ELECTRONICS AND ELECTRICAL ENGINEERING
ISSN 1392 – 1215 2012. No. 6(122)
ELEKTRONIKA IR ELEKTROTECHNIKA

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

Automatic Assessment of Pronunciation Quality of Children within Assisted Speech Therapy

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

crossref http://dx.doi.org/10.5755/j01.eee.122.6.1813

Introduction

Assistive speech therapy is a general term that includes assistive, adaptive, and rehabilitative techniques for people with pronunciation disorders. One of the most important and provocative characteristics of a Computer Based Speech Therapy System (CBST) is to provide real time feedback – an ability traditionally reserved for humans (i.e. Speech and Language Therapists – SLT). That is why, together with emotion recognition, automatic assessment of pronunciation quality is considered to be the key for reducing the gap between traditional and computer assisted speech therapy [1].

Our researches on assisted therapy of speech disorders has been started since 2005, when we designed Logomon – the first CBST for Romanian language. This software consists of four modules as follows: *Monitor Program* (management of therapy), *3D Articulator Model* (indicates the correct positioning of cheeks, lips, teeth, and language for different sounds), *Homework Manager* (extends speech therapy outside the logopaedic clinic), and *Fuzzy Expert System* (suggests the parameters of personalized therapy) [2, 3].

The importance of automatic pronunciation feedback in speech disorders has been highlighted by numerous studies [4]. In addition with the utilisation of the software without the presence of a SLT – *portability*, there is another main advantage of this approach – *objectivity* (i.e. the opportunity to obtain scores that are not influenced by the subjectivity). Together, these benefits justify the investments in intelligent interfaces that are able to perform real time analysis of speech production and to react accordingly.

In this paper we focus on phoneme-level scoring of pronunciation of preschooler and young-schooler with dyslalia – a speech disorder that involve the mispronunciation of one or many sounds. Each step of the process is presented (i.e. acquisition of data, human scoring, Hidden Markov Models – HMM training and classification) and the results (i.e. the performances of our system) are summarized in a separate section.

Literature review

Although they seem similar, the speech recognition and the pronunciation scoring are different research field within speech processing. In the first case, the system has to find the best match (e.g. between an uttered word and a dictionary) and a good recognition rate requires a good quality of pronunciation. In the second case, the next user' utterance is indicated by the system and a similarity measure between received and expected speech sequence is calculated [5].

Also, discrimination between pronunciation scoring for speech therapy (CBST) and for language instruction (Computer Assisted Language Learning - CALL) has to be performed [6]. While in the second case an overall speaker-level score is usually preferred, in the first case a phoneme-speaker-level is more appropriate due to localisation of pronunciation problems to specific sounds. However, besides this subtle difference, both CBST and CALL require the same techniques and the obtained results are comparable.

Research on automatic assessment of pronunciation quality has been carried out for phoneme, word, sentence, and speaker level [7, 8]. Moreover, there are systems designed for a specific language or that can be adapted for several language and also speaker-independent or speaker-dependent systems [7, 9].

Several features of input speech can be combined in order to obtain a number that reflect the quality of pronunciation. In the case of text-independent scenario (i.e. the system don't know what will be the next user's utterance), only parameters of input utterance are available and, in consequence, only general acoustic features such as duration, energy, rhythm, pauses length and pitch frequency can be used [10, 11]. On the other hand, in a usual speech therapy scenario the subject repeat after teacher's voice so the system is text-dependent. In this situation both an input utterance and an acoustic model (e.g. HMM) are available and more powerful features can be calculated (e.g. phoneme model likelihood, posterior probability of phonemes) [12].

Using a hybrid measure for phoneme quality (e.g. the average logarithm of the frame-based posterior probability and the normalized logarithm of the phoneme-based posterior probability) in a text-dependent system could lead to a correlation coefficient between scores provided by the human raters and scores provided by the machine up to 0.85 [13].

These performances require however a large phonemes database, that are not currently available for Romanian language. For example, in the case of a common phoneme-level mispronunciation detection system, a phonetically transcribed database of 130,000 phones uttered in continuous speech sentences by 206 speakers was used [14].

Another approaches to detect and measure phoneme mispronunciations is based on SVMs (Support Vector Machines) as classifier and log-likelihood ratios calculated from HMMs as feature vectors [15, 16]. Several variations on this theme could be obtained using one or two HMM models (trained with native or/and non-native speaker data) and selecting different types of features. Further, the scores obtained from HMMs act as input data for a secondary classifier.

As far as we know, there is no CBST able to perform mispronunciation detection and scoring for Romanian language. The age of the subjects and the pronunciation specificities caused by the dyslalia are another two important points that make us believe that this research is a step forward.

Methods

Data Acquisition. In order to obtain correct and mispronunciation utterance, we have recorded 60 children from Regional Logopaedic Center of Suceava, Romania. We turn our attention on three consonants (R /r/, S /s/, Ş /ʃ/) that are, according to statistics, the most frequent mispronunciated phonemes in Romanian language. These phonemes was uttered in different acoustic context (i.e. position in utterance and neighboring phonemes), as is showed in Table 1. The entire utterance corpus *U* consisted of 3551 items including 1428 correct pronounced phonemes and 2123 mispronounced phonemes.

Table 1. Acoustic context of tracked phonemes

name	example for phoneme /r/
isolate	/r/
before a vowel	/ra/, /rə/, /re/, /ri/, /ro/, /ru/
after a vowel	/ar/, /ər/, /er/, /ir/, /or/, /ur/
between vowels	/ara/, /ere/, /iri/, /oro/, /uru/
consonant combinations	/pra/, /vre/, /bri/

These items were recorded using an original method [17] so that several requirements to be meet:

- Minimal impact on child behaviour;
- The speech therapist's voice has to be ignored;
- After recording is necessary to easily (quasi automatic) split the stream into utterances.

Because of using a hand-switch that open/close microphone electrical circuit, the recorded stream contains only child phonemes separated from a "silence" zone. Moreover, each "silence" zone begins and ends with an easily to detect "marker", produced from switch-hand press/release, as it is shown in the bottom area of Fig. 1.

Human Scoring. In order to establish the pronunciation quality for each utterance, three acoustic experts (i.e. speech and language therapists) were involved in our research. Each of them rate each utterance with an integer score: 0 – *unintelligible*, 1 – *intelligible but poor*, 2 – *moderate*, 3 – *good*. In addition, without being informed, an expert could randomly receive the same utterance twice so that an internal consistency to be checked. If they meet external and internal consistency, then the human assessment results can be seen as a "benchmark" for automatic scoring of pronunciation quality.

In order to compute the similarity between each human expert's evaluations and the others' assessments, we have used a modified correlation coefficient (1) (2):

$$C_k = \frac{\sum_{i=1}^{n}\left((x_k)_i - \mu_{X_k}\right)\left((\bar{x}_{j\neq k})_i - \mu_{X_{j\neq k}}\right)}{\sqrt{\sum_{i=1}^{n}\left((x_k)_i - \mu_{X_k}\right)^2 \sum_{i=1}^{n}\left((\bar{x}_{j\neq k})_i - \mu_{X_{j\neq k}}\right)^2}}, \quad (1)$$

$$(\bar{x}_{j\neq k})_i = \frac{\sum_{j\neq k}^{m}\left((x_j)_i\right)}{m-1}, \quad (2)$$

where *m* – number of human raters; *j, k* – index of a specific rater; *n* – number of utterances; *i* – index of a specific utterance; X_k – the set of scores indicated by rater *k;* $(x_k)_I$ – the score indicated by rater *k* for utterance *I;* μ_{Xk} – the mean of scores from the set X_k; $X_{j\neq k}$ – the sets of scores indicated by „non *k*" raters; $(\bar{x}_k)_i$– the average score indicated by raters "non *k*" for utterance *I;* $\mu_{Xj\neq k}$ – the mean of scores from the sets $X_{j\neq k}$.

The consistency within raters was measured using the classic correlation coefficient (i.e. Pearson product-moment correlation coefficient) between the scores of utterances that had been evaluated twice by the same expert.

The results have shown that there is a high correlation both between each evaluator and the others and between scores associated by the same rater for the same utterance. However, detailed data are placed in Results section.

Automatic Pronunciation Scoring. In Fig. 1 is presented the architecture of automatic pronunciation scoring system. In automatic pronunciation scoring it is assumed that HMM model (3) and acoustic vectors are already known (Fig. 1). We first establish what the subject has to say so we focus on a specific HMM acoustic model. Then the subject pronounces the indicated word and the system generate the correspondent acoustic vectors (Preprocessing Segmentation and Feature Extraction). Finally, the system computes the probability that the model to generate those observations. This probability can be interpreted as a similarity measure between the model and the observation (i.e. utterance).

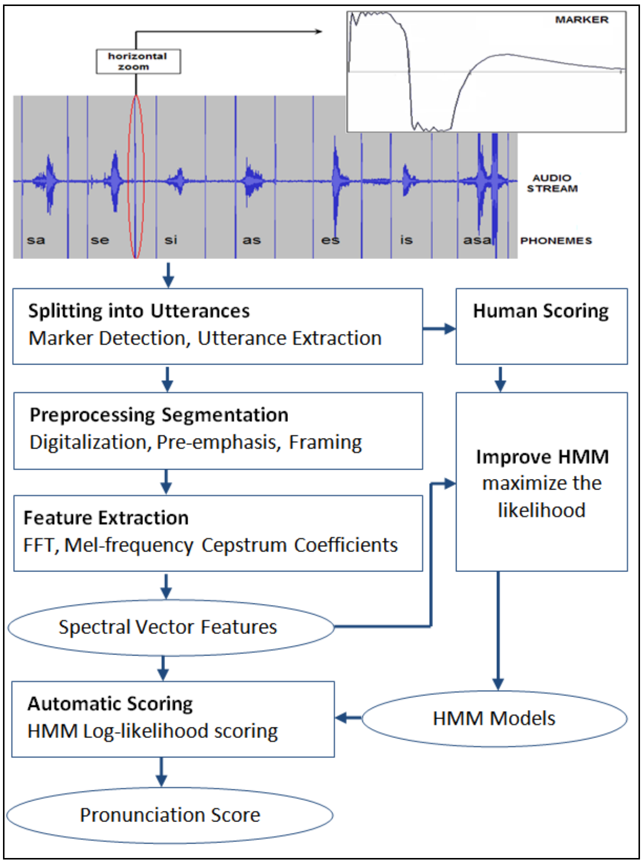
horizontal zoom
MARKER
AUDIO STREAM
sa
se
si
as
es
is
asa
PHONEMES
Splitting into Utterances
Marker Detection, Utterance Extraction
Human Scoring
Preprocessing Segmentation
Digitalization, Pre-emphasis, Framing
Improve HMM
maximize the likelihood
Feature Extraction
FFT, Mel-frequency Cepstrum Coefficients
Spectral Vector Features
Automatic Scoring
HMM Log-likelihood scoring
HMM Models
Pronunciation Score

Fig. 1. The architecture of automatic pronunciation scoring system

For each utterances class C_i a reference acoustic model (benchmark) is obtained based on correct pronunciation phonemes and using an iterative Forward-Backward algorithm (i.e. Baum-Welch)

$$\lambda_i = \{S, O, \prod, A, B\}\,, \qquad (3)$$

where i – index of utterance class;   – 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.

The number of hidden states of each model was assigned based on number of constituent phonemes in order to account for spectral dynamics (coarticulation). Because of this "coarticulation", states are sometimes context dependent which means that the same phoneme pronunciation is dependent on neighboring (preceding and/or following) phonemes.

The observations sets were obtained using 10 ms chunks and cepstrum coefficients were calculated for each frame.

The normalized (i.e. independent from the duration of utterance) log-likelihood score of utterance u from utterance class C_i is defined as

$$L_u = [\sum_{t=t_0}^{t_0+d-1} \log p(o_t \mid \lambda_i)]/d\,, \qquad (4)$$

$p\ (o_t \mid \lambda_i)$ – probability of current frame as observation vector o_t; d – number of frames of utterance; t_0 – index of current frame.

Results and discussions

The results obtained for both human and automatic scoring are presented in Table 2. First column contains all five utterances classes and an overall (i.e. average) evaluation. In the next columns, are written correlation coefficients for inter-rater, intra-rater and human-machine evaluation.

Table 2. Inter-rater, intra-rater and human/machine correlations across utterances classes

consistency type **utterance class**	**inter-rater**	**intra-rater**	**human-machine**
isolate	0.71	0.80	0.63
before a vowel	0.76	0.84	0.53
after a vowel	0.75	0.84	0.52
between vowels	0.76	0.86	0.50
consonant combinations	0.79	0.87	0.61
overall	**0.76**	**0.84**	**0.54**

First of all, the overall correlation coefficients reveal that there are significant differences between the three types of evaluations. As we expected, there is a relative strong relation between the scores indicated by the same human expert for the same utterances (column number three) and a significant weaker relation in the case of scores indicated by all raters (column number 2). However, the second result (i.e. 0.76) can be seen as the expected "upper bound" on the performance of an ideal automatic scoring system.

Referring to the differences between correlation coefficients for utterances classes in the case of human evaluation, we found that the accuracy of evaluation depends on the duration/length of spoken items (the same patterns were obtained for all the three sounds we build our research on.). This may be explained by the fact that human raters are used to longer speech productions that offer valuable contextual information.

Not the same thing we can say about automatic evaluation where, as it is presented in column number four, the correlation coefficients seems to vary inversely proportional to the utterances' length. The explanation has to do with the differences between human and machine "perception". While human raters focused on consonants (affected phonemes) and score all utterance based on this segment, automatic system scores items as a whole and, in the case of utterance that contains vowels (usually less susceptible to misprocunciation), computes higher scores. This fact is supported by the "break" of above mentioned inverse proportionality relation in the case of *consonant combinations*.

Conclusions

In this paper we focus on human and automatic scoring of pronunciation of children with speech disorders. Consequently, we present both theoretical and practical related issues such as: acquisition of data, human scoring, Hidden Markov Models training and classification, and the performances of our system.

The relative small utterance corpus has led to a relatively low level of correlation between automatic and human evaluation. However, taking into consideration difficulties related with low age of children, with environmental noise and with particularities implied by speech disorders, we consider our findings 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: Spanda Press, Iaşi, 2001. – 206 p.
2. **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.
3. **Zaharia M. H., Leon F.** Speech Therapy Based on Expert System // Advances in Electrical and Computer Enginering. – University of Suceava, 2009. – No. 1(14). – P. 74–77.
4. **Pentiuc S. G., Tobolcea I., Schipor O. A., Danubianu M., Schipor D. M.** Translation of the Speech Therapy Programs in the Logomon Assisted Therapy System // Advances in Electrical and Computer Enginering. – University of Suceava, 2010. – No. 2(10). – P. 48–52.
5. **Schipor O. A., Pentiuc S. G., Schipor M. D.** 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. DOI: 10.5755/j01.eee.109.3.181.
6. **Levy M.** Computer–assisted language learning. – Clarendon Press, 1997.
7. **Cincarek T., Gruhn R., Hacker C., Noth E., Nakamura S.** Automatic pronunciation scoring of words and sentences independent from the non–native's first language // Computer Speech & Language. – Elsevier, 2009. – No. 1(23). – P. 65–88.
8. **Dimitrakakis C., Bengio S.** Phoneme and Sentence–Level Ensembles for Speech Recognition // Journal on Audio, Speech, and Music Processing 2011.
9. **Levi S. V., Winters S. J., Pisoni D. B.** Speaker–independent factors affecting the perception of foreign accent in a second language // Journal of Acoustical Society of America, 2007. – No. 4(121). – P. 2327–2338.
10. **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.
11. **Kemesis P., Ridzvanavicius J., Stasiunas A.** Speech Perception Analyzer // Electronics and Electrical Engineering. – Kaunas: Technologija, 1998. – No. 3(16). – P. 12–15.
12. **Lileikyte R., Telksnys L.** Quality Measurement of Speech Recognition Features in Context of Nearest Neighbour Classifier // Electronics and Electrical Engineering. – Kaunas: Technologija, 2012. – No. 2(118). – P. 9–12. DOI: 10.5755/j01.eee.118.2.1165.
13. **Dong B., Fengpei G., Fuping P., Shui–duen C.** Automatic Scoring of Pronunciation Quality with Hybrid Measure // Proceedings of International Conference on Artificial Intelligence and Computational Intelligence. – IEEE, 2009. – P. 381–384.
14. **Franco H., Bratt H., Rossier R., Gadde V. R., Shriberg E., Abrash V., Precoda K.** EduSpeak®: A speech recognition and pronunciation scoring toolkit for computer–aided language learning applications // Language Testing. 2010. – No. 3(27). – P. 401–418.
15. **Wei S., Hu G., Hu Y., Wang R.** EduSpeak®: A new method for mispronunciation detection using Support Vector Machine based on Pronunciation Space Models // Speech Communication. 2009. – No. 10(51). – P. 896–905.
16. **Yoon S., Hasegawa M., Sproat R.** Automated Pronunciation Scoring using Confidence Scoring and Landmark–based SVM // Proceedings of Intespeech. – Brighton, UK, 2009. – P. 1903–1906.
17. **Schipor O. A., Schipor M. D., Nestor M., Pentiuc S. G.** Automatic Parsing of Audio Records of Children with Dyslalia // SAACS–07. – Iasi, Romania, 2007. – P. 267–270.

Received 2012 03 19
Accepted after revision 2012 05 12

O. A. Schipor, S. G. Pentiuc, M. D. Schipor. Automatic Assessment of Pronunciation Quality of Children within Assisted Speech Therapy // Electronics and Electrical Engineering. – Kaunas: Technologija, 2012. – No. 6(122). – P. 15–18.

In this paper we present our results in automatic evaluation of pronunciation quality of children with dyslalia (mispronunciation of specific phonemes). Our aim is to offer real-time, quality feedback so that to reduce the gap between human assisted and artificial speech therapy. We present both theoretical and practical related issues such as: acquisition of data, human scoring, Hidden Markov Models training and classification, and the performances of our system. The obtained results encourage us to continue the development of Logomon – the first computer based speech therapy system for Romanian language. Ill. 1, bibl. 17, tabl. 2 (in English; abstracts in English and Lithuanian).

O. A. Schipor, S. G. Pentiuc, M. D. Schipor. Automatinis vaikų tarsenos kokybės vertinimas pagalbinio kalbėjimo terapijoje // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2012. – Nr. 6(122). – P. 15–18.

Pateikiami vaikų, turinčių dislaliją (klaidingas specifinių fonemų tarimas), automatinio tarties kokybės įvertinimo rezultatai. Tikslas – pasiūlyti kokybišką realaus laiko grįžtamąjį ryšį mažinant trukmę tarp pagalbinės (žmogaus) ir dirbtinės kalbos terapijos. Pateikiami teoriniai ir praktiniai duomenų rinkimo, vertinimo, paslėptų Markovo modelių mokymo ir klasifikavimo bei sistemos našumo aspektai. Gauti rezultatai skatina tęsti darbus kuriant Logomoną – pirmąją kompiuterinę kalbos terapijos sistemą rumunų kalbai. Il. 1, bibl. 17, lent. 2 (anglų kalba; santraukos anglų ir lietuvių k.).