Odyssey 2010
The Speaker and Language Recognition Workshop
28 June – 1 July 2010, Brno, Czech Republic

# Detection target dependent score calibration for language recognition

*Raymond W. M. Ng*[*], *Cheung-Chi Leung*[†], *Tan Lee*[*], *Bin Ma*[†] *and Haizhou Li*[†‡]

[*]Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong
[†]Institute for Infocomm Research, Singapore
[‡]Department of Computer Science and Statistics, University of Eastern Finland, Finland
[*]{wmng,tanlee}@ee.cuhk.edu.hk, [†]{ccleung,mabin,hli}@i2r.a-star.edu.sg

## Abstract

Based on the conventional score calibration techniques with gaussian backend and logistic regression of the relative likelihood scores, this paper proposes a method of score calibration specific to a subset of related languages. Detection scores to two related languages are considered as two sources with similar and complementary information. In the proposed score calibration, an optimal linear combination of these two sources is derived. Experiments to NIST LRE 2009 with the proposed method give an equal error rate of 3.33%, which is a 25.2% relative reduction compared with the results from globally calibrated scores. Errors in differentiating two related languages can also be reduced by some modifications in parameter optimization.

## 1. Introduction

Spoken language detection is the process of automatically detecting the presence of target language(s) in a speech segment. In NIST Language Recognition Evaluation (LRE), the task is single target detection repeated with many target languages. One has to determine whether a given speech segment is spoken in a hypothesized language.

In a language detection system, the scores from the detectors of different languages may have different score distributions. Score adjustment would be desirable. Although it may be ideal to adjust the detection scores of different target languages separately, in the general approaches a global score transformation is performed [1]. For instance, if every language pair is treated as a separate case, in a detection problem with $N$ languages there will be $\binom{N}{2}$ objectives to be fulfilled. It is difficult for this kind of score adjustment to bring an overall reduction of detection errors.

A global score adjustment and calibration method often assumes the least prior knowledge. Gaussian backend scores and likelihood ratios are commonly adopted measures by LRE systems [2][3]. In those measures, target class scores are mixed with the non-target class scores to give a relative score indicating target likelihood. Score calibration has been studied in the problem of speaker recognition, and recently also in language recognition [1][4][5]. An effective information-preserving score calibration is proposed which uses a *scaling* factor and multiple *translational* factors to maximize the total posterior probability of the scores [4].

In NIST LRE 2009, detection targets include some pairs of related languages [6]. Public results reveal higher recognition errors in these related pairs [7]. Detection to these related languages becomes a bottleneck in a state-of-the-art language recognition system. Intuitively, if some error reduction techniques specific to these related languages are introduced, there is hope to reduce the global error. In this paper, we start with a set of language recognition results which is well calibrated in the global level. Further calibration specific to those related language pairs will be proposed. Our goal is to reduce the global language recognition error, as well as the confusion among these related languages.

## 2. Language detection

### 2.1. Task specification

Suppose there are $K$ speech segments to be tested with $N$ language hypotheses, a language recognition system maps every segment $k \in [1, 2, \ldots, K]$ to a score vector $\vec{s}$ of length $N$. Score calibration is performed [4], and a relative measure of detection log likelihood ratio is derived. The $n^{\text{th}}$ log likelihood ratio $\lambda^n_{\neg n}$ ($n \in [1, 2, \ldots, N]$) is calculated by dividing the $n^{\text{th}}$ score by the sum of all other scores in $\vec{s}$, then take the logarithm. For every speech segment $k$, there results a vector of log detection likelihood ratio:

$$\begin{bmatrix} \lambda^1_{\neg 1}(k) & \lambda^2_{\neg 2}(k) & \cdots & \lambda^n_{\neg n}(k) & \cdots & \lambda^N_{\neg N}(k) \end{bmatrix}^T \quad (1)$$

In a closed-set detection problem, the number of target classes (i.e. possible languages) is known to be $N$. A set of $N$ hypotheses is postulated about the language of segment $k$. To make a decision on accepting or rejecting speech segment $k$ as in language $n$, $\lambda^n_{\neg n}(k)$ is evaluated. When the ratio is higher, it is more likely that speech segment $k$ belongs to target language $n$. The $N$ hypothesis tests are repeated for each of the $K$ speech segments in the test set. $N \times K$ trials are processed in total. A single threshold $\theta$, independent of $k$ and $n$, is chosen for the decision-making process:

$$\lambda^n_{\neg n}(k) - \theta \geq 0 \mapsto \text{accept } k \text{ belongs to class } n \quad (2)$$

$$\lambda^n_{\neg n}(k) - \theta < 0 \mapsto \text{reject } k \text{ belongs to class } n \quad (3)$$

### 2.2. Evaluation metrics

Cost performance, $C_{\text{Avg}}$, is a common evaluation metric to a language recognition system with $N$ target classes. Adopting the application-dependent parameters in NIST LRE 2009 closed-set tests [6], $C_{\text{Miss}} = C_{\text{FA}} = 1$, $P_{\text{Target}} = 0.5$. $C_{\text{Avg}}$ is given by:

$$C_{\text{Avg}} = \frac{1}{N} \sum_{n_t=1}^{N} C_{\text{detect}}(n_t) \quad (4)$$

$$C_{\text{detect}}(n_t) = \frac{1}{2} P_{\text{Miss}}(n_t) + \sum_{n_n \neq n_t} \frac{1}{2} \frac{P_{\text{FA}}(n_t, n_n)}{N-1} \quad (5)$$

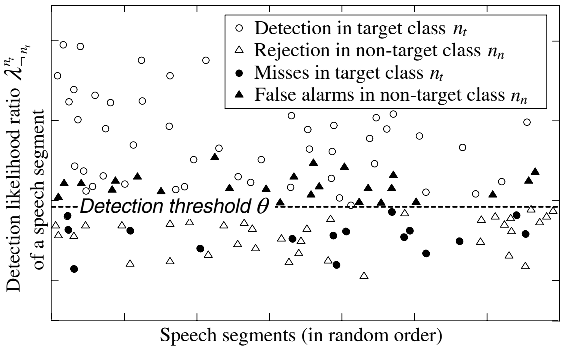

Figure 1: *Likelihood ratio $\lambda^{n_t}_{\neg n_t}$ for $n_t$ detection in a data set with two classes: $n_t$, $n_n$*

Eq.(4) and (5) say that in each single-class detection, there is one *miss* term and $N-1$ *false alarm* terms together to contribute to the average cost. Calculation of the two types of errors can be traced in the decision making function in Eq.(2) and (3). An acceptance of a speech segment in the non-target language constitutes a *false alarm* in Eq.(2). An rejection of a speech segment in the target language constitutes a *detection miss* in Eq.(3). After the $N \times K$ trials in a detection experiment, the probability of false alarm ($P_{\text{FA}}$) and miss ($P_{\text{Miss}}$) can thus be found:

$$\begin{aligned} P_{\text{FA}}(n_t, n_n) &= P(\lambda^{n_t}_{\neg n_t} - \theta \geq 0 | c = n_n) \\ &= \frac{P(\lambda^{n_t}_{\neg n_t} - \theta \geq 0, c = n_n)}{P(c = n_n)} = \frac{\| \mathcal{F}(n_t, n_n) \|}{\| \mathcal{I}(n_n) \|} \end{aligned} \tag{6}$$

$$\begin{aligned} P_{\text{Miss}}(n_t) &= P(\lambda^{n_t}_{\neg n_t} - \theta < 0 | c = n_t) \\ &= \frac{P(\lambda^{n_t}_{\neg n_t} - \theta < 0, c = n_t)}{P(c = n_t)} = \frac{\| \mathcal{M}(n_t) \|}{\| \mathcal{I}(n_t) \|} \end{aligned} \tag{7}$$

In the above equation, $c(k)$ is the true class (language) of the speech segment indexed $k$. $\mathcal{I}(n_t)$ contains the indices of speech segments whose true class is $n_t$. (i.e. $\mathcal{I}(n_t) : k \in [1, 2, \ldots, K] | c(k) = n_t$). $\mathcal{M}(n_t)$ is the subset of $\mathcal{I}(n_t)$ where the indexed speech segments are falsely rejected from class $n_t$. $\mathcal{F}(n_t, n_n)$ is the subset of $\mathcal{I}(n_n)$ where the indexed speech segments are falsely accepted as class $n_t$. $\| \cdot \|$ denotes set cardinality. Physically, $\| \mathcal{M}(n_t) \|$ and $\| \mathcal{F}(n_t, n_n) \|$ count the number of misses and false alarms in the experimental data set. An example of detection likelihood of a two-class data set with target class $n_t$ and non-target class $n_n$ is plotted in Figure 1, in which $\| \mathcal{M}(n_t) \|$ can be obtained by counting the number of filled circles, while $\| \mathcal{F}(n_t, n_n) \|$ is the number of filled triangles.

The dominance of *detection misses* or *false alarms* in a detection experiment is affected by the detection threshold $\theta$. By fixing different values of $\theta$, a performance curve can be plotted for each detector. The capability of a detector system with $N$ targets are summarized by the averaged performance (Eq.(4)). Two operating points are of our interest.

(1) $C_{\text{min}} = \min_{\theta} C_{\text{Avg}}$ is the minimum global average cost

(2) $C_{\text{eer}} = \operatorname{eer}_{\theta} C_{\text{Avg}}$ is the cost of *equal error rate* at the operating point where the weighted sum of $P_{\text{Miss}}$ of all languages has the smallest difference with the weighted sum of $P_{\text{FA}}$ of all language pairs.

# 3. Score calibration with related languages

In NIST LRE 2009, there are 23 target languages to detect. Among them five related language pairs shown below are generally considered to be mutually intelligible [6].

- Russian-Ukrainian
- Hindi-Urdu
- Farsi-Dari
- Bosnian-Croatian
- English(American)-English(Indian)

## 3.1. Cost minimization by likelihood ratio adjustment

Let $n_1$ and $n_2$ represent two related languages. They are mutually intelligible. Detection among $n_1$ and $n_2$ is believed to give many *misses* and *false alarms*. Here we make two hypotheses:

**Hypothesis 1:** Cost minimization specific to $n_1$ and $n_2$ would be beneficial to the reduction of the global cost performance $C_{\text{min}}$ and $C_{\text{eer}}$.

**Hypothesis 2:** The log likelihood ratios for $n_1$ and $n_2$ contain similar and complementary information.

According to Hypothesis 1, we propose to minimize the cost terms $C_{n_1,n_2}$ and $C_{n_2,n_1}$ where:

$$\begin{aligned} C_{n_1,n_2} &= P_{\text{Miss}}(n_1) + \frac{1}{N-1} P_{\text{FA}}(n_1, n_2) \\ C_{n_2,n_1} &= P_{\text{Miss}}(n_2) + \frac{1}{N-1} P_{\text{FA}}(n_2, n_1) \end{aligned} \tag{8}$$

Eq.(8) is a rewritten form of Eq.(5), retaining only the cost components related to classes $n_1$ and $n_2$. Note the cost for a single detection miss is $N-1$ times of the cost for a single false alarm. This ratio is inherited from the $C_{\text{detect}}$ definition in Eq.(5).

In the following, the minimization of $C_{n_1,n_2}$ is illustrated as an example. Let $n_t = n_1$ be the *target language* and $n_r = n_2$ is the *related language*. Referring to Eq.(6) and (7), we can choose to adjust the threshold $\theta$ and/or the likelihood ratio $\lambda^{n_t}_{\neg n_t}$ for a smaller $C_{n_t,n_r}$. Because this cost minimization is specific to $n_t$ and $n_r$ only, we fix the global parameter $\theta$ and adjust $\lambda^{n_t}_{\neg n_t}$.

Another issue is that target class specific cost minimization should be performed to the in-class data in $n_t$ or $n_r$ only, while this information is generally unavailable in the testing set. The workaround is to use a rough estimate of target class. Let $\tilde{\mathcal{I}}(n)$ be the estimated indices of speech segments in language $n$ (i.e. estimate of $\mathcal{I}(n)$). $\tilde{\mathcal{I}}(n)$ is derived heuristically. By evaluating the vector of detection likelihood ratios of speech segment $k$ (Eq.(1)), $k$ is put in $\tilde{\mathcal{I}}(n)$ if $\lambda^{n}_{\neg n}(k)$ is found to be among the largest three ratios.

In cost minimization, the goal is to have an adjusted $\lambda'^{n_t}_{\neg n_t}$ such that both sets $\mathcal{M}(n_t)$ and $\mathcal{F}(n_t, n_r)$ shrink. According to Hypothesis 2, $\lambda^{n_t}_{\neg n_t}$ and $\lambda^{n_r}_{\neg n_r}$ contain similar and complementary information. We propose the following adjustment:

$$\begin{aligned} &\lambda'^{n_t}_{\neg n_t}(k, \alpha_{n_t,n_r}) = \lambda^{n_t}_{\neg n_t}(k) + \tilde{\tau}_{n_t,n_r}(k, \alpha_{n_t,n_r}), \text{ where} \\ &\tilde{\tau}_{n_t,n_r}(k, \alpha_{n_t,n_r}) = \begin{cases} \alpha_{n_t,n_r} \lambda^{n_r}_{\neg n_r}(k) & \text{if } k \in \{\tilde{\mathcal{I}}(n_t) \cup \tilde{\mathcal{I}}(n_r)\} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{9}$$

Literally, Eq.(9) says that in the detection of language $n_t$, the log likelihood ratio for a subset of speech segments indexed $k \in \{\tilde{\mathcal{I}}(n_t) \cup \tilde{\mathcal{I}}(n_r)\}$ has to be adjusted as a linear combination of $\lambda^{n_t}_{\neg n_t}(k)$ and $\lambda^{n_r}_{\neg n_r}(k)$. $\lambda'^{n_t}_{\neg n_t}$ is the adjusted likelihood ratio. $\alpha_{n_t,n_r}$ is the weight for likelihood ratio combination.

After $\lambda^{n_1}_{\neg n_1}$ is adjusted for the minimization of cost $C_{n_1,n_2}$, minimization of $C_{n_2,n_1}$ can be done in the same manner. By substituting $n_t = n_2$, $n_r = n_1$, and repeating the operation in Eq.(9), $\lambda'^{n_2}_{\neg n_2}$ is found from $\lambda^{n_2}_{\neg n_2}$ and $\tilde{\tau}_{n_2,n_1}$.

### 3.2. Detection target dependent score calibration

The likelihood ratio adjustment for two detection targets $n_1$ and $n_2$ is illustrated in Fig 2. Score calibration for each detection target is effective to a subset of speech segments indexed by $k \in \{\tilde{\mathcal{I}}(n_t) \cup \tilde{\mathcal{I}}(n_r)\}$. The weight for likelihood ratio combination, $\alpha_{n_t,n_r}$, is unique in the adjustment in each detection target. The proposed score calibration is referred to as *detection target dependent score calibration*.

### 3.3. Optimal parameters for score calibration

For the likelihood ratio adjustment of each *target language* $n_t$ with its *related language* $n_r$, we use the development data set to find the optimal parameters such that the errors indicated by Eq.(6) and Eq.(7) are minimized. Instead of minimizing the sets $\|\mathcal{F}(n_t, n_r)\|$ and $\|\mathcal{M}(n_t)\|$, we propose to minimize the *total erroneous deviations* of likelihood ratios in the development data set. *Erroneous deviations* can be easily visualized in Figure 1. For a detection miss, it is the vertical distance from the detection threshold $\theta$ to the filled circle. For a false alarm, it is the vertical distance from the filled triangle to $\theta$. Mathematically, the minimization of *total erroneous deviations* is formulated as follows:

$$\begin{aligned}
&\min_{\upsilon,\alpha_{n_t,n_r}} \quad \sum_{k=1}^{K} \max\left(y_{n_t}(k) \times f(k, \alpha_{n_t,n_r}, \upsilon), 0\right) \\
&\text{subject to (s.t.)} \quad |\alpha_{n_t,n_r}| \leq 1, \\
&f(k, \alpha_{n_t,n_r}, \upsilon) = \lambda'^{n_t}_{\neg n_t}(k, \alpha_{n_t,n_r}) - (\theta + \upsilon), \\
&y_{n_t}(k) = \begin{cases} 1 & \text{if } c(k) \neq n_t \\ -(N-1) & \text{if } c(k) = n_t \end{cases}
\end{aligned} \tag{10}$$

$\lambda'^{n_t}_{\neg n_t}$ is the adjusted likelihood ratio defined in Eq.(9). $f(\cdot)$ is the *deviation* of $\lambda'^{n_t}_{\neg n_t}$ from a reference point $(\theta + \upsilon)$. This point is an $\upsilon$-shifted detection threshold. $y_{n_t} \times f(\cdot)$ returns positive values for erroneously detected segments and negative values for appropriately detected ones. The $\max(\cdot)$ operation removes *deviations* which are not erroneous. Among the positive-valued *deviations* there are two error types: misses and false alarms. $y_{n_t}$ scales the two error types with the default $N-1 : 1$ ratio. Every $\alpha_{n_t,n_r}$ is bounded such that $\lambda'^{n_t}_{\neg n_t}$ lies in a suitable range. The objective function is convex on $\alpha_{n_t,n_r}$. Thus, with a fixed $\upsilon$, a globally optimal solution of $\alpha_{n_t,n_r}$ can be found [8]. The objective function in Eq.(10) tries to push the likelihood ratios of detection misses and false alarms towards the reference point $(\theta + \upsilon)$.

The polarity of $\upsilon$ indicates the optimization goal towards fewer misses or fewer false alarms. Referring to Figure 1, a positive $\upsilon$ pushes the dashed line (reference point) upwards. Thus there will be more filled circles (missed targets) included in the optimization in Eq.(10), and the parameter for likelihood ratio adjustment, $\alpha_{n_t,n_r}$, will be optimized towards the goal of having fewer misses. Oppositely, a negative $\upsilon$ will lead to an optimal parameter which favours fewer false alarms.

## 4. Experiments

### 4.1. Original scores with frontend calibration

We use a phonotactic-prosodic fusion system to run the NIST LRE 2009 closed-set language detection task with 30-second test utterances. The phonotactic system adopts a parallel phone recognition followed by vector-space-model (PPRVSM) approach and is one of the subsystems in the Institute for Infocomm Research submission in NIST LRE 2009 [9]. The prosodic system uses a comprehensive set of prosodic features with vector-space models for language recognition [10]. Frontend calibration and fusion are not discussed in details in this paper. These processes are part of the language detector system in Figure 2. In brief, the phonotactic and prosodic scores are calibrated separately with the same Gaussian backend [2]. Following the maximum-a-posteriori (MAP) criterion, each system is calibrated before score fusion is done with FoCaL [4]. Frontend calibration and fusion parameters are trained using a separate development set. For notation simplicity, scores and error costs after the frontend calibration and fusion are referred to as *original* scores and *original* costs hereinafter.

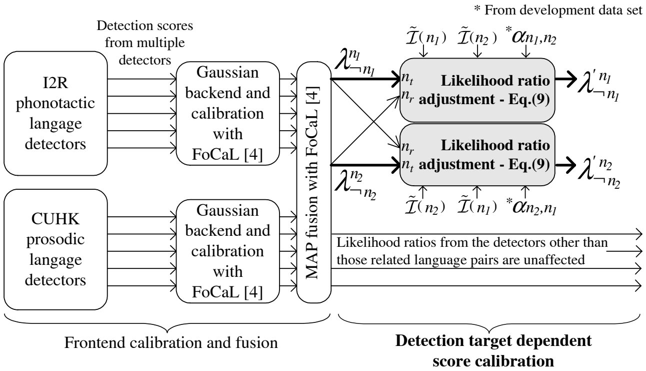
* From development data set
Detection scores from multiple detectors
I2R phonotactic langage detectors
CUHK prosodic langage detectors
Gaussian backend and calibration with FoCaL [4]
Gaussian backend and calibration with FoCaL [4]
MAP fusion with FoCaL [4]
$\lambda^{n_1}_{\neg n_1}$
$\lambda^{n_2}_{\neg n_2}$
$\tilde{\mathcal{I}}(n_1)$ $\tilde{\mathcal{I}}(n_2)$ $^*\alpha_{n_1,n_2}$
$n_t$ $n_r$ Likelihood ratio adjustment - Eq.(9)
$n_r$ $n_t$ Likelihood ratio adjustment - Eq.(9)
$\tilde{\mathcal{I}}(n_2)$ $\tilde{\mathcal{I}}(n_1)$ $^*\alpha_{n_2,n_1}$
$\lambda'^{n_1}_{\neg n_1}$
$\lambda'^{n_2}_{\neg n_2}$
Likelihood ratios from the detectors other than those related language pairs are unaffected
Frontend calibration and fusion
Detection target dependent score calibration

Figure 2: *System diagram of score calibration*

### 4.2. Detection target dependent calibrated scores

Detection target dependent calibration is carried out for each of the five related language pairs listed in Section 3. These language pairs are highlighted in the NIST LRE 2009 task specification [6] and we do not propose any methods in finding out these pairs.

Figure 2 shows the system diagram of the complete language detection system, in which the two shaded blocks are the modules of "detection target dependent calibration" for one pair of related target languages $n_1$ and $n_2$. A pair of adjusted likelihood ratios, denoted by $\lambda'^{n_1}_{\neg n_1}$ and $\lambda'^{n_2}_{\neg n_2}$, is derived. There are totally ten target languages (in five pairs) whose likelihood ratios are adjusted following Eq.(9). In each adjustment, the optimal $\alpha_{n_t,n_r}$ is found from a development data set, with the objective function in Eq.(10). A convex optimization tool, cvx, is used [11]. The optimal $\alpha_{n_t,n_r}$ parameters are then substituted in Eq.(9) with the NIST LRE 2009 evaluation set. $C_{\text{min}}$ and $C_{\text{eer}}$ with the evaluation data is reported.

The development set in this experiment includes telephone speech data from NIST LRE 2007 evaluation data, and telephone bandwidth broadcast radio speech from NIST LRE 2009 training data. The total number of speech segments in the development set is 6041. For evaluation, NIST LRE 2009 evaluation data is used. We select 10635 test segments from a pool of data with nominal durations of 3, 10 and 30 seconds. Those 10635 selected test segments have actual durations matching the 30-second nominal test durations.

To test the best reference point $(\theta + \upsilon)$ as mentioned in Section 3.3, the experimental procedures described above are repeated with different values of $\upsilon$. Recall that a positive $\upsilon$ favours fewer misses and a negative $\upsilon$ favours fewer false alarms.

# 5. Results

### 5.1. Reference point for erroneous deviation minimization

In training the score calibration parameters $\alpha_{n_t,n_r}$, a universal reference point $(\theta + \upsilon)$ is first assumed. A sequence of $\upsilon$ from $-6$ to 6, spaced 0.5 apart, is tested. $C_{\text{min}}$ and $C_{\text{eer}}$ with the evaluation data are plotted in Figure 3.

From Figure 3, a clear trend of increasing errors can be observed if the value of $\upsilon$ is too large or too small. Both $C_{\text{min}}$ and $C_{\text{eer}}$ attain the lowest values when $\upsilon$ equals 3.5. This positive value implies optimization of parameter $\alpha$ in Eq.(10) should prefer fewer detection misses. It is reasonable since the error cost for a detection miss is $N - 1$ times of the cost for a false alarm, as defined in Eq.(8).

The exact value of $\upsilon$ depends on the *erroneous deviations* of likelihood ratios, which are the vertical distances between detection threshold $\theta$ and filled circles/triangles in Figure 1 (also mentioned in Section 3.3). By inspecting the scores with the development data set, *erroneous deviations* of likelihood ratios are generally found to be smaller than 6. Therefore $\upsilon$ is tried in the range from $-6$ to 6. Unique optimal $\upsilon$ can also be trained in the optimization for different $n_t$, $n_r$ pairs. Nevertheless, a reasonable guess of a universal $\upsilon$ already leads to $C_{\text{min}}$ and $C_{\text{eer}}$ reduction, compared with the original error terms as shown in the horizontal lines in Figure 3.

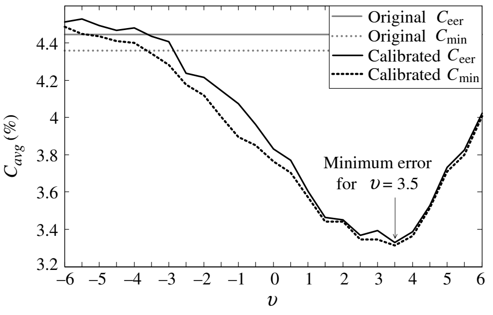

Figure 3: *Original and calibrated $C_{min}$, $C_{eer}$ with optimal $\alpha_{n_t,n_r}$ under different values of $\upsilon$*

### 5.2. Global detection errors

Figure 4 shows the detection errors for the evaluation data with original and calibrated scores. Original scores are those globally calibrated with FoCaL [4]. Calibrated scores are obtained via the proposed calibration method in this paper. $\upsilon$ is chosen to be 3.5. $C_{\text{min}}$ and $C_{\text{eer}}$ for the original scores over 23 target languages is 4.36% and 4.45% respectively. After the proposed calibration, $C_{\text{min}}$ and $C_{\text{eer}}$ over 23 target languages are reduced to 3.31% and 3.33% respectively. A relative EER reduction of 25.2% is achieved.

Table 1 shows the error statistics at the global $C_{\text{min}}$ and $C_{\text{eer}}$ operating points for the five related language pairs (listed in Section 3) versus the other 13 languages whose detection likelihood ratios are untouched in this experiment. As expected, the major contribution in error reduction comes from the five related language pairs, as these languages have high detection error rates initially and the score calibration proposed in this paper is specific to these languages. On the other hand, for the other 13 untouched languages, reduction in $C_{\text{min}}$ and $C_{\text{eer}}$ can also be observed.

There are opinions that the use of a common scale for multiple detections of different targets is not desirable [1][12]. If unique detection thresholds are determined for different targets, the equal error rate before and after "detection target specific calibration" is 3.18% and 2.78% respectively.

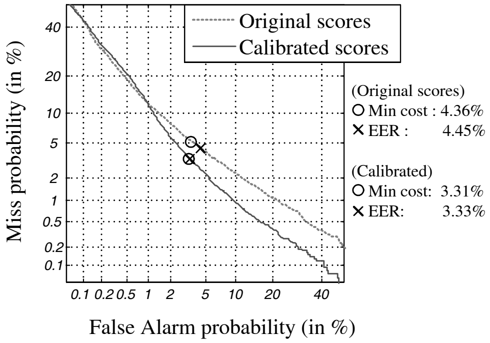

Figure 4: *DET curve for the original and the calibrated scores with $\upsilon = 3.5$*

Table 1: *Error statistics before and after score calibration*

| | With $\lambda_{\neg n_t}^{n_t}$ * | | With $\lambda_{\neg n_t}^{'n_t}$ * | |
|---|---|---|---|---|
| | $C_{\text{min}}$ | $C_{\text{eer}}$ | $C_{\text{min}}$ | $C_{\text{eer}}$ |
| Average(Avg.) on 23 languages | 4.36% | **4.45**% | 3.31% | **3.33**% |
| Avg. of 5 related language pairs | 7.70% | 7.69% | 5.40% | 5.42% |
| Avg. of other 13 languages | 1.79% | 1.95% | 1.71% | 1.72% |

* $\lambda_{\neg n_t}^{n_t}$ is the original scores, $\lambda_{\neg n_t}^{'n_t}$ is the calibrated scores derived in Eq.(9)

### 5.3. Detection errors in different target languages

Table 2 describes the $C_{\text{detect}}$ metric (Eq.(5)) in the five related language pairs at the operating points corresponding to global $C_{\text{eer}}$ before and after calibration. It is observed that all target languages except Russian have error reductions after calibration, indicated by smaller $C_{\text{detect}}$.

Referring to Eq.(5), the component terms of $C_{\text{detect}}(n_t)$ are recorded in Table 2 for analysis. These terms include the miss rate to a target language ($P_{\text{Miss}}(n_t)$), the false alarm rate specific to a related language pair ($P_{\text{FA}}(n_t, n_r)$) and the overall false alarm rate for the calculation of $C_{\text{detect}}(n_t)$ in Eq.(5). It is reminded at the global $C_{\text{eer}}$ operating point, the corresponding miss and false alarm probabilities in a single target language do not have to satisfy the *equal error* criterion. For instance, with original scores, $P_{\text{Miss}}$ and $P_{\text{FA}}$ for Bosnian is 35.49% and 1.58% at the global $C_{\text{eer}}$ operating point, giving $C_{\text{detect}}$ of 18.54% (Table 2).

The optimal parameter $\alpha_{n_t,n_r}$ found by Eq.(10) is also recorded in Table 2. This parameter specifies the proportion of the related language likelihood ratio ($\lambda_{\neg n_r}^{n_r}$) to be added to the target language likelihood ratio ($\lambda_{\neg n_t}^{n_t}$) in calibration (Eq.(9)). By looking at the parameter $\alpha_{n_t,n_r}$, two scenarios can be observed.

In the first scenario, a negative $\alpha_{n_t,n_r}$ is found to be optimal. Take Russian detection as an example and refer to Eq.(9), such an adjustment subtracts $\lambda_{\neg n_r:\text{Ukrainian}}^{n_r:\text{Ukrainian}}$ from the original $\lambda_{\neg n_t:\text{Russian}}^{n_t:\text{Russian}}$ likelihood ratio. The subtraction operation suppresses the high scores in $\lambda_{\neg n_t:\text{Russian}}^{n_t:\text{Russian}}$ in case of a false alarm in Ukrainian, and compensates the low scores in case of a detection miss in Russian. Recall that the error cost for a detection

Table 2: *Errors for different targets at $C_{eer}$ operating point*

| $n_t$:Target | $n_r$:Related | @$C_{\text{eer}}$ with $\lambda^{n_t}_{\neg n_t}$ * | | | | @$C_{\text{eer}}$ with $\lambda'^{n_t}_{\neg n_t}$ * | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| language | language | $P_{\text{Miss}}(n_t)$ | $P_{\text{FA}}(n_t, n_r)$ | $\sum \frac{P_{\text{FA}}(n_t,n_n)}{N-1}$ | $C_{\text{detect}}(n_t)$ | $\alpha_{n_t,n_r}$ | $P_{\text{Miss}}(n_t)$ | $P_{\text{FA}}(n_t, n_r)$ | $\sum \frac{P_{\text{FA}}(n_t,n_n)}{N-1}$ | $C_{\text{detect}}(n_t)$ |
| Bosnian | Croatian | 35.49% | 23.94% | 1.58% | 18.54% | 0.76 | 12.68% | 71.28% | 3.57% | 8.12% |
| Croatian | Bosnian | 8.78% | 74.93% | 5.07% | 6.92% | 0.43 | 8.78% | 79.72% | 4.18% | 6.48% |
| Dari | Farsi | 14.91% | 14.32% | 3.22% | 9.07% | 0.34 | 11.05% | 35.29% | 3.01% | 7.03% |
| Farsi | Dari | 0.26% | 72.49% | 7.08% | 3.67% | −0.30 | 0.51% | 47.81% | 4.79% | 2.65% |
| Eng(Ame) | Eng(Ind) | 2.08% | 55.50% | 5.92% | 4.00% | 0.05 | 3.40% | 45.52% | 3.81% | 3.61% |
| Eng(Ind) | Eng(Ame) | 2.54% | 38.93% | 6.51% | 4.53% | 0.13 | 3.05% | 37.50% | 4.54% | 3.79% |
| Hindi | Urdu | 4.20% | 80.74% | 12.67% | 8.43% | 0.62 | 1.80% | 97.89% | 9.13% | 5.46% |
| Urdu | Hindi | 2.11% | 85.76% | 11.12% | 6.61% | 0.67 | 2.11% | 96.85% | 8.58% | 5.35% |
| Russian | Ukrainian | 0.00% | 52.06% | 10.43% | 5.21% | −0.27 | 0.19% | 43.56% | 10.52% | 5.35% |
| Ukrainian | Russian | 19.07% | 3.25% | 0.73% | 9.90% | 0.76 | 10.82% | 34.03% | 1.98% | 6.40% |

* $\lambda^{n_t}_{\neg n_t}$ is the original scores, $\lambda'^{n_t}_{\neg n_t}$ is the calibrated scores in Eq.(9).
All $C_{\text{detect}}$ and $P_{\text{FA}}$ terms are calculated at the operating points for global $C_{\text{eer}}$

miss is $N-1$ times of the cost for a false alarm (Eq.(8)). So the biggest concerns are those Russian speech segments having large scores in $\lambda^{n_r:\text{Ukrainian}}_{\neg n_r:\text{Ukrainian}}$, which will incur detection misses of Russian after the subtraction operation in Eq.(9). As a result, the prerequisite for a negative $\alpha$ to be optimal is a low false alarm rate in the detector of the related language. In Table 2, $P_{\text{FA}}(n_t\text{:Ukrainian}, n_r\text{:Russian})$ is only 3.25%. A subtraction will not incur detection misses of Russian. Similarly, scores of the Dari detector have relatively low false alarm rate in Farsi (14.32%), and it is subtracted from the scores of the Farsi detector.

The second scenario occurs for the detector $n_t$ where false alarm rate is high in the detector of the related language. The optimal $\alpha_{n_t,n_r}$ parameters found by Eq.(10) are non-negative. This is because subtraction of scores would incur a significant number of detection misses, which means a high cost $C_{\text{detect}}(n_t)$ contributing to the global error. In the score adjustment of American and Indian English, the optimal value of $\alpha_{n_t,n_r}$ are found to be around zero. For other detectors, optimal values of $\alpha_{n_t,n_r}$ are positive. Essentially the adjusted score is a weighted sum of scores from $n_t$ and $n_r$ detectors. The two related languages are less differentiated, in return for fewer detection misses $P_{\text{Miss}}(n_t)$, and/or fewer false alarms irrelevant to the related language pairs $P_{\text{FA}}(n_t, n_n | n_n \notin \langle n_t, n_r \rangle)$.

Table 3: *Confusion costs of specific pairs of related languages*

| $n_t$:Target | $n_r$:Related | Original | Calibrated:23 lang | | Calibrated:2 lang | |
|---|---|---|---|---|---|---|
| language | language | $\underset{\theta_{n_t}}{\text{eer}}\, C_{\text{cf}}(n_t)$ | $\alpha_{n_t,n_r}{}^*$ | $\underset{\theta_{n_t}}{\text{eer}}\, C_{\text{cf}}(n_t)^*$ | $\alpha_{n_t,n_r}{}^\sharp$ | $\underset{\theta_{n_t}}{\text{eer}}\, C_{\text{cf}}(n_t)^\sharp$ |
| Bosnian | Croatian | 30.10% | 0.76 | 35.16% | −0.17 | 29.82% |
| Croatian | Bosnian | 31.33% | 0.43 | 32.97% | −0.01 | 31.05% |
| Dari | Farsi | 14.87% | 0.34 | 16.92% | −0.49 | 12.31% |
| Farsi | Dari | 12.05% | −0.30 | 11.79% | −0.55 | 11.54% |
| Eng(Ame) | Eng(Ind) | 16.10% | 0.05 | 16.24% | −0.52 | 16.04% |
| Eng(Ind) | Eng(Ame) | 16.38% | 0.13 | 17.24% | −0.74 | 15.04% |
| Hindi | Urdu | 28.28% | 0.62 | 32.21% | −0.59 | 28.77% |
| Urdu | Hindi | 30.31% | 0.67 | 32.98% | −0.85 | 29.05% |
| Russian | Ukrainian | 14.71% | −0.27 | 11.31% | −0.60 | 10.32% |
| Ukrainian | Russian | 11.54% | 0.76 | 16.47% | −0.81 | 9.77% |

* $\alpha_{n_t,n_r}$ and error terms in the middle columns are for 23 languages, identical with those obtained in previous sections, with $\upsilon = 3.5$,$N = 23$ in Eq.(10).
♯ $\alpha_{n_t,n_r}$ and error terms in the rightmost columns is for 2 languages, obtained with $\upsilon = 0$, $N = 2$ in Eq.(10).

### 5.4. Confusion among related language pairs

By looking at $P_{\text{FA}}(n_t, n_r)$ before and after calibration in Table 2, it is suggested that detection target dependent calibration towards a lower global error somehow sacrifices the differentiation between a target language $n_t$ and its related language $n_r$. A follow-up experiment is conducted to show that it is possible to improve the classification of confusing language pairs. As we are only interested in the confusion between the target language $n_t$ and the related language $n_r$, in every detector we look at speech segments of the two related languages only (i.e. We look at $\lambda^{n_t}_{\neg n_t}(k)$ where $k \in \{\mathcal{I}(n_t) \cup \mathcal{I}(n_r)\}$, $\mathcal{I}(n)$ is the indices of speech segments whose true language is $n$). The total number of target classes is two and Eq.(5) is revised to give a *confusion cost*, $C_{\text{cf}}(n_t)$, between $n_t$ and $n_r$ in detecting $n_t$.

$$C_{\text{cf}}(n_t) = \frac{1}{2} P_{\text{Miss}}(n_t) + \frac{1}{2} P_{\text{FA}}(n_t, n_r) \tag{11}$$

Instead of using a single detection threshold $\theta$, detection of language $n_t$ will use a target language specific detection threshold $\theta_{n_t}$. In the following, we will look at the *confusion costs* of different target languages at the *equal error rate* operating points, where $P_{\text{Miss}}(n_t)$ and $P_{\text{FA}}(n_t, n_r)$ have the smallest difference. This cost is denoted as $\underset{\theta_{n_t}}{\text{eer}}\, C_{\text{cf}}(n_t)$.

Table 3 shows the confusion costs $C_{\text{cf}}$ at equal error rates for the five related language pairs. The leftmost column of figures is the original confusion costs without "detection target dependent score calibration". The Bosnian-Croatian pair and the Hini-Urdu pair show the biggest confusion among five pairs.

In the middle, the revised metric, $C_{\text{cf}}$, illustrates how classifications between two related languages suffer with detection target dependent score calibration towards a lower *global* error cost. $\alpha$ parameters are the same as those in Table 2. The confusion cost, $C_{\text{cf}}$, is increased when $\alpha$ is positive. It is reminded that Table 3 records an equal error rate specific to only two languages $n_t$ and $n_r$; while in Table 2, a different operating point with a minimum global equal error rate constraint is picked.

Finally, calibration algorithm with Eq.(10) is modified, subject to lower error costs specific to only the two related languages. New $\alpha$ parameters are derived with Eq.(10) substituting $\upsilon = 0$ and $N = 2$. $\upsilon$ is chosen such that optimization in Eq.(10) does not have bias in reducing misses or false alarms. $N = 2$ as only two detection targets $n_t$ and $n_r$ are of interest. The optimal $\alpha$ parameters and new confusion cost $C_{\text{cf}}$ are recorded on the rightmost columns in Table 3. The newly found

$\alpha$ parameters are all negative. It follows our discussion in Section 5.3 that a negative $\alpha$ reduces the confusion between related language pairs. The confusion costs in all target languages except Hindi demonstrate different degrees of reduction.

## 6. Conclusions

In this paper, detection target specific score calibration for some related languages is proposed for language detection. With an optimal weight, the linear combination of scores between related classes effectively reduces the final detection errors in a global sense. The calibration method can be extended to different scenarios such as minimizing the confusion within a subset of related languages. Given limited information to differentiate the related languages in language recognition, the results indicate that overall improvement of detection errors is still possible.

## 7. Acknowledgment

This research is partially supported by the General Research Funds (Ref: CUHK 414108) from the Hong Kong Research Grants Council.

## 8. References

[1] N. Brümmer and D. A. van Leeuwen, “On calibration of language recognition scores,” in *Proc. IEEE Odyssey - The Speaker and Language Recognition Workshop*, pp. 1-8, 2006.

[2] E. Singer, P.A. Torres-Carrasquillo, T.P. Gleason, W.M. Campbell and D.A. Reynolds, “Acoustic, phonetic, and discriminative approaches to automatic language identification,” in *Proc. Interspeech*, pp. 1345-1348, 2003.

[3] M. F. BenZeghiba, J.-L. Gauvain and L. Lamel,“Language score calibration using adapted Gaussian back-end,” in *Proc. Interspeech*, pp. 2191-2194, 2009.

[4] N. Brümmer and J. du Preez, “Application-indepedent evaluation of speaker detection” in *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230-275, 2006.

[5] D. Zhu, H. Li, B. Ma and C.-H. Lee, “Optimizing the performance of spoken language recognition with discriminative training,” in *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 16, no. 8, pp. 1642-1653, 2008.

[6] The 2009 NIST Language Recognition Evaluation Plan (LRE09). [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/lre/2009/LRE09_EvalPlan_v6.pdf

[7] The 2009 NIST language recognition evaluation results. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/lre/2009/lre09_eval_results

[8] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[9] C.-C. Leung, R. Tong, B. Ma and H. Li, “A lattice-based phonotactid language recognition system with CMLLR adaptation and its implementation issues,” in *Proc. IALP*, pp. 285-288, 2009.

[10] R.W.M. Ng, C.-C. Leung, T. Lee, B. Ma and H. Li, “Prosodic attribute model for spoken language identification,” in *Proc. ICASSP*, pp. 5022-5025, 2010.

[11] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming. [Online]. Available: http://standford.edu/ boyd/cvx. June 2009.

[12] A. Martin, G.Doddington, T. Kamm, M. Ordowski, M. Przybocki, “The DET curve in assessment of detectioni task performance,” in *Proc. Eurospeech*, pp. 1895-1898, 2007.