KEYWORD SPOTTING IN MEETING DATA

Ing. Igor Szöke, Doctoral Degree Programme (3)
Dept. of Computer Graphics and Multimedia, FIT, BUT
E-mail: szoke@fit.vutbr.cz

Supervised by: Dr. Jan Černocký

ABSTRACT

This paper describes several approaches to keyword spotting (KWS) for informal continuous speech. We compare acoustic keyword spotting, spotting in word lattices generated by large vocabulary continuous speech recognition and a hybrid approach making use of phoneme lattices generated by a phoneme recognizer. The systems are compared on carefully defined test data extracted from ICSI meeting database. The acoustic and phoneme-lattice based KWS are based on a phoneme recognizer making use of long temporal split context feature extraction and posterior estimation using neural nets. The advantages and drawbacks of different approaches are discussed.

1 INTRODUCTION

Keyword spotting (KWS) systems are used for detection of selected words in speech utterances. Searching for various words or terms is needed in spoken document retrieval which is a subset of information retrieval. KWS in spoken speech differs from searching in written text by the ambiguity, and we have to count on inaccuracies of recognition systems. Therefore, the estimation of *confidence* of the found keyword is of crucial importance.

The search of keywords and computation of confidence can be done in several ways:

- acoustic KWS, where the model of the keyword is composed of phoneme models at the time the keyword is entered and can provide the on-line keyword spotting.
- an approach making use of transcription of speech into discrete units (words or phonemes) and its storing in structure of parallel hypothesis (called lattices). Lattices can be indexed and quickly searched off-line.

This paper deals with the comparison of these three approaches to KWS and their evaluation on informal continuous speech (recordings of meetings) within AMI project.

2 CONFIDENCE ESTIMATORS FOR KWS

The state-of-the-art approaches of keyword confidence estimation are described in [3]. The keyword hypothesis (given by a speech recognizer) can be compared to other

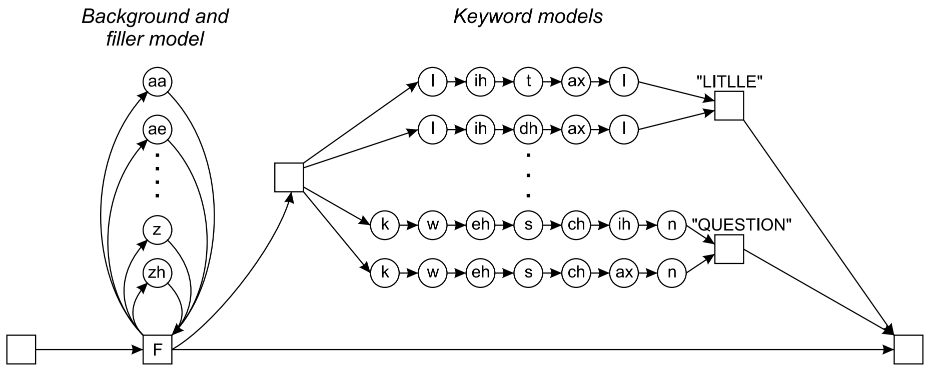

Figure 1: Keywords spotting network using mono-phones.

(non-keyword) hypothesis in same time. Mathematical statistics calls this a *hypothesis testing*. An optimal solution of hypothesis testing is Neyman-Pearson Lemma. The most powerful test is the likelihood ratio, when the Neyman-Pearson Lemma can be rewritten to

$$LR(kw) = \frac{p(H_0(kw))}{p(H_1(kw))}, \tag{1}$$

where $LR(kw)$ is the likelihood ratio test for the keyword, $p(H_0(kw))$ and $p(H_1(kw))$ are probability density functions of null and alternative hypothesis respectively. The *null hypothesis* means that the keyword exists and is correctly recognized in a portion of speech. The *alternative hypothesis* means that there is no keyword or the keyword is incorrectly recognized. Our goal is to test the null hypothesis against the alternative hypothesis. The null hypothesis is accepted if

$$LR(kw) > Th, \tag{2}$$

where Th is a threshold. Probability density function of null hypothesis can be modelled by for example Gaussian Mixtures combined with Hidden Markov models (GM/HMM). Precise modelling of alternative hypothesis is more difficult and it is still not clear.

2.1 LIKELIHOOD RATIO AS THE KEYWORD CONFIDENCE

The null hypothesis is represented by the keyword model and the alternative hypothesis is usually represented by a background model. The model of keyword is concatenated from phoneme models (we allow also for pronunciation variants) and the background model can be a phoneme loop or a set of anti-keywords.

Figure 1 shows an example of on-line acoustic keyword spotting network which is an implementation of evaluation of the likelihood ratio. The left filler and background models are phoneme loops. The filler model is used for “catch” non-keyword speech. As these systems were developed for real-time operation, we have not used any right filler model. The likelihood of the keyword is taken from the last state of keyword model and immediately compared with the likelihood at the output of background model (node F).

2.2 POSTERIOR PROBABILITY AS THE KEYWORD CONFIDENCE

In this case, the complete output of speech recognizer is available in form of a graph of parallel hypothesis (lattice), the posterior probability of the keyword can be computed.

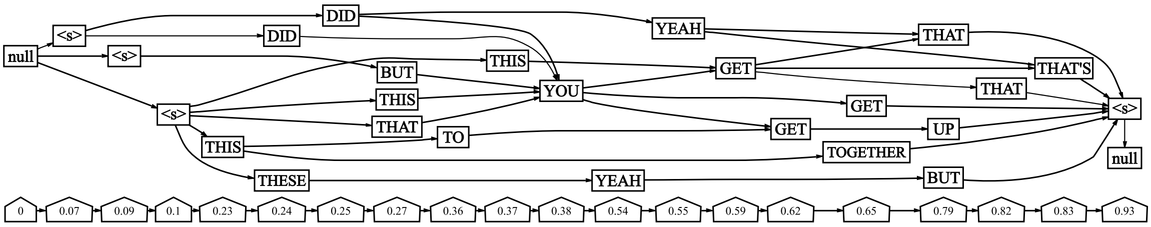

Figure 2: A word lattice - the output of LVCSR system.

Lattice is forward acyclic graph, time is represented by nodes and words or phonemes by arcs[1]. Starting or ending node of the arc means the beginning or the ending time of the word respectively. Each word has also attached the likelihood (acoustic and language) given by the recognizer. See figure 2 for an example of the lattice.

Let's define the joint likelihood from the beginning of a lattice to a node N. It is denoted $L^{\alpha}(N)$ and called *forward likelihood*. The joint likelihood from a node N to the end of a lattice and is denoted $L^{\beta}(N)$ and called *backward likelihood*. The meaning of likelihoods L^{α} and L^{β} is, how likely it is to go from the beginning of the lattice to the node, or from the node to the end of the lattice. We should take in account all possible paths from the beginning or from the end respectively. Mathematical definition is the following:

$$L^{\alpha}(N) = \sum_{p \in P_b^N} \prod_{A_i = A_1^p}^{A_n^p} L(A_i) \tag{3}$$

$$L^{\beta}(N) = \sum_{p \in P_N^e} \prod_{A_i = A_1^p}^{A_n^p} L(A_i) \tag{4}$$

where P_b^N and P_N^e is the set of all possible paths from lattice beginning to node N and from node N to lattice end. A_1 is the first arc of the path and A_n is the last arc of the path.

The meaning of posterior probability of the keyword is: how likely is to go through the keyword from the beginning to the end of the lattice divided by how likely is to go through the lattice. Mathematical definition of posterior probability of keyword kw in lattice $latt$ is:

$$PP^{latt}(kw) = \frac{L^{\alpha}(N_b(kw))L(kw)L^{\beta}(N_e(kw))}{L^{\alpha}(N_e(latt))} \tag{5}$$

where $N_b(kw)$ and $N_e(kw)$ is the beginning or ending node of the keyword respectively. The $L(kw)$ is the likelihood of the keyword and $N_e(latt)$ is the last node of the lattice. In case the keyword is composed of a string of paths (in the case of phoneme lattice), the keyword likelihood is the product of all arc's likelihood:

$$L(kw) = \prod_{u \in U(kw)} L(u) \tag{6}$$

where U is set of units belonging to keyword kw and $L(u)$ is the likelihood of unit u.

[1] Another possibility is to represent a word or a phoneme as end node of an arc — in this case the arc represents only timing information and likelihoods. Both these formats are equal.

Type	System	Models	FOM
Acc. KWS	LCRC NN ICSI40h	mono-phones	**67.43**
Acc. KWS	GM/HMM NIST2005	tri-phones	**89.00**
Phn Latt. KWS	LCRC NN ICSI40h	(lattice)	**65.00**
Word Latt. KWS	GM/HMM NIST2005	(lattice)	**91.00**

Table 1: The results of different KWS systems.

3 RECOGNIZERS AND EVALUATION

Phoneme posterior estimator used for on-line acoustic keyword spotting and phoneme lattice keyword spotting (denoted as **LCRC NN ICSI40h**) is based on split temporal split and a cascade of 3 neural networks [5]. The temporal context of critical band spectral densities is split into left and right context (LC-RC) parts. Both parts are processed by DCT to de-correlate and reduce dimensionality. The feature vector created by concatenation of vectors over all filter bank energies is fed to NN. Two NNs are trained to generate phoneme-state posterior probabilities for left- and right-context parts respectively. Third NN functions as a merger and produces final set of posteriors.

A large GM/HMM based LVCSR recognizer using advanced techniques for acoustic modelling such as vocal tract length normalization (VTLN), minimum phoneme error training (MPE) and a powerful language model [1] is used for word lattice keyword spotting (denoted as **GM/HMM NIST2005**).

Our keyword spotting systems were tested on a large database of informal continuous speech of ICSI meetings [2] (sampled at 16 kHz). Attention was paid to the definition of fair division of data into training/development/test parts with non-overlapping speakers. It was actually necessary to work on speaker turns rather than whole meetings, as they contain many overlapping speakers. We have balanced the ratio of native/nonnative speakers, balanced the ratio of European/Asiatic speakers and moved speakers with small portion of speech or keywords to the training set. The amounts of data in the training, development and test parts are 41.3 h, 18.7 h and 17.2 h respectively. The development part was used for tuning of the system.

In the definition of keyword set, we have selected 17 of the most frequently occurring words (each of them has more than 95 occurrences in each of the sets) but checked, that the phonetic form of a keyword is not a subset of another word nor of word transition. The percentage of such cases was evaluated for all candidates and words with high number of such cases were removed.

Our experiments are evaluated using *Figure-of-Merit* (FOM) [4], which is the average percentage of correct detections per 1, 2, . . . 10 false alarms per hour. We can approximately interpret it as the accuracy of KWS provided that there are 5 false alarms per hour.

4 RESULTS

The results of acoustic keyword spotting, phoneme lattice and word lattice keyword spotting are in table 1. More detailed experiments were published in [6] and [7].

The best accuracy is provided by system using searching in LVCSR word lattices and keyword confidence computation using likelihood ratio. The usefulness of LVCSR-KWS is however limited - the keyword must be contained in the LVCSR's vocabulary. This scenario is on the other hand well handled by the other two approaches. The phoneme-lattice based KWS is not reaching the accuracies of the acoustic KWS, but can be used for a fast pre-selection of candidates. Especially in case of searching in archives containing hundreds or thousands hours of speech, the speed of search is as important as the accuracy. An indexation and search system for fast match of a query with a set phoneme lattices is currently under developed in our group.

ACKNOWLEDGEMENTS

This work was partially supported by EC project Augmented Multi-party Interaction (AMI), No. 506811 and Grant Agency of Czech Republic under project No. 102/05/0278.

REFERENCES

[1] T. Hain et al. The 2005 AMI system for the transcription of speech in meetings. In Proceedings of Rich Transcription 2005 Spring Meeting Recognition Evaluation Workshop, Edinburgh, July 2005.

[2] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. The ICSI meeting corpus. In International Conference on Acoustics, Speech, and Signal Processing, 2003. ICASSP-03, Hong Kong, April 2003.

[3] H. Jiang. Confidence measures for speech recognition: A survey. In Speech Communication, volume 45.

[4] J. R. Rohlicek, W. Russell, S. Roukos, and H. Gish. Continuous hidden markov modeling for speaker-independent word spotting. In International Conference on Acoustics, Speech, and Signal Processing, 1989. ICASSP-89, volume 1, Glasgow, UK, May 1989.

[5] P. Schwarz, P. Matějka, and J. Černocký. Hierarchical structures of neural networks for phoneme recognition. In Proceedings of ICASSP 2006, Toulouse, France, May 2006.

[6] I. Szöke, P Schwarz, P. Matějka, L. Burget, M. Karafiát, M. Fapšo, and J. Černocký. Comparison of keyword spotting approaches for informal continuous speech. In Proceedings of Eurospeech 2005, Lisaboa, Portugal, September 2005.

[7] I. Szöke, P Schwarz, P. Matějka, L. Burget, M. Karafiát, and J. Černocký. Phoneme based acoustics keyword spotting in informal continuous speech. In Proceedings of TSD 2005, Karlovy Vary, Czech Republic, September 2005.