OpenVOC - Open Platform for Multilingual Vocabulary Training Integrating Speech Technology Components

Michael Hofmann, Oliver Jokisch

Laboratory of Acoustics and Speech Communication
Dresden University of Technology, 01062 Dresden, Germany
{michael.hofmann3,oliver.jokisch}@mailbox.tu-dresden.de

Abstract

The language acquisition consists of several parts which can not be learned independently: grammar, pronunciation, vocabulary and others. Whereas the first has to be learned interactively, lead by tutor or teacher to be efficient (e.g. by taking part in language training courses offered by language schools), vocabularies and their pronunciation can be partly learned autonomously. Language training therefore should cover all these parts by different means of practice material and support: dialogs simulating typical situations which 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. An "automatic teaching program" may appear inferior to a human language teacher, but it does have some potential advantages: it can be used an unlimited amount of time and students get less embarrassed than in a classroom. Although there are various language training programs available, there is a lack of systems which use speech technology. This paper presents a framework for such a system to ease the evaluation of speech technology (used in educational systems) and to boost the application of speech synthesis and recognition in general.

All core components for a learning application are implemented: A database for lesson and user data, synthesis support using a simple command line interface, a platform independent graphical user interface which runs on a large scale of operating systems. The whole system is distributed under a free license to minimize the barriers imposed on the user and to boost the speech technology use in this field.

1. Current state of computer-assisted language teaching

1.1. Common approaches and existing systems

Several commercial training systems exist which use auditory output to present correct pronunciation and speech recognition to evaluate speech quality, intonation and accentuation. They focus either on the learning of the basic vocabularies or on specific topics like business expressions. Most of them provide the possibility to present prerecorded spoken versions of the lecture material, some of them also a possibility to evaluate a recorded version of a students utterance. For a detailed discussion of the available programs and features see [1].

There are also a number of educational systems developed by universities all worldwide. The ISLE project [2] aims at the improvement of English as a second language for Italian and German learners. It uses an HMM-based recognizer trained on non-native speech to align student and reference utterance and language-specific mispronunciation rules to detect mistakes by the speaker. The SRI EduSpeak System [3] uses HMMs specifically adapted to non-native speakers using Bayesian adaption techniques. It was also shown [4] that the training with non-native speech improves recognition rates, thus making it easier to judge pronunciation errors.

A Japanese-only variant of a system using a synthetic or a natural reference and forced alignment was presented in [5]. It uses formant synthesis and can impose the correct prosody on the students speech. Another English pronunciation system for Japanese Students [6] models many common error patterns to detect erroneous phoneme segments. Other research is done on the Fluency project [7], the SPECO project [8] and the virtual language tutor [9].

Several free programs for language and vocabulary learning exists (KVocTrain, FlashKard, Langdrill, LingoTeach), though none of them is known to provide support for speech technology.

1.2. Evaluation and scientific issues

The perfect automatic evaluation of speech by a non-native speaker regarding pronunciation is a still unsolved problem. Different ways for scoring the pronunciation quality using e.g. temporal measures and whether they correlate with expert ratings are described in [10, 11, 12].

Approaches exist that compare the recorded speech with several HMM reference models available for the utterance, calculating distances to them based on e.g. spectral or segmental similarity. Sometimes models are trained of the wrong patterns to identify errors.

The derivation and presentation of reliable scores for the distinction between correct and wrong utterances has to consider problems like user acceptance and self confidence of the learner as well as the experience of the student with pronunciation training and the total amount of correction information that should be shown in order to actually improve pronunciation and not instead cause confusion.

2. The OpenVOC approach

2.1. Targets and underlying speech technology

One of the major problems second language learners have to cope with is insufficient vocabulary knowledge. About 2,000 to 3,000 word families are needed to enable language use, whereas foreign students reading university texts need to have 10,000 to 11,000 word families at their disposal [13, 14].

The acquisition of such a large amount of vocabularies can be supported by computer programs and the use of speech tech-

nology. The presented system concentrates on the efficient rehearsal of single words and phrases that can be defined by the user according to his/her needs.

The task of learning vocabularies can be divided in different levels of knowledge that have to be acquired to enable every day use. Mastering words means that one is able to read them, write them, understand their different meanings and speak them with correct pronunciation.

Therefore the application has to support all these steps of learning. It has to provide means to achieve more intense exposure to the foreign language with the effect of learning new vocabularies in an efficient way. It should be possible to quickly define and rehearse larger amounts of new vocabularies.

2.2. Text-to-speech synthesis

In contrast to all known commercial training software the program uses arbitrary external synthesis packages to provide the output. This enables the program to render acoustic versions of all requested words and phrases that are supported by the synthesizer, making it independent from preexisting lecture material.

Speech synthesis is said to be not good enough to be used in applications that require accurate pronunciation. In the opinion of the authors currently available packages with carefully crafted voices provide enough quality to be used in educational software for single words and phrases. Although the quality of sentences and longer texts may suffer from the known effects of monotonic prosody and insufficient expressiveness, these constraints do not apply in the limited domain of a vocabulary training application that uses learning material which mainly consists of single words and short phrases.

As a pronunciation reference and for teaching purposes, speech synthesis output is already successfully used in some on-line dictionaries such as the LEO dictionary for translations from English and French to German [15].

The LexDRESS system [16] is focusing on word domain synthesis (for German) by using a special allophone set and a word-based prosodic model. In [17], a bilingual synthesis approach (for Russian and German) is introduced which is leading to consistent system resources for a dictionary.

Independent from limited word domain approaches, state-of-the-art corpus synthesizers achieve an intelligibility of almost 100% in a known domain and Mean Opinion Scores (MOS) of about 3.8 [18] (compared to MOS of natural speech of about 4.7). Consequently, these synthesizers can be already used for vocabulary training (at least as a first acoustic reference). Nevertheless, for a qualified pronunciation teaching to reduce foreign accent, the synthesis technology requires further improvement.

2.3. Synthesis support

Widely available speech synthesis packages include Mbrola [19] and Festival [20], 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 complete TTS system. The Mbrola program is distributed under non-free license terms allowing non-commercial 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 for a wide range of languages. For support of German and Russian a common TTS-synthesis prototype is used [17].

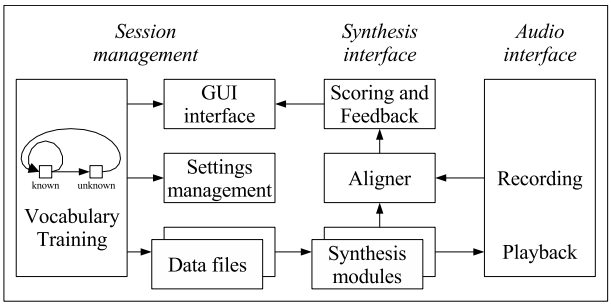

Figure 1: *Application structure*

2.4. Automatic speech recognition

Although there are a several 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 program, the method of recognition-by-synthesis is used to prepare the speech signal patterns to be evaluated. This only requires the same speech synthesis package as used beforehand. To align the synthesized reference and the student utterance, Dynamic Time Warping (DTW) is used. Various characteristics of the reference and student signal are then calculated and compared.

2.5. Intellectual property issues

Nowadays speech technology is readily available in many different languages, for various operating systems and under free or restricted license terms. The presented program makes it possible to use nearly all available speech synthesis packages that provide a command line or system library interface. This makes the system very versatile and allows easy integration of new or experimental synthesis packages as the become available.

The program is released under the General Public License (GPL) of the Free Software Foundation, granting the freedom to run the program for any use, to study how it works, to change and to redistribute it. The authors hope to in this way lower the barriers and the costs normally associated with such a training software and to foster community development and widespread use by not imposing unnecessary limitations.

3. System architecture

3.1. Application framework

The system is divided in session management and synthesis related parts. The latter is also responsible for the platform dependant audio interface. The program is written using object-oriented C++ source code and uses the Gimp Toolkit (GTK) for the user interface. Training material, program files and user data are separated and individual settings are saved between sessions. The students performance is monitored and information on vocabulary history and difficulties is used to adopt the training process accordingly.

```
<lesson title="Fruits">
  <vocabulary id="883694554">
    <translation lang="deutsch">
      Erdbeere {w.}
    </translation>
    <translation lang="english">
      strawberry
    </translation>
    <translation lang="русский">
      земляника; клубника
    </translation>
  </vocabulary>
</lesson>
```

Figure 2: *Corresponding XML data.*

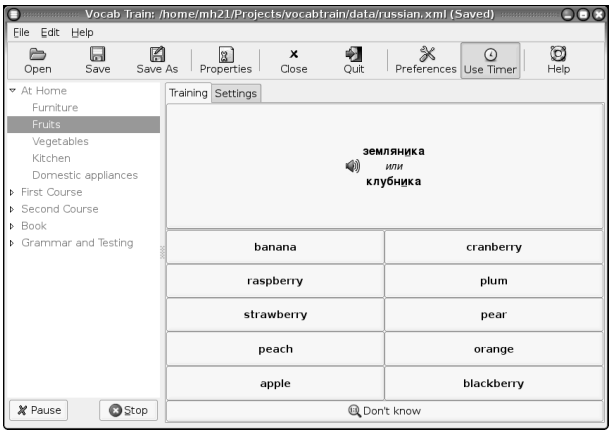

Figure 3: *The training program in action.*

3.2. Data formats

All stored information is formated in the Extensible Markup Language (XML) using Unicode UTF-8 encoding. This enables the storage of arbitrary characters like Cyrillic letters or Japanese symbols. It can be easily converted from and to other training systems, using e.g. XSLT transformations.

3.3. User interface

The implemented user interface provides easy access to available data files and enables the student to enter their own or to modify already existing material in a very fast way. Additional information like gender, meaning description and alternative translations can be saved together with the vocabularies and will be displayed and used in an appropriate manner. Lessons can be trained in arbitrary order and number. Acoustic versions of every word in languages with synthesis support can be obtained and stress positions are displayed where available.

3.4. Speech synthesis interface

To adopt to different synthesis systems, a simple interface is developed. External wrapper programs are called that handle the different engines thus making it possible to adapt nearly all systems with very low effort. Currently available is support for Festival and Mbrola synthesis as well as synthesis engines supported by the Microsoft Windows Speech API.

4. Functional concept

4.1. Learning modes

4.1.1. Concept

At the moment the program provides the following 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 [21].
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. This will improve the students ability to recognize trained words and phrases in conversation.
- 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.

4.1.2. Repetitio mater memoriae

"Repetition is the mother of memory" shows the main part of vocabulary training: Repetition until the vocabularies can be memorized completely. The training system therefore currently focuses on this 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. 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*.

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.

4.2. Evaluation of the learning process

To evaluate the total 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.

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

Similar scoring methods are used for energy and spectral similarity. The obtained values are aligned to the recorded utterance and displayed to the user, no evaluation is done at the moment regarding the correctness or intelligibility of the utterances.

5. Conclusions

This paper presented a training application for language learning supported by speech technology. The authors assume that available speech synthesis systems are suitable enough for educational use. The shown system should ease deployment of readily available synthesis technology in this area. Further research is needed on language independent measurement that indicate the intelligibility of utterances and on ways of adjusting the kind and amount of presented errors to the user. Possible further evaluation modules will take, e.g., pitch contour deviation and other prosodic parameters into account.

Source code, informations about the program and further information are available at [22].

6. Acknowledgments

OpenVOC was developed with support of Boris Lobanov, Andrej Davydov and Dimitrij Zhadinets (United Institute of Informatics Problems in Minsk, Belarus) within the cooperation project "Development of a Multi-Voice and Multi-Language Text-To-Speech and Speech-To-Text Conversion System for the languages Belorussian, Polish and Russian" (European project INTAS 04-77-7404).

7. References

[1] A. Neri, C. Cucchiarini, H. Strik, and L. Boves, "The pedagogy-technology interface in computer assisted pronunciation training," *Computer Assisted Language Learning*, vol. 15, no. 5, pp. 441–447, 2002.

[2] W. Menzel, D. Herron, P. Bonaventura, and R. Morton, "Automatic detection and correction of nonnative english pronunciation," in *Proceedings of InSTILL'00*, Dundee, Scotland, 2000, pp. 49–56.

[3] H. Franco, V. Abrash, K. Precoda, H. Bratt, R. Rao, and J. Butzberger, "The SRI EduSpeak(TM) System: Recognition and Pronunciation Scoring for Language Learning," in *Proceedings InSTILL'00*, Dundee, Scotland, 2000.

[4] S. Goronzy, M. Sahakyan, and W. Wokurek, "Is nonnative pronunciation modelling necessary?" in *Proceedings Eurospeech'01*, vol. 1, Aalborg, Denmark, 2001, pp. 309–312.

[5] M. Yoram and K. Hirose, "Language training system utilizing speech modification," in *Proceedings ICSLP'96*, vol. 3, Philadelphia, PA, 1996, pp. 1449–1452.

[6] Y. Tsubota, T. Kawahara, and M. Dantsuji, "Practical use of english pronunciation system for japanese students in the call classroom," in *Proceedings ICSLP'04*, Jeju Island, Korea, 2004, pp. 1689–1692.

[7] M. Eskenazi, S. Hansma, J. Corwin, and J. Albornoz, "User adaptation in the Fluency pronunciation trainer," in *Proceedings Eurospeech'99*, 1999, pp. 847–850.

[8] K. Vicsi, P. Roach, A.-M. Öster, Z. Kacic, F. Csatári, A. Sfakianaki, R. Veronik, and G. Gordos, "A multilingual, multimodal. speech training system, speco," in *Proceedings Eurospeech'01*, 2001, pp. 2807–2810.

[9] O. Engwall, P. Wik, J. Beskow, and B. Granström, "Design strategies for a virtual language tutor," in *Proceedings ICSLP'04*, Jeju Island, Korea, 2004, pp. 1693–1696. [Online]. Available: http://www.speech.kth.se/ctt/publications/papers04/

[10] C. Cucchiarini, H. Strik, D. Binnenpoorte, and L. Boves, "Towards an automatic oral proficiency test for dutch as a second language: Automatic pronunciation assessment in read and spontaneous speech," in *Proceedings InSTILL'00*, Dundee, Scotland, 2000.

[11] C. Cucchiarini, H. Strik, and L. Boves, "Automatic pronunciation grading for dutch," in *Proceedings STILL'98*, Marholmen, Sweden, 1998, pp. 95–98.

[12] F. de Wet, C. Cucchiarini, H. Strik, and L. Boves, "Using likelihood ratios to perform utterance verification in automatic pronunciation assessment," in *Proceedings Eurospeech'99*, vol. 1, Budapest, Hungary, 1999, pp. 173–176.

[13] R. Waring, "Second language vocabulary acquisition, linguistic context and vocabulary task design," The British Council Conference in St Andrews, Scotland, Sept. 1995. [Online]. Available: http://www1.harenet.ne.jp/˜waring/papers/BC.html

[14] P. Nation and R. Waring, "Vocabulary size, text coverage and word lists," in *Vocabulary: Description, acquisition and pedagogy*, N. Schmitt and M. McCarthy, Eds. Cambridge: Cambridge University Press, Sept. 1997, pp. 6–19. [Online]. Available: http://www1.harenet.ne.jp/˜waring/papers/cup.html

[15] "LEO - link everything online," web service of TU Munich, Germany. [Online]. Available: http://dict.leo.org/?lang=en

[16] R. Hoffmann, O. Jokisch, U. Hirschfeld, and A. L., "Lexdress - speech synthesis for a speaking pronunciation dictionary," in *Proceedings ESSV'04*, Cottbus, Germany, 2004, pp. 183–190.

[17] B. Lobanov, R. Hoffmann, A. Ivanov, A. Kubashin, T. Levkovskaja, J. Helbig, and O. Jokisch, "A bilingual German / Russian text-to-speech system," in *Proceedings SPECOM'98*, St. Petersburg, 1998, pp. 327–330.

[18] Y. Alvarez and M. Huckvale, "The reliability of the itu-t p.85 standard for the evaluation of text-to-speech systems," in *Proceedings ICSLP'02*, Denver, 2002, pp. 329–332.

[19] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. van der Vreken, "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, 1996, pp. 1393–1397.

[20] A. Black and P. Taylor, *The Festival speech synthesis system*, University of Edinburgh, 1997.

[21] M. Hofmann and B. Lobanov, "Vocabulary training program using TTS and speech recognition technologies," in *Proceedings of the Second Baltic Conference on Human Language Technologies*, Tallinn, 2005.

[22] "OpenVOC: Open platform for multilingual vocabulary training integrating speech technology components," Project homepage. [Online]. Available: http://mh21.piware.de/openvoc/