SpeechIndexer: A Flexible Software for Audio-Visual Language Learning

Ulrike Glavitsch, Klaus Simon
EMPA, Swiss Federal Laboratories for Materials Science and Technology
Media Technology Laboratory
Überlandstrasse 129, 8600 Dübendorf, Switzerland
[ulrike.glavitsch, klaus.simon]@empa.ch

and

Jozsef Szakos
Department of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
jozsef.szakos@inet.polyu.edu.hk

ABSTRACT

This paper presents SpeechIndexer as a software tool to create teaching and learning material for language courses and as an e-learning program to train the oral comprehension and speech production. We introduce the player function where students can follow speech and text closely and the role play function that allows the learner to get involved in a dialog. Teachers can create material from the vast amount of speech recordings available (audio books, radio and TV podcasts, language learning CDs, etc.) that specifically match the knowledge level and interest of individual students or the whole class. The software may complement regular language courses or may serve to teach languages where teaching and learning material scarcely exist, e.g. endangered languages that have a pure oral tradition.

Keywords: e-learning, language learning, teaching and learning material creation, oral comprehension and speech production training

1. INTRODUCTION

Current language learning software contains a variety of different functional blocks such as read speech, grammar modules and vocabulary exercises [1, 2]. These programs are highly interactive, the learning progress is constantly evaluated and the student can advance at her own pace. Some of the programs even provide a pronunciation training where the user speaks words into the microphone and a built-in automatic speech recognition component checks the utterance for correctness. However, state-of-the-art language learning programs have a very static design. They contain a control logic that presents the functional blocks in some order, accept inputs from the user and check it with fixed stored values and they contain jumps into predefined subroutines, e.g. in the case of automatic checks of pronunciation correctness. There is no way to adapt or extend the data content of this e-learning software, i.e. it is not possible to add material to it, namely, additional speech files or new vocabulary. Yet, a teacher in face-to-face instruction observes the level of knowledge in his class and provides additional training when needed.

In this paper, we consider the SpeechIndexer software as a platform to flexibly create both teaching and learning material and to serve as a language learning program. We present two functions of the software that specifically support the audio-visual learning of a language. These functions are described here for the first time in detail.

The SpeechIndexer software has been developed at ETH Zurich and EMPA in Switzerland by two of the authors. The original goal was the documentation of endangered

aboriginal Formosan languages that have a pure oral tradition and where large archives of recorded speech exist [3]. The language teaching aspect, however, has always been kept in mind during the development since endangered languages need also be taught to keep them alive [4].

The structure of the paper is as follows. We give an overview of the SpeechIndexer software in Section 2. Then, we describe how a learner trains his listening comprehension and speech production using SpeechIndexer's player and role play function in Section 3 and 4. We present the flexible creation of teaching and language material in Section 5. Finally, we draw conclusions and give an outlook to future work in Section 6.

2. SPEECH INDEXER OVERVIEW

SpeechIndexer has been developed for the indexing and retrieval of speech files. The core component of SpeechIndexer is its semi-automatic indexing where segments of speech are correlated with their corresponding text segments. The textual transcription for a given speech file is entered via the program or is loaded from outside. A built-in pause finder automatically divides the speech file into pause and speech segments. A speech segment denotes a flow of speech uttered without breathing and contains a few words, a phrase, a sentence or even more than that. The user manually combines corresponding speech and text segments to form the correlations called indices [5]. Text belonging to an index (indexed text) is marked and appears red, bold and underlined. The speech segment behind an indexed text can be played directly. Fig. 1 shows the SpeechIndexer main window that contains the audio window with the speech signal and the text window. The text window shows some unindexed text (black) and indexed sections (red, bold and underlined). The audio file is a section of Churchill's famous speech "A United States of Europe" [6]. Furthermore, the user can mark speech segments of individual speakers with different colors to see at first glance which parts of speech belong to the same speaker. Fig. 2 shows an indexed speech file with two speakers (Brian and Jane) that are marked differently – each of which with a different color. The speech file in this example is from the audio CD of an English language course [7]. The user can define marking colors of speakers individually. The speaker profiles are saved under a so called extras file. All loaded and created files (audio, segmentation, text, indices and extras file) are captured under a project name. It is possible to load all files of a project at once and to save a set of files as a project. In addition, SpeechIndexer provides two search functions: (1) searching within the same file and (2) searching across collections of speech recordings. In both cases, the search results are listed and each result can be played immediately.

A number of helpful visualizations have been implemented. First, the speech signal is always visible, i.e. the user sees where the signal has low- or high-energy and where there are pauses in the signal. Secondly, a cursor in the speech signal follows the waveform when the signal is played, and more importantly, the indexed text currently played is highlighted. Finally, the aforementioned speaker coloring scheme provides a better overview in an indexed speech file with multiple speakers.

Fig. 1: SpeechIndexer main window with signal and text window. Indexed text appears red, bold and underlined.

3. LISTENING COMPREHENSION TRAINING

Given a fully indexed speech file students train their listening comprehension in the following way. They play the speech file and follow the played text. The currently played text is highlighted for clarification, i.e. it appears in a different color that the user may specify. (Behind the scenes, the program checks for each play position whether there is an index that contains this audio position and it highlights the corresponding text segment. The played text is set to its original state as soon as playing the corresponding speech segment has finished.) Students can always stop the recording and repeat playing a previous

segment if they did not understand it. They can look up unknown words offline. After doing this listening exercise again and again, the learner understands the speech recording without looking at the text. In fact, this is the final goal of listening comprehension training, namely, being able to fully understand a piece of authentic speech.

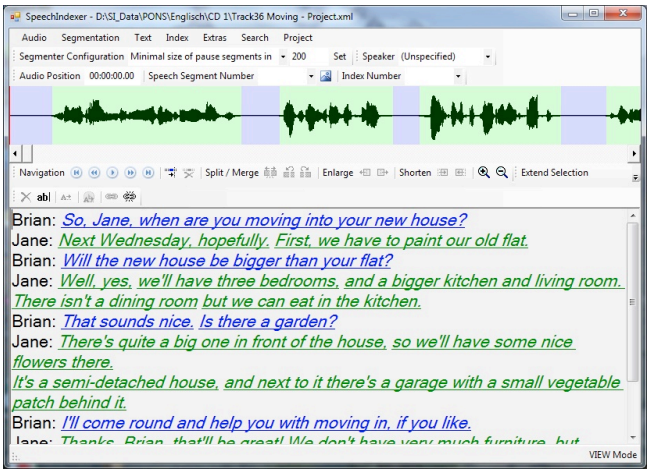

Fig. 2: SpeechIndexer main window with speakers marked individually.

4. SPEECH PRODUCTION TRAINING

Speech production is trained with SpeechIndexer's role play function. The role play function operates on indexed dialogs as they often occur on CDs that are delivered with language learning books. The readings of language books are typically dialogs in everyday situations (in the supermarket, at the bus station, in the restaurant, etc.). Those speech recordings are indexed by the teacher and the speakers are marked individually. The role play function lets SpeechIndexer take over the role of one dialog partner while the learner speaks the role of the other part. For this purpose, the program plays the speech of one dialog partner and mutes the speech of the other. The student can always compare his given utterance by listening to the speech part he is supposed to take over in a second round. This way, the student learns to speak whole sentences in typical life contexts.

The role play function is activated by declaring one of the speakers as muted speaker. As a consequence, all text of this speaker is made invisible. When playing the file, the speech of the muted speaker is played soundless. Fig. 3 shows SpeechIndexer with the role play function activated. The text of speaker Jane that is visible in Fig. 2 is cleared in Fig. 3 and Jane's speech is muted. The learner is supposed to speak Jane's role.

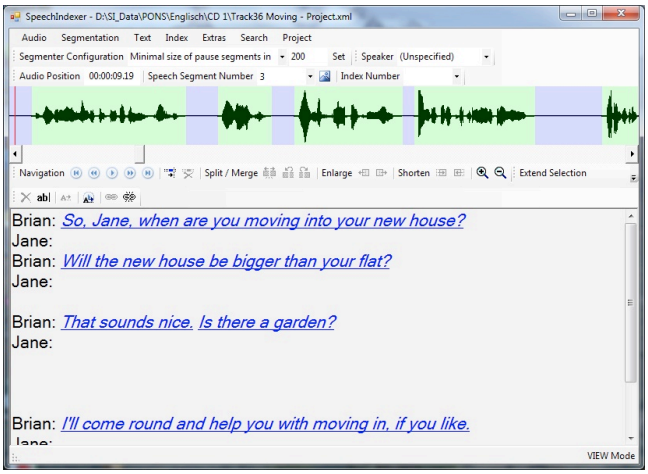

Fig. 3: SpeechIndexer with role play function activated.

In fact, we believe that the role play function of SpeechIndexer is a better way to train the speech production of a learner than an automatic speech recognition component that lets a student speak single words only. First, state-of-the-art automatic speech recognition component often make errors and will recognize utterances as correctly pronounced if there are not and vice versa. But more importantly, the proposed role play function lets a learner to become involved in a dialog and trains a natural way of speaking.

5. LEARNING FROM AUTHENTIC DATA

With SpeechIndexer the teacher flexibly creates teaching and learning material suited for the needs of the class or for individual students. Speech recordings nowadays can be gained from various sources, e.g. from CDs associated with language books, from audio books and radio or TV podcasts. The audio tracks of these media files are converted into the WAVE file format by freely available programs and loaded into SpeechIndexer. Some radio stations provide transcripts of their broadcasts and also audio books often come with the texts. Given transcripts can be converted into UTF8 files and loaded into SpeechIndexer. The teacher enters the text of an audio section within the program if no electronic transcript is available. The final step is the synchronization of the audio file with the texts by creating the indices and the marking of the speakers if

needed. The set of files can be given to students for listening comprehension and speech production training.

It is a particular advantage of SpeechIndexer that it allows the creation of teaching and learning material from current, authentic recordings, e.g. from news broadcasts of yesterday. Thus, teaching and learning material for the present-day use of the language is created. This is specifically helpful in cases where the teaching and learning material is outdated or when it hardly exists. For instance, there is only very little language teaching and learning material available for endangered languages that mostly have a pure oral tradition.

6. CONCLUSIONS AND OUTLOOK

We have considered SpeechIndexer as a flexible tool to create language teaching and learning material from authentic speech recordings. From its original design goal it lets the learners train their oral fluency, namely, their listening comprehension and speech production.

SpeechIndexer gives the language learner a new view on the spoken language and makes it more transparent. It shows aspects of the spoken language that cannot be seen otherwise. It shows the speech signal and the position played, it highlights the text currently played and lets the text segments of different speakers appear in different marking colors. This way, the learner perceives spoken language as concrete. We believe that it is also a well-suited e-learning tool for primary and secondary school children since it has an uncomplicated user-interface and is easily learned.

Currently, SpeechIndexer is intended to complement existing language learning material to specifically train the oral fluency and comprehension. However, we plan to evaluate SpeechIndexer in language classes in order to get feedback on its usability and effectiveness. Key questions are whether teachers create teaching and learning material in a user-friendly and effective way and how well learners improve their oral fluency when using SpeechIndexer. A corresponding project with the Swiss National Science Foundation and various partners, e.g. with the University of Teacher Education Central Switzerland, is in discussion.

7. REFERENCES

[1] Tell Me More® V10 English, www.tellmemore.com

[2] Babbel, www.babbel.com

[3] J. Szakos, U. Glavitsch. Seamless Speech Indexing and Retrieval: Developing a New Technology for the Documentation and Teaching of Endangered Formosan Aboriginal Languages. Proc. Intl. Conference on Education and Information Systems: Technologies and Applications (EISTA' 04), Orlando, Florida, July 21 – 25, 2004.

[4] J. Szakos, U. Glavitsch, O. Hess. From Speech Corpora to Textbook Generation - Extending software technology to non-European languages, 7th Teaching and Language Corpora Conference (TaLC7), Paris, France, July 2 - 4, 2006.

[5] J. Szakos and U. Glavitsch. SpeechIndexer in Action: Managing Endangered Formosan Languages. In Proceedings of Interspeech, 2007.

[6] Never give in! No. 1: Winston Churchill's Greatest Speeches.

[7] PONS Power-Sprachkurs Englisch in 4 Wochen, Rebecca Davies, PONS Stuttgart.