Phoneme Based Acoustics Keyword Spotting in Informal Continuous Speech *

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

Faculty of Information Technology, Brno University of Technology, Czech Republic
szoke@fit.vutbr.cz

Abstract. This paper describes several ways of acoustic keywords spotting (KWS), based on Gaussian mixture model (GMM) hidden Markov models (HMM) and phoneme posterior probabilities from FeatureNet. Context-independent and dependent phoneme models are used in the GMM/HMM system. The systems were trained and evaluated on informal continuous speech. We used different complexities of KWS recognition network and different types of phoneme models. We study the impact of these parameters on the accuracy and computational complexity, and conclude that phoneme posteriors outperform conventional GMM/HMM system.

1 Introduction

Acoustic keyword spotting (KWS) systems are widely used for the detection of selected words in speech utterances. Searching for various words or terms is needed in applications such as spoken document retrieval or information retrieval. An advantage of acoustic keyword spotting is in the possibility to spot out-of-vocabulary words, which are dropped in LVCSR systems. The paper deals with comparison of different KWS systems and their evaluation on informal continuous speech (recordings of meetings) within AMI project.

The paper first discusses training and testing data sets. Metrics of evaluation are defined later. The configuration of the GMM/HMM and the FeatureNet phoneme posterior estimator is discussed next. Description of several types of recognition networks of acoustic KWS system follows. Results are discussed and conclusions are drawn at the end of the paper.

A modern acoustic keyword spotter was proposed in [4] and it is based on maximum likelihood approach [1]. General KWS network using phoneme models is shown in Figure 1. Parts denoted A and C are filler models (phoneme loop) which model non-keyword parts of utterance. Part B is linear model for given keyword. Part D is a background model (phoneme loop) which models the same

* 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. Jan Černocký was supported by post-doctoral grant of Grant Agency of Czech Republic No. GA102/02/D108.

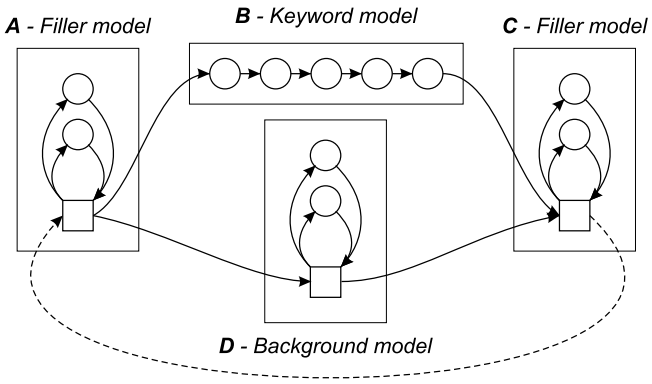

Fig. 1. General model of keyword spotting network.

part of the utterance as the keyword model. The confidence L_R of detected keyword is computed as likelihood ratio $L_R = L_{ABC}/L_{ADC}$, where L_{ADC} is likelihood of the best path through the model not containing the keyword and L_{ABC} is likelihood of a path through this same model containing the keyword.

2 The Data and Evaluation

Our keyword system was tested on a large database of informal continuous speech of ICSI meetings [3] (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 training/development/test parts division is 41.3, 18.7 and 17.2 hours of speech respectively. Development part is used for phoneme insertion penalty tuning.

In the definition of keyword set, we have selected 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. The final list consists of 17 keywords: `actually, different, doing, first, interesting, little, meeting, people, probably, problem, question, something, stuff, system, talking, those, using`.

Our experiments are evaluated using *Figure-of-Merit* (FOM) [4], which is the average 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.

Realtime coefficient (computational cost) was measured for some experiments. That was done at one computer with Intel P4 2.5 GHz HT processor. Computational cost experiment was run on 0.67 h of test set which equals to 2405 s. There was no other

load on test computer. The RT coefficient means ratio between *total time spent by CPU* and *total time of utterances*.

$$RT = \frac{T_{CPU}}{T_{utterances}} \tag{1}$$

3 Acoustic Keyword Spotting System

Presented KWS system is based on that described in [4], but we did some simplifications to run our system on-line. The after-keyword filler model C is not used. Background model D and front filler model A are grouped – they are the same phoneme loops. An example of our recognition network (for context-independent phonemes) is shown in Figure 2. The network has two parts: *keyword models* and *filler and background model*. Each keyword model contains concatenated phoneme models, we allow also for pronunciation variants. After a token goes through the keyword model to the end node, corresponding token is taken from phoneme loop (node `F`). Then likelihood ratio of these two is computed. Background and filler model contains no language model.

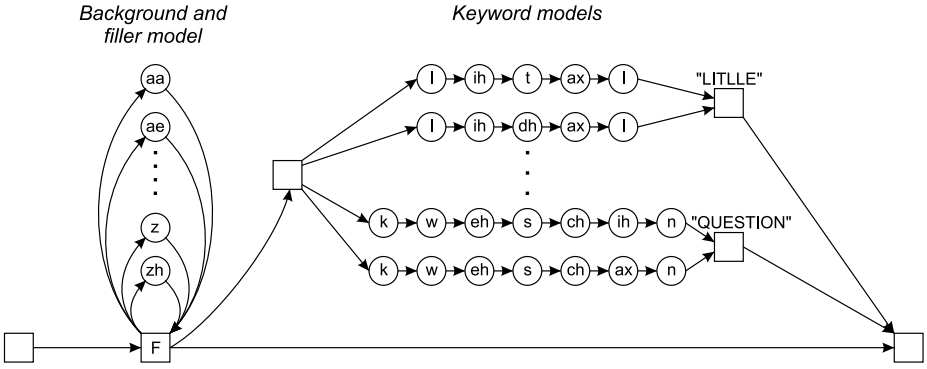

Fig. 2. Keywords spotting network using context-independent phonemes.

3.1 GMM/HMM System

Gaussian mixture model hidden Markov model based system is trained on 10 h long subset of the training set (denoted as **ICSI10h**). Raw data was parameterized using 13 Mel-frequency cepstral coefficients with Δ and $\Delta\Delta$. Two different sets of models were trained. Context-independent phonemes (*phonemes*, 43 units) and context-dependent phonemes (*triphones*, 77659 units). Standard training technique for GMM/HMM based on HTK was used.

Another set of experiments were done using triphone models trained on conversational telephone speech (*CTS*) database. CTS database contains about 277 hours of speech (mainly from the the Switchboard database), see Table 3.1. Raw data was parameterized using 13 perceptual linear prediction (*PLP*) coefficients [2] with Δ and $\Delta\Delta$. Parameters were normalized by cepstral mean and variance normalization. Models

trained on CTS database are denoted as **CTS277h-noad**. ICSI database was down-sampled to 8 kHz and PLP parameterized in the same way as CTS database. Then the *CTS277h-noad* models were adapted using MAP adaptation on full ICSI train set (adapted models are denoted as **CTS277h-adap**).

Database	Time
Switchboard 1	248.52 h
Switchboard 2 - Cellular	15.27 h
Call Home English	13.93 h
Sum	277.72 h

Table 1. Definition of training set of CTS database.

3.2 LCRC FeatureNet System

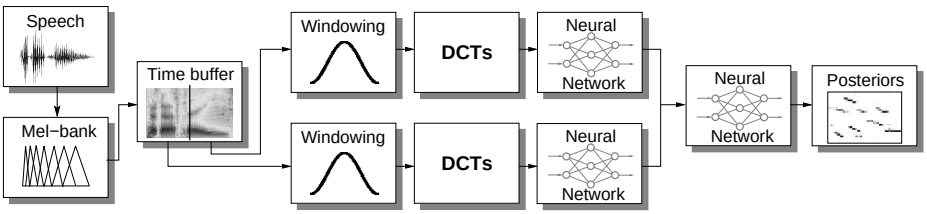

Fig. 3. Phoneme recognizer with split temporal context

Another approach to get acoustics keyword spotting is to use our LCRC FeatureNet phoneme recognizer [5]. It is a hybrid system based on Neural Networks (NN) and Viterbi decoder without any language model. An unconventional feature extraction technique based on long temporal context is used: The temporal context of critical band spectral densities is split into left and right context (*LCRC*). This allows for more precise modelling of the whole trajectory while limiting the size of the model (number of weights in the NN). Both parts are processed by DCT to de-correlate and reduce dimensionality. The feature vector which is feed to NN is created by concatenation of vectors over all filter bank energies. Two NNs are trained to produce the phoneme posterior probabilities for both context parts. Third NN functions as a merger and produces final set of posterior probabilities. We use 3-state models and the NN produces 3 posteriors for the beginning, the center and the end of a phoneme.

The LCRC FeatureNet system is trained on the same 10 h long subset of the training set as GMM/HMM system (denoted as **LCRC10h**) and on full training set 41.3 h (denoted as **LCRC40h**). System produces posteriori probability only for context-independent phoneme set (*phonemes*, 43 units).

Posteriori Probability Transformation Logarithmized histogram of distribution of linear posteriori probability is plotted in Figure 4. Two maxima at 0 and 1 are caused by the fact that NN is always too sure that a phoneme is or is not present (it is trained to discriminate). For the decoding, it is useful to "soften" these maxima.

The posterior probabilities are logarithmized. Logarithm function has ability to "increase resolution" close to 0, so the probabilities will not be so concentrated there. We can see the histogram of logarithmized posterior probabilities in Figure 5. The distribution of log-probabilities is nearly gaussian. But there is still one peak close to 0, which is caused by the value of linear probabilities around 1.

We design a posterior transformation to soften out both maxima. The transformation function (so-called *PostTrans*) contains two logarithms:

$$PostTrans_{(I,L,R)}(P) = \begin{cases} \log_L(\frac{P}{I}), & P < I \\ \log_R(\frac{P}{1-I}), & P \geq I \end{cases} \quad (2)$$

where L, R and I are constants. We experimentally tuned the constants to $I = 0.1$, $L = 50$ and $R = 1.1$. A new *LCRC40h* system with posteriori transformation $PostTrans_{(0.1,50,1.1)}$ is denoted as **LCRC40hPostTrans**.

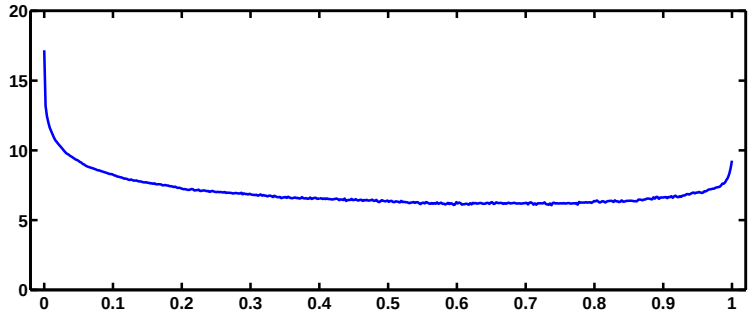

Fig. 4. Logarithmized histogram of linear posteriori probability distribution.

3.3 Recognition Networks

Experiments were done with the following KWS recognition networks. All recognition networks contain no language model.

- A network consisting of phonemes (denoted as **CI**, Figure 2).
- A network consisting of reduced set of triphones (denoted as **CDred**). Background and filler model of *CDRed* network contains only triphones which appears in keyword models. No context sensitive links are used (eg. triphone `A-B+C` has a link to triphone `D-E+F`) in *CDred* network.
- A network consisting of reduced set of triphones and phonemes (denoted as **CI&CDred**). It is *CDRed* network with added phonemes to the filler and background model.

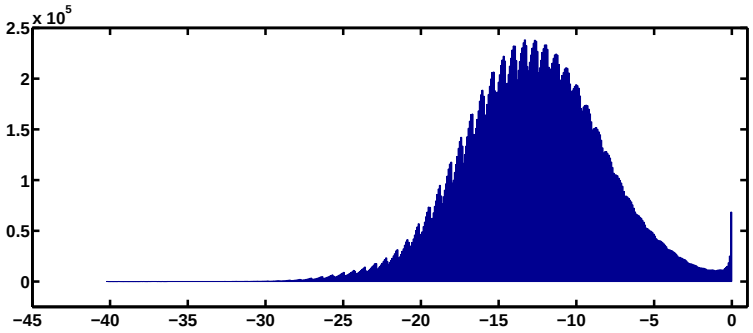

Fig. 5. Histogram of logarithmized posteriori probability.

– A network consisting of full set of triphones (denoted as **CD**, Figure 6). Links among triphones in *CD* network are context sensitive (eg. there is link between triphones `A-B+C` and `B-C+D` but not between `A-B+C` and `D-E+F`). The first and the last phoneme of keyword is expanded to all context possibilities in triphone networks (*CDred*, *CI&CDred* and *CD*). *CD* network was also optimized using algorithm for finite state automaton minimization (denoted as **CDopt**).

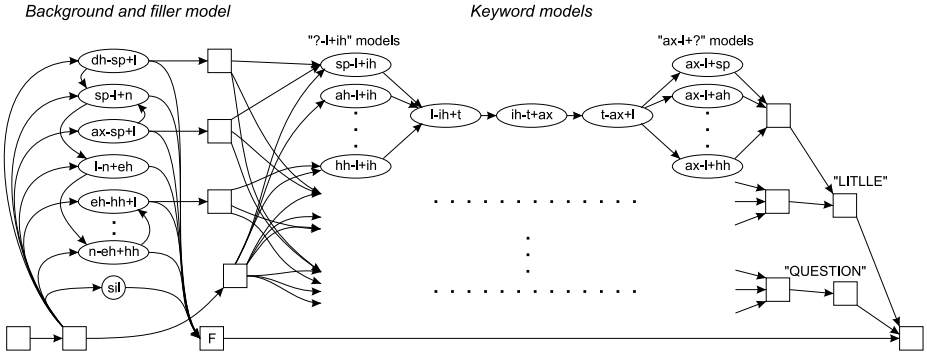

Fig. 6. Keyword spotting network using full set of triphones.

4 Results

Results of GMM/HMM based acoustic keyword spotting system are listed in Table 2. The *ICSI10h* experiment set shows, that using full set of triphones (*CD* network) gives the best FOM. Disadvantage of this approach is huge network size and slow decoding. *CDred* is a good compromise between speed and accuracy. Comparison between *CTS277h-noad* and *CTS277h-adap* with *CDred* network shows 3% improvement caused

by adaptation of models. Network optimization has important impact on the speed. The system is about 8× faster. The best GM/HMM based keyword spotting system which we have is the *CTS277h-adap* using *CDopt* network with 63.66% FOM.

Model	Network	#HITs	#FAs	#KWs	FOM	Realtime coefficient	Net size nodes, links
ICSI10h	CI	3142	2877867	3289	**47.77**	0.51	264, 339
ICSI10h	CDred	3177	2774259	3289	**57.15**	1.07	4375, 8461
ICSI10h	CI&CDred	3164	2904486	3289	**57.52**	1.50	4417, 8545
ICSI10h	CD	3173	2914897	3289	**61.88**	56.62	102k,3508k
CTS277h-noad	CDred	3189	2752492	3289	**56.39**	–	7637, 14742
CTS277h-adap	CDred	3159	2927968	3289	**59.39**	–	7637, 14742
CTS277h-adap	CD	3147	3032251	3289	**63.66**	73.03	119k, 4256k
CTS277h-adap	CDopt	3147	3032251	3289	**63.66**	8.50	28k, 83k

Table 2. The results of different acoustic keyword spotting systems based on GM/HMM.

The results of LCRC FeatureNet based acoustic keyword spotting system are listed in Table 3. All the LCRC FeatureNet systems work with *CI* network. This makes the decoding (the posteriors can be pre-computed) very fast – the realtime coefficient is 0.021. The best LCRC FeatureNet based keyword spotting system is the one using posteriori transformation function *LCRC40hPostTrans* with 64.46% FOM. The ROC curves are ploted in Figure 7.

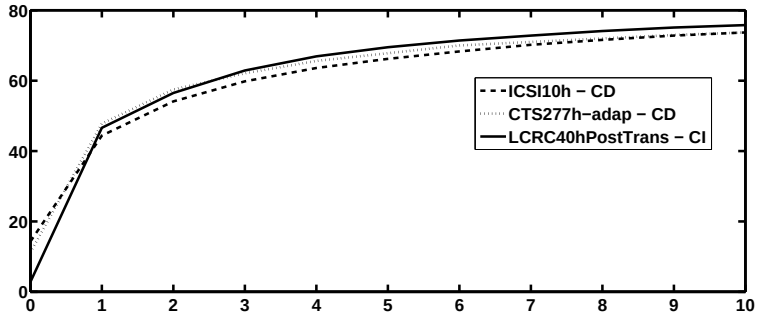

Fig. 7. ROC curves of three best systems. X-axis is the number of FAs per hour, Y-axis is the percentage of detected keywords.

Model	Network	#HITs	#FAs	#KWs	FOM	Realtime coefficient	Net size nodes, links
LCRC10h	CI	3145	3065025	3289	**61.39**	0.021	263, 337
LCRC40h	CI	3153	3062984	3289	**62.46**	0.021	263, 337
LCRC40hPostTrans	CI	3148	3031465	3289	**64.46**	0.021	263, 337

Table 3. The results of different acoustic keyword spotting systems based on LCRC FeatureNet.

5 Conclusion

The paper deals with a comparison of different KWS systems and their evaluation on informal continuous speech (recordings of meeting). We measured the accuracy of GMM/HMM based and FeatureNet based systems. *Figure-of-merit* scoring method was used to compare the performance of systems. The test data-set (about 17 h of speech and 17 searched keywords) was designed for statistically reliable results.

The best system using Gaussian mixtures hidden Markov models approach is triphone models trained on 277.72 h of narrow-band conversational telephone speech corpus adapted to target ICSI database. FOM of the system is 63.66%. Triphone system trained on 10 h of wide-band ICSI database has reached 61.88% FOM. The best LCRC FeatureNet system was trained on 41.3 h of ICSI database. It generates phoneme posteriori probabilities which are transformed using proposed function. FOM of the LCRC FeatureNet system is 64.46% and outperform conventional GMM/HMM system. The system uses *CI* network in comparison to *CDopt* network for the best GMM/HMM system, which makes it simple and fast.

References

1. L. R. Bahl, F. Jelinek, and R. L. Mercer. A maximum likelihood approach to continuous speech recognition. In *IEEE Trans. Pattern Analysis and Machine Inteligence, PAMI-5(2)*.
2. H. Hermansky. Perceptual linear predictive (PLP) analysis for the speech. In *Journal of the Acoustical Society of America, 1990. JASA-90*, pages 1738–1752, 1990.
3. 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.
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. Towards lower error rates in phoneme recognition. In *Proc. TSD 2004*, number ISBN 87-90834-09-7, pages 465–472, Brno, Czech Republic, September 2004.