VOCABULARY TRAINING PROGRAM USING TTS AND SPEECH RECOGNITION TECHNOLOGIES

Michael Hofmann[*], **Boris Lobanov**[†]
[*]University of Technology, Dresden, Germany
[†]United Institute of Informatics Problems, Minsk, Belarus

Abstract

Learning the right pronunciation is one of the most difficult tasks in mastering a foreign language. Common (commercial) PC based training applications provide only limited support for accentuation and listening or speaking. The aim of this paper is to present a vocabulary learning application that uses freely available speech synthesis and recognition technology. Various programs are analyzed whether and how they can be integrated in such a system. Different modes for the evaluation of recorded speech are presented and analyzed in their ability to judge pronunciation and accentuation correctness. It is shown that while synthesis support is readily available, the used approach of recognition-by-synthesis imposes severe limits to the ability of the system to generate meaningful pronunciation rates.

Keywords: text to speech, memory, synthesis, pronunciation, assessment

1. Introduction

1.1. Language learning

The knowledge of a foreign language consists of several parts that can't be learned independently (Neri et al. 2002): grammar, pronunciation, vocabulary and others. Whereas the first has to be learned interactively lead by a tutor or teacher to be efficient, e.g. by taking part in language training courses offered by a university or language school, vocabularies and their pronunciation can in part be learned autonomously.

Language training therefore should cover all this parts through different means of practice material and support: dialogs simulating every-day situations that appear in real life, the interactive construction of sentences to practice grammar, Listen and repeat exercises to train understanding and pronunciation, vocabulary drills for memorizing.

1.2. Task definition

To assist the learning process of a foreign language, a vocabulary training program with the support of freely available speech synthesis and recognition systems is developed.

Assistance is given in all of the following steps in learning words and phrases in a foreign language: listen to them, be able to understand them, to write them down, to translate them into the native language and to speak them with correct pronunciation.

To achieve this, synthesized words and phrases have to be recognized and translated by the student. The right accentuation is obtained from a database and presented to the student while learning. Afterwards the student is requested to speak the phrase, which is recorded and evaluated regarding pronunciation. Training sessions can be easily created and multiple training modes are supported. Support for German, English and Russian synthesis is available. The program can be run on current Unix and Windows operating systems.

Special attention is paid to the use of freely and widely available components. To achieve reasonable portability and good performance, the program is written in C++ using common platform-independent class libraries and toolkits. To implement the Graphical User Interface (GUI), the GTK+-Toolkit was used.

2. Learning process

The here presented training system focuses on a single task: to enable a student of a foreign language to learn and memorize vocabularies and their pronunciation effectively.

Among the different human memory systems the long term memory is responsible for the storage of declarative knowledge like vocabularies. Because the content is subject to the natural forgetting process, repeated recalls are necessary for information to last for longer.

Repetitio mater memoriae—Repetition is the mother of memory. Therefore the most important part of vocabulary training is the repetition of them in different ways until it can be assumed that the student memorized them completely.

With the increasing consolidation of the trained vocabularies, the time intervals in which material is reviewed can be increased over time using a technique called *spaced repetition*.

3. Program structure

3.1. Features

The program features an easy to use interface to be able to enter new training data and to train a number of vocabularies repeatedly. Acoustic versions of the entered vocabularies and stress positions are automatically provided if support for this feature in the target language is found. The repetition rate and training mode can be adjusted to the students needs. The knowledge level for each word or phrase is saved between sessions.

Experimental support for feedback of pronunciation quality and the level of correctness is available.

3.2. Synthesis support

In contrast to commonly available language learning programs, the auditory input to the student is not retrieved from a store of prerecorded examples. To allow the use for different languages and environments, speech synthesis is used for the generation of an acoustical and phonemic representation of the training material for the current exercise.

Widely available speech synthesis packages include Mbrola (Dutoit et al. 1996) and Festival (Black and Taylor 1997), which provide support for many different languages.

The Festival package is provided under the revised BSD license. It provides support for English and Spanish voices and offers all functions of a TTS system.

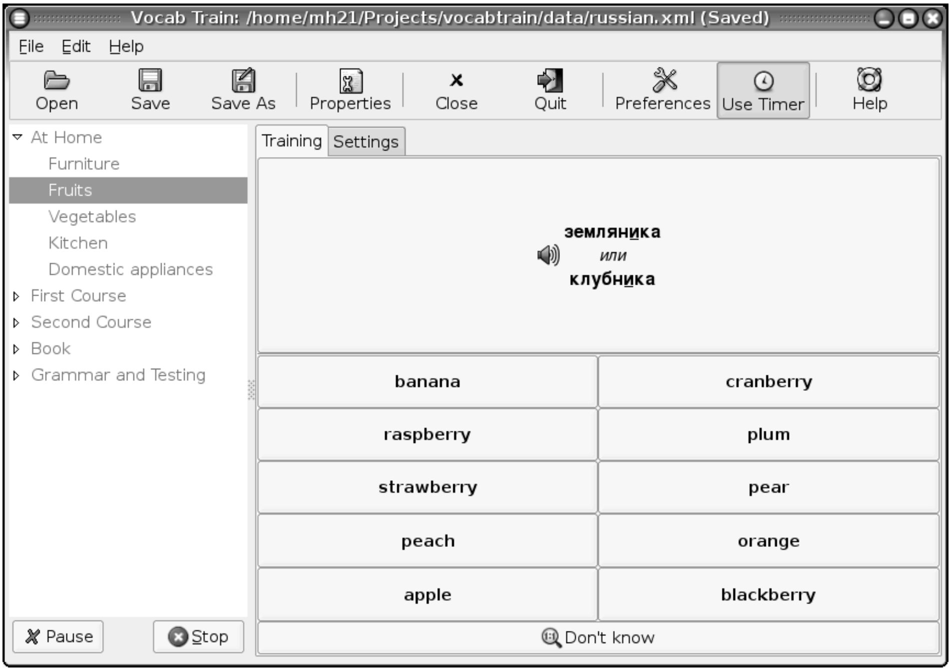

Figure 1: The training program in action

The Mbrola program is distributed under non-free license terms allowing non-comercial and non-military use only. It provides only the phone to wave conversion and can be combined with different text to phone converters to create a full text to speech system. It can be used with a wide range of languages.

For support of German and Russian a common TTS-synthesis prototype is used (Lobanov et al. 1998).

The attainable speech quality from nowadays available speech synthesizers is far good enough to be used in training programs. Although also longer training material is imaginable, the problems of current speech synthesis technology with more or less monotonic speech patterns because of inadequate prosodic control do not affect the training on word and phrase level.

4. Learning modes

The program provides several possibilities to train vocabulary knowledge. Both the input mode (how the current exercise is presented to the student) and the answer mode (the way in which the student can supply the answer) can be selected (table 1).
The input for the student can be provided in several ways:

- Written. The question is printed to the screen. Additionally trains the reading for languages in different alphabets, e.g. Russian or Greek.
- Acoustic. The vocabulary to be translated is converted to auditory output and played. The playback can be repeated if desired. Will improve the students ability to recognize trained words and phrases in conversation.

Table 1: Possible training modes

Question	Answer	Training effects
Acoustic	Select other language	Understanding, translation
Acoustic	Select same language	Understanding, letter-phoneme relation
Acoustic	Write other language	Understanding, translation, spelling
Acoustic	Write same language	Understanding, letter-phonome relation, spelling
Written	Select other language	Reading, translation
Written	Write other language	Reading, translation, spelling

- Combined. This gives additional training effects in languages like French or English, where no direct letter-phoneme relation exists.

The student can choose between three different modes to answer the question:

- Select among a certain number of alternatives. In this way, a large number of vocabularies can be reviewed in a short time.
- Provide the answer by entering it directly in the same language as the question. Combined with the speech synthesis, it is possible to train only the understanding or disambiguation of words.
- Same as before, only use the other language.

5. Recognition and Scoring

To evaluate the pronunciation of the exercises by the student, the following four steps must be performed (Ambra Neri et al. 2003):

- Speech recognition: The incoming signal has to be recorded, processed and recognized.
- Scoring: Pronunciation evaluation based on different properties of the recognized speech.
- Error detection: The program isolates certain phones or parts of the utterance that do not match the stored representation with a certain confidence.
- Error diagnosis: The type of error made is identified.

The information obtained in the last three steps is then presented to the student. Care has to be taken that the student is actually able to understand the displayed information to correct his pronunciation accordingly.

5.1. Speech recognition

Although there are a couple of speech synthesis packages with a large number of different voices available, very few speech recognition systems can be obtained freely. Worse, most of these systems have to be trained to a certain speaker and are very limited in the number of languages with which they can be used.

To minimize the dependencies on speech technology software used by the system, the method of recognition-by-synthesis is used to prepare the speech signal patterns to be recognized. This only requires the same speech synthesis package as used beforehand.

For the recognition to work independently from the used audio equipment, the characteristics of the channel: student – microphone – sound card, has to be corrected to match the characteristics of the synthesis. The assumption is made that that characteristics of the given channel are time-invariant.

The spectrum of a longer utterance by the student in his native language is averaged over the whole time and compared to a similar obtained reference from the speech synthesizer. Each of these averaged spectra is regarded as the frequency response of the corresponding channel. The difference between them in the log domain is used for a band filter to calculate the attenuation for each separate band (Young 1996).

To align the synthesized reference and the student utterance, silence at the beginning and the end of each signal is stripped using a threshold of -40 dB.

Both signals are then converted to a sequence of feature vectors. Implemented algorithms are a Fourier transformation, a mel spaced filterbank and MFCC coefficients. Dynamic time warping is used to fit the students utterance to the reference signal.

5.2. Error scoring and feedback

A lot of work is done in the field of automatic pronunciation assessment (Cucchiarini et al. 1998; Neumeyer et al. 2000; Teixeira et al. 2000). The term pronunciation covers a wide range of speech properties from segmentation to word stress. Erroneous pronunciation can mean a deviation in any of the following (Cucchiarini et al. 1997): the fluency of the utterance, the used syllable structure, word stress, kind of intonation and the segmental quality.

Not all of them can be directly measured. Some observable factors are the overall speech rate, the phonetic segmentation, the phone selection by the speaker and the pitch contour.

In the context of the training program and the used recognition method, possibilities are limited: the overall speech rate and the pitch contour are meaningless because of the only use of words and short phrases. Phone selection can only be judged from the similarity of reference and student signal, not by comparison with other similar sounding, but wrong phones, which makes scoring more difficult.

To evaluate the deviation of the segment duration from the reference, the following equation is used. L_{user} and L_{ref} denote the total length of the user and reference signal and l_i the length of the corresponding phone segments.

$$S = \sum_{i}^{N} \left| \ln \left(\frac{l_{i,user}/L_{user}}{l_{i,ref}/L_{ref}} \right) \right|$$

Neri et al. (2002) examines different available Computer Assisted Pronunciation Training (CAPT) courseware on the way they provide feedback for pronunciation errors. It is concluded that feedback should be limited to a simple grade of correctness and a highlighting of the incorrect areas in the utterance.

5.3. Summary and problems

This paper presented a flexible configurable vocabulary training application that can be configured to work with various available synthesizing systems.

Although synthesis support is easy to implement and aids in learning a foreign language, the reliable evaluation of the students own pronunciation is very desirable.

The used approach of recognition-by-synthesis provides a possibility to judge the phonemic segmentation of utterances by the student in comparison to the synthesizer. Problems like the spectrum distortion by slow amplitude changes, tolerances of the found phone borders and difficulties in distinguishing successive consonants need to be solved to generate meaningful scores.

References

Ambra Neri; Catia Cucchiarini; Strik, Wilhelmus 2003. Automatic speech recognition for second language learning: How and why it actually works. In: *Proceedings of 15th International Congress of Phonetic Sciences*, Barcelona, Spain. 1157–1160

Black, A.; Taylor, P. 1997. The Festival speech synthesis system. University of Edinburgh

Cucchiarini, C.; de Wet, F.; Strik, H.; Boves, L. 1998. Assessment of dutch pronunciation by means of automatic speech recognition technology. In: *Proceedings ICSLP '98*, Sydney, Australia. 751–754

Cucchiarini, C.; H. Strik, H.; Boves, L. 1997. Automatic evaluation of Dutch pronunciation by using speech recognition technology.. In: *Proc. of IEEE ASRU*, Santa Barbara. 622–629

Dutoit, T.; Pagel, V.; Pierret, N.; Bataille, F.; van der Vreken, O. 1996. The MBROLA project: towards a set of high-quality speech synthesizers free of use for non-commercial purposes. In: *Proceedings ICSLP '96*: Vol. 3, Philadelphia, PA. 1393–1397

Lobanov, B.; Hoffmann R.; Ivanov, A.; Kubashin, A.; Levkovskaja, T.; Helbig, J.; Jokisch, O. 1998. A bilingual German / Russian text-to-speech system. In: *Proceedings of the 3nd International Workshop "Speech and Computer" SPECOM'98*, St. Petersburg. 327–330

Neri, A.; Cucchiarini, C.; Strik, H.; Boves, L. 2002. The pedagogy-technology interface in computer assisted pronunciation training. In: *Computer Assisted Language Learning* **15(5)**, 441–447

Neumeyer, L.; Franco, H.; Digalakis, V.; Weintraub, M. 2000. Automatic scoring of pronunciation quality. In: *Speech Communications* **30(2-3)**, 83–94

Teixeira, C.; Franco, H.; Shriberg, E.; Precoda, K.; Sonmez, K. 2000. Prosodic features for automatic textindependent evaluation of degree of nativeness for language learners

Young, Steve 1996. A review of large-vocabulary continuous speech recognition. In: *IEEE Signal Processing Magazine* **13(5)**, 45–57

MICHAEL HOFMANN is student of the University of Technology, Dresden, Germany. He is an invited researcher at the United Institute of Informatics Problems Nat. Ac. of Sc., Belarus.

BORIS LOBANOV is head of the Speech Recognition and Synthesis Laboratory at the United Institute of Informatics Problems Nat. Ac. of Sc., Belarus and Professor at the University of Bialystok, Poland. He has a Dr. Sc. degree in text-to-speech synthesis (1984). He is member of the European Speech Communication Association since 1995. His sphere of interests include TTS-synthesis, speech analysis and recognition and speech technology applications. He has written more then 200 publications in the area of speech synthesis and recognition.