This is the submitted version of this conference paper. Published as:


© Copyright 2010 IEEE
OPTIMISING FIGURE OF MERIT FOR PHONETIC SPOKEN TERM DETECTION

Roy Wallace, Robbie Vogt, Brendan Baker, Sridha Sridharan

Speech and Audio Research Laboratory, Queensland University of Technology (QUT), Brisbane, Australia

{rg.wallace,r.vogt,bj.baker,s.sridharan}@qut.edu.au

ABSTRACT

This paper introduces a novel technique to directly optimise the Figure of Merit (FOM) for phonetic spoken term detection. The FOM is a popular measure of STD accuracy, making it an ideal candidate for use as an objective function. A simple linear model is introduced to transform the phone log-posterior probabilities output by a phone classifier to produce enhanced log-posterior features that are more suitable for the STD task. Direct optimisation of the FOM is then performed by training the parameters of this model using a non-linear gradient descent algorithm. Substantial FOM improvements of 11% relative are achieved on held-out evaluation data, demonstrating the generalisability of the approach.

Index Terms—spoken term detection, speech processing, speech recognition, information retrieval

1. INTRODUCTION

The emergence of vast collections of recorded speech is driving an urgent need for technologies to enable access to the information in these collections. Consequently, there has been recent interest in Spoken Term Detection (STD), the task of detecting occurrences of search terms rapidly and accurately in audio archives [1].

STD generally consists of the two distinct phases of indexing and searching. The role of indexing is to process the raw audio data into a form that is suitable for rapidly searching for query terms. Indexing is performed once in an off-line process, while many searches are then performed on this index.

The performance of an STD system is characterised by both search accuracy and search speed. Accuracy, particularly, relates to the usefulness of the results produced by a search. The Figure of Merit (FOM) is a well-established evaluation metric of STD accuracy [2] based on the expected rate of detected search term occurrences over the low false alarm rate operating region.

This paper introduces a novel technique for improving the accuracy of a phonetic-based STD system by directly maximising the FOM. For the system presented in this study, a phonetic posterior-feature matrix is generated during indexing and searched with a fast Viterbi decoding pass.

The Figure of Merit is directly optimised through its use as an objective function to train a transformation of the posterior-feature matrix. This direct optimisation of FOM is a form of discriminative training, as the FOM can be formulated in terms of the separation of scores attributed to true search term occurrences and false alarms.

Discriminative methods have been previously proposed in the context of STD, however, often these approaches do not seek to directly maximise the STD metric. In [3], for example, a Minimum Classification Error (MCE) criterion is used to improve the word error rate (WER) of the initial word transcript, with no assurance that optimising MCE will lead to optimal STD accuracy.

Maximum FOM training has been applied in other detection problems such as language [4] and topic [5] identification. Few studies have, however, aimed at the direct maximisation of FOM for STD [6, 7]. The technique used in [7] trains importance weights for a small number of feature functions without much justification of their selection or quantification of their contributions to overall performance. The task in [7] is also best described as utterance retrieval, that is, the detection of utterances containing the term, as opposed to the detection of individual term occurrences as in STD.

In [8], the concept of enhanced phone posteriors is introduced where a posterior-feature matrix is used for phone and word recognition, but trained to optimise per-frame phone classification. In the work presented in this paper, the phone posteriors are once again adjusted but with FOM as the objective function (rather than phone error rate) in both training and evaluation.

In Section 2, the Figure of Merit is defined, followed by details of the baseline STD system in section 3. Section 4 describes the technique used to optimise the Figure of Merit, followed by experimental results in Section 5 and conclusions in Section 6.

2. FIGURE OF MERIT

In general, STD accuracy is measured in terms of both detected term occurrences (hits) and false alarms. The Figure of Merit (FOM) metric measures the percentage of correctly detected term occurrences averaged over all operating points between 0 and 10 false alarms per term per hour [2]. This work uses the term-weighted FOM, obtained by averaging the detection rate at each operating point across a set of evaluation search terms, to avoid introducing a bias towards frequently occurring terms [1].

The FOM can be formally defined in terms of a set of STD results. Given a set of query terms, $q \in Q$, search is first performed on $T$ hours of data, resulting in a set of events, $e \in E$, where $e$ is either a hit or false alarm. Each event $e$ has the attributes $q_e$, $l_e$, $s_e$ where $q_e$ is the query term to which the event refers, the label $l_e = 1$ if the event is a hit or 0 for a false alarm and $s_e$ is the score of the event.

Also, for each term, $q \in Q$, define the set of hits as $E_q^* = \{e \in E : l_e = 1 \land q_e = q\}$ and the set of false alarms as $E_q^- = \{e \in E : l_e = 0 \land q_e = q\}$. The FOM can then be defined as

$$FOM = \frac{1}{A} \sum_{e_k \in E^+} h_{e_k} \max\left(0, A - \sum_{e_j \in E^-} (1 - H(e_k, e_j))\right)$$

where $A = 10T|Q|$, $h_{e_k} = \frac{1}{|Q||E|}$ and

$$H(e_k, e_j) = \begin{cases} 1 & s_{e_k} > s_{e_j} \\ 0 & \text{otherwise.} \end{cases}$$

(1)
In this formulation, each hit, $e_k$, contributes a value of between 0 and $h_{e_k}$, depending on the number of false alarms which outscore it. The value is 0 when the event is outscored by $A$ false alarms or more. Such events have no effect on FOM, so it is possible to re-define the FOM in terms of truncated results sets, $R^- \subseteq E^-$, containing the top scoring false alarms with $|R^-| \approx A$, and $R^+ \subseteq E^+$, containing the top scoring hits which outscore all $e \in (E^- - R^-)$.

By summing over these subsets, (1) can thus be re-written as

$$
FOM = \frac{1}{A} \sum_{e_k \in R^+} h_{e_k} \sum_{e_j \in R^-} H(e_k, e_j).
$$

Equation (2) can be interpreted as the weighted proportion of correctly ranked pairs of hits and false alarms, similar to the definition of the AUC (area under the ROC curve) in [7].

3. PHONE POSTERIOR-FEATURE MATRIX STD SYSTEM

The indexing and search approach adopted for the STD system is based on that described in [9]. The indexing phase produces a posterior-feature matrix, as described below in Section 3.1. Search is then performed in this matrix, as described in Section 3.2.

The use of separate indexing and searching phases allows search to be performed at 50 times faster than real-time, and the system is also suitable for use as a post-processor for candidates selected by other (faster) STD techniques, if necessary [9].

3.1. Indexing

Indexing involves the generation of a posterior-feature matrix, as follows. As in [9], the LCRC FeatureNet phoneme classifier [10] is first used to produce phone posterior probabilities for each frame of audio. These output posterior probabilities are then logarithmised.

The resulting log-posteriors form the contents of the posterior-feature matrix, $X = [x_1, x_2, \ldots, x_U]$, with $x_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,N}]^T$, and $x_{i,p}$ referring to the log-posterior of phone $p$ at frame $t$, in an utterance of $U$ frames. These log-posteriors provide the emission probabilities for each state in the decoding network used in search (see Section 3.2).

In contrast to [9], phones are modelled with a single state only, to reduce index size and the number of parameters to be trained. To give an indication of the quality of the phone classifier, using the log-posteriors directly for open-loop phone recognition gives 46% accuracy on the evaluation set described in Section 5.1.

3.2. Search

Once the index has been constructed, the system can accept search terms in the form of a word or phrase. A pronunciation lexicon is used to convert the search term into a sequence of phones. Letter-to-sound rules may be used for out-of-vocabulary terms.

Search then consists of a modified Viterbi decoding on the posterior-feature matrix. As in [9], a network is constructed from context-independent phones, with two connected parts; term model and background model. The term model consists of the sequence(s)

1In this work, however, we aim to detect individual term occurrences, rather than classify whole utterances. The AUC characterises average performance over all utterance retrieval operating points, whereas the FOM refers to average performance over STD operating points between 0 and 10 false alarms per term per hour.

2The alternative transformations suggested in [9] did not improve STD accuracy in empirical trials.

of phones constituting the search term’s pronunciation(s). The background model is an open loop of all phones. Classical Viterbi decoding is then carried out, with the emission probability of a phone $p$ in the network at frame $t$ being directly provided by the corresponding log-posterior probability, $x_{t,p}$. A phone insertion penalty, tuned on development data, is used to counteract a tendency to favour a short phone duration.

At each frame, a score is computed for each search term. This is calculated as the log likelihood ratio between the score of the token exiting the term model and the score of the token in the background model. For an event beginning at frame $b$ and having a length of $n$ frames, this score is

$$
s = \sum_{t=b}^{b+n-1} (p_t^T x_t - g_t^T x_t) = \sum_{t=b}^{b+n-1} (p_i - g_i)^T x_t
$$

where $p_i$ is a mask vector representing the state occupancy at time $t$ of the hypothesised Viterbi alignment for the search term. That is, the elements of $p_i$ are

$${\text{if }} i \text{ is the index of the current phone } \text{ then } {p_i = 1, \text{ otherwise } p_i = 0}.$$  

Similarly, $g_i$ represents the hypothesised alignment for the background model at time $t$.

Whenever the score for a term at a particular time is greater than a threshold (and greater than potential overlapping candidates), the event is output as a putative term occurrence.

4. OPTIMISING THE FIGURE OF MERIT

The contribution of this paper is a novel method for direct optimisation of FOM for an STD system. This is achieved by introducing an extra layer of modelling that provides a mechanism for transformation of the posterior-feature matrix. In the remainder of this section, this mechanism is first described, followed by a description of the optimisation algorithm used to maximise FOM on the training set.

4.1. Enhanced posterior-feature linear model

A simple linear model is introduced to transform the phone log-posterior probabilities output by the phone classifier, $X$, to produce enhanced log-posterior features, $X'$, that are more suitable for the STD task. The linear transform is decomposed into a decorrelating transform, $V$, and an enhancement transform $W$, giving

$$x'_t = WVx_t. \quad (4)$$

The decorrelating transform, $V$, is obtained through principal component analysis (PCA) of the log-posterior features. Performing PCA additionally provides the opportunity for dimensionality reduction by retaining only the $M \leq N$ directions of highest variability. Dimensionality reduction reduces susceptibility to over-fitting of the proposed model by reducing the number of free parameters to train in $W$, and can also have the benefit of suppressing low-energy directions that may be dominated by noise.

The weighting matrix, $W$, is an $N \times M$ transform that produces a set of enhanced posterior-features from the decorrelated features. The goal of the novel training algorithm is to optimise the weights in $W$ that maximise FOM directly. While the original posterior-features, $X$, were optimised for phone classification, it is hypothesised that a discriminative algorithm optimising FOM directly will place additional emphasis on differentiating phones that provide the
most useful information in an STD task. The algorithm for optimising \( W \) is described in Section 4.2.

Using the enhanced posteriors directly in the searching, the score for an event is given by

\[
s = \sum_{t=b}^{b+n-1} (p_t - g_t)^T W V x_t = \sum_{t=b}^{b+n-1} (p_t - g_t)^T x_t. \tag{5}
\]

The use of (5) instead of (3) actually requires no change to the searching phase. Compared to the baseline approach, this is implemented simply as a transformation of the posterior-feature matrix index. Alternatively, rather than storing the \( N \)-dimensional features \( x_t' = W V x_t \) in the index, the \( M \)-dimensional decorrelated features \( V x_t \) could instead be stored, with final multiplication by the weighting matrix \( W \) performed at search time. In the case that \( M < N \), this approach provides for index compression, which is desirable in some applications.

### 4.2. Optimisation algorithm

This section details the method for training the weights, \( W \). The goal of the optimisation algorithm is to maximise FOM, however, the FOM is not a continuously differentiable function. Therefore, this section introduces a continuously differentiable objective function, \( f \), that is a close approximation to FOM.

The function \( f \) is defined by replacing the step function, \( H(e_k, e_j) \) in (2) with a sigmoid, \( \varsigma(e_k, e_j) \). That is,

\[
f = -\frac{1}{A} \sum_{e_k \in R^+} \sum_{e_j \in R^-} h_{ek} \varsigma(e_k, e_j) \tag{6}
\]

\[
\varsigma(e_k, e_j) = \frac{1}{1 + \exp(-\alpha(s_{ek} - s_{ej}))}.
\]

The parameter \( \alpha \) is a tunable constant controlling the slope of the sigmoid. A value of \( \alpha = 1 \) was found to be reasonable in preliminary experiments, and is used in this work.

The optimisation approach taken here is that of gradient descent. Similar to [11, 12], whereby the weights, \( W \), are found that correspond to a minimum of \( f \). Specifically, the algorithm used is the Polak-Ribière variant of nonlinear conjugate gradients (CG) with the Newton-Raphson method and backtracking line search [13].

Prior to gradient descent, the weights are initialised such that \( W = V^T \), that is, the inverse of the decorrelating PCA transform. In the case that \( M = N \), this initialisation ensures that (5) is equivalent to (3) before optimisation. In this way, a reasonable starting point is assured, and any adjustment of the weights away from this point during gradient descent constitutes an adaptation away from the baseline configuration.

Given (5) and (6), the derivative of \( f \) with respect to the weights to be trained, \( W \), is found to be

\[
\frac{\partial f}{\partial W} = -\frac{\alpha}{A} \sum_{e_k \in R^+} \sum_{e_j \in R^-} h_{ek} \varsigma(e_k, e_j) \left(1 - \varsigma(e_k, e_j)\right) d(e_k, e_j) \tag{7}
\]

\[
d(e_k, e_j) = \frac{\partial s_{ek}}{\partial W} - \frac{\partial s_{ej}}{\partial W} \tag{8}
\]

\[
\frac{\partial s}{\partial W} = \sum_{t=b}^{b+n-1} (p_t - g_t) (V x_t)^T \tag{9}
\]

Intuitively, (9) shows that the change in an event’s score is related to the difference in alignments between the term model, and the background model. Furthermore, (8) and (7) show that a change in \( W \) will increase \( f \) if such a change generally causes the scores of hits to increase relative to the scores of false alarms.

For computation of the Newton-Raphson step size, the Hessian is approximated with the diagonal of the Hessian, and is forced to be positive-definite by adding a multiple of the identity matrix when necessary [13]. The diagonal of the Hessian is found to be

\[
\frac{\partial^2 f}{\partial W^2} = \frac{\alpha^2}{A} \sum_{e_k \in R^+} \sum_{e_j \in R^-} \varsigma(e_k, e_j) \left(1 - \varsigma(e_k, e_j)\right) \left(1 - 2 \varsigma(e_k, e_j)\right) D(e_k, e_j) \tag{10}
\]

with the elements of \( D(e_k, e_j) \) given by

\[
D_{i,j}(e_k, e_j) = d_{i,j}^2(e_k, e_j).
\]

### 5. Experimental results

#### 5.1. Training and evaluation data

The data used for training and evaluation is conversational telephone speech selected from the Fisher corpus [14]. Selected conversations are annotated as having high signal and conversation quality, from American English speakers and not made via speaker-phone.

Training of the LCRC FeatureNet phoneme classifier uses 100 hours of speech with force-aligned 10ms frame phone labels, with each frame modelled with 15 log mel-filterbank channel outputs. The phone set consists of 43 phones, including a pause/silence phone.

The training set for the gradient descent algorithm consists of 10 hours of speech and 400 eight-phone search terms (referred to as \( \text{training terms} \)) with 1041 occurrences. The evaluation set consists of 8.7 hours of speech and 400 eight-phone search terms (\( \text{evaluation terms} \)). Search terms are selected randomly from those with at least one true occurrence in the reference transcript. Speakers and search terms do not overlap across training/evaluation sets.

#### 5.2. FOM optimisation results

This section presents the improvement in FOM achieved on evaluation data by using the enhanced posterior-features with the weights \( W \) trained according to the algorithm in Section 4.2. Figure 1 shows the FOM achieved on training and evaluation sets using the weights obtained after each of the first fifty CG iterations. Continued but small improvements were observed on further iterations.
It is evident that the algorithm is not overly tuned to the training search terms because searching for these terms in the evaluation audio does not provide a greater relative gain (+10% c.f. +11%). Rather, there is an indication of slight dependence on the training audio as, for both term lists, the relative gain achieved on the training audio is greater than that achieved on the evaluation audio (+13% c.f. +11% for evaluation terms). Overall, the technique appears to generalise well to search terms and audio not used during training.

Future work could involve improving the algorithm for training the parameters of the linear model. Also, incorporating the parameters of the neural network-based phone classifier in the optimisation process is expected to yield further improvements.

6. CONCLUSIONS

This paper proposed a novel technique for direct optimisation of the Figure of Merit for phonetic STD. A gradient descent algorithm was shown to be effective for this purpose.

By modelling vectors of log-posterior probabilities and training weights to exploit this information to optimise an objective function that is a close approximation to the FOM, the resulting system offers substantial improvements over the baseline that uses log-posterior probabilities directly.

7. REFERENCES


Table 1: FOM achieved before FOM optimisation (Initial FOM, with \( W = V^T \)) and after training (Max FOM), and relative improvement compared to initial FOM, for different values of \( M \), the number of dimensions retained after PCA.

<table>
<thead>
<tr>
<th>( M ) (( M = N ))</th>
<th>Training set (FOM)</th>
<th>Eval. set (FOM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>43</td>
<td>0.569 0.703 (+24%)</td>
<td>0.547 0.606 (+11%)</td>
</tr>
<tr>
<td>40</td>
<td>0.566 0.699 (+24%)</td>
<td>0.534 0.597 (+12%)</td>
</tr>
<tr>
<td>35</td>
<td>0.566 0.686 (+21%)</td>
<td>0.522 0.586 (+12%)</td>
</tr>
<tr>
<td>25</td>
<td>0.437 0.678 (+55%)</td>
<td>0.390 0.572 (+47%)</td>
</tr>
</tbody>
</table>

Table 2: FOM achieved before and after training, and relative improvement compared to baseline, evaluated when searching using various combinations of the audio and search term sets.

<table>
<thead>
<tr>
<th>Training terms (FOM)</th>
<th>Eval. terms (FOM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training audio</td>
<td>0.569 0.703 (+24%)</td>
</tr>
<tr>
<td>Eval. audio</td>
<td>0.597 0.654 (+10%)</td>
</tr>
</tbody>
</table>

Before FOM optimisation (iteration 0), search is effectively performed directly on the phone log-posterior probabilities as in [9]. This baseline approach results in an FOM of 0.547 on the evaluation set.

Table 1 shows that the FOM optimisation approach results in substantially improved FOM for both the training and evaluation sets in all situations. Figure 1 illustrates the optimisation algorithm for the first fifty iterations when all dimensions of PCA (\( M = N \)) are retained, corresponding to the first row of Table 1. In this case, the FOM on the evaluation set continues to improve along with the training set, up to 0.606 (+11%), without over-training. Performance on the training terms and data, not surprisingly, enjoys a higher relative gain (+24%).

As mentioned in Section 4.1, the number of feature dimensions retained after PCA, \( M \), is a tunable parameter that influences the dimensions of \( W \) (an \( N \times M \) matrix), and consequently the number of free parameters of the data. Table 1 shows that it is advantageous to retain all dimensions, that is, to use \( M = N = 43 \). For example, using \( M = 25 \) initially severely degrades FOM (from 0.547 to 0.390). However, it can be seen that FOM optimisation provides the highest relative gain on the evaluation data in this case (+47%). In fact, by using \( M = 25 \), an index compression factor of about 0.6 may be achieved with a relative decrease in FOM of less than 6% (from 0.606 to 0.572). Depending on the application, this may be a desirable compromise.

To further explore the generalisation characteristics, the dependence on terms and/or audio was investigated by evaluating the FOM achieved in two further situations — that is, searching for the evaluation terms in the training audio and secondly, searching for the training terms in the evaluation audio. Table 3 indicates the number of terms from each term list that occur in each block of audio. Note that the FOM optimisation algorithm uses only terms and audio from the training set. These further results, included in Figure 1 and Table 2, illustrate that FOM is substantially improved for all combinations.

### Table 3: Number of search terms occurring at least once and number of term occurrences in the training and evaluation sets.

<table>
<thead>
<tr>
<th>Terms</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training audio</td>
<td>400 1041</td>
</tr>
<tr>
<td>Eval. audio</td>
<td>60 512</td>
</tr>
</tbody>
</table>

It is evident that the algorithm is not overly tuned to the training search terms because searching for these terms in the evaluation audio does not provide a greater relative gain (+10% c.f. +11%). Rather, there is an indication of slight dependence on the training audio as, for both term lists, the relative gain achieved on the training audio is greater than that achieved on the evaluation audio (+13% c.f. +11% for evaluation terms). Overall, the technique appears to generalise well to search terms and audio not used during training.

Future work could involve improving the algorithm for training the parameters of the linear model. Also, incorporating the parameters of the neural network-based phone classifier in the optimisation process is expected to yield further improvements.

<table>
<thead>
<tr>
<th>( M )</th>
<th>Training set (FOM)</th>
<th>Eval. set (FOM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>43</td>
<td>0.569 0.703 (+24%)</td>
<td>0.547 0.606 (+11%)</td>
</tr>
<tr>
<td>40</td>
<td>0.566 0.699 (+24%)</td>
<td>0.534 0.597 (+12%)</td>
</tr>
<tr>
<td>35</td>
<td>0.566 0.686 (+21%)</td>
<td>0.522 0.586 (+12%)</td>
</tr>
<tr>
<td>25</td>
<td>0.437 0.678 (+55%)</td>
<td>0.390 0.572 (+47%)</td>
</tr>
</tbody>
</table>

Table 1: FOM achieved before FOM optimisation (Initial FOM, with \( W = V^T \)) and after training (Max FOM), and relative improvement compared to initial FOM, for different values of \( M \), the number of dimensions retained after PCA.

<table>
<thead>
<tr>
<th>Training terms (FOM)</th>
<th>Eval. terms (FOM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training audio</td>
<td>0.569 0.703 (+24%)</td>
</tr>
<tr>
<td>Eval. audio</td>
<td>0.597 0.654 (+10%)</td>
</tr>
</tbody>
</table>

Table 2: FOM achieved before and after training, and relative improvement compared to baseline, evaluated when searching using various combinations of the audio and search term sets.