Phoneme based acoustics keyword spotting in informal continuous speech *

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

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

This paper describes several ways of keywords spotting (KWS), based on Gaussian mixture (GM) hidden Markov modelling (HMM). Context-independent and dependent phoneme models are used in our system. The system was trained and evaluated on informal continuous speech. We used different complexities of KWS recognition networks and different types of phoneme models. The impact of these parameters on the accuracy and computational complexity is investigated.

1 Introduction

Acoustic keyword spotting (KWS) systems are widely used for 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.

The paper first discusses structure of standard acoustic KWS system and the metrics of evaluation. Section containing brief description of experiments and definition of test set follows. Results are discussed and conclusions are drawn at the end of the paper.

2 Acoustic Keyword Spotting

Modern acoustic keyword spotter was proposed in [5] 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 background model (phoneme loop) which models the same part of utterance as the keyword model.

We recognize (decode) the utterance using model ABC concatenated of models A, B and C, and using model ADC concatenated of models A, D and C. We have to do an assumption, that models B and D will recognize exactly the same part of utterance. So the final likelihood of model ABC (L_{ABC}) and ADC (L_{ADC}) should differ only because of models B and D. If the part of utterance beneath model B is not a keyword, the likelihood of the model B will be low (bad match to the keyword model), but the likelihood of the model D will be high (good match to the phoneme loop). In the other case, when the part of utterance beneath model B is the keyword, the likelihood will be high and so will the likelihood of model D (the likelihoods will be the same in ideal case). We can compute ratio of likelihoods ADC and ABC $L_R = L_{ADC}/L_{ABC}$. It is clear, that if there is

*This work was partially supported by EC project Multi-modal meeting manager (M4), No. IST-2001-34485, EC project Augmented Multi-party Interaction (AMI), No. 506811 and Grant Agency of Czech Republic under project No. 102/05/0278. Jan Cernocky was supported by post-doctoral grant of Grant Agency of Czech Republic No. GA102/02/D108.

a keyword beneath model B, the ratio L_R will approach 1 and be lower for non-keywords. The ratio L_R is a confidence of a detected keyword.

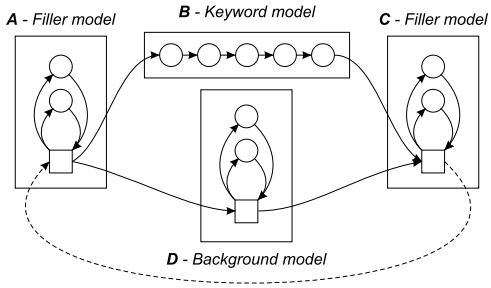

Figure 1: General model of keyword spotting network.

2.1 Performance Measure

L_R can be tresholded. If the L_R is above the threshold we raise alarm. Several cases can occur. Alarm was raised and there is keyword in utterance at the same time (a **HIT**). Alarm was raised and there is no keyword in utterance (a false alarm – **FA**). There is keyword in utterance but no alarm was raised (a **MISS**)

The level of threshold can be set. We receive more FAs and more HITs (less MISSes) by lowering the threshold and less FAs and less HITs (more MISSes) by increasing. We need to find a trade-off between HITs and FAs. The KWS is evaluated using *Figure-of-Merit* (**FOM**) proposed in [5], 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.

2.2 Our KWS System

Presented KWS system is based on that presented in [5], which needs full utterance to detect putative keywords. We did some simplifications for on-line KWS detection. The after-keyword filler model is not used. 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 pronunciation variants. The filler and background model are the same simple phoneme loops. After a token goes through a keyword model to the end node, corresponding token is taken from phoneme loop (node `F`). Then likelihood ratio of these two is computed.

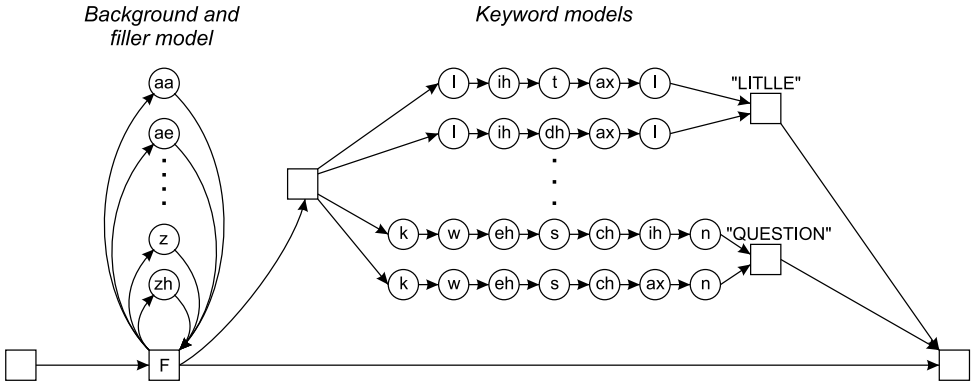

Figure 2: Keywords spotting network using context-independent phonemes.

3 Experiments

Our keyword system was tested on a large database of informal continuous speech of ICSI meetings [3] (sampled at 16 kHz). In the definition of experimental data-sets, 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 contain division is the 41.3, 18.7 and 17.2 hours of speech respectively.

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.`

As was said above, the KWS system uses phoneme units. Two different sets of units were trained. Context-independent phonemes (*phonemes*, 43 units) and context-dependent phonemes (*triphones*, 77659 units). Each of the sets was trained on the same 10 h long subset of the training set (denoted as **ICSI10h**). Standard training technique for Gaussian mixture hidden Markov model were used. Raw data was parameterized using 13 Mel-frequency cepstral coefficients with Δ and $\Delta\Delta$.

Context-dependent models [4] trained on conversational telephone speech (*CTS*) database were adapted for comparison. CTS database contains about 300 hours of speech, the features were 13 perceptual linear prediction (*PLP*) coefficients [2] with Δ and $\Delta\Delta$ parametrization. These non-adapted original models are denoted as **CTS300h-noad**. ICSI database was downsampled to 8 kHz and PLP parameterized. Then the CTS300h-noad models were adapted using full ICSI train set (adapted models are denoted as **CTS300h-adap**).

Experiments were done with the following KWS recognition networks. Context-independent phonemes network (denoted as **CI**, Figure 2). Reduced context-dependent phonemes network (denoted as **CD**). Reduced context-dependent and context-independent phon. network (denoted as **CI&CD**). Full context-dependent phonemes network (denoted as **CDfull**, Figure 3).

The first and the last phoneme of keyword is expanded to all context possibilities in context-dependent networks (Figure 3). Because of many triphones (up to 80 K), reduced network contains only triphones which appear in keyword models. This considerably reduces the network complexity and computational time. **CD** and **CI&CD** network have full connected phoneme loop containing reduced set of triphones or reduced set of triphones and phonemes respectively (Figure 2). Triphones of the **CDFull** network are connected depending on their contexts (Figure 3).

4 Results and Conclusion

The FOM performances and computational speed were measured for all types of presented recognition networks using ICSI10h models. The realtime factor was measured on Intel P4 2.5GHz HT CPU with 512MB RAM. The results are listed in Table 1.

The results clearly show, that triphones (**CD** and **CDfull**) models are more precise than context-independent models. On the other hand, the size of network with all triphones (**CDfull**) is huge and the decoding takes about 100 times more time than the **CI** network. Combination of phonemes and reduced set of triphones (**CI&CD**) could be a good compromise. Comparison of *ICSI* and *CTS* models is interesting: the **CTS300h-noad** models give only slightly worse results

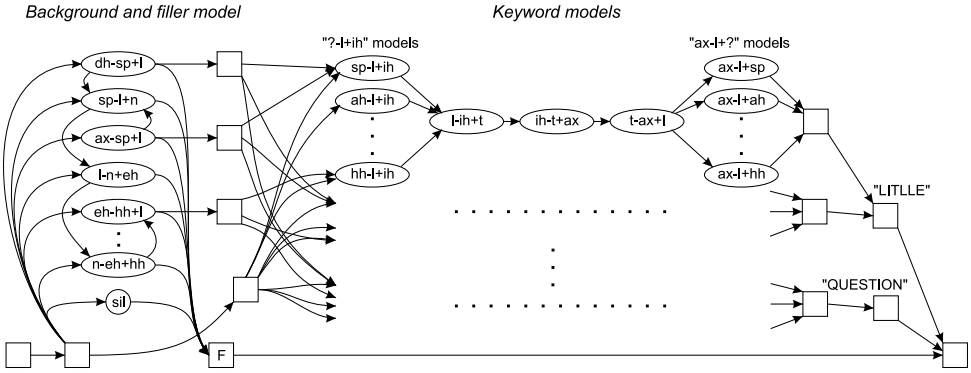

Figure 3: Keyword spotting network using full set of context-dependent phonemes.

Model	Network	#HITs	#FAs	#KWs	FOM	Realtime factor	Net size nodes links
ICSI10h	CI	3142	2877867	3289	**47.77**	0.51	264 339
ICSI10h	CD	3177	2774259	3289	**57.15**	1.07	4375 8461
ICSI10h	CI&CD	3164	2904486	3289	**57.52**	1.50	4417 8545
ICSI10h	CDfull	3173	2914897	3289	**61.88**	56.62	102k 3508k
CTS300h-noad	CD	3189	2752492	3289	**56.39**	–	7637 14742
CTS300h-adap	CD	3159	2927968	3289	**59.39**	–	7637 14742
CTS300h-adap	CDfull	3147	3032251	3289	**63.66**	–	119k 4256k

Table 1: The results of different acoustic keyword spotting systems.

than **ICSI10h** ones. We can see, that the influence of more training data is remarkable and almost overrides the channel mismatch between models and data (even if there is degradation due to downsampling to 8 kHz). After CTS models were adapted (**CTS300h-adap**), we obtained about +2% better performance compared to **ICSI10h**.

Our next work will be aimed at searching for keywords in phoneme and word lattices. This technique combined with presented KWS system should be used as the engine of a spoken document retrieval system for meeting and lecture indexation and searching.

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. 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] D. Povey and P. C. Woodland. Large-scale MMIE training for conversational telephone speech recongition. College Park, MD, 2000.

[5] 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.