

# First Broadcast News Transcription System for Khmer Language

Sopheap Seng\*, Sethserey Sam\*', Laurent Besacier\*, Brigitte Bigi\*, Eric Castelli'

\* LIG 220, rue de la chimie, B.P. 53 38041 Grenoble Cedex 9

' MICA 1 Dai Co Viet, Hanoi, Vietnam

e-mail: sopheap.seng@imag.fr

## 1. Introduction

In this paper we present an overview on the development of a large vocabulary continuous speech recognition (LVCSR) system for Khmer, the official language of Cambodia, spoken by more than 20 million people. As an under-resourced language, develop a LVCSR system for Khmer is a challenging task. We describe the difficulties and our methodologies for quick language data collection and processing for automatic speech recognition. Furthermore, the approaches and tools (mostly Open Source) used for the development of our system are documented and made publicly available on the web. We hope this will contribute to accelerate the development of LVCSR systems for a new language, especially for an under-resourced language of developing countries where resources and expertise are limited.

## 2. Khmer language characteristics

With respect to speech recognition, the Khmer language bears challenging characteristics: (1) the lack of language resources (text and speech corpora) in numeric form, (2) the writing system without explicit word boundary, which calls for automatic segmentation approaches to make statistical language modeling feasible and (3) the acoustic and phonologic characteristics are not yet well studied.

The statistical nature of the approaches used in automatic speech recognition requires a great quantity of language resources in order to perform well. Such large resources are available for languages like English, French, German, Chinese or Japanese. For under-resourced languages which are mostly from developing countries, those resources are available in a very limited quantity because of its economic interest and the lack of standardized automatic processing tools (character encoding, word processing software). In this situation, language data collection is a challenging task and requires innovative approaches and tools. Similar to Chinese and Japanese, Khmer is written without spaces between words. A sentence in Khmer ពណ៌សម្តេចថាពណ៌ខ្មៅ could be segmented into ពណ៌|ស|ម្តេច|ថា|ពណ៌|ខ្មៅ (color|white|why|say|color|black) or ពណ៌|សម្តេច|ថា|ពណ៌|ខ្មៅ (color|king|say|color|black). A correct segmentation of a sentence into words requires the full knowledge of the vocabulary and the semantics of the sentence. The state-of-the-art automatic segmentation method which is generally based on a vocabulary can give only around 95% of correct word segmentation because of the ambiguities during segmentation. This makes text data processing for word n-gram language modeling more complicated and other segmentation approaches and modeling units must be investigated. Note that a text in Khmer language could also be segmented into syllables, characters cluster (groupe of inseparable characters) or characters. The characters cluster could be a potential modeling unit as its segmentation is trivial and has no ambiguities.

## 3. Language data Acquisition

### 3.1. Text Corpus

Create a statistical language model consists in estimating from a text corpus the probability of word n-gram. A large amount of in domain text data (several hundred millions words) is needed in order to obtain accurate probability estimation. As our system is targeted to automatic broadcast news transcription, the classic way to get in domain text data is to take content from daily newspapers. Method for text collection from the web is becoming more and more popular as the web allows to obtain freely and quickly a large quantity of text. Recently, several research

works proposed techniques to exploit the resources from the web for natural language processing. In [1], a web robot that retrieves text from the Internet to build a text corpus is proposed. From some given starting points on the web, the robot can reach and retrieve recursively text documents and html pages. However, we must control the robot in order to get only the text in the targeted language and domain. Another approach in [2] consists in calculating words n-gram probabilities using the search engine. The probabilities are estimated from the number of pages found using a given search engine. Those kinds of methods applied well to languages which have already a significant coverage on the Internet. For an under-resourced language like Khmer, the number of websites and the speed of Internet connections are often limited. There are only around 340,000 websites registered in Cambodian domain name *.kh* (results from Google) and most of them propose contents in English instead of Khmer.

In our case, retrieving the Khmer pages from some well selected daily news websites allows us to get big quantity of text more rapidly than using a robot to crawl many sites on the net as proposed in [1]. Once html pages are retrieved, further process are needed in order to build a text corpus:

- Filtering in order to extract only text from html pages
- Converting legacy character encoding to standard Unicode encoding
- Segmenting text into phrases and words using automatic segmentation tools
- Converting special signs and numbers to text
- Structuring the text corpus

By using our tool *ClipsTextTK* [3], this process could be partly done rapidly by adapting the language independent tools of the toolkit such as filtering and structuring. We developed tools for the conversion of encoding (from legacy ad-hoc code to Unicode), the conversion of special characters and numbers to text and the segmentation of text into phrases and words. The word segmentation tool is developed using an algorithm which segments a text into words based on a vocabulary of 36000 words obtained from the official Khmer dictionary (*Chhoun Nat* dictionary) with an optimization criteria: *longest matching*. Our segmentation tool, estimated on some heldout data, gives 95% of correct word segmentation.

The collection of Khmer text from the Internet allows us to get 25130 pages (448Mbytes) of *html* pages from 5 selected news websites. After filtering, converting and structuring process, a text corpus of 0.5 millions phrases, which lead to 15.5 millions words (249Mbytes) is obtained. In these 15.5 millions words, 15 % of out of vocabulary (OOV) words are found compared to the original 36k vocabulary. A non negligible part of the OOV words is probably due to the word segmentation errors, while another part corresponds to real OOV words.

### **3.2. Speech Corpus**

To train the acoustic models for our system, a speech corpus is needed. Speech corpora can be created by recording a well selected text read by professional readers in a studio. The recording task is however very time and resource consuming as we need to prepare the text data and scenarios and fully control the recording process. To obtain speech signal quickly and freely, we tried several techniques. The first consists in searching the web, the sites that propose the radio broadcast news in Khmer language. Many organizations such as Voice of America, Radio Australia and Radio FreeAsia have broadcast program in Khmer language and put on their website the entire broadcast news for public download and most of the time the scripts is also available. From those sites, we can retrieve quickly a big quantity of speech signal but with a poor quality (narrowband) because of compression rate used to make the file size smaller to download. In order to obtain a good quality speech signal, with help from our partner, Institut de Technologie du Cambodge (ITC) in Cambodia, we built a recording system from basic equipments: a computer with a radio receiver card installed, a recording program that we scheduled to record several hours of broadcast news of different radio stations in Phnom Penh, Cambodia. From this operation, we got recordings of 30h of good quality speech signal of radio broadcast news in Khmer language.

A manual transcription campaign of the recording speech signals was organized at ITC. Twenty volunteers (students at ITC) who were motivated to contribute to the development of the language resources for Khmer were recruited and trained to do the manual transcription. By using an Open Source tool *Transcriber* [4], 6h30mn of speech signal were manually transcribed in Unicode Khmer script (only speech read in the studio was transcribed and without extra detailed information). This 6h30mn of transcription contains 2950 phrases of 47000 words pronounced by 16 different speakers (4 women). 172 phrases (2 speakers) are then taken to serve as test corpus during the evaluation of our LVCSR system.

#### 4. Designing a pronunciation dictionary

The pronunciation dictionary is a key component for acoustic modeling. However, a standard Khmer pronunciation dictionary is not yet available. While a manually generated pronunciation dictionary gives a good quality, this task is time consuming and requires extended knowledge on the acoustic and phonology systems of the language in question. There were several techniques found in the literature for modeling a pronunciation dictionary. Among them we can mention [5] which proposed a modeling technique based on pronunciation rules. This method requires knowledge on the target language and also of its phonetization rules. A more automatic method was proposed in [6] and requires a good quality acoustic-phonetic decoder. Grapheme based modeling has been successfully addressed for different languages [7,8]. It has the advantage of being straightforward and fully automatic. As the letter-to-sound relation in Khmer language is relatively close, the process of pronunciation dictionary generation could be primarily done based on grapheme. For Khmer language, our grapheme based dictionary is generated by converting the words in its Unicode representation to a Roman representation using a simple Romanization technique.

#### 5. Choice of tools for ASR system development and experimentation results

We chose Open Source tools as our objective is also to be able to contribute a manual documentation that describes recipes (steps and tools) used to build the system for under-resourced language. For the text collection over the Internet, we use *wget*, a widely available tool comes with most of Linux distribution. This tool allows us to automate the task of web resources retrieval. As mentioned previously, our *ClipsTextTK* [3] is an Open Source toolkit, easily adapted to a new language and aims at building a general purpose text corpus. On the other hand, we use *Transcriber* [4] to make the manual transcription of our speech corpus.

For the language model, we used *Srilm* [9], a freely distributed tool for training language model. The tri-gram model was trained from our text corpus. The corpus is segmented into words, in our first system, to estimate word n-gram probabilities.

*Sphinx* [10] is an Open Source speech recognition framework. It was used to develop our ASR system. Note that in *Sphinx3.6* there is a *SphinxTrain* package which allows to train the acoustic model to use during the decoding phase with *Sphinx* decoder. With *SphinxTrain*, we were able to create a first context independent acoustic model from our transcribed speech corpus.

The evaluation of our system on our test corpus of 172 phrases give a first baseline results WER of 50.5%. This score is obtained when a re-segmentation is applied to the hypothesis before comparing to the references. For language without explicit word boundaries, the system error rate is generally given at the character level (CER). If we estimate the Characters Cluster Error Rate (CCER) of our first system, we obtained 33.6%. This latter unit is a good reference unit to use in system evaluation, better than word.

#### 6. Conclusion

Building a LVCSR system for Khmer has to deal with several challenging tasks such as the lack of language resources and the text segmentation problem. We proposed methodologies for rapid language data collection and processing based mainly on Open Source tools to build a speech

recognition system for Khmer language. Our second contribution is a publicly available manual of development of a speech recognition system for a new under-resourced language. Our future work consists in improving our ASR system for Khmer by trying to exploit the combination of other lexical and sub-lexical units instead of words, as the units for statistical language modeling.

## References

- [1] D. Vaufreydaz, “Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue”, *Thèse de doctorat de l'Université J. Fourier - Grenoble I*, France, 226 pages, January 2002.
- [2] X. Zhu, R. Rosenfeld, “Improving Trigram Language Modelling with the World Wide Web”, *ICASSP'01*, pp. 533-536, Salt Lake City, USA, Mai 2001.
- [3] <http://www-clips.imag.fr/geod/User/brigitte.bigi/logiciel.html>
- [4] C. Barras, et al, “Transcriber: development and use of a tool for assisting speech corpora production”, *Speech Communication special issue on Speech Annotation and Corpus Tools*, Vol 33, No 1-2, January 2000.
- [5] X. Huang, et al, “Spoken Language Processing – A Guide to Theory, Algorithm, and System Development”, *Practice Hall*, 2001.
- [6] T. Sloboda, A. Waibel, “Dictionary learning for spontaneous speech recognition”, *ICSLP'96*, Philadelphia, PA, USA, 1996.
- [7] Billa J. et al, “Audio indexing of Arabic broadcast news”, In Proceedings of the *IEEE International Conference on Acoustique, Speech and Signal Processing*. 2002 Orlando, FL, PP. 5-8
- [8] Bisani M., Ney H., “Multigram-based grapheme-to-phoneme conversion for LVCSR” In Proceedings of the *EUROSPEECH*. 2003 Geneva, Switzerland, pp.933-936
- [9] A. Stolcke. “SRILM -- an extensible language modeling toolkit”, In Proc. *Intl. Conf. on Spoken Language Processing*, 2002.
- [10] <http://cmusphinx.sourceforge.net/html/cmusphinx.php>