A NEW KEYWORD SPOTTING ALGORITHM WITH PRE-CALCULATED OPTIMAL THRESHOLDS

J. Junkawitsch, L. Neubauer, H. Höge, and G. Ruske

Technical University of Munich, Arcisstr. 21, D-80290 Munich, Germany
Siemens Corp., Otto-Hahn-Ring 6, D-81730 Munich, Germany

ABSTRACT

Keyword spotting is a very forward-looking and promising branch of speech recognition. This paper presents a HMM-based keyword spotting system, which works with a new algorithm.

The first discussion topic is the description of the search algorithm, that needs no representation of the non-keyword parts of the speech signal. For this purpose, the computation of the HMM scores and the Viterbi algorithm had to be modified. The keyword HMMs are not concatenated with other HMMs, so that there is no necessity for filler or garbage models. As a further advantage, this algorithm needs only low computational expense and storage requirement.

The second discussion topic is the determination of a optimal decision threshold for each keyword. In order two decide between the two possibilities “keyword was spoken” and “keyword was not spoken”, the scores of the keywords are compared with keyword specific decision thresholds. This paper introduces a method to fix decision thresholds in advance. Starting with measured phoneme distributions, the score distributions of whole keyword models can be calculated. Furthermore, these keyword distributions form the basis of the computation of decision thresholds.

Tests with spontaneous speech databases yielded 73.9% Figure-Of-Merit when using context-dependent HMMs. The detection rate at 10 fa/kw/h comes to 80%.

1. INTRODUCTION

Keyword spotting becomes a very important branch of speech recognition. Latest research experiences show that it is nearly impossible to design a recognizer that covers all words uttered during the practical application. Therefore the approach of only detecting keywords within any other words, sounds and noise without modeling these non-keyword parts of the utterance can be considered very promising. In contrary to other keyword spotting systems, which need some out-of-vocabulary representation, we renounce from explicit modeling the background.

The presented speaker-independent keyword word spotting system is able to indicate and classify predefined words within continuous speech. There are no restrictions concerning keyword occurrences: each utterance may contain any number of keywords. This keyword spotter, which is part of the big German speech understanding and translation project Verbmobil, is a phoneme based recognition system, so that all keywords are composed of phoneme Hidden Markov Models (HMM).

The paper presents a new algorithm which uses normalized scores during the Viterbi search. By a special treatment of the HMM scores, in this algorithm neither garbage nor filler nor silence models are necessary to represent the out-of-vocabulary parts of the utterance. Every keyword is allowed to start and to finish anywhere within the spoken sentence at any position. Since this method enables the recombination of different paths with different length within the Viterbi search, no additional computing load occurs. Backtracking is not necessary. The output of the algorithm yields normalized scores indicating the matching of the keywords at distinct time positions.

In order to decide between the two possibilities “keyword was spoken” and “keyword was not spoken” the normalized score of each word must be compared with a word specific threshold. A new technique is presented to calculate optimal values for these thresholds. Starting from the score distributions of the individual phonemes contained in a keyword, the total score distribution of the whole word is computed by length-weighted convolution of the single phoneme distributions. The score distributions are modeled by Gaussians, thus the threshold for each individual keyword can be calculated analytically.

This method enables the determination of the thresholds without the need of ever having spoken the keywords at all! This can be a decisive condition in practical applications where (like in our VERBMOBIL project) not enough training material exists for the individual keywords, whereas the phoneme models can be trained very well. Additionally, this new method allows a-priori estimation of the suitability and usefulness of each keyword in advance (without having spoken these words).

2. SYSTEM OVERVIEW

The keyword spotting system consists of three fundamental modules. The first module is concerned with signal analysis and feature extraction. The speech signal is sampled with a frequency of 16 kHz, and 400 samples at a time are combined to frames with a frame period of 10 ms. A total number of 30 mel-filtered cepstral coefficients and their time derivatives of first and second order are calculated, giving a feature vector with 64 components. This vector is optimized by the use of a linear discriminant analysis. The second module computes the emission probabilities for all states of all
HMMs. The third module contains the recognizer itself. Fig. 1 shows the basic structure of this module.

![Figure 1: basic structure of the recognition module](image)

The basic unit of the recognizer is a keyword model with a subsequent decision instance. For all keywords such basic units are set up, but no keyword can influence another one because all basic units are completely independent. Each keyword model is realized as a concatenation of phoneme HMMs, but no filler, silence or garbage models are attached at the beginning or the end. The first state of the keyword model has no predecessor state and the last state has no successor state. The search is done applying a modified Viterbi algorithm using specific normalized scores. The decision instance observes the score of the last state of a keyword model and provides classification by comparing local score minima with a decision threshold.

### 3. MODIFIED VITERBI ALGORITHM

For our purposes the standard Viterbi algorithm and the computation of scores have to be modified. The algorithm described below is based on local scores, which can be interpreted as a distance measure from a certain state to the best state with the highest emission probability. The term $LSc_{t,s_j}$ means the local score of state $s_j$. This local score is defined as the difference between the negative logarithm of the emission probability $p(x_t|s_j)$ of state $s_j$ and the value of the best state at that time. So the best state always has $LSc_{t,s_j} = 0$, while all other states have scores greater than zero.

$$LSc_{t,s_j} = -\log p(x_t|s_j) - \min_{s_j}(-\log p(x_t|s_j))$$

Since the first state of a keyword HMM is not concatenated with any predecessor, new search paths are allowed to start at any point in time. These paths don’t take over any previous scores and begin with zero. Nevertheless, paths with different starting points are allowed to recombine in a proper way. For this purpose it is necessary to compare paths with different length within the search. This is achieved by using length normalized scores. The total amount of accumulated scores and penalties is devided by the number of states the path has passed. So two variables must be handled for each state $s_j$ and each time instant $t$: the normalized score $NSc_{t,s_j}$ and the length of the path $L_{t,s_j}$.

When $A_{ij}$ is the penalty from state $s_i$ to state $s_j$, the search algorithm can be notated in the following recursive form:

$$NSc_{t,s_j} = \min \left( NSc_{(t-1),s_j} \left( 1 - \frac{1}{1 + L_{(t-1),s_j}} \right) + \frac{1}{1 + L_{(t-1),s_j}} (A_{ij} + LSc_{t,s_j}) \right)$$

$$L_{t,s_j} = 1 + L_{(t-1),s_j}$$

At any time instant $t$ a new path may start with

$$L_{t,s_j} = 1$$

and $NSc_{t,s_j} = LSc_{t,s_j}$

These equations can be interpreted as computing all possible paths leading to state $s_j$ and selecting the best one. To calculate the normalized score of such a path, first reduce the weight (i.e. its length) of the predecessor score and then add the weighted local score of state $s_j$. All weights depend on the length of the path. When the best predecessor state $s_k$ is determined, its incremented length becomes the length of state $s_j$.

This algorithm needs low computational expense and storage requirements. So it’s sufficient to allocate memory only for one column of the trellis. Tracing back thru the Viterbi matrix is not necessary. No HMMs for the non-keyword parts of the speech signals must be trained. Moreover this algorithm easily enables a time synchronous implementation of a keyword spotting system.

As a result of this calculation scheme, the normalized score of the last state of a keyword model, which serves as input for the decision instance, no longer is a monotonously increasing function in time. The score decreases when the HMM represents the speech signal well and increases when there is only bad matching. So the decision instance has to look for the minima of the score and evaluate them.

### 4. DECISION THRESHOLDS

The purpose of any decision instance is to determine which class the input belongs to. In the case of the keyword spotter the two possibilities “keyword was spoken” and “keyword was not spoken” have to be decided by observing the score of the last state of the keyword HMM. The decision rule is implemented by using a decision threshold. If the value of a local score minimum remains under this threshold, the occurrence of the keyword is indicated, else the keyword is rejected. In contrary to using the same decision threshold for each keyword, experience shows that using individual thresholds improves recognition performance substantially.

In the following a method is presented to fix optimal thresholds for particular keywords automatically. The score of the last state of a HMM is considered to be a random variable. Its conditional probability density functions (cpdf) depend on the corresponding class of the sample. If these cpdfs are known for the two cases “keyword was spoken” $o_l$ and “keyword was not spoken” $o_b$ decision thresholds can be calculated using distinct approaches.

#### 4.1. Computation of cpdfs of keywords

In a phoneme based HMM recognition system a word model is composed of its phoneme models. The normalized score $Y$ of a keyword HMM is interpreted as a random variable. Since each keyword HMM is composed of a sequence of phonemes, this
keyword score can be calculated using phoneme scores \( X \). These phoneme scores are also thought to be random variables. Because of various mean phoneme lengths, weighting coefficients \( \alpha_i \) must be used to take this effect into account.

\[
Y = \alpha_1 \cdot X_1 + \alpha_2 \cdot X_2 + \ldots, \quad \text{with} \quad \alpha_i = L_i / L_{\text{total}}
\]

\( L_i \) is the mean duration of phoneme \( i \) measured in its number of frames, while \( L_{\text{total}} \) is the sum of all mean phoneme durations, representing the mean duration of the keyword.

The \( \text{cpdf} \) of the keyword score \( Y \) can be calculated as the convolution of the scaled \( \text{cpdf} \)'s of the involved phonemes:

\[
f_Y(y) = \frac{1}{\alpha_1} f_{X_1}(\frac{1}{\alpha_1} y) \otimes \frac{1}{\alpha_2} f_{X_2}(\frac{1}{\alpha_2} y) \otimes \ldots
\]

In order to reduce computing expense and to get analytic solutions, all \( \text{cpdf} \)'s are approximated by Gaussians \( N(\mu, \sigma) \). In this way the keyword score \( Y \) is a Gaussian, too. Its mean value and standard deviation can be simply calculated using the means \( \mu_i \) and the standard deviations \( \sigma_i \) of the phonemes:

\[
Y = N((\alpha_1 \mu_1 + \alpha_2 \mu_2 + \ldots), \sqrt{\alpha_1^2 \sigma_1^2 + \alpha_2^2 \sigma_2^2 + \ldots})
\]

These few parameters \( \mu_i \) and \( \sigma_i \) must be measured once before calculating the \( \text{cpdf} \)'s of any keywords. Figure 2 shows an example of experimentally measured and Gaussian approximated \( \text{cpdf} \)'s of a phoneme for the conditions “phoneme was spoken” and “phoneme was not spoken”. A comparison of calculated and experimentally measured \( \text{cpdf} \)'s of keywords yielded quite good accordance of the means and standard deviations. It is remarkable that in general long keywords have small variances, whereas short keywords have great variances. Thus it is very promising to use word specific threshold strategies like the Bayes’ decision rule, the Neyman-Pearson rule, etc. But best results were achieved by prescribing a fix detection rate \( \alpha \). Since the \( \text{cpdf} \) is a Gaussian, the equation can be solved and \( Y_{th} \) calculated (\( \Phi^{-1} \) means the inverse Gaussian distribution):

\[
\alpha = \int \limits_0^{Y_{th}} p(Y | \omega_1) \, dY \Rightarrow Y_{th} = \mu + \sigma \cdot \Phi^{-1}(\alpha)
\]

5. RESULTS

The keyword spotting system was evaluated with speech data bases dealing with the scheduling of date appointments for meetings. These English dialogues are spontaneous speech, spoken from native male and female speakers, containing a lot of silence parts and noise.

Two kinds of HMM were used to test the system: context-independent and context-dependent HMMs. Both were trained on a approximatly 12 hours speech data base consisting of Multicom 94.1, Multicom 94.2, Verbmbol CD 6.0 and Verbmbol CD 8.0.

The tests were done with the Multicom 94.4 corpus. This data base contains 614 utterances of 2.13 hours duration. The list below shows the test vocabulary. When using these 25 keywords, in Multicom 94.4 there are a total number of 941 keyword occurrences.

Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday, morning, afternoon, tomorrow, evening, lunch, breakfast, meeting, office, schedule, appointment, vacation, calendar, seminar, conference, secretary, together, available, holiday

Figure 3 shows the receiver-operating-characteristics (ROC) for both context-independent and context-dependent HMMs using an equal decision threshold for each keyword. The figure of merit, a commonly used performance score, is defined as the average detection rate from 0 to 10 fa/kwh (false alarms per keyword per hour). This evaluation yielded a FOM of 58.5% for context-independent and 73.9% for context-dependent HMMs.

![Figure 2: cpdf's for the two classes “phoneme was spoken” and “phoneme was not spoken”. The measured densities are approximated by Gaussians.](image-url)
Evaluation of recognition performance using word specific thresholds was performed with context-independent HMMs. Table 1 shows some results, prescribing various detection rates $\alpha$. The experimentally yielded detection rates are better than previously prescribed and all measured operating points lie above the ROC for the context-independent case.

<table>
<thead>
<tr>
<th>prescribed $\alpha$</th>
<th>detection rate</th>
<th>$fa/kw/h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>30%</td>
<td>34.0%</td>
<td>0.319</td>
</tr>
<tr>
<td>35%</td>
<td>39.7%</td>
<td>0.750</td>
</tr>
<tr>
<td>40%</td>
<td>48.0%</td>
<td>1.41</td>
</tr>
<tr>
<td>45%</td>
<td>53.3%</td>
<td>2.53</td>
</tr>
</tbody>
</table>

Table 1: Recognition performance using pre-calculated keyword specific decision thresholds and context-independent HMMs.

### 6. DISCUSSION

A new keyword spotting algorithm is presented that allows keyword detection without representation of the non-keyword parts of a utterance by silence, garbage or filler models. Every point in time a new Viterbi path is allowed to start. In order to be able to compare paths of different length, specific normalized scores are used within the search. A new approach is introduced, which allows calculation of word specific decision thresholds in advance. Starting with score distributions of phonemes, the cpdf’s of keywords can be calculated and then applying different strategies decision thresholds can be fixed. Using only equal-length keywords, word specific decision thresholds don’t become very effective. The advantageous application of word specific decision thresholds is the greater the more the vocabulary contains both short keywords and very long keyword-phrases.

In contrary to utilizing Bayes’ decision rule for settling the boundaries, fixing a definite detection rate has proven to give best results for practical applications. Tests carried out with the Multicom 94.4 spontaneous continuous speech corpus yielded a FOM of 58.5% for context-independent and 73.9% for context-dependent HMMs. This notable result is attained with low computational and storage expense, since no out-of-vocabulary models are used. Moreover, this keyword spotting system doesn’t need any language model. However, it may be promising to apply a specific kind of post-processing in order to avoid keyword overlaps and to consider statistically achieved regularities of the language.

### 7. ACKNOWLEDGEMENT

This work was funded by the German Federal Ministry for Research and Technology (BMFT) in the framework of the Verbmobil Project under Grant 01 IV 102/C. The responsibility for the contents of this study lies with the authors.

### 8. REFERENCES


