The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/94406

Please be advised that this information was generated on 2019-09-27 and may be subject to change.
Evaluation of i-vector Speaker Recognition Systems for Forensic Application

Miranti Indar Mandasari, Mitchell McLaren, David van Leeuwen

Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
{m.mandasari,m.mclaren,d.vanleeuwen}@let.ru.nl

Abstract

This paper contributes a study on i-vector based speaker recognition systems and their application to forensics. The sensitivity of i-vector based speaker recognition is analyzed with respect to the effects of speech duration. This approach is motivated by the potentially limited speech available in a recording for a forensic case. In this context, the classification performance and calibration costs of the i-vector system are analyzed along with the role of normalization in the cosine kernel. Evaluated on the NIST SRE-2010 dataset, results highlight that normalization of the cosine kernel provided improved performance across all speech durations compared to the use of an unnormalized kernel. The normalized kernel was also found to play an important role in reducing miscalibration costs and providing well-calibrated likelihood ratios with limited speech duration.

Index Terms: i-vector, speaker recognition, forensics, calibration, short utterances

1. Introduction

One of the potential application areas of automatic speaker recognition is investigation and evidence reporting in forensics. In a typical case scenario, a victim may have received a threatening phone call. The recording of this phone call (referred to as the trace) may then be compared to a database of speech samples of known criminals to produce a ranked list of potential suspects. In this scenario the speaker recognition system is used for investigation purposes. When a suspect of the crime is found (by the aforementioned search or other means), an automatic speaker recognition system can be used to establish the degree of support that the suspect is the author of the recorded sample — this is referred to as evidence reporting. The speaker recognition system in this case must be well-calibrated and should report the strength of evidence as a likelihood ratio so as to adhere to modern fact finding conventions in court [1].

The forensic scenario is very challenging for speaker recognition for several reasons. The quality of the trace (e.g., signal-to-noise ratio) cannot be controlled and is not known, and the duration of the speech sample can vary from a few seconds to several hours. Further, the recording conditions are often not precisely known making the calibration of the speaker recognition system difficult. These circumstances typically vary from case to case such that finding speech data for system calibration that is representative of the trace conditions becomes a laborious process. An ideal system would be able to produce well-calibrated likelihood ratios without sensitivity to factors such as trace quality and duration, thus allowing each forensic case to be treated with the same calibrated system.

A speaker recognition system that was reported to exhibit good calibration characteristics during the recent NIST Speaker Recognition Evaluations (SRE) [2] was the state-of-the-art i-vector framework [3]. An i-vector is a compact representation of an utterance extracted from a low-dimensional total variability subspace trained via factor analysis. I-vectors are subject to inter-session compensation before performing speaker detection using a cosine kernel.

In this paper, we investigate the effects of speech duration on the calibration of the i-vector framework for speaker recognition. Focus is given to the classification performance and calibration costs of the i-vector system that has been developed and calibrated using the homogeneous duration speech dataset. Analysis is expected to highlight where the i-vector framework is sensitive to variations such as speech duration and its mismatch to the dataset used in calibration. Such sensitivities may trivially be dealt with by conditioning the calibration data on exactly the same duration characteristics as the trial at hand [4, 5]. This may be relatively easy for the duration factor studied in this paper, but will be less trivial for factors like signal to noise ratio and room acoustics. As to alleviate the need for this potential laborious process, we hope that by characterizing the duration-dependence of the i-vector system, it will be possible to design methods for dealing with this type of calibration issue.

This paper structured as follows. Section 2 defines calibration and its role in the context of forensics. Section 3 details the speaker recognition system, speech data sources and experimental setup. The results and analysis are given in Section 4.

2. Calibrating Similarity Scores

For forensic evidence reporting, scores from an automatic speaker recognition system must have the interpretation of a likelihood ratio (LR) in the forensic sense,

\[
\text{LR} = \frac{P(E|H_p,I)}{P(E|H_d,I)} \quad (1)
\]

where \( E \) is the trace (incriminating recording), \( H_p \) and \( H_d \) represent the prosecutor and defense hypothesis respectively, and \( I \) denotes other circumstances relevant to the case. Likelihood ratios can be used in court by the fact finder (judge or jury) to compute the posterior odds,

\[
\frac{P(H_p|E,I)}{P(H_d|E,I)} = \text{LR} \cdot \frac{P(H_p|I)}{P(H_d|I)} \quad (2)
\]

where the second factor is the prior odds determined by the court after considering other evidence. Calibration (i.e., converting scores to likelihood ratios) is a difficult task, but the most common way is to use a linear transformation of the
scores [6]. Calibration then involves optimizing this transforma-
tion using a development set of scores to minimize
\[
C_{llr} = \frac{1}{N_{H_1}} \sum_{i=1}^{N_{H_1}} \log_2(1 + \frac{1}{LR_i}) + \frac{1}{N_{H_2}} \sum_{j=1}^{N_{H_2}} \log_2(1 + LR_j)
\]
where \(N_{H_1}\) and \(N_{H_2}\) are the number of comparisons in the two hypoth-
oses respectively. The \(C_{llr}\) provides an estimation of the cali-
bration error over all priors. Readers are directed to [7] and [8] for further details on calibration and \(C_{llr}\).

3. Experimental method and data

3.1. Speaker recognition system

Features were extracted from audio samples by calculating 19 MFCCs plus log energy from speech audio signals every 10 ms using a 20 ms analysis window. These features were augmented with delta and double-delta coefficients. Speech activity detec-
tion was implemented using the same dataset used for i-vector space based on a set of held-out i-vectors. Normaliza-
tion was applied. Along with performance properties, the characteristics of the target and non-target score distributions were analyzed.

The last set of experiments concern system calibration. Lin-
ear calibration was performed using the FoCal toolkit [6], where scores for training the calibration parameters were sourced from the calibration portion of the full-full train-test duration combi-
ation (as defined in Section 3.2). The calibration parameters (an offset and a linear scaling) were learned using logistic re-
gression, and applied to the evaluation half of the trials of all the different duration conditions. Thus, there was no overlap between target speakers in calibration training and evaluation, but there was a mismatch between the duration of the segments used for calibration and evaluation.

3.2. Evaluation and calibration data

We use NIST SRE-2010 evaluation data [2] to characterize the performance of our i-vector system. Although our system has been developed to deal with a variety of recording, speaking style and vocal effort conditions [9], in this work we concen-
trate on the telephone trials (a.k.a. 'condition 5') as this con-
dition appears most relevant to the forensic cases. We use the 'extended core test' trial list distributed by NIST after the of-

ficial submission deadline, which consists of 416119 trials, as this leads to better error measurement statistics than the original trial list that has far fewer trials.

The core condition consists of 5-minute conversations, where each conversation side was typically found to contain 80 seconds of active speech. In order to study the effect of du-
ration, we synthesized test and train data sets of 5, 10, 20, 40

seconds by truncating the feature streams after SAD\(^1\). For con-

3.3. Experimental setup

Experiments were carried out using evaluation data with a speech duration of \(d = 5\) s, \(10\) s, \(20\) s, \(40\) s as well as full length utterances. An exhaustive set of duration combinations were tri-

alled so as to adequately analyze the effect of duration mismatch in the i-vector system. Here, we also varied whether or not co-

tsine kernel normalization was applied. Along with performance properties, the characteristics of the target and non-target score distributions were analyzed.

4. Results

In this section, we present and analyze the performance of our i-vector system with respect to varying speech duration in terms of classification performance, corresponding effects on score distributions and miscalibration cost. The use of both cosine kernel and normalized cosine kernel scoring is investigated.

4.1. Basic performance results

As an initial starting point, we investigate the effect of vary-
ing utterance duration on the performance of the i-vector based

\(^{1}\)Note that we did not use the data form the ‘10 second’ training and test conditions from NIST, because these have not been distributed in the ‘extended’ version, and moreover, we wanted to study duration dependence in a wider range of durations.
speaker recognition system. The corresponding EER values for both cosine kernel conditions are presented in Figure 1. It can be observed that the error rate increased as the train and test duration was reduced, which characterizes the typical behavior of most classifiers in speaker recognition [4, 5]. Noteworthy is the symmetry of the EERs around the point of matched train and test duration in Figure 1—that is, the i-vector system provided comparable EER when the train and test durations are swapped. This characteristic is distinct from other classifiers that typically treat the train speech segments different to the test speech segments. The symmetry of the i-vector system, therefore, contributes a simplistic configuration in that a distinction between train or test utterance does not have to be made when dealing with speech samples of mismatched duration.

Observing EER trends in Figure 1, we see that the i-vector system with a normalized kernel has, in general, better performance across all duration combinations than the unnormalized kernel. This is of interest as the cosine kernel normalization was based solely on a full-length utterance dataset. Other classifiers such as support vector machines and joint factor analysis (JFA) require that the dataset used for score normalization be matched to the evaluation condition in order to maintain reasonable classification performance [4, 5]. Although not shown here, similar trends between the unnormalized and normalized system were also found in terms of \( C_{\text{min}} \) and EER values from several duration combination trials in normalized i-vector system.

To better analyze the system performance at specific train and test durations, \( C_{\text{min}} \) and EER values from several duration combinations are presented in Table 1. From the matched train-test duration combination trials, it can be observed that the system performance was reduced by close to a factor of two as the duration of speech was halved. Full-length utterance training, on the other hand, provided a more graceful reduction in the system performances. These trends along with the symmetrical behavior of the i-vector system are of particular interest in forensic evidence reporting where long speech samples can be collected from a suspected speaker in an interview scenario, while the trace may be of uncontrolled duration.

### 4.2. Score distributions

We performed experiments to analyze the system score distributions by looking at the mean and standard deviation of the target and non-target scores with respect to varying train and test durations. Figure 2 shows the mean of both scores distributions from these experiments when using the unnormalized and normalized cosine kernels. It can be seen from Figure 2(a) that there was a negative shift in both score distributions of the unnormalized system as the speech length was reduced. In contrast, Figure 2(b) indicates that the normalized system provided a relatively stable non-target score mean along with a more uniform separation between the target and non-target scores.

Table 2 details the mean and standard deviation of the standard deviations of the target and non-target score distributions when using both kernels in the i-vector system. Limited fluctuation occurred in the standard deviation of the score distributions as indicated by the low \( \sigma \) relative to the average standard deviation \( \mu \). Thus, the standard deviation of the i-vector scores have a limited sensitivity to the length of speech duration and the application of the cosine kernel normalization.

### 4.3. Calibration experiments

A well-calibrated system is required when using an automatic speaker recognition system in forensic application for presenting evidence to court. Here we analyze the calibration in terms of the extra costs \( C_{\text{det}} \) due to miscalibration.

The values of \( C_{\text{det}} \), representing several duration conditions...
the speech signal (in this case utterance duration) to the perfor-
mance of a modern speaker recognition system in terms of dis-
rimination and calibration. We used fixed calibration param-
eters trained on scores from full-length utterance trials. This
showed that duration variation has quite an influence on the
quality of the LR, in some cases producing $C_{llr}$ costs larger
than one which indicates that such a system should not be
used in those duration conditions. Normalization of the cosine
costs larger so for short durations this calibration would be con-
sidered very poor. The miscalibration costs for both cosine ker-
els are depicted graphically in Figure 3. By comparing Fig-
ure 3(a) and Figure 3(b), it can be observed that the miscal-
ibration cost is reduced dramatically w.r.t. normalization in the
cosine kernel scoring. Nonetheless, even in the case of the nor-
malized kernel, the full-utterance calibration produced damag-
ing log likelihood-ratios for the shorted durations (cf. Table 3).

5. Conclusions

In this paper we analyzed the effect of a quality factor from
the speech signal (in this case utterance duration) to the perfo-
rmance of a modern speaker recognition system in terms of dis-

Table 2: Statistics of the standard deviation of the target and
non-target score distributions across all duration combinations.

<table>
<thead>
<tr>
<th>Score Distribution</th>
<th>Cosine $\mu$</th>
<th>Cosine $\sigma$</th>
<th>Norm. Cosine $\mu$</th>
<th>Norm. Cosine $\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td>0.084</td>
<td>0.002</td>
<td>17.5</td>
<td>1.16</td>
</tr>
<tr>
<td>Non-target</td>
<td>0.082</td>
<td>0.004</td>
<td>14.8</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Table 3: $C_{llr}$ and $C_{llr}^{\min}$ values for female scores with respect to
train and test speech duration for the i-vector system using both
unnormalized and normalized cosine kernels.

<table>
<thead>
<tr>
<th>Duration (sec)</th>
<th>Cosine $C_{llr}$  </th>
<th>$C_{llr}^{\min}$</th>
<th>Norm. Cosine $C_{llr}$  </th>
<th>$C_{llr}^{\min}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>8.45</td>
<td>0.70</td>
<td>1.54</td>
<td>0.73</td>
</tr>
<tr>
<td>10</td>
<td>5.48</td>
<td>0.51</td>
<td>1.02</td>
<td>0.50</td>
</tr>
<tr>
<td>20</td>
<td>4.09</td>
<td>0.32</td>
<td>2.60</td>
<td>0.30</td>
</tr>
<tr>
<td>40</td>
<td>1.54</td>
<td>0.14</td>
<td>1.02</td>
<td>0.13</td>
</tr>
<tr>
<td>60</td>
<td>5.48</td>
<td>0.14</td>
<td>1.02</td>
<td>0.13</td>
</tr>
</tbody>
</table>

There are ways to deal with this calibration phenomenon.
One is to re-calibrate for every possible duration condition
using development data of matching duration, which can be labo-
rious and does not generalize trivially to other quality factors
or very long duration conditions. A better way would be to
include the quality factor in the calibration model as ‘side infor-
mation,’ which still needs calibration data of similar conditions,
but hopefully the calibration model can interpolate for unseen
duration conditions. Our future work includes the investiga-
tions on such calibration models, and extending the analysis to other
quality factors such as signal to noise ratio and room acoustics.

6. References

mouchel, “Support vector machines versus fast scoring in the low-
[6] N. Brummer, FoCal-II: Toolkit for calibration of multi-
[8] D. van Leeuwen and N. Brummer, “An introduction to application-independent evaluation of speaker recognition sys-
weighted LDA for robust speaker recognition using i-vectors,” in Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Process-
ing, 2011, pp. 5456–5459.
sine similarity scoring without score normalization techniques,” in Proc. Odyssey Workshop, 2010.