

On TFSR (semi)automatic systems supportability: novel instruments for analysis and compensation

Francesco Borchi, Monica Carfagni, and Matteo Nunziati

Department of Mechanics and Industrial Technologies (DMTI), University of Florence, Florence, Italy

{francesco.borchi|monica.carfagni|matteo.nunziati}@unifi.it

Given a voice sample belonging to a known speaker, and a second sample, belonging to an unknown speaker, the likelihood ratio (LR) is defined as the ratio between the probability that the second sample belongs to the same person or to a different person. The aim of (semi)automatic Technical Forensic Speaker Recognition (TFSR) is to estimate the LR, obtaining the lowest possible level of false positives and false negatives. This is an important but difficult goal and although many improvements into (semi) automatic speech recognition have been introduced case work often requires a rapid response to a speaker identification question. It is not possible to wait until an optimal system has been developed therefore, while improvements will be progressively attained through research, real cases must employ the available systems even if they are not perfect. Of course, the accuracy level of currently available systems can not be ignored: the accuracy level must to be assessed and its effects on recognition scores must be compensated.

The paper presents a theoretical TFSR system classification based on the behaviour of the systems. The classification has been carried out by means of commonly available instruments and enables the identification of all major problems which can occur using (semi)automatic TFSR systems. By analysing the responses provided by some experimental systems developed at DMTI, it has been possible to define two major theoretical problems here referred to as the “under/over estimating” problem and the “supportability” problem. The former is related to the experimental evidence of under or over scoring behaviour of a system (i.e. the tendency of the system to produce lower or higher scores than expected), causing the system itself to be inconsistent with respect to the commonly adopted LR scale. The latter is related to the level of false positives and false negatives and their persistence at high scores. Both problems are identified by testing the systems against a set of known recordings.

After the definition of such problems, a new function, Supportability of System (SoS), is proposed to score the amount and distribution of both false positives and negatives. The function is obtained by generalising the well known LR test index. SoS scores will be employed to *compress* all obtained LRs, reducing the size of both false positive and false negative tails, on the basis of a theoretically motivated mathematical model. Additionally a basic model for under/overestimation compensation will be presented. Using the proposed models, scores will be decremented, which leads to more robust systems and eliminates or reduces the risk of wrong scores though at the cost of discriminative power.

The applied compensation increases with the amount of false scores, therefore a poorly performing system will show a low discriminative power. For that reason a smooth scale will be proposed to define how much it is possible to compensate a system while maintaining a good discriminative power and, thus, to estimate the level of supportability of a system. The Degree of Supportability (DoS) scale, defines a score which decreases with the amount of compression required by the system in order to be compensated. It will be used to assess whether a system is employable in real cases or not.

Since the main aim in this research has been to start a discussion about a common evaluation framework specifically developed for the forensic field, the proposed methodology maintains the widest possible compatibility with the common interpretation logics (e.g. Evett table), currently approved and accepted by the international forensic phonetic and acoustic community.