



# SpeechIndexer in Action: Managing Endangered Formosan Languages

Jozsef Szakos<sup>1</sup>, Ulrike Glavitsch<sup>2</sup>

<sup>1</sup> Department of Indigenous Languages and Communication, National Dong Hua University, Hualian, Taiwan

<sup>2</sup> Department of Computer Science, ETH Zurich, Zurich, Switzerland

szakosdr@ms38.hinet.net, ulrike.glavitsch@inf.ethz.ch

## Abstract

Among the most endangered languages of the World, Formosan Austronesian vernaculars occupy a sadly prominent place. Two decades spent in language documentation would only weakly repair the broken transmission chain between the generations of speakers, if there had been no chance to develop SpeechIndexer, a novel software combining portability, adaptability for processing large amount of speech data and its first transcription. This presentation introduces the situation of the languages under investigation and shows examples of SpeechIndexer used for marking up and retrieving spoken corpus data, archived for preservation.

**Index Terms:** speech indexing, endangered languages, Austronesian, Formosan, language documentation, retrieval.

transcribed, the others under work, about a dozen are indexed with SpeechIndexer.

Saaroa: Ethnic population: 300 (2000 UNESCO Red Book) [2]. There are still about 20 speakers. Our materials after 12 years of field-work include 569 hours of recordings, half of them sentence by sentence translations of fieldwork stories. Again, about a dozen tapes marked up with SpeechIndexer.

Kanakanavu: Ethnic population: 250 (UNESCO) [3]. There are still around 10 elderly speakers. Our speech recordings over a span of 12 years include 370 hours of recordings. The SpeechIndexer indexing and transcription work is in progress. The above mentioned materials are all digitized in .wav and MP4 formats.

## 1. Introduction

This presentation is more practical in nature, based on the past years of language data collection among indigenous tribes, some of them left less with very few fluent speakers. Although Taiwan is regarded as the homeland of the Austronesian language family (extending over the Pacific and comprising more than 1200 languages), all efforts at language preservation can barely counterbalance the rapid disappearance of more than half of the more than 20 languages still spoken about a hundred years ago. Remnants of about 40 dialects still can be located at this moment, but in about one generation, predictably, only six or seven major languages will be actively used, or only in a bilingual setting. Even these will have undergone modernization, adaptation and reduction. It is therefore imperative that we preserve and document, archive authentic language data, with minimal outside contamination from dominant languages like Mandarin Chinese, Taiwanese, Hakka, also spoken in Taiwan.

This wide range of applications expands the functions and expectations towards a supporting software. While the original recordings should be archived, unchanged, left unmodified, unsegmented, the large amounts should be distinctly retrievable and easily exchangeable into forms for linguistic research, language revitalization and language materials preparation, even for language testing purposes.

## 2. The situation of the languages involved

Tsou: Ethnic population: 2,127 [1]. More than two thirds still speak the language, elementary textbooks available, Bible translation in progress. The first author's own research has been going on for 20 years, with grammar, texts, dictionary as a result. There exist more than a thousand hours of recordings, about two thirds of which come from intensive field-work, others from local radio stations. About one third is fully

With the recent rise of indigenous awakening in Taiwan, there came an indigenous TV channel, broadcasting news and some programs in local languages. We have collected about 150 DVD materials, containing 900 hours of broadcast, in good sound quality, but owing to the nature of broadcasting style the languages are all mixed (Chinese, Taiwanese and also lots of music). The indexing of these recordings, together with broad transcriptions is also under way. The languages here are Amis, Bunun, Paiwan, Rukai, Atayal, Yami, These are 'major' languages, with thousands of speakers.

Corpora collected at the Institute of Linguistics, National Taiwan University, at Providence University (Yami) and Academia Sinica serve the linguistic community, but follow a different approach from ours [4]. Theirs is server centered, users are web-clients, while our purpose is to provide 'half-baked' authentic, indexed but otherwise unmodified linguistic materials to the researchers.

### 3. Functions of SpeechIndexer in action

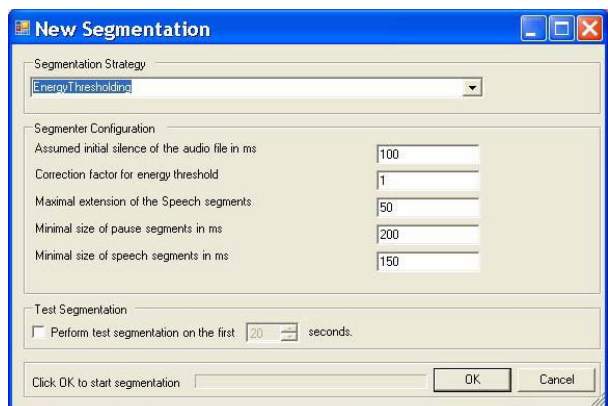
#### 3.1. File management: Opening the audio and text, loading the index:



The above screen-shot shows SI with the audio and text files loaded. Stressing the primacy of speech, audio must be loaded first, followed by text and indices.

#### 3.2. Suggested pre-segmentation according to linguistic units

This is the most innovative part of the software, creating a segmentation initially, which is counterchecked by the researcher. If necessary, the segmenter configuration can be changed, to provide the best approximation for the length of utterances to be indexed. Then the full segmentation is performed automatically.



Here, the segmentation after testing different possibilities is defined.

#### 3.3. Transcription and editing

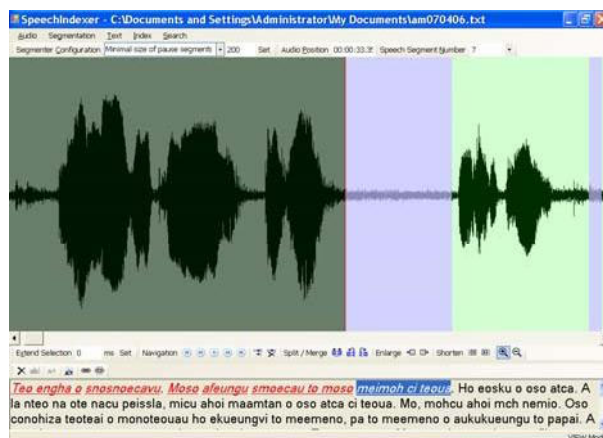
On the one hand we can navigate the audio file, edit the neighboring segments (increasing, decreasing), we can zoom in/out our selection length.



At this point we can insert the text in the text part of the screen. We can then select part of the text and audio to link them in the indexation. We may either work on an already transcribed speech or conduct the transcription during hearing the record segments.

#### 3.4. Retrieval and use of data for phoneticians, morphologists

An example of zooming in on a particular phrase shows how phoneticians, phonologists can use the software.



The parts marked can be copied to other files and investigated using other (phonetic) software.

#### 3.5. Retrieval and use of data for discourse analysis

Some recordings involve dialogues. Depending on the indexing style of the user, an initial broad transcription is enough, then he/she can concentrate on the narrow transcription of significant parts. Another type of indexing is also conceivable for example for research on code-switching.

#### 3.6. Using the data for language testing

In the case of language teaching and transcription training where the hearing abilities of students, examinees need to be tested, any authentic recording can be loaded and there is a realistic assessment as to the speed and precision of the transcribing students, when the results are compared with a standard transcribed text of the sound file. This provides a

quantifiable realistic situation of many professional linguistic applications (courtroom, etc.)

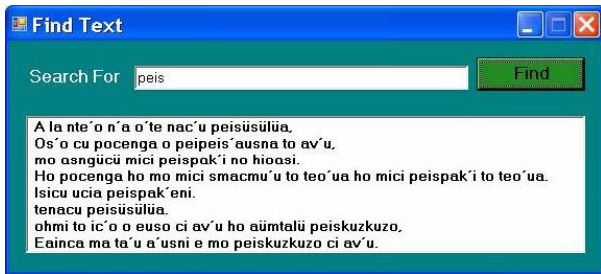
### 3.7. Data quality from field recordings and from indigenous mass media sources: Overview

Field recordings are very different in quality, depending on the weather conditions, noise, recording skills and number of people involved in the speech act. The same applies to many interview situations broadcast on the indigenous TV station. SpeechIndexer's capacity to individually regulate the pausefinder levels is a great help in selecting the right level of segmentation, even with noisy recordings.

### 3.8. Using SpeechIndexer for enhancing field-work

Field researchers can save precious verification time if they can transcribe their recordings by themselves and mark the unclear places to be verified. Working with such a file when in the field again, the efficiency of transcription can be raised.

### 3.9. Speech concordancing



The above example in the Tsou language shows how the virtually segmented units can be searched and displayed according to the transcribed morphemes. The list can be saved, together with the corresponding sound files.

## 4. Outlook

This software is supposed to be used together with other phonetic, and database applications. In future developments upward and downward compatibility will be preserved.

New features to be implemented are the concatenation of sound files up to memory limit of PC, and the corresponding connection of transcription and index files.

If the transcription file is in XML format, also containing the indexing information, there is no need for an extra index file, and it can be made compatible with XARA and other corpus linguistics applications.

In our experience, despite initial format problems (computers with Chinese) the speed and accuracy of transcription has been increased and we hope to open this possibility to other interested linguists.

## 5. Acknowledgements

Special thanks go to the speakers of the aboriginal languages in the mountains of Taiwan who bring all the sacrifices to preserve their heritage. The authors express gratitude to Professor Jurg Gutknecht at the ETH Zurich who greatly supported the project. Greatest thanks go to the graduate students Oliver Hess and Christian Singer who devoted so much original planning and software programming to create this handy tool for less resourced languages [5][6].

## 6. References

- [1] //www.ethnologue.org/show\_language.asp?code=tsu
- [2] //www.ethnologue.org/show\_language.asp?code=sxr
- [3] //www.ethnologue.org/show\_language.asp?code=xnb
- [4] //formosan.sinica.edu.tw/
- [5] Hess O., SpeechIndexer – Halbautomatische Indexierung von Sprachdaten, diploma thesis at the Department of Computer Science, ETH Zurich, 2006.
- [6] Singer C., Entwicklung eines Pausenfinders und dessen Einbettung in SpeechIndexer, diploma thesis at the Department of Computer Science, ETH Zurich, 2006.