Review of Dissertation Thesis

Hybrid word-subword spoken term detection

submitted by Igor Szöke
at the Faculty of Information Technology, BUT in Brno

The thesis was completed in 2010. It is written in English and has137 pages.

Scientific background of the thesis:
The thesis deals with the task known as Spoken Term Detection (STD), or less formally as Key-word/phrase spotting. The goal of the STD is to find possible occurrences of given words or phrases in an acoustic signal that contains speech. The task may be very complex, especially when the signal has low quality, speech is informal and spontaneous, and more than one speaker's speech is present, e.g. in case of the STD in a telephone conversation. Solving the task may significantly help the police or security agencies when they need to search few words in large amounts of speech recordings. That is why the investigations focused on the STD belong to the important and hot topics in the speech technology research.

Main objectives of the work:
As there already exist several established approaches to the STD, the author decided to focus on the most promising one that is based on the employment of a large vocabulary continuous speech recognition (LVCSR) system. To make the task harder, he investigates how the LVCSR algorithms can be adapted to allow for the search of words that are not in the system's vocabulary. This leads to the design of a system that is hybrid in the way that it combines the in-vocabulary words with phoneme strings representing the out-of-vocabulary ones.

Methods and methodology:
It is evident that the author has very good knowledge of the topic. In chapters 2, 3 and 4, he presents a good overview of the theoretical background, an extensive analysis of known methods and approaches, and a detailed survey of evaluation metrics used in the STD task. In these chapters, he benefits from his rich experience gained during his participation in the international STD evaluation campaign organized by NIST in 2006. Therefore, the used methodology and experiment assessment are compatible to those used in other advanced research teams.

Results and achievements:
The original contributions of the thesis include (among others): several proposed and experimentally verified methods for the combination of word and subword units in the LVCSR-based STD search, the application of multigrams in the phone-based STD framework, or a definition of a new evaluation metric named Upper-Bound-Term-Weighted-Value. The experimental results achieved by the proposed methods are comparable to those offered by the other state-of-the-art algorithms. The fact that none of the author's methods brings significantly better results does not lessen the importance of his thesis.

Thesis and author's publication activity:
The list of author's publications contains more than 20 items, mainly contributions in conference proceedings. Eight of them are indexed in the ISI Web of Knowledge database. This is not a bad outcome of a 6-year PhD study period. I just wonder why the latest paper included in the list was published in 2008 and why there is no recent publication on the topic.

Conclusion:
The author of the thesis proved to have an ability to perform research and to achieve original scientific results. I recommend the thesis for presentation with aim of receiving the degree of Ph.D.

Some comments and questions:

1. In Table 2.1, the author compares several approaches to the STD and lists their advantages and drawbacks. The acoustic STD method is labeled as being slow in both the recognition and search phases. This statement is not fully correct. The speed of the decoder itself is actually faster than in the other approaches and also the (repeated) search can be made very fast, if we store a relatively small set of pre-computed values, i.e. if we adopt the same "pre-compute and store" strategy that is implicitly considered in evaluating the other two techniques.
2. In the thesis, the author reports on many experiments he has done during his research. He compares many different methods as well as their various modifications, explores the impact of many free parameters (e.g. pruning thresholds, language model factors, etc). In the text, it is often said that these free parameters have been tuned to get the best performance in each particular type of experiment. Can these tuning procedures be explained in more details, especially with respect to the data they were used in them?
3. In Chapter 7 and also in Appendix B, the author describes the basics of the grapheme-to-phoneme conversion technique and its application in the STD system. I would like to know, if the system utilizes alternative pronunciation variations, namely those that can be predicted because they may result from assimilation, co-articulation, etc.

In Liberec, August 14th, 2010

Jan Nouza
Institute of Information Technology and Electronics
Technical University of Liberec