PREDICTION OF KEYWORD SPOTTING PERFORMANCE BASED ON PHONEMIC CONTENTS

David A. van Leeuwen

TNO Human Factors Research Institute
P.O. Box 23
3769 ZG Soesterberg,
The Netherlands

Wessel Kraaij and Rudie Ekkelenkamp

TNO-TPD
Stieltjesweg 1
2600 AD Delft,
The Netherlands

ABSTRACT

In word spotting, one of the main difficulties is the false alarms, especially for small words. A model is presented for predicting the false alarm rate on the basis of the phonemic content of a word. This model is tested for a word spotter that has been used in the TREC Spoken Document Retrieval (SDR) track. Finally, results are presented for the retrieval task.

1. INTRODUCTION

Keyword spotting is a form of continuous speech recognition that allows the monitoring of connected discourse (*e.g.*, radio broadcasts or communication channels) for the occurrence of specific words. The technique can also be applied for searching audio archives for specific words or phrases. Specifically for Spoken Document Retrieval (SDR), keyword spotting can be used as an efficient basic technology, although indexing the words from a large vocabulary word recognition system has led to better results in TREC SDR tasks[1, 2].

The advantage of using word spotting in spoken document retrieval is that at *retrieval time* any word or short phrase can be used as search keys, if a phonetic transcription of the keywords can be found or generated. Retrieval systems, based on classic large vocabulary recognition systems, have the disadvantage that the query words can be Out Of Vocabulary (OOV) with respect to the recognizer's vocabulary which is defined at recognition/indexing time. This may not seem an important feature in the current TREC SDR tasks[1], but it can be important for other 'delayed spotting' tasks, such as the search for proper names in acoustic databases. One of the main problems in utilizing word spotting for SDR is the high false alarm rate found, specifically for small keywords [3]. In this paper we describe experiments that allow prediction of the effectiveness of keywords, based on their phonetic contents. This means that we are investigating the performance of a word spotting system as a function of the keywords. Usually, keyword spotting systems are evaluated for the set of keywords as a whole, but we will show that there is a strong performance dependence on the keyword.

2. A MODEL FOR WORD SPOTTING

Despite the fair amount of research in the area of word spotting, a solid theory for the operation of word spotting systems seems lacking. This may be due to the fact that there are many different technological approaches to the problem of word spotting. We will present here a simple model for the occurrence of false alarms at a given recognition rate.

The problem of word spotting can be stated such that words w_i, $i = 1, \ldots, n_{\mathrm{kw}}$ must be found and identified in a continuous speech audio channel. The fraction of words correctly found is called the hit rate, or accuracy a, and is usually presented as a percentage. The words that are reported as hits, but do not correspond to the actual spoken words, are called false alarms. The expected number of false alarms found increases linearly in the time that the audio channel is monitored, and therefore the correct measure is the *false alarm rate*, r_f. When more keywords are monitored, more false alarms are expected, and therefore r_f is usually expressed in terms of false alarms per keyword per unit time.

Many word spotting systems have an acoustic confidence measure that can be given for all the words spotted. After the word spotting run, a threshold level d can be chosen, above which candidate words are selected as spotted words. This threshold will determine $r_f(d)$ and $a(d)$. The full specification of the word spotting system is then given by the Receiver Operating Characteristic, ROC, which is a parametric plot of r_f versus a with threshold d as parameter. In order to specify a concise characteristic of the ROC, the Figure Of Merit (FOM) has been introduced, as a mean of $a(d)$ for standard values of $r_f(d)$ [4]. The trade-off between hits and false alarms, as summarized in the ROC, means that the performance of either of the two quantities must be specified with the value of the

other. In this paper the hit rate is considered constant, *i.e.*, the word spotting parameters are tuned in a way that a given a is reached.

2.1. Basic framework

We will here try to model r_f for constant a, as a function of the keyword that is spotted. A strongly simplified model of speech is made here. The continuous discourse of the speech channel to be monitored is described as a random sequence of phones p_t

$$\cdots p_{t-1}\, p_t\, p_{t+1} \cdots$$

Here, t is a time index. The speed at which the phones occur is on average r phones per second. The phones p are not uniformly distributed, e.g., the 'schwa' /ax/ has a higher probability than the /j/. Let us denote the probability that phone i occurs at any given time u_i, the *unigram* phone probability.

The word spotter is modeled as using acoustic models for phones $i = 1, \ldots, n_{\rm ph}$, the same base units of speech that we used in order to describe the continuous discourse. Because the models are not perfect, phones can be recognized incorrectly. The phone confusion matrix M_{ij} describes the probability that phone i is recognized as phone j. This includes the 'special' indices $i = 0$ for a phone insertion and $j = 0$ for a phone deletion. If a phone j is found in recognition, the probability e_j that a different phone has been uttered is

$$e_j = \sum_{i \neq j} u_i M_{ij}. \tag{1}$$

This immediately gives us an expression for the expected false alarm rate $\hat{r}_f$ for a monophone word consisting of the sole phone j:

$$\hat{r}_f = r e_j.$$

We can generalize this result for words consisting of more phones, by assuming independence of the probabilities. If a word w is modeled by the word spotter as a sequence of phones $f_1, f_2, \ldots, f_l$, then this simple model predicts the false alarm rate

$$\hat{r}_f(w) = r \prod_{j=1}^{l} e_{f_j}. \tag{2}$$

This expected false alarm rate does not include false alarms, that are a correct recognition of sub-words of other words or phrases. These 'linguistic ambiguities' are neglected as a minor fraction of the false alarms.

2.2. Linear approximation

The parameters $\{e_j\}$ of eq. 1 are dependent on the word spotter used, but of course also on the acoustic domain for which the spotter is tested, the speakers, etc. They can, in principle, be obtained from the phone confusion matrix M_{ij}, but it is not trivial to measure this matrix. Because word spotting intrinsically involves continuous speech, phone deletions and insertions pose the problem of *alignment* of reference and hypothesis phone strings. Consider phone string '1234' which is aligned to the hypothesis string '154': which of two successive reference phones '23' is assigned the deletion, and which the confusion with the found phone '5'?

Another approach for estimating $\{e_j\}$ can be made by measuring r_f for many words $\{w_i\}$, and fitting the over-specified set of equations similar to eq. 2 for all w_i. This can be accomplished by taking the logarithm of eq. 2, for all words w_i

$$\log r_f(w_i) = c + \sum_{j=1}^{l_i} \log e_{f_j} + E_i. \tag{3}$$

Here, the error term E_i is introduced as a parameter to be minimized in the fitting procedure, and $c = \log r$. Eq. 3 can be solved in least squares sense, *i.e.*, minimizing $\sum_i E_i^2$. This may be appreciated by rewriting the set of equations introducing the phone count N_j^i as the number of times phone j occurs in word w_i:

$$\log r_f(w_i) = c + \sum_{j=1}^{n_{\rm ph}} N_j^i L_j + E_i, \tag{4}$$

where we have written $L_j = \log e_j$. Eq. 4 is a matrix equation that is readily solved in least squares sense for $\{c, L_j\}$, a process known as multiple linear regression.

2.3. The number of phones in a word

If we make an even stronger simplification by assuming L_j independent of the phone j, eq. 4 can be further reduced to

$$\log r_f(w_i) = c + L N_{\rm ph}^i + E_i \tag{5}$$

where $N_{\rm ph}^i = \sum_j N_j^i$ is the number of phones in word w_i. The many assumptions made above suggest a linear dependence on $N_{\rm ph}^i$, but higher order terms are expected to give a non-linear dependence. The importance of eq. 5 is that the total number of phones in a word $N_{\rm ph}^i$ is the lowest order predictor of the false alarm rate.

3. DATABASE AND WORD SPOTTER

A number of experiments have been performed on the TREC SDR-7 test data [5]. The acoustic data, both for training and testing, consisted of American English speech, recorded from North-American Broadcast News shows [6]. For the experiments, we used the Abbot speech recognition system for acoustical classification [7, 8]. The acoustical models were kindly provided by Dr. Tony Robinson. For acoustical training,

100 hours of speech was available, of which approximately 70 hours were used (leaving out commercials and data having low acoustical confidence after Viterbi alignment). A finite state grammar decoder (`fsgd`) was used to build the word spotter. The grammar consisted of n_{ph} phones with unigram weights, parallel to n_{kw} keywords with uniform weighting. A global parameter α controlled the relative total weights of phones and keywords, thereby defining the operating point in the ROC.

The SDR-7 test database consisted of another 100 hours of similar data. For the false alarm prediction experiment, only part of this data has been used. One CDrom (approximately 2.5 hours) was used for estimating the parameters $\{c, L_j\}$, another CDrom for evaluating the prediction. The TREC SDR task has been re-run on the complete 100 hours of test data, effectively only re-scoring retrieval results that have been obtained in the TNO SDR-7 track [9].

3.1. Word spotter evaluation

In order to evaluate the word spotter, the spotted words had to be time-aligned to the reference transcription. This was made possible by forced alignment of the reference transcription. Words not occurring in the available dictionary, a 20 000 word subset of the CMU dictionary [10], were dealt with by defining a 'reversed word spotter' grammar. This is a grammar with reference words in order, and with recurrent parallel phone states at the positions of OOV words. This way, approximately 70 % of the transcriptions could be time-aligned. The other 30 % gave search errors, and were not used in the experiment.

Keywords were evaluated as 'correct' if their time of occurrence agreed with the time aligned reference transcription within a small margin. In all other cases, the word was considered a 'false alarm.' This is quite a strict definition, *e.g.*, consider the keyword 'president' in 'the president's wife.' By the automatically applied rules this is a false alarm, but for most applications this should be considered a hit. Detailed analysis of approximately one hour of test data, for keywords containing 10 phones or more, indicates that half of the false alarms would be considered relevant in a subjective evaluation.

4. EXPERIMENT

Several experiments have been carried out with the SDR-7 acoustic data and the word spotter. The influences of n_{kw}, α, stop word lists, and unigram word probabilities have been studied. For this paper, only the experiments using a single keyword per wordspot run are reported on. For all the words occurring in the reference transcriptions, a frequency sorted word list was made. For each of the words w, all pronunciations according to the CMU dictionary were found. Each of these pronunciations $w \rightarrow f_1 f_2 \cdots f_l$ was used in a one-keyword wordspot run in the acoustic database. Because of the many individual pronunciations (17661), this involves a considerable processing time. Only one CDrom (typically 2.5 hour of time aligned data) was used for test experiments. The word spotter is fast,[1] approximately $1/220$ real time.

The false alarm rate r is reported in false alarms per hour, since the number of keywords is always one in these experiments. The parameter α is arbitrarily chosen as 0.5, distributing the *a priori* weight of keyword and filler phone models evenly. By fixing α, the accuracy is more or less independent of the keyword.

4.1. Dependence on the number of phones

One would expect that the false alarm rate would depend strongly on the length of the keyword [3], or better, the number of phones in the keyword. In order to test eq. 5, we averaged r_f for $N^i_{\mathrm{ph}} = 1, 2, \ldots$. In figure 1 the dependence is shown. The logarithmic scale is suggested by eq. 5, but the dependence is not quite linear. One of the reasons might be that speech is *not* a sequence of random phones, as is assumed in the model, but has many linguistic correlations. Interestingly enough, the standard deviation follows the mean quite closely. This suggests that the distribution of r_f for fixed N^i_{ph} is exponential, *i.e.*, of the form $f(x) = \lambda e^{-\lambda x}$. [11] A variance stabilizing transformation for r_f is $r_f \rightarrow \log(r_f + r_0)$. It has been verified that such a transformation makes the variance more or less constant indeed.

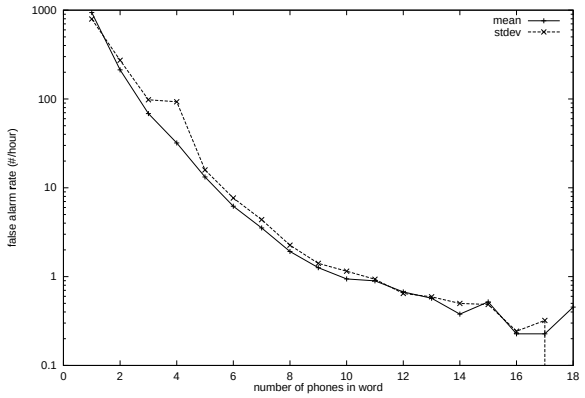

Figure 1: The dependence of the false alarm rate on the number of phones. The mean and standard deviation of all words w_i having the same number of phones N^i_{ph} is shown.

[1] Using an Alpha processor with GCC compiled code under the Linux operating system, we realize the temporary validity of the statement.

4.2. Multiple linear regression

The same data described in the previous section can be used to solve eqs. 4. We used a subset of the ICSI phone set, resulting in $n_{\mathrm{ph}} = 51$ used phones. In order to deal with measured values $r_f = 0$, we added the constant $r_0 = 0.1$ FA/hour before taking the logarithm. This might be interpreted as 'unseen false alarms' because only a finite time is measured. The mean log-weight $\{L_i\}$ is -0.833, with standard deviation 0.50. This mean corresponds roughly to the slope found in fig. 1. Using the fitted weights $\{L_i\}$, we tested the predictions to another part of the speech database (the 2*nd* CDrom). In fig. 2, the results of the predicted versus the measured false alarm rates are shown. The linear relationship is clearly visible, but the variance still is quite high.

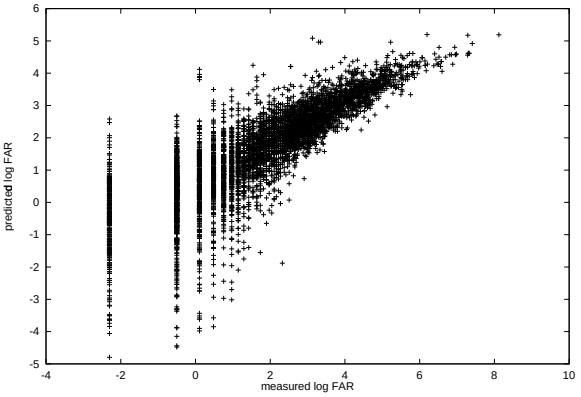

Figure 2: Predicted false alarm rate $\log(\hat{r}_f + r_0)$ versus measured $\log(r_f + r_0)$. The 'columns' at low r_f result from discrete number of false alarms (0, 1, 2, ...) in the measurement.

4.3. Database Retrieval

The false alarm prediction parameters $\{L_i, c\}$ can be used in a spoken database retrieval query, as a confidence measure for all of the keywords in the query. TNO competed in the TREC SDR-7 task under the name 'TwentyOne', for the first time in this track. A simple word spotting system was used then. One of the main problems at the time was the high number of false alarms, specifically for short words. [3] The false alarms have the effect of decreasing the precision of the retrieved documents. We re-scored our retrieval results based on the predicted false alarm rate $\hat{r}_f$. This was carried out by weighing the retrieved documents with $f(\hat{r}_f)$, using several weighing functions f. Unfortunately, it was not possible to increase the average precision of the retrieval results significantly. For some of the topics a much better retrieval result was obtained, but for other topics the retrieval disimproved, yielding a negligible net effect.

5. DISCUSSION

For a specific implementation of a word spotter (Abbot) we have seen that the false alarm rate can be predicted, based on the phonemic contents of the keyword. We believe that similar results would be obtained for other word spotting architectures, *e.g.*, Hidden Markov Model (HMM) based recognition systems. HMM based systems usually have an acoustic confidence measure of the spotted word, but as a general rule, the confidence will scale with the number of phones in a word.

Although the *mean* of predicted values behaves regularly, as may be appreciated from figures 1 and 2, the variability of the prediction is still very high. This means that the accuracy of the prediction is quite low for the individual key word. Only when large tests (*i.e.*, many keywords) are performed, the prediction becomes more meaningful. Reasons for this high variability can be:

speech The inherent variability of speech. Some of the words might be more clearly pronounced than others, for instance because of their higher risk of confusability.

dictionary The dictionary, providing the link between the spoken word and the spotted word, might vary in applicability for the words in the test set.

linguistics The model does not contain information about the order of phones in speech, which are given by words and grammar. Simple statistical techniques, such as the application of phone bigram probabilities, may lower the variability.

The values $\{L_j\}$ fitted using multiple regression could be interpreted as estimations for phone error probabilities $\{e_j\}$. Despite the high dimensionality of the fit (52 parameters!) the values for $\{e_j\}$ are reasonably ranged. The lowest e_j values are 0.12, 0.17, 0.18 for /em/, /aw/, /er/, and the highest values are 0.98, 0.99, 1.13 for /dh/, /b/, /p/, respectively. Only one value is impossibly high, so this is not a bad result given the high number of fitting parameters.

Unfortunately, we were not able to use the predicted false alarm rates to increase the retrieval precision of the TREC SDR-7 task. A reason for this might be a large variance in the predicted r_f, in combination with a low number of words in each of the queries. On average, 7 query words per query were used for a word spot run. The predicted $\hat{r}_f$ can only re-order the weights of this relatively low number of query words. Given the large variances, the probability for improving the retrieved document order is low.

5.1. Further research

It would be interesting to improve the model by including linguistic knowledge about the phones, such as

statistical bigram models, in order to reduce the variability in the prediction of the false alarm rate.

Another subject of further research is the correlation of the measured $\{L_i\}$ with the phone confusion matrix M_{ij}. This needs an algorithm for solving the phone alignment problem. We have already experimented with such an algorithm, and we expect to report on the results shortly.

For document retrieval purposes, query expansion might improve the statistics on which the re-ordering of retrieved documents is based.

6. REFERENCES

[1] S. Renals and D. Abberley, "The THISL spoken document retrieval system," in *TWLT 14* (D. Hiemstra, F. de Jong, and K. Netter, eds.), pp. 129–139, 1998.

[2] D. Abberley, S. Renals, and G. Cook, "Retrieval of broadcast news documents with the THISL system," in *ICASSP*, 1998. to appear.

[3] W. Kraaij, J. van Gent, R. Ekkelenkamp, and D. van Leeuwen, "Phoneme based spoken document retrieval," in *Language Technology in Multimedia Information Retrieval* (D. Hiemstra, F. de Jong, and K. Netter, eds.), vol. TWLT 14, pp. 141–152, 1998.

[4] NIST, "The road rally word-spotting corpora (RDRALLY1)." NIST Speech disc 6-1.1, September 1991.

[5] E. M. Voorhees and D. K. Harman, eds., *The seventh Text REtrieval Conference*, 1998.

[6] "`http://www.ldc.upenn.edu/ldc/about/broadcast97.html`." URL.

[7] A. J. Robinson, "The application of recurrent nets to phone probability estimation," *IEEE Trans. Neural Networks*, vol. 5, pp. 298–305, 1994.

[8] T. Robinson, M. Hochberg, and S. Renals, *Automatic Speech and Speaker recognition—Advanced Topics*, ch. The use of recurrent networks in continuous speech recognition, chapter 10, pp. 233–258. Kluwer Academic Publishers, 1996.

[9] R. Ekkelenkamp, W. Kraaij, and D. A. van Leeuwen, "TNO TREC7 site report: SDR and filtering," in *Proceedings Text Retrieval Conference*, no. 7, 1998.

[10] "`ftp://ftp.cs.cmu.edu/project/fgdata/dict/cmudict.0.4.Z`." URL.

[11] M. B. Priestley, *Spectral analysis and time series*, ch. 2, p. 65. Academic Press, 1981.