Abstract

Confidence measure plays an important role in keyword spotting. To enhance the effectiveness of the confidence measure, we propose a novel method which improves the performance of keyword spotting by directly maximizing the area under the ROC curve (AUC). Firstly, we approximate the AUC as an objective function with the weighted mean confidence measure. Then, we optimize the objective function by training the weighting factors with the generalized probabilistic descent algorithm. Compared with the current method based on minimum classification error (MCE) criterion, the proposed method makes a global enhancement of ROC curve and does not need to train any threshold. The experiments conducted on the King-ASR-023 database show that the proposed method outperforms both the method averaging phone-level confidences and the method based on MCE.

Index Terms: keyword spotting, confidence measure, receiver operating characteristics (ROC) curve, area under curve (AUC)

1. Introduction

Keyword spotting (KWS) refers to the detection of occurrences of selected words or phrases in unconstrained speech. It plays an important role in the speech processing without formal grammar such as casual conversations.

It is well-known that the confidence measure is critical for keyword spotting. The optimal performance for the confidence measure can be viewed as the assignment of high confidence to correct hypothesized keywords and low confidence to incorrect hypothesized keywords in a consistent way. Several methods have been proposed in previous research, for example, online dynamically filler model [1], support vector machine (SVM) based method [2], and confusion garbage model [3].

Phone is often used as a basic unit in KWS, and the phone-level confidence is usually employed to compute the word-level confidence [4]. The phones with different distributions have different impact on human spoken word perception. Several methods have been proposed [4-6] to model this fact, and the results show that better performances can be achieved by weighting different phone-level confidence. To train the weighting factors, the minimum classification error (MCE) criterion has been adopted. The MCE loss function can reduce two types of errors (namely, the miss-hit errors and false alarm errors) to improve the performance of a KWS system. In the KWS task, MCE may not be an appropriate criterion for training the weighting factors. Firstly, the loss function for misclassification measure sums two types of errors as one metric, which cannot describe the tradeoff between the hit rate and false alarm rate in different conditions. Secondly, the MCE based method excessively depends on a confidence threshold which is used to determine whether a sample is correctly recognized or not. Each sample will be compared with the threshold to estimate the MCE loss in the training procedure. If the threshold is changed, the loss function based on MCE does not keep optimal any more. Thirdly, only one threshold is trained for MCE, and it can hardly satisfy all possible cases in the practical applications, since the method based on MCE trains the parameters for one specific operating point rather than for all operating points.

In binary classification, the receiver operating characteristics (ROC) curve is frequently used to illustrate the classification capability of a classifier. The ROC curve can depict the tradeoff between true positive and false positive rates for all actual operating points. To quantify the performance of a classifier with ROC curve, the area under the ROC curve (AUC) is employed. The AUC is also a general and reliable measure for the discrimination of a classifier [7]. A higher AUC indicates a better performance. Therefore, it has worked as a metric in the evaluation of KWS performance [8].

The Wilcoxon-Mann-Whitney (WMW) statistic [7] is adopted to approximate the AUC. Recently, this approximation has been introduced to the recognition and detection tasks for training parameters, such as language recognition [9] and voice activity detection [10]. And the representation of WMW statistic is similar to the definition of the Figure of Merit (FOM), a metric for the evaluation of spoken term detection [11].

Motivated by the research with the WMW statistic, we propose a strategy to improve the performance of confidence measure for KWS by maximizing the AUC directly. In this method, the AUC is represented as a continuous and differentiable objective function in the framework of the weighted mean confidence measure. The objective function is optimized through the generalized probabilistic descent (GPD) algorithm, and the weighting factors of phone-level confidence are trained during the optimization. Experiments are conducted to evaluate the proposed method, and the results show that the proposed method achieves a better performance over two current methods.

2. Confidence measures for KWS

The widely used method for KWS is based upon a two-stage strategy. In the first stage, speech recognition is performed to detect and segment hypothesized keywords. Hidden Markov Model (HMM) has been adopted prominently for the recognition task in KWS, and the keyword-filler network [6] is often created to model the keywords and non-keywords. After recognition, the hypothesized phone label is given for each time-frame in the duration of the hypothesized keyword, and the correlative acoustic likelihood is computed. In the second stage, the verification is carried out according to the confidence measure. The confidence measure is calculated from the information obtained in the previous stage. Then, the word-level confidence measure is compared with a pre-defined threshold to decide on accepting or rejecting the hypothesized keyword.

For computing the word-level confidence of a hypothesized
keyword, the phone-level confidence is acquired at first. Suppose that the hypothesized keyword \( w \) consists of \( N_w \) hypothesized phones, and \( ph_t^w \) is denoted as the \( t \)th hypothesized phone, the phone-level confidence \( CM(\text{ph}_t^w) \) is estimated from the frame-level logarithm posterior probabilities [4]:

\[
CM(\text{ph}_t^w) = \frac{1}{t_e(\text{ph}_t^w) - t_s(\text{ph}_t^w) + 1} \sum_{t = t_e(\text{ph}_t^w)}^t \log p(q_i | o_t)
\]

\[
= \frac{1}{t_e(\text{ph}_t^w) - t_s(\text{ph}_t^w) + 1} \sum_{t = t_e(\text{ph}_t^w)}^t \log \frac{p(o_t | q_i) p(q_i)}{p(o_t)}
\]

where, \( t_e(\text{ph}_t^w) \) and \( t_s(\text{ph}_t^w) \) represent the beginning and the end of the hypothesized phone \( ph_t^w \), respectively. At time \( t \), \( o_t \) is the acoustic observation, and \( q_i \) is the aligned HMM state in the phone model. \( p(o_t | q_i) \) represents the acoustic likelihood as the conditional probability of observation \( o_t \) on the condition of the state \( q_i \). \( p(q_i) \) is the state prior probability of \( q_i \). \( p(o_t) \) is estimated by accumulating the acoustic likelihoods of all possible active states when processing \( o_t \).

The confidence measure of the whole hypothesized keyword can be obtained by averaging phone-level confidences as:

\[
CM_{\text{avg}}(w) = \frac{1}{N_w} \sum_{i=1}^{N_w} CM(\text{ph}_i^w)
\]

It is assumed that phone-level confidence should not contribute equally to the final word-level confidence [4], and different linear transformation for different phone-level confidence can have a positive effect on the word-level confidence. Compared with Eq. (2), the confidence score can be modified as the weighted mean confidence measure:

\[
CM_{\text{wm}}(w) = \frac{1}{N_w} \sum_{i=1}^{N_w} (\alpha o_i \cdot CM(\text{ph}_i^w) + \beta \text{ph}_i^w)
\]

The weighting factors comprise the weight \( \alpha o_i \cdot \) and bias \( \beta \text{ph}_i^w \), they are employed to adjust the impact of hypothesized phone \( ph_i^w \) on the confidence of hypothesized keyword \( w \) which contains \( ph_i^w \). Both \( \alpha o_i \cdot \) and \( \beta \text{ph}_i^w \) can be trained by MCE criterion [4, 6]. In next section, a novel method is proposed for training these parameters.

3. Parameter training using AUC optimization

3.1. Approximation of AUC

In the binary classification problem, the ROC curve depicts the performance of a classifier by plotting the true positive rate (rate of positive samples being labeled as positive) against the false positive rate (rate of negative samples being labeled as positive) [7]. If the ROC curve rises toward the upper left corner, the AUC value would be higher, and performance is better.

The AUC value can be denoted by the normalized Wilcoxon-Mann-Whitney (WMW) statistic [7]. Let \( x_1, \ldots, x_{N^+} \) be the outputs of \( N^+ \) positive examples and \( y_1, \ldots, y_{N^-} \) be those of \( N^- \) negative examples. The expression of AUC is illustrated by Eq. (4), which represents the expression of the probability in the discrete case.

\[
A = \frac{1}{N^+ \cdot N^-} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} I(x_i - y_j)
\]

where the function \( I(x) \) is defined as:

\[
I(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}
\]

AUC would have values between 0 and 1. An approximation for \( I(x) \) is the sigmoid function, which makes the zero-one function continuous and differentiable:

\[
S(x) = \frac{1}{1 + e^{-\gamma x}}
\]

where \( \gamma \) is a positive real number used to control the slope of the sigmoid. Then the expression of AUC is turned into:

\[
\hat{A} = \frac{1}{N^+ \cdot N^-} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} S(x_i - y_j)
\]

3.2. AUC expression for KWS

The dimensions of ROC curve for KWS are slightly different from those in the binary classification problem. For KWS task showed in Figure 1, the hit rate is the fraction of target keywords present in the test data correctly detected, and the false alarm rate is the number of false alarms (false positives) per keyword per hour of test data.

Note that the verification performance is affected by the first part process in two-stage KWS, and not all the target keywords in the test data can usually be spotted. As a result, the maximum of hit rate \( \theta_{\text{max}} \) is not perfect, in other words, the point with the highest hit rate on the ROC curve is not the point on the upper right corner. However, there is only a fix ratio between the hit rate and true positive rate, as well as a difference of unit measure between the false alarm rate and false positive rate. So WMW statistic is feasible to describe the performance in KWS.

To approximate the AUC with WMW statistic, the hypothesized keyword set \( H \) generated from the first stage of KWS can be separated into two groups \( H^+ \) and \( H^- \), according to the labeled transcription. \( H^+ \) is the set of the positive examples with correct hits, and \( H^- \) is the set of the negative samples with false alarms. The AUC can be calculated with all elements in \( H^+ \) and \( H^- \) by employing the weighted mean confidence measure as:

\[
\hat{A} = \frac{\theta_{\text{max}}}{|H^+| \cdot |H^-|} \sum_{u \in H^+} \sum_{v \in H^-} S CM_{\text{wm}}(u) - CM_{\text{wm}}(v)
\]

\[
= \frac{\theta_{\text{max}}}{|H^+| \cdot |H^-|} \sum_{u \in H^+} \sum_{v \in H^-} S D(u, v)
\]

![Figure 1: ROC curve of KWS.](image-url)
where \( \theta_{\text{max}} \) is a fix value as the maximum of hit rate to adapt the AUC value for KWS, and it can be estimated by training data. And \( D(u, v) \) is defined as:
\[
D(u, v) = C_{\text{wrm}}(u) - C_{\text{wrm}}(v)
\]
where, \( C_{\text{wrm}}(\cdot) \) is acquired by Eq. (3).

### 3.3. Training procedure

AUC is used as an objective function for training the weighting factors \( \alpha_{ph} \) and \( \beta_{ph} \) of the weighted mean confidence measure. Let \( C \) denote the set of all phones that can represent any keywords in the task. For each \( ph \in C \), it is observed that \( \alpha_{ph} \) should be a positive number. In order to satisfy this constraint, \( \alpha_{ph} = \exp(\tilde{\alpha}_{ph}) \) is defined, and \( \tilde{\alpha}_{ph} \) is the transformation parameter.

The maximization of \( \tilde{A} \) is carried out through iteratively using the GPD algorithm. \( \{ \tilde{\alpha}_{ph}(n)| ph \in C \} \) and \( \{ \tilde{\beta}_{ph}(n)| ph \in C \} \) denote the sets of parameters estimated at the \( n \)th iteration. They are updated for each \( ph \in C \) as:
\[
\tilde{\alpha}_{ph}(n + 1) = \tilde{\alpha}_{ph}(n) + \varepsilon_n \frac{\partial \tilde{A}}{\partial \tilde{\alpha}_{ph}}|_{\tilde{\alpha}_{ph}=\tilde{\alpha}_{ph}(n)}
\]
\[
\tilde{\beta}_{ph}(n + 1) = \tilde{\beta}_{ph}(n) + \varepsilon_n \frac{\partial \tilde{A}}{\partial \tilde{\beta}_{ph}}|_{\tilde{\alpha}_{ph}=\tilde{\beta}_{ph}(n)}
\]
where \( \varepsilon_n \) is a small positive learning rate as the iteration step. The gradients of \( \tilde{A} \) with respect to its parameters are computed using the back-propagation algorithm as:
\[
\frac{\partial \tilde{A}}{\partial \tilde{\alpha}_{ph}} = \frac{\theta_{\text{max}}}{|H+|-|H-|} \sum_{u \in H^+} \sum_{v \in H^-} \frac{\partial S(D(u, v))}{\partial \tilde{D}(u, v)} \frac{\partial \tilde{D}(u, v)}{\partial \tilde{\alpha}_{ph}}
\]
\[
\frac{\partial \tilde{A}}{\partial \tilde{\beta}_{ph}} = \frac{\theta_{\text{max}}}{|H+|-|H-|} \sum_{u \in H^+} \sum_{v \in H^-} \frac{\partial S(D(u, v))}{\partial \tilde{D}(u, v)} \frac{\partial \tilde{D}(u, v)}{\partial \tilde{\beta}_{ph}}
\]

where
\[
\frac{\partial S(D(u, v))}{\partial \tilde{D}(u, v)} = \gamma(1 - S(D(u, v)))S(D(u, v))
\]
and the partial derivatives of \( D(u, v) \) w.r.t. \( \tilde{\alpha}_{ph} \) and \( \tilde{\beta}_{ph} \) are given respectively by:
\[
\frac{\partial D(u, v)}{\partial \tilde{\alpha}_{ph}} = \frac{\partial (C_{\text{wrm}}(u))}{\partial \tilde{\alpha}_{ph}} - \frac{\partial (C_{\text{wrm}}(v))}{\partial \tilde{\beta}_{ph}}
\]
\[
\frac{\partial D(u, v)}{\partial \tilde{\beta}_{ph}} = \frac{\partial (C_{\text{wrm}}(u))}{\partial \tilde{\beta}_{ph}} - \frac{\partial (C_{\text{wrm}}(v))}{\partial \tilde{\beta}_{ph}}
\]

Finally, the last uncertain parts of the partial derivatives above are acquired as:
\[
\frac{\partial \tilde{\alpha}_{ph}^w}{\partial \tilde{\alpha}_{ph}} = \begin{cases} 1 & \text{if } ph^w = ph \\ 0 & \text{if } ph^w \neq ph \end{cases}
\]
\[
\frac{\partial \tilde{\beta}_{ph}^w}{\partial \tilde{\beta}_{ph}} = \begin{cases} 1 & \text{if } ph^w = ph \\ 0 & \text{if } ph^w \neq ph \end{cases}
\]

It is also showed that a hypothesized keyword \( w \) in the training process only contributes to the phones which constitute \( w \).

Here, the gradients of \( \tilde{A} \) w.r.t. \( \tilde{\alpha}_{ph} \) and \( \tilde{\beta}_{ph} \) are computed from Eq. (9) to Eq. (20) completely. It is seen that the proposed method does not employ any confidence threshold in the training procedure.

### 4. Experiments

Experiments are conducted to evaluate the proposed confidence measure based on AUC optimization. For comparison, another two current confidence measures are also carried out. The first one is averaging phone-level confidences described by Eq. (2) as a baseline. The second one is the weighted mean confidence measure with the weighting factors trained by MCE [6].

The experimental data are from a Mandarin telephone speech database, King-ASR-023 [12]. This database contains the sentences spoken by 650 native Chinese speakers. The voices are recorded via public telephone lines at the sample rate of 8 kHz. We use 66,492 sentences pronounced by 540 speakers (270 females, 270 males) to train acoustic model, and the total length of the training data is about 92 hours. The data set for training the weighting factors contains 9,978 sentences with the length of 14 hours, and this set is intended for both the MCE and AUC optimization based methods. The test set consists of one hour speech, including 710 sentences pronounced by another 34 speakers (17 females, 17 males).

In the front-end, the length and shift of analysis frame are 25ms and 10ms respectively. The used feature is 12th-ordered Mel-frequency cepstrum coefficients (MFCCs) and the normalized short-time energy, appending their first and second order derivatives (39-dimensional feature). The acoustic models are tied-state tri-phone continuous density HMMs. Each HMM has 3 left-to-right states, and the number of the Gaussian mixture components is 16 for each state. The filler models are adopted to filter out the non-keywords. All syllables in Mandarin are employed as the fillers. The keyword network and the filler network share the same tri-phone HMM acoustic models.

To prepare for the training weighting factors, we employ a revised training scheme [6] under keyword spotting scenario. The KWS system runs many times to provide adequate training samples. For each time, the keywords are dynamically selected from high-frequency words to cover as many tri-phones as possible. And the size of the keyword list is fixed and consistent with that of the test process. The system works under the same HMM configuration. For each hypothesized keyword in the results, the information e.g. keyword content, starting time, end time, tri-phones, and phone-level confidences are all recorded. Then the results are compared with the time labeled transcription obtained from the keyword segmentations of Viterbi alignment. The correct hypothesized keywords in the correct time segments are labeled as correct hits (positive samples), and the others are labeled as false alarms (negative samples). Total 10,126 correct hits and 58,343 false alarms of hypothesized keywords are detected. Evaluated from training data, the maximum of hit rate \( \theta_{\text{max}} \) is 91.6%. During the optimization of the
objective function, the parameter $\gamma$ is selected as 2.0, while the iteration step $\varepsilon_n$ is set as $\varepsilon_n = 0.5 \cdot (1 - n/1000)$ with maximal iteration of $n$ set to 1000 for MCE and 300 for AUC optimization respectively [10].

For the test, 50 words are selected as keywords, including 40 short words and 10 long words. Each short word consists of two syllables (6 phones), and each long word consists of three syllables (9 phones). These keywords appear 326 times in all test utterances.

Figure 2 presents the AUC values achieved on training and test sets using the proposed method during 300 GPD iterations. And the AUC value can also describe the performance of ROC curve for KWS task into the framework of the weighted mean confidence measure. The AUC used as the objective function is optimized in training the weighting factors of phones. The parameters are embedded into the approximation of AUC and trained by a GPD algorithm.

This strategy is attractive because it offers an approach to improve the performance of KWS directly with evaluation metric of interest. The proposed method optimizes the whole ROC curve for all operating points. Compared with the method based on MCE, the proposed method does not need to train any threshold. The experimental results show that the optimization of AUC with GPD algorithm is effective. And the proposed method for confidence measure achieves a better performance in KWS task over both the method averaging phone-level confidences and the method based on MCE.

5. Conclusions

In this paper, we have proposed a novel method for integrating the performance of ROC curve for KWS task into the framework of the weighted mean confidence measure. The AUC used

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.845</td>
</tr>
<tr>
<td>MCE</td>
<td>0.855</td>
</tr>
<tr>
<td>AUC Optimization</td>
<td>0.861</td>
</tr>
</tbody>
</table>

Table 1: AUC comparison for the three methods.

6. Acknowledgements

This work is supported by the grant from the National Basic Research Program of China (973 Program), No. 2007CB311100.

7. References


