An Experimental Evaluation of Keyword-Filler Hidden Markov Models

A. Jansen* and P. Niyogi†

April 13, 2009

Abstract

We present the results of a small study involving the use of keyword-filler hidden Markov models (HMM) for spotting keywords in continuous speech. The performance dependence on the amount of keyword training data and the choice of model parameters is documented. Also, we demonstrate a strong correlation between individual keyword spotting performance and median duration of that keyword. This dependence highlights the inadequacy of reporting system performance in terms of averages over arbitrary keyword sets, which is typically done for this task.

1 Introduction

This paper presents an exposition of several subtleties in the implementation and experimental evaluation of a basic keyword-filler hidden Markov model (HMM) to serve as a baseline for another study. We found that the vast majority of keyword spotting studies in the literature restrict themselves to reporting an average figure-of-merit performance over an arbitrary set of keywords and gloss over some important deviations from the standard theory that were necessary to achieve high detection rates. To the best of our knowledge, an examination at the present level of detail has not been presented for this technology, so we believe a record of our experiences are worth documenting in this technical report.

2 Keyword-Filler Hidden Markov Models

The current state-of-the-art keyword-filler hidden Markov model dates back nearly two decades to the seminal papers of Rohlicek et al. (1989) and Rose and Paul (1990). The basic idea is to create one hidden Markov model of the keyword and a separate hidden Markov of the filler (i.e., non-keyword) regions. These two models are joined to form a composite keyword-filler HMM that is used to perform a Viterbi decode of the speech

*Department of Computer Science, University of Chicago
†Departments of Computer Science and Statistics, University of Chicago.

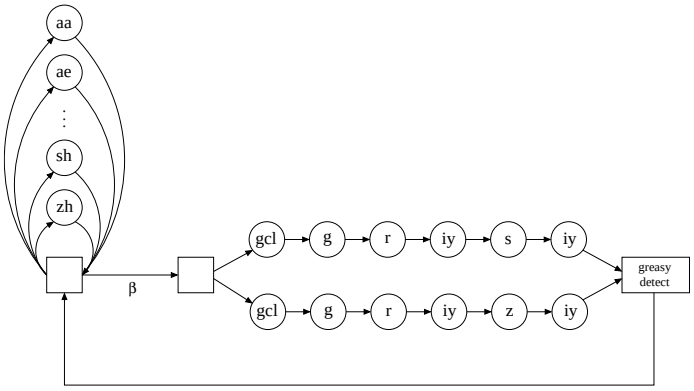

Figure 1: Keyword-filler hidden Markov model for the keyword "greasy" with two pronunciation paths. The parameter β specifies the filler to keyword transition probability.

(see Figure 1). Putative detections arise whenever the Viterbi path passes through the keyword portion of the model. Such detections can be scored according to the ratio between the likelihood of the Viterbi path that passed through the keyword model and the likelihood of an alternate path that passes solely through the filler portion.

Adjusting the operating point of such a keyword spotter can be accomplished two parameters commonly introduced into the model. The obvious lever is to simply adjust the threshold on the putative detection likelihood ratio scores. However, if the Viterbi decode is unable to enter the keyword model in the first place, threshold adjustment will not allow us to sample the entire receiver operating characteristic (ROC) curve. Thus, the second lever that may be adjusted is the value of the parameter β (see Figure 1), which we define as the transition probability from the filler model to the keyword model.[1] Artificially increasing this parameter can be used to force the Viterbi decode into the keyword model, increasing the number of putative detections. This transfers the responsibility limiting false alarms to the threshold on the likelihood ratio score.

The keyword model is commonly constructed from a simple left-to-right topology with one state per constituent phone of the keyword. Multiple pronunciations may be modeled separately as parallel paths in the keyword model. Examples of the keyword are used at the very least to estimate the state transition probabilities, thus modeling the phonetic state durations. It is also possible to adapt the observation probability Gaussian mixture models (GMM) using the keyword data as well. The most common filler model implemented is a phone loop comprised of a fully connected set of phonetic states.

We implement the system of Szöke et al. (2005), using a basic context independent acoustic model (monophone) with a single state per phone to construct both the

[1]The model specification of Rose and Paul (1990) introduces this nuisance parameter in the transition probability from the final keyword state to the filler model. This is functionally equivalent to the method used here.

Table 1: The list of 48 phones used in our experiments and the corresponding TIMIT labels included for each (reproduced from Lee and Hon (1989)).

Phone	Example	Incl	Phone	Example	Incl
iy	b*ea*t		en	butt*on*	
ih	b*i*t		ng	si*ng*	eng
eh	b*e*t		ch	*ch*urch	
ae	b*a*t		jh	*j*udge	
ix	ros*e*s		dh	*th*ey	
ax	th*e*		b	*b*ob	
ah	b*u*tt		d	*d*ad	
uw	b*oo*t	ux	dx	bu*tt*er	
uh	b*oo*k		g	*g*ag	
ao	ab*ou*t		p	*p*op	
aa	c*o*t		t	*t*ot	
ey	b*ai*t		k	*k*ick	
ay	b*i*te		z	*z*oo	
oy	b*oy*		zh	measure	
aw	b*ough*		v	*v*ery	
ow	b*oa*t		f	*f*ief	
l	*l*ed		th	*th*ief	
el	bott*le*		s	*s*is	
r	*r*ed		sh	*sh*oe	
y	*y*et		hh	*h*ay	hv
w	*w*et		cl (sil)	(unvoiced closure)	{p,t,k}cl
er	b*ir*d	axr	vcl (sil	(voiced closure)	{b,d,g}cl
m	*m*om	em	epi (sil)	(epenthetic closure)	epi
n	*n*on	nx	sil	(silence)	h#, pau

keyword and filler models. The construction details for this monophone model are provided in the following section.

3 TIMIT Monophone Acoustic Model Construction

To construct our monophone acoustic model, we consider the standard 48 phone set of Lee and Hon (1989). The definition of this set in terms of TIMIT labels is shown in Table 1. For the acoustic front end, we employed the rastamat package (Ellis, 2005) to compute a traditional 39-dimensional mel-frequency cepstral coefficient (MFCC) feature set for 25 ms windows sampled every 10 ms. This included 13 cepstral coefficients computed over the full frequency range (0-8 kHz), as well as 13 delta and 13 delta-delta (acceleration) coefficients. Cepstral mean subtraction was applied on the 13 original coefficients, and principal component diagonalization was subsequently performed for the resulting 39 dimensional vectors.

The TIMIT phonetic transcription allows us to create a set $\{(x_i, p_i\}_{i=1}^{L}$ of MFCC feature vectors-phone pairs, where each $x_i \in \mathbb{R}^{39}$ and each p_i is an element of Table 1. Our monophone model consists of one conditional probability distribution $P(x|p)$ for

each p, which can be estimated from the training examples $\{x_i|p_i = p\}$. In our implementation, we used the assume a C-component GMM of the form

$$P(x|p) = \sum_{c=1}^{C} \omega_{pc} \mathcal{N}(\vec{\mu}_{pc}, \mathbf{\Sigma}_{pc})(x), \quad (1)$$

where $\omega_{pc} > 0$ and $\sum_{c=1}^{C} \omega_{pc} = 1$ for each p; and $\mathcal{N}(\vec{\mu}, \mathbf{\Sigma})$ is a normal distribution with mean $\vec{\mu}$ and full covariance matrix $\mathbf{\Sigma}$. The maximum likelihood estimate of these GMM parameters are found using the expectation-maximization (EM) algorithm.

4 Evaluation Procedure

Both the TIMIT and BURadio corpora provide a time aligned word transcription. This transcription may be used to determine a set of intervals $I_w = \{[a_i, b_i]\}_{i=1}^{N_w}$ that contain the keyword w. While keyword spotting literature has relied on multiple performance metrics in the past, we employ a community standard figure of merit (FOM) score in our evaluation. Given a set $\mathcal{D}_w$ of detections of keyword w, the FOM score is defined as the mean detection rate when we allow $1, 2, \ldots, 10$ false positives per keyword hour. This metric is a means to summarize the high precision performance of the detectors; this performance may be graphically characterized by the initial region of operating curves measuring the relationship between detection rates vs. false alarms per keyword, per hour, as threshold is varied. In computing this figure of merit, we consider a keyword detection $t \in \mathcal{D}_w$ to be "correct" if there exists an interval $[a, b] \in I_w$ such that $t \in [a - \Delta, b + \Delta]$, where Δ is a short tolerance (20 ms for TIMIT and 100 ms for BURadio) that is set according to the precision of the word transcription.

5 Results

In this section, we consider the performance of the keyword-filler HMM model described above for the task of spotting instances of a given keyword in unconstrained speech. All experiments were conducted using the TIMIT (Garofolo et al., 1993) and Boston University radio news (Ostendorf et al., 1995) speech corpora. TIMIT data is used for training the acoustic monophone model and for training and testing in the toy keyword spotting experiments described below in Section 5.1. Boston University Radio News (BURadio) was used exclusively for a larger scale keyword spotting performance evaluation described below in 5.2.

5.1 "greasy" Experiments

In this section, we consider various forms of the keyword-filler HMM architecture on the task of spotting the keyword "greasy" in the TIMIT database. In particular, we test three models:

- Model 1: Fix keyword state GMMs according to monophone model of standard pronunciations ([gcl g r iy s/z iy]) and estimate β from data

- Model 2: Fix keyword state GMMs according to monophone model of standard pronunciations and tune β
- Model 3: Adapt keyword state GMMs using keyword examples and tune β

In each case, the background/filler portion of the model is trained on all sx/si sentences in the TIMIT corpus.

Table 5.1 displays the FOM dependence of each HMM method on the number of training examples of the keyword.[2] Here, we use a setting of $\beta = 0.05$ for Models 1 & 2. Several trends emerge from these results:

1. The absolute best performance is achieved if we adapt the keyword state GMM parameters using all 462 keyword examples. This is not surprising as the resulting acoustic model of each state will match the phonetic context perfectly, rather than relying on a set of general purpose monophone models. However, the obvious downside is that this approach requires a significant amount of keyword training instances to achieve good estimates. Indeed, even when provided as many as 100 training examples, performance takes a significant hit. If given 10 examples or less, the models break down completely.

2. Choosing a suitable β and fixing the emit probability distributions results is remarkably stable performance as we decrease the number of training examples. This is a consequence of the extremely small number of parameters (12 for the two six-state pronunciations of "greasy") involved in this model. However, the best performance of this falls significantlyshort of that for the other two models.

3. Estimating the individual phone to keyword transition probabilities from data (Model 1) outperforms the optimal β for all phones (Model 2) down to 50 training examples. This is largely a result of the non-uniformity of the estimated phone-keyword transition probabilities across the phone set, since "greasy" always occurs in the same context in TIMIT. In particular, the probability of transitioning into the keyword states are concentrated in the source phone [n], which becomes likely enough to detect a large number of the keyword instances. However, Model 1 model would fail if we attempted to detect "greasy" in a different context. Even in this setting of fixed context, when in the minimal supervision regime (5-10 examples) the estimated to-keyword transition probabilities becomes too small to produce an adequate number of candidate detections and performance drops significantly.

Figure 2 displays the FOM dependence on the setting of β, the background to keyword transition probability, when we fix the acoustic model and learn only the keyword and background transition probabilities from data. Here, we train each keyword model with all 462 "greasy" instances and train the background/filler model on all TIMIT sx/si training sentences.

[2]It is important to note that when we provide very few training examples, the performance of the word model depends on which particular training examples are used. Thus the figure-of-merit values displayed in Table 5.1 are averages over several random selections of training examples for each number of training speakers value.

Table 2: Figure-of-merit performance dependence on the number of keyword training examples.

N_{train}	Model 1 FOM (%)	Model 2 FOM (%)	Model 3 FOM (%)
462 (all)	94.6	91.1	97.4
200	93.7	91.1	95.2
100	92.1	91.2	79.6
50	91.2	90.6	32.2
25	89.9	90.3	4.3
10	81.1	90.3	0.0
5	67.9	88.5	0.0

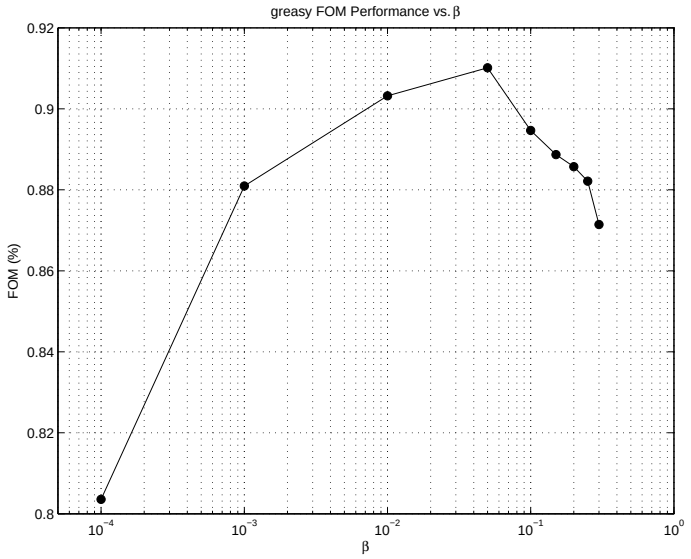

Figure 2: Figure-of-merit performance for the keyword “greasy” as a function of the filler to keyword transition probability.

Table 3: Keywords used in the BURadio experiments, along with the number of training/testing instances and the median duration (computed with training instances) for each keyword.

Keyword	# Train	# Test	Median T
about	116	87	250 ms
Boston	272	122	470 ms
by	337	250	180 ms
city	41	54	320 ms
committee	41	37	380 ms
congress	13	16	550 ms
government	43	55	440 ms
hundred	121	98	310 ms
Massachusetts	334	102	710 ms
official	7	89	410 ms
percent	80	47	450 ms
president	52	33	490 ms
program	44	99	510 ms
public	68	122	340 ms
seven	39	60	370 ms
state	273	312	300 ms
thousand	56	54	490 ms
time	82	88	320 ms
year	144	163	230 ms
yesterday	90	53	550 ms

5.2 Boston University Radio News Experiments

We now consider a larger scale keyword spotting testbed to investigate keyword-dependent performance effects. BURadio is clean, 16 kHz/16 bit read speech, making it a tolerable acoustic match for our monophone model. BURadio consists of 7 speakers (4 males and 3 females), each reading on the order of one hour of speech for a total of 7.3 hours. We partitioned the speakers into a training group, consisting of the two males and two females (f1a,f3a,m1b,m2b), and a testing group of the remaining speakers (f2b,m3b,m4b).

Table 3 lists the 20 keywords (18 content, 2 function) used in our experiments, along with the median duration and number of occurrences in each division of the data. These words were chosen to cover a wide range of word complexities in both duration and numbers of phones and syllables.

Model architecture selection for BURadio experimentation was not a trivial extension of the lessons learned in the previous section. First, since the keywords now occur in arbitrary context, using the estimated value of β from training data (Model 1) did not produce a sufficient number of candidate keyword detections to achieve competent spotting performance. Second, notice that many of the keywords in Table 3 have less than 50 keyword training examples. Indeed, this was not sufficient to train keyword models with adapted GMM parameters (Model 3). Thus, the only acceptable approach for BURadio evaluation was to fix GMM parameters a choose a suitable value for β

Table 4: Figure-of-merit performance for each keyword using the whole word Poisson process models.

Keyword	FOM (%)	Keyword	FOM (%)
program	98.1	public	60.6
Massachusetts	98.0	hundred	49.5
yesterday	89.0	seven	46.5
Boston	85.7	congress	45.0
thousand	85.7	city	38.0
official	79.3	about	30.2
government	72.6	state	23.3
president	71.8	time	21.8
committee	70.3	year	20.3
percent	65.9	by	9.8

Average FOM: 58.1%

(Model 2). However, the optimal value of β shifted from that found in the TIMIT "greasy" experiments; we determined an optimal value of $\beta = 0.3$ for the keyword "Boston" and applied this across the entire keyword set.

Each BURadio keyword model is trained on all instances of the target word in the training group. Each keyword detector is evaluated on at least one hour of test group speech containing all of the instances of both the keyword and words that contain that keyword.[3] Table 4 lists the figure of merit performance using whole word models, for each keyword.

Table 3 lists the resulting keyword-filler HMM figure-of-merit scores. The individual keyword FOM values show correlations of $r = 0.77$ with the number of constituent syllables, $r = 0.75$ with the number of constituent phones (including closure silences), and $r = 0.83$ with median keyword duration. Figure 3 plots figure-of-merit vs. median keyword duration. A linear relationship is evident, with the best fit line is shown in black.

6 Conclusions

We have evaluated the performance of a standard keyword-filler HMM on the task of spotting keywords in TIMIT and BURadio. We found that using estimated values of all keyword-filler HMM transition probabilities does not produce a reliable keyword spotter. Appropriate tuning of the filler to keyword transition probability, β, overcomes this problem, though β must be retuned for different corpora.

Adapting the keyword state GMMs using keyword examples can improve detection rates when given access to 100 or more keyword examples. However, the EM adaptation of these models fails for smaller numbers of examples. Finally, we find the keyword spotting performance is strongly correlated with the median duration of an

[3]Note that care is taken to manage the imperfect correspondence between embedded keyword strings and embedded keyword utterances. For example, "timely" and "bipartisan" are treated as containing a positive examples of the keywords "time" and "by", respectively; "sentiment" and "abysmal" are not

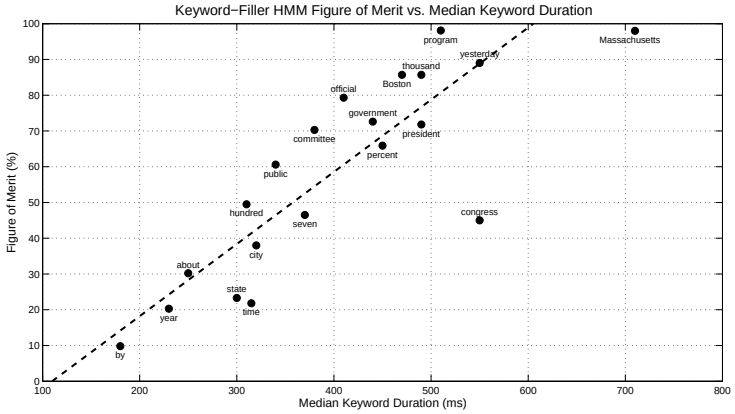

Figure 3: Figure-of-merit performance of the keyword-filler HMM plotted against median duration for each of the 20 BURadio keywords.

individual keyword. This indicates that the standard practice of reporting an average FOM score for a set of arbitrary keywords is meaningless unless this duration effect is in some way accounted for.

References

Daniel P. W. Ellis. PLP and RASTA (and MFCC, and inversion) in Matlab, 2005. URL `http://www.ee.columbia.edu/ dpwe/resources/matlab/rastamat/`. (online web resource).

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Linguistic Data Consortium, Philadelphia, PA, 1993.

Kai-Fu Lee and Hsiao-Wuen Hon. Speaker-independent phone recognition using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(11):1641–1648, 1989.

M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel. The Boston University Radio News Corpus. Technical Report ECS-95-001, Electrical, Computer and Systems Engineering Department, Boston University, 1995.

J. R. Rohlicek, W. Russell, S. Roukos, and H. Gish. Continuous hidden markov modeling for speaker-independent wordspotting. In *Proc. of ICASSP*, 1989.

R. C. Rose and D. B. Paul. A hidden markov model based keyword recognition system. In *Proc. of ICASSP*, 1990.

Igor Szöke, Petr Schwarz, Pavel Matějka, Lukáš Burget, Martin Karafiát, and Jan Černocký. Phoneme based acoustics keyword spotting in informal continuous speech. In *Lecture Notes in Computer Science - TSD 2005 (V. Matousek et al.)*, pages 302–309. Springer-Verlag, Berlin, 2005.