Comparison of Keyword Spotting Approaches for Informal Continuous Speech

Igor Szöke, Petr Schwarz, Lukáš Burget, Martin Karafiát, Pavel Matějka, Jan Černocký

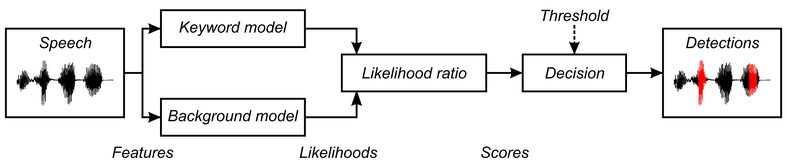
Brno University of Technology, Faculty of Information Technology
e-mail: szoke@fit.vutbr.cz

Abstract

This poster 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 advantages and drawbacks of different approaches are discussed. The acoustic and phoneme-lattice based KWS are based on a phoneme recognizer making use of temporal-pattern (TRAP) feature extraction and posterior estimation using neural nets. We show its superiority over traditional HMM/GMM systems. We also propose a posterior masking algorithm to speed-up acoustic keyword spotting.

Introduction

To know/have the Information is important. The goal for Keyword Spotting (KWS) is to find the keyword (the information) in speech data.



We have three different methods of keyword detection:

1. Acoustic KWS
2. KWS based on word (LVCSR) lattice searching
3. KWS based on phoneme lattice searching

All three techniques use likelihood ratio approach for keyword score calculation:

$$S_{KW} = L(KW)/L_{bkg}$$

Evaluation

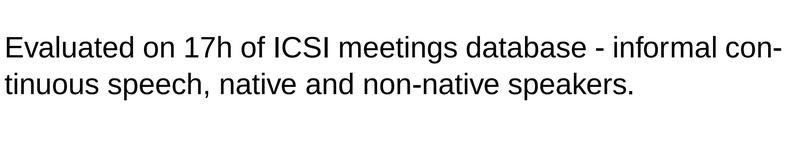
Evaluated on 17h of ICSI meetings database - informal continuous speech, native and non-native speakers.

Test name	num. of words	pron. var.	note
Test 17	17	33	most frequent
Test 1	2310	3514	rare, appear 1×
Test 5	4104	6537	rare, at most 5×
Test 10	4710	7567	rare, at most 10×

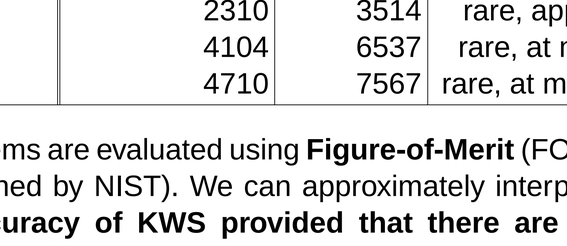
All systems are evaluated using **Figure-of-Merit** (FOM) metric (defined by NIST). We can approximately interpret it as **the accuracy of KWS provided that there are 5 false alarms per hour**.

Speech recognizers

Our phoneme recognizer (TRAPNN-LCRC) is based on **temporal patterns** (TRAPs) and **neural networks** (NN) with split context (left and right).



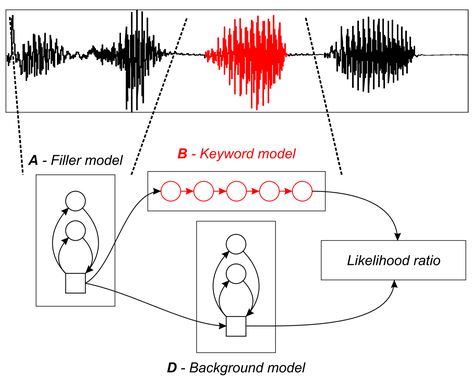
It produces **a matrix of phoneme posterior probabilities**. Each phoneme is divided into **3 "states"**.



This **posterior matrices** are used as features for **acoustic KWS** and for **phoneme lattices** generation.
Word (LVCSR) lattices are generated using **standard LVCSR GMM/HMM** based system. Both systems for lattices are trained on 10h of ICSI meetings, acoustic KWS system is trained on 40h of ICSI meetings.

Acoustic KWS

"Listening" of speech signal. Special recognition network and decoder are used for calculation of likelihood ratio.



Computed score is thresholded and alarm is raised each time there is local minimum of score below the threshold.

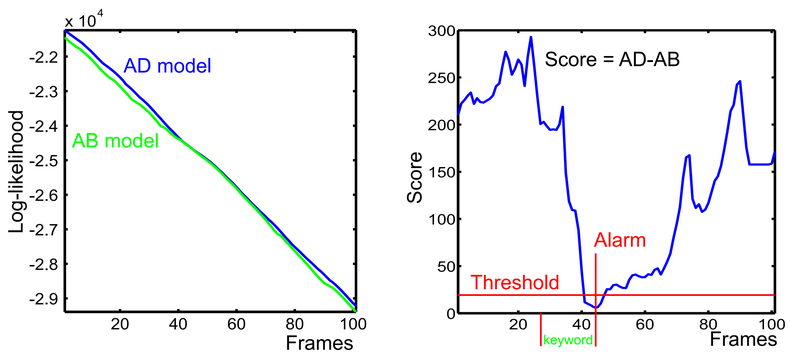
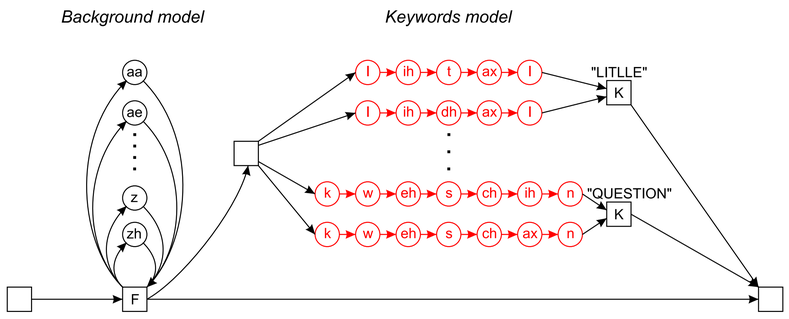


Illustration of monophone keyword spotting network.



System	Phonemes	FOM	Net size
GMM/HMM ICSI10h	CI	**47.77**	263, 337
GMM/HMM CTS277h-adap	CD	**63.66**	24k, 68k
TRAPNN-LCRC	CI	**64.46**	263, 337

Word and phoneme lattice KWS

Lattice based keyword spotting system searches for the keyword in word or phoneme lattices.

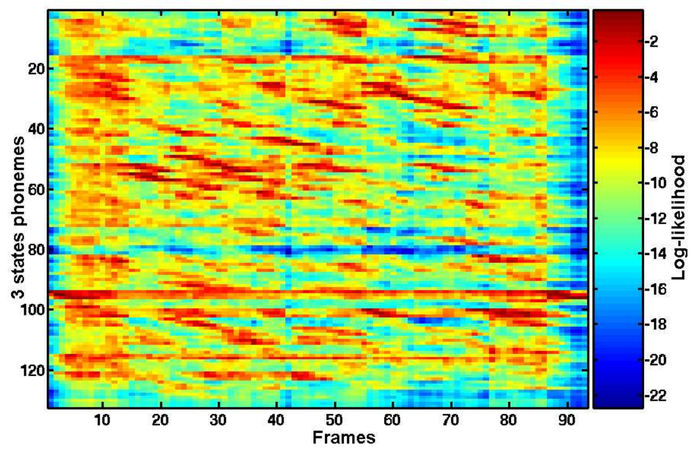
- Searching in **word lattice** - just to find the keyword (single link) and compute the score.
- Searching in **phoneme lattice** - search for sequence of links. Due to lower phoneme recognition accuracies, searching algorithm needs to allow for substitutions and insertions. Penalization of substitutions and insertions uses phoneme confusion matrix.

Results for **phoneme lattice keyword spotting** depending on branching factor and searching method.

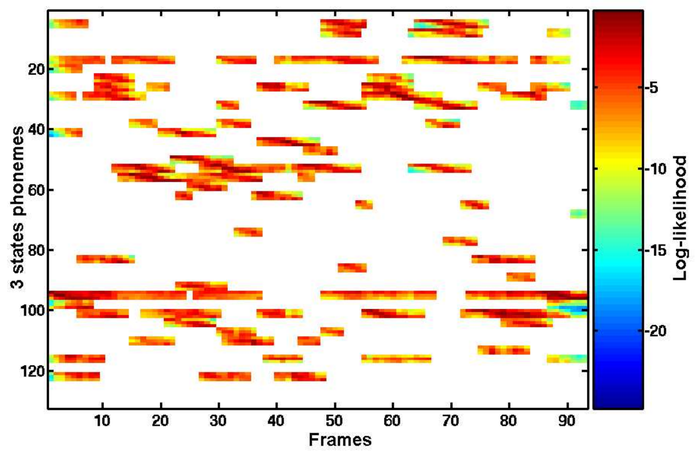
Method	Branching factor					
	2	3	4	5	6	7
	32.66	43.85	49.91	53.45	55.30	56.33
Ins	32.86	44.44	50.39	54.09	55.86	56.89
Subs	36.91	**51.25**	55.66	**57.45**	**58.32**	**58.90**
Ins Subs	**37.03**	51.08	**55.76**	57.20	58.22	58.84
Size ratio	1.0	4.9	10.6	17.6	25.7	35.4

Masking of posterior matrix

Method for **optimization of acoustic KWS**. Using of phoneme lattice for masking of posterior probability matrix.



White place was deleted. Only parts (phonemes) which appear in phoneme lattice remain.



- Uses less space (up to 2/3 smaller).
- Faster searching (decoding) (up to 1/3 faster).

Branching	2	3	4	5	6	7	Baseline
Phn. KWS	36.91	51.25	55.66	57.45	58.32	**58.90**	-
Ac. KWS	37.85	51.87	58.04	60.38	61.29	**61.46**	61.53
Place saved	93%	89%	85%	80%	77%	**73%**	0%

Results and conclusion

Test set	Acoustic	Word lattice	Phoneme lattice
Test 17	64.46	**66.95**	60.03
Test 10	72.49	66.37	∼(62)
Test 5	74.11	64.71	∼(63)
Test 1	**74.95**	61.33	∼(**66**)

- Merging of word lattice and phoneme lattice and/or acoustic keyword spotting is needed to get good KWS system (fast, accurate and without dictionary limitation).
- Word lattice KWS provides very good results with common words.
- Superiority of acoustic and phoneme lattice KWS, over word lattice KWS is noticable in the test involving rare words.

Property	Acoustic KWS	Word KWS	Phn. KWS
Searching speed	Slow	Fast	Fast
Searching in	Acoustic	Text	Text
Vocabulary	Open	Close	Open
Accuracy	Average	Higher	Lower
Hits/FAs ratio	Average	Lower	Higher
Mode	On-line or Off-line	Off-line	Off-line

Future work

- Improvement of phoneme recognizer (front-end to acoustic and phoneme lattice KWS).
- Indexation of word and phoneme lattices to get realtime reply in KWS system working with >1000h of data.
- Merging of word and phoneme lattice and acoustic KWS to get fast and reliable KWS system.