Review of the Thesis of Ing. Igor Szöke: Hybrid Word-Subword Spoken Term Detection

Herbert Gish
BBN Technologies
Cambridge, MA 02138

The primary problem addressed in this thesis is that of finding keywords in speech. In particular it focuses on the very challenging problem of finding keywords that do not occur in the lexicon of a speech recognizer and it is usually referred to as the out of vocabulary problem or OOV problem.

One approach to the OOV problem is to use a speech recognizer for sub-word units. However this approach loses the power brought to speech processing by an LVCSR system. Another approach is to employ both types of recognizers and post process their outputs but this approach has some computational issues. The approach that Ing. Szöke follows integrates both types of recognizers into a hybrid spoken term detector. While this hybrid approach has been investigated previously it is still a challenging problem worthy of pursuit and the candidate brings to it a variety of new ideas definitely in keeping with the state-of-the-art of the technology in this field. In fact, we can readily say, that some of his previously published work has advanced the state-of-the-art in this field.

In his pursuit of his design and analysis of a hybrid system I have been very impressed by the thoroughness and depth of understanding he brings to the problem. For example, the candidate considered in depth the various methods for evaluating the performance of keyword spotting systems. This lead him to choose and modify a NIST defined performance measure to better suit his purposes. The modification removed the effect of threshold normalizations as a factor affecting performance.

One aspect of the original contribution of the thesis lies in the new ways that the hybrid system has been approached. In particular the approach focuses on both performance as given by a word spotting metric as well as the performance as measured by the complexity of the system that achieves the particular performance level. In addition an important contribution is made by the multigram modeling employed for the sub-word units. The author trains and utilizes a multigram model for sub-word recognition which gives the sub-word recognition some of the benefits of word recognition.

Ing. Szöke's multi-dimensional view of the performance of a hybrid recognition system includes the computational aspects of the systems, such as memory and cpu requirements as well as the performance metric. This aspect of his work as well as the work on the scoring function and multigram subword modeling in the context of the hybrid recognition system represents original and useful contributions to our understanding of OOV word recognition.

In examining the publications of the candidate I very satisfied with both the quantity and quality of the work. The work covers various aspects of speech processing and information extraction from speech as well as core components of the thesis and is at a level that readily supports his candidacy for the Ph.D. degree. The publications imply the candidate has excellent research skills.

Ing. Szöke does a very reasonable job in reporting on the history of keyword spotting (Section 2.4.1) and since I played a role in some of the early work I think I can provide some additional insight into some of this history. In particular Ing. Szöke traces some of the history of the use different recognizer outputs for the purposes of keyword spotting pointing out the first use of N-best lists [Wei95] and lattices [Jam95]. This recounting of the history omits the work of Jeanrenaud et al. [1] in which a keyword scoring metric is employed that calculates the posterior probability of the final state of a keyword at a time *t*. The computation of this posterior probability utilizes the forward and backward scores of the Baum-Welch algorithm and is in effect a lattice computation that utilizes the most comprehensive of all lattices. Weintraub [Wei95] cites an earlier version of this approach to scoring (in Szöke's reference [RRRG89]) and recognizes its applicability to the keyword scoring of recognizer outputs but I believe that he doesn't quite interpret the scoring correctly.

I have omitted the details of the computation and recommend that the candidate check out reference 1 below and evaluate my assessment.

In Section 2.4.6 on confidence measures the author notes that the original purpose of confidence measures was to reject misrecognized words or utterances for dialog or speech understanding systems [PSG98]. I would suggest that the original purpose, using posterior probabilities as the word confidence measures as in [RRRG89] or in reference 1 below, was scoring for keyword occurrences. The next application, quite possibly, was for topic discrimination – see reference 2 below.[2]

I consider the issues I have raised relating to the history of keyword spotting and confidence measures to be relatively minor but worth the author's attention.

In conclusion, after having examined Ing. Szöke's doctoral thesis I am of the opinion that it meets the requirements of the proceedings leading to the Ph.D. title conferment.

---

[1] P. Jeanrenaud, K. Ng, M. Siu, J.R. Rohlicek and H. Gish, "Phonetic-based word spotter: Various configurations and application to event spotting," *Eurospeech* 1993, pp. 1057-1060.

[2] J. McDonough, H. Gish, J. R. Rohlicek, K. Ng, P. Jeanrenaud, Topic discriminator using posterior probability or confidence scores, United States Patent 5625748, 1997.