A Computer-Assisted Prosody Pronunciation Teaching System

Dávid Szthahó*, Gábor Kiss*, László Czap**, Klára Vicsi*

*Budapest University of Technology and Economics
Department of Telecommunication and Media Informatics
{sztaho, kiss.gabor, vicsi}@tmit.bme.hu

**University of Miskolc
Department of Automation and Infocommunication
czap@mazsola.iit.uni-miskolc.hu

Abstract

Work in the last decade shows, that Computer-Assisted Pronunciation Teaching (CAPT) systems are useful, flexible tools for giving pronunciation instructions and evaluating at subject’s speech. This paper describes a newly developed CAPT system that intends to address appropriate teaching of such supra-segmental parameters as intonation, stress and speech rhythm. Two modules are implemented: (1) intonation and stress teaching, and (2) rhythm teaching using dynamic time warping. The automatic feedback of the system is evaluated by using speech samples from hard of hearing children. The automatic assessment methods give automatic feedback that is consistent with the subjective decisions of teachers. Visual feedback was also proposed which is based on the dynamic time warping algorithm and gives simple and understandable visualization of the intonation and rhythm of the subject’s utterance.

Index Terms: speech prosody, intonation, speech recognition, speech aid

1. Introduction

Work in the last decade shows that the Computer-Assisted Pronunciation Teaching (CAPT) systems are useful, flexible tools for giving pronunciation instructions and evaluating at subject’s speech. However, pronunciation teaching has many issues. There are pedagogical, technological questions. From a technological viewpoint it is hard to provide understandable, accurate feedback to pronunciation mistakes. It is almost impossible to give a one hundred percent accurate and automatic diagnosis of speech production errors.

In [1] the authors state that the applications of CAPT systems need to fulfill three main necessary requirements: (1) provide enough understandable information with accurate articulatory instructions, (2) give goals and motivation to the subjects to follow the role-oriented courses and realistic practices, and (3) ensure immediate useful feedback, mostly about functions related to speech intelligibility. During the teaching process supra-segmental functions, such as intonation, pronunciation duration, stress, must be involved [3].

Bradlow et al. [5] showed that development in perception leads to increase in production. Hirahata [4] shows that the opposite is also true: development in production leads to increase in perception.

CAPT visual feedback is mostly realized using display of artificial face [6, 7], waveforms, pitch, spectrograms [8, 9, 10] and automatic speech recognition (ASR) [11]. This helps in recognizing words and understanding temporal information. From a pedagogical point of view the usage of waveforms and spectrograms is contentious. Understanding of these figures takes too much time, and is not worth the effort. However, spectrograms can help emphasize the intensity change and duration of syllables in speech. Visualization of pitch is easier and more comprehensible. It is still not clear which type of visual feedback is the most conductive to intelligibility. There is also no agreement in how prosody should be measured and taught.

The automatic diagnosis of pronunciation errors is an expected feature of a CAPT application, but a general, one hundred per cent accurate diagnosis system is not realistic. However, it is required that the feedback should correlate with the human assessment.

Earlier, in the framework of the European SPECO project (Contract no. 977126), a product-oriented multilingual CAPT system was developed for speech handicapped persons, utilizing visually displayed acoustic properties of speech, which are correct from an acoustic phonetic point of view, but easy to understand and interesting for children. [12, 13, 14]. The system uses mainly the segmental description of speech.

This paper describes a newly developed CAPT system that intends to address appropriate teaching of such supra-segmental parameters as intonation, stress and speech rhythm. The software is intended for use in the teaching of prosody to children with a hearing impairment or with a cochlear implant, but the system will also be useful for Computer-Assisted Language Learning purposes. It tries to implement an intelligible, useful and successful learning method with visual and automatic feedback that helps to maintain the motivation of children. Two modules are implemented: (1) intonation and stress teaching, and (2) rhythm teaching. The automatic feedback of the system is evaluated by using speech samples from hard of hearing children. The automatic answer of the system was compared with subjective assessment of teachers.

The system is still using the Hungarian databases for the training of the Hungarian prosody, but it is easy to adapt to other languages.

The paper is structured as follows. In Section 2 the general goal of the proposed CAPT application is described, followed in Sections 3 and 4 by detailed description of evaluation methods, automatic and visual feedback. In Section 5 the output of the CAPT application is compared to the subjective assessment of teachers using real sound material of hard of hearing children.

2. General functionality of the CAPT application

The proposed CAPT application has two modules. One addresses automatic feedback for intonation, that is how the fundamental frequency changes over time. In Hungarian,
sentence modalities are expressed mostly by typical sentence intonation curves (fundamental frequency curves). Children who have issues producing correct intonation must learn how to intone these modalities, otherwise their speech is much less intelligible. The list of the typical sentence intonation curves of modalities included in the application is shown in Table 1.

The other module of the application deals with speech rhythm. With its help, the subjects (such as hard of hearing children) can develop their pronunciation by receiving automatic evaluation of their timing mistakes. The assessment of rhythm is based on two parameters: (1) vowel durations, and (2) temporal distance of neighbouring vowels. An automatic speech segmentation method [15] is used with force alignment in order to specify the exact timing information of the spoken phonemes.

### Table 1. List of typical intonation curves of sentence modalities

<table>
<thead>
<tr>
<th>Type of intonation</th>
<th>Type of sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descending</td>
<td>declarative sentence</td>
</tr>
<tr>
<td>Falling</td>
<td>question to be complemented</td>
</tr>
<tr>
<td>Ascending-falling</td>
<td>yes-no questions</td>
</tr>
<tr>
<td>Falling-descending</td>
<td>imperative and exclamation sentences</td>
</tr>
<tr>
<td>Floating</td>
<td>clauses (not closing)</td>
</tr>
<tr>
<td>Rising</td>
<td>one word questions</td>
</tr>
</tbody>
</table>

### 3. Automatic assessment of intonation

In order to allow effective learning of correct intonation, a proper fundamental frequency calculation method, an automatic assessment and visual feedback must be implemented. The following sections describe the methods used in the proposed CAPT application.

The fundamental frequency was calculated using autocorrelation for every 10 ms with 100 ms calculation window length. The resulting pitch curve was filtered for octave jumps and was smoothed with mean filtering.

The correctness of the intonation depends on the direction of change in speech melody. Therefore, the automatic assessment must consider only the information of the derivative of the intonation curve parts, not the absolute value.

The proposed algorithm for the automatic assessment of intonation compares the subject’s speech sample to a reference intonation curve of an Etalon speaker. We want to show the direction of change in speech melody. Therefore, the sentences were warped linearly. Then the algorithm splits the reference intonation curve into parts, according to time points, where sudden change occurs in fundamental frequency function. A time point of sudden change \( t_{\text{change}} \) is defined by local maxima of the second derivative of the fundamental frequency function.

The algorithm compares the direction of changes in the parts of the two speech samples (subject’s and reference) defined by the \( t_{\text{change}} \) time points of the reference speech sample. The direction of the speech signal part \( i \) is given by

\[
\text{dir}_i = \text{sign} \left( \sum_{t_{i-1}}^{t_{i+1}} \frac{\partial f(t)}{\partial t} \right) \frac{t_{i+1} - t_i}{t_{i+1} - t_i + 1},
\]

where \( t_i \) is the time point \( i \) of sudden change of the pitch curve. An intonation curve part of the subject’s utterance is defined as correct if the \( \text{dir}_i \) of the paired fundamental frequency curve parts (subject’s and reference) are equal, otherwise it is defined as incorrect. The final intonation score is given by the ratio of the correct and incorrect intonation parts.

### 4. Rhythm

The correct rhythm is a primary requirement of fluent speech. Too long or short phoneme durations can reduce intelligibility or result in complete unintelligibility. Therefore learning proper phoneme timing is very important. In the proposed CAPT system two timing information assessments are implemented: (1) duration of vowels, and (2) time interval lengths between vowels. An automatic speech segmentation method [15] extended with force alignment method was applied to determine the temporal boundaries of phonemes.

#### 4.1. Automatic speech segmentation method with force alignment

A general goal of ASR systems is to recognize continuous speech. Therefore, they are not designed to locate exact timing positions of phonemes, which is a primary requirement in any speech teaching application. In [15] a recognition system is proposed which, instead of phonemes, can determine phonemes grouped into categories with different acoustic characterization. These are: low vowel, high vowel, voiced and unvoiced resonant consonants, spirant and stops. These categories, with some simplifications, fit all European languages, therefore this classification system is almost considered a language independent method. In order to locate exact timing information force alignment is applied with a prior knowledge of the sequence of spoken phonemes.

The main goal of the proposed force alignment is to find the position of phonemes in speech signal on the basis of a known phoneme sequence. The known phoneme sequence is first transcribed into the related phoneme categories, because the recognizer deals with phoneme categories according to acoustic characteristics. Let \( h_S = h_{h_1}^{h_n} \) be a sequence of probabilities where each \( h_{h_1}^{h_n} \) element is the probability of the occurrence of the paired fundamental frequency curve parts \( h_1 \) to \( h_n \) and \( n \) is the total number of phonemes in the known phoneme category sequence. If we consider all possible combinations of \( f_0, f_1, ..., f_n \in \mathbb{J} \) time indexes with the constriction \( 0 \leq f_0 \leq f_1 \leq ... \leq f_n \leq \tau_{\text{max}} \) the goal is to find the \( S \) sequence with the highest probability:

\[
S_{\text{best}} = \arg \max_S h_S = \arg \max_S h_{h_1}^{h_n} \cdot h_{h_2}^{h_n} \cdot ... \cdot h_{h_{n-1}}^{h_n}.\]

The \( h_{h_n}^{h_{n-1}} \) probabilities are calculated by:

\[
h_{h_n}^{h_{n-1}} = \prod_{i=t_{h_n}}^{t_{h_{n-1}}} a_{h_n,h_i},
\]

where \( h_i \) is the recognized phoneme class at time \( t \) by the segmentation method in [15] and \( a_{h_n,h_i} \) is a priori state transition probability from phoneme category \( h_n \) to \( h_i \).

After the force alignment process, the original phoneme sequence is rewritten based on the best phoneme category sequence. The accuracy of segmentation within a given tolerance is shown in Figure 1 in the function of tolerance.
4.2. Automatic assessment of rhythm

Based on the temporal locations of phonemes that are generated by the segmentation method, two types of measurement score are calculated that are addressed to measuring the rhythm of the subject’s speech sample. Both scores use reference phoneme durations obtained from the Hungarian Reference Speech Database [16].

The first measurement score is based on vowel durations. For every vowel \( v_m \) of the subject’s speech sample the following \( sc_{vm} \) score is calculated:

\[
s_{cvm} = \begin{cases} 
1 & \text{if } d_{\text{min}} \leq d_{vm} \leq d_{\text{max}} \\
0 & \text{if } d_{vm} < w \cdot d_{\text{min}} \text{ or } d_{vm} > \frac{d_{\text{max}}}{w} \\
1 - \frac{d_{\text{max}} - d_{vm}}{d_{\text{max}}/w} & \text{if } d_{\text{max}} < d_{vm} < \frac{d_{\text{max}}}{w} \\
1 - \frac{d_{vm} - d_{\text{min}}}{d_{\text{min}} \cdot w} & \text{if } d_{\text{min}} \cdot w < d_{vm} < d_{\text{min}} 
\end{cases}
\]

where \( d_{vm} \) is the duration of vowel \( v_m \); \( d_{\text{min}} \) and \( d_{\text{max}} \) are respectively the minimum and maximum duration measured in the reference database for \( v_m \). \( w \) is a constant (between 0 and 1) for setting the strictness of the assessment (set to 0.8). The final score for vowel durations for a speech sample is given by

\[
s_c = \frac{\sum v_s sc_{vm}}{n_v}, \quad (5)
\]

where \( n_v \) is the total number of vowels.

The second measurement score is based on the time interval lengths between vowels. A similar score is calculated as (4) for each interval using the consonants that are placed between the vowels. For each interval the following score is given by

\[
s_{cm} = \begin{cases} 
1 & \text{if } d_{\text{min}} \leq d_{cm} \leq d_{\text{max}} \\
0 & \text{if } d_{cm} < w \cdot d_{\text{min}} \text{ or } d_{cm} > \frac{d_{\text{max}}}{w} \\
1 - \frac{d_{\text{max}} - d_{cm}}{d_{\text{max}}/w} & \text{if } d_{\text{max}} < d_{cm} < \frac{d_{\text{max}}}{w} \\
1 - \frac{d_{cm} - d_{\text{min}}}{d_{\text{min}} \cdot w} & \text{if } d_{\text{min}} \cdot w < d_{cm} < d_{\text{min}} 
\end{cases}
\]

where \( d_{cm} \) is the duration of the interval between vowel \( v_m \) and \( v_{m+1} \). \( d_{\text{min}} \) and \( d_{\text{max}} \) are the minimum and maximum duration measured in the reference database for all consonants in the interval. \( w \) is the same constant as in (4). The final score for interval lengths between vowels for a speech sample is given by

\[
s_c = \frac{\sum c_s sc_{cm}}{n_c}, \quad (7)
\]

where \( n_c \) is the total number of intervals between vowels.

5. Evaluation of automatic assessment

The automatic assessment scores were evaluated using real speech samples from 19 hard of hearing children. Each child read three sentences from all categories. Recordings were made in “Török Béla” School of Hard of Hearing Children in Budapest. 48 sentences (randomly selected, evenly distributed among the categories listed in Table 1) were subjectively assessed by three teachers according to correctness of pronunciation, classifying them into three categories: good, medium and bad, with sample number 10%, 40% and 50% of the database respectively. The teachers were experts in the field of pedagogy of speech impaired children. The same sentences were recorded with healthy children, where all speech samples were assessed as good. The subjective decisions were compared to the automatic assessment measurements scores described in previous sections (intonation, with the intonation of healthy children used as reference; vowel durations; time interval lengths between vowels) in two ways: the (a) lowest and the (b) average of the three types of automatic scores were calculated and compared to the subjective assessment. Figure 2 shows the relation between the subjective and the automatic assessment.

The scores that were given by the CAPT application correspond to the three subjective categories. Although both assigned values (lowest and average of automatic scores) are distinct in the three subjective categories, the lowest of the automatic scores represent a stricter judgment, which is more in line with the subjective assessment of the teachers. This implies that one incorrect parameter of the pronunciation determines the correctness of the total perception.

6. Visual feedback

Proper visual feedback is a general requirement of any CAPT application. It must maintain the attention and motivation of the subject and must be simple, but effective in order to teach.

Two types of visual feedbacks are proposed: comparison of intonation curves, and rhythm of the reference and subject’s speech sample. Because the two speech samples are different in duration, a dynamic time warping method is applied. The rhythm scores are calculated by (4) and (6) for each vowel and interval between vowels. For each vowel that (4) gives 1.0, the duration is set to that of the vowel in the same position in the reference. Similarly, for each interval between vowels that (6) gives 1.0, the interval length is set to that which is found in the same position in the reference. If the value of (4) and (6) are different from 1.0, the temporal values of the vowels and intervals are left unchanged. The result of this is that if the speech sample of the subject is assessed as good, the temporal parameters of his or her recording will be the same as the reference. Otherwise, if incorrect rhythm is detected, the exact
position of the mistake will be apparent to the subject. Differently assessed utterances are displayed in Figure 3.

![Visual feedback of rhythm](Image)

Figure 3. Visual feedback of rhythm. (a) good rhythm; (b) example for incorrect vowel duration and interval length between vowels. The temporal positions of vowels are represented by rectangles (phonemes noted by Sampa characters). In each figure, the upper line contains the vowels of the reference speech sample, the bottom line contains the rhythm of the subject’s speech sample. The place and direction of mistakes in pronounced rhythm are marked with arrows.

The visual feedback of the intonation is based on the information of how the intonation curves change over time. The intonation curves are normalized by subtracting the average of the first three fundamental frequency values for each curve.

![Visual feedback of intonation](Image)

Figure 4. Visual feedback of intonation. (a) good intonation; (b) example for incorrect intonation. Colour notations: blue – reference, red – subject’s speech sample. The lines represent the temporal position of vowels.

The normalized fundamental frequency values of reference and the subject’s speech sample are displayed by applying the same dynamic time warping method as in the case of rhythm. Whether the timing parameters (temporal positions, durations) are correct or incorrect in the subject’s utterance will be reflected in the visual feedback. Examples of good and incorrect intonation are displayed in Figure 4.

Our visual feedback method where the intonation is presented together with the rhythm (and their correctness) is novel and very expressive. Based on early field experiences with hard of hearing children, it offers an easy way to learn prosody of utterances.

7. Discussion

According to the evaluation in Section 5, the described automatic assessment methods of intonation and rhythm are consistent with the subjective decisions of teachers. However, there are many open questions that emerge when subjects (most importantly children) are using a CAPT application, such as maintaining interest of the children and the tasks included.

Figures 3 and 4 shows simple but easily understandable elements, which are necessary for children if they are not to lose interest when using a teaching system.

In addition to the proper visual and automatic feedback the application must contain didactically well-constructed tasks that hold the child’s attention and facilitate the development of pronunciation. Such tasks have been designed in cooperation with teachers. Many training practices will help the children to learn correct intonation and rhythm of sentences by incorporating various bound and free dialogues into the system which improve the children’s pronunciation.

In “Török Béla” School of Hard of Hearing Children in Budapest hard of hearing children use the program successfully and they like it. But the official evaluation of the effectiveness of the teaching system together with the visual feedback will be carried out in the next few months.

8. Conclusions

This paper describes a newly developed CAPT system that intends to address appropriate teaching of such supra-segmental parameters as intonation, stress and speech rhythm. Two modules are implemented: (1) intonation and stress teaching and (2) rhythm teaching using dynamic time warping. The automatic feedback of the system is evaluated by using speech samples from hard of hearing children. The scores of the automatic assessment system correspond exactly to the three categories which were given subjectively by the teachers.

An evaluation was carried out in order to test if the automatic assessment scores describe properly the subjective decisions of teachers. The results show that a strict (lowest value of the three automatic scoring method) assessment is the most in line with the subjective judgments.

Visual feedback was proposed that is based on the dynamic time warping algorithm and gives simple and understandable visualization of the intonation and rhythm of the subject’s utterance. The application, with designed training practices, will be deployed in “Török Béla” School of Hard of Hearing Children in Budapest and the progression of the subjects will be measured over several months.

9. Acknowledgements

This work was partially supported by the European Union and the European Social Fund through project FuturICT.hu grant no.: TAMOP-4.2.2.C-11/1/KONV-2012-0002. and grant no.: TAMOP-4.2.2.C-11/1/KONV-2012-0013 organized by VIKING Zrt, Balatonfüred.

10. References


