapy on Expert System // Advances in Electrical and Computer Enginering. – University of Suceava, 2009. – No. 1(9). – P. 74–77.
4. **Schipor O. A., Pentiuc S. G., Schipor M. D.** Improving Computer Based Speech Therapy Using a Fuzzy Expert System // Computing and Informatics. – Slovak Academy of Sciences, 2010. – No. 2(29). – P. 303–318.
5. **Pentiuc S. G., Tobolcea I., Schipor O. A.,Danubianu M., Schipor M. D.** Translation of the Speech Therapy Programs in the Logomon Assisted Therapy System // Advances in Electrical and Computer Enginering. – University of Suceava, 2010. – No. 4(10). – P. 48–52.
6. **Jensen J. H., Ellis D. P. W.** Quantitative Analysis of a Common Audio Similarity Measure // IEEE Transactions on Audio, Speech, and Language Processing. – IEEE, 2009. – No. 4(17) – P. 693–703.
7. **Paola A., Gaglio S., Lore G., Ortolani M.** An ambient intelligence architecture for extracting knowledge from distributed sensors // Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, Seoul, Korea, 2009. – P. 104–109.
8. **Rudžionis A., Ratkevičius K., Rudžionis V.** Speech in Call and Web centers // Electronics and Electrical Engineering. – Kaunas: Technologija, 2005. – No. 3(59). – P. 58–63.
9. **Polzehl T., Sundaram S., Ketabdar H., Wagner M, Metze F.** Emotion Classification in Children's Speech Using Fusion of Acoustic and Linguistic Features // Proceedings of INTERSPEECH-2009, Brighton, U.K., 2009. – P. 340–343.
10. **Broek E. L., Schut M. H. Author A. B., Else S., More S.** Unobtrusive Sensing of Emotions (USE) // Journal of Ambient Intelligence and Smart Environments. – IOS Press, 2009. – No. 1(3) – P. 287–299.
11. **Paulikas Š., Karpavičius R.** Application of Linear Prediction Coefficients Interpolation in Speech Signal Coding // Electronics and Electrical Engineering. – Kaunas: Technologija, 2007. – No. 8(80). – P. 39–42.
12. **Yang B., Lugger M.** Emotion recognition from speech signals using new harmony features // Signal Processing. – Elsevier, 2010. – No. 5(90) – P. 1415–1423.
13. **Kemesis P., Ridzvanavicius J., Stasiunas A.** Speech Perception Analyzer // Electronics and Electrical Engineering. – Kaunas: Technologija, 1998. – No. 3(16). – P. 12–15.
14. **Chen J. Y., Olsen P. A. Hershey J. R.** Word confusability - measuring hidden Markov model similarity // Proceedings of INTERSPEECH'2007. – Antwerp, Belgium, 2007. – P. 2089–2092.
15. **Schipor D. M., Pentiuc S. G., Schipor O. A.** End-User Recommendations on LOGOMON - a Computer Based Speech Therapy System for Romanian Language // Advances in Electrical and Computer Enginering. – University of Suceava, 2010. – No. 4(10). – P. 57–60.

Received 2010 09 29

O. A. Schipor, S. G. Pentiuc, M. D. Schipor. The Utilization of Feedback and Emotion Recognition in Computer based Speech Therapy System // Electronics and Electrical Engineering. – Kaunas: Technologija, 2011. – No. 3(109). – P. 101–104.

In this paper we present two appropriate ways for the Computer Based Speech Therapy (CBST) improvement. Real-time quality feedback in free environments and emotion recognition could produce a better similarity between human and artificial speech therapist. Using these extended approaches, a CBST system can become a better speech therapist assistant and that is why we intend to implement these technology improvements on our CBST system - Logomon. Ill. 3, bibl. 15 (in English; abstracts in English and Lithuanian).

O. A. Schipor, S. G. Pentiuc, M. D. Schipor. Grįžtamojo ryšio ir emocijų taikymas kompiuterinėje logopedinėje programoje // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2011. – Nr. 3(109). – P. 101–104.

Pateikti du būdai, kaip patobulinti kompiuterinę logopedinę programą. Realaus laiko kokybiškas grįžtamasis ryšys ir emocijų atpažinimas gali padidinti žmogaus ir kompiuterinėje programoje taikomo dirbtinio intelekto tarpusavio bendravimo efektyvumą. Taikant šias priemones galima gerokai patobulinti jau naudojamą kompiuterinę programą Logomon. Il. 3, bibl. 15 (anglų kalba; santraukos anglų ir lietuvių k.).