An automatic speaker recognition (ASR) system must produce reliable likelihood ratios (LR) in order to be used for evaluating and presenting speech evidence to court. The LR is only reliable if it produced from a well-calibrated ASR. A study by Rodriguez (2007) showed that the LR calculated from the un-calibrated system was often misleading, while the calibrated system produced more reliable LRs. These results illustrate the necessity of performing LR calibration when using the ASR system for forensic applications. Several calibration techniques for ASR systems, such as the linear calibration method as presented in (Brümmer, 2006) and the novel line-up calibration method which was developed by van Leeuwen and Brümmer (2011) have been proposed in order to calibrate the LR produced by ASR.

This paper reports on the LR calibration performance of the i-vector based ASR. Calibration of LRs was firstly conducted using the line-up approach which resembles voice line-up task in the forensic phonetic cases. The resulting calibration performance was then contrasted against the previous findings when using linear calibration presented in (Mandasari, 2011).

**Line-up method for likelihood ratio calibration**

Voice line-up is one of the most frequent forensic phonetic tasks in preparing evidence reports for the court (Butcher, 2002). In this voice-line up task, the witness listens to a recording where the voice of the suspect is lined up together with a number of ‘foils,’ and then they are asked whether the can recognize the perpetrator or not. The line-up calibration method that was developed by van Leeuwen and Brümmer (2011) was motivated by the witness line-up scenario in forensic tasks. In their paper, each speaker from the evaluation dataset is “lined-up” with all foils speakers from the development dataset so as to resemble the voice line-up task in forensics. The rank of the suspect within the line-up of foil speakers, when compared to the trace recording, is then used to compute the LR. Here, the expressions for LR have been derived to depend only on the ASR performance and rank, under the assumption that recordings of the foils are chosen such that they do not differ considerably from the suspect recording.

**i-vector based speaker recognition system**

The experiments presented in this abstract focus on the i-vector framework for ASR which was recently developed by Dehak (2010). In general, i-vector is a representation of speech utterance
extracted from a low-dimensional space referred to as the total variability space. The latest findings from the present authors in (Mandasari 2011) give several valuable remarks regarding the application of the i-vector based ASR to the task of forensic speaker recognition. By utilizing a normalized cosine-kernel, the i-vector system was found to offer good calibration performance while exhibiting some robustness to shorter durations of speech samples.

**Experiment and results**

The performance of LR calibration was evaluated using mis-calibration ($C_{llr}-C_{llr}^{\text{min}}$) and the general system performance was determined in terms of Equal Error Rate (EER). Experiments were conducted using the recent NIST speaker recognition evaluation (SRE) 2010 dataset which was divided into two sets of disjoint speakers. One half of this dataset was used as the calibration dataset and the other half as the evaluation dataset. An exhaustive evaluation of suspect and trace speech sample duration combinations was conducted using speech durations 5, 10, 20, 40 and 80 seconds. The performance parameters of the line-up calibration experiment were then compared to the linear calibration experiment results for the same i-vector ASR system as presented in (Mandasari, 2011).

Results indicated that the $C_{llr}^{\text{min}}$ values were similar between the linear and the line-up calibration method. It was found, however, that the $C_{llr}$ value in the line-up calibration was up to 50% lower than the $C_{llr}$ in the linear calibration experiment. Interestingly, the general system performance (as measured by EER) was found to be slightly better for the line-up calibration. These findings lead to the conclusion that the line-up calibration method appears to act as a score normalization, improving the general system performance, while the scores can indeed be interpreted as likelihood ratios. The formalism of line-up calibration method still needs to be improved in order to completely replace the traditional linear calibration with the new line-up calibration method.

**References**


