

# A Robust Framework for Forensic Speaker Verification

Yosef A. Solewicz (1,2), Moshe Koppel (1), Saad Sofer (2)

(1) Dept. of Computer Science, Bar-Ilan University, Ramat-Gan, Israel

(2) Israel National Police, Division of Identification and Forensic Science, Jerusalem, Israel

solewicz@013.net.il

## Abstract

This paper discusses the application of automatic speaker verification systems in forensic casework. A framework for reporting the system outcome is proposed. Specific system requirements to properly cope with forensic idiosyncrasies are analyzed through a series of simulations. Results suggest that the design of a forensic speaker verification system not necessarily match the settings of current state-of-the-art systems.

## 1. Introduction

Automatic speaker verification (ASR) as a biometric is still a long way from fingerprints or DNA fields. Speech is essentially behavioral, time-variant, noisy and, therefore, its application in the forensic context is controversial. For this reason, the validity of traditional examination methods like aural tests or visual comparison of spectrograms (erroneously known as ‘voiceprints’) has been severely criticized [1]. Moreover, besides the intrinsic noisy nature of speech, both aural and visual examinations are almost completely subjective, introducing further inconsistency to the verification process.

Automatic speaker verification has been recently introduced to the forensic field [2] as a new or complementary approach to older identification techniques. The goal is to obtain an incriminating weight for the voice evidence in question. The common idea is to wrap ASR systems in a statistical/Bayesian framework [3,4]. In this form, the system output could be comfortably interpreted as an indication of the suspect’s culpability in terms of statistical significances, likelihood ratios, confidence measures, etc.

Nevertheless, projecting an automatic speaker verification system for forensic purposes involves distinct considerations from those found in other applications. Forensic ASR should pursue robustness, keeping consistency in the inference scale. This should be attained even at the expenses of high accuracy in results. Thus, a system which offers limited support to court, but it is claimed to be confident in a wide range of recording environments is largely preferable to systems tuned to perform well in some specific benchmark. Moreover, the system should be consistent in scaling results throughout diverse environments.

Unfortunately, at present, no standard verification protocols exist for forensic ASR and the ‘culpability level’ obtained is dependent on the systems’ underlying

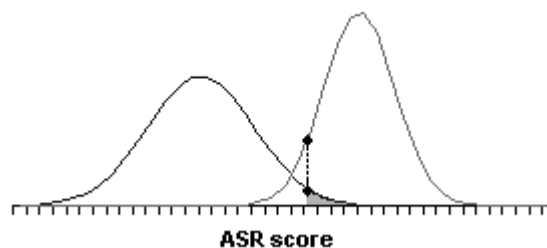
inference mechanism, meaning that different systems could yield different outcomes. Moreover, the validity of the evidential value obtained through the proposed systems is still uncertain. Nevertheless, we believe that ASR can still be a valuable tool in forensic casework.

This paper discusses, proposes and analyzes ways to better amalgamate current ASR technology in a forensic context, minimizing part of the challenging issues and is organized as follows: Section 2 briefly discusses assessment methodologies. Nuances of forensic ASR are discussed in section 3. Section 4 presents a series of validating experiments, and the results are reported in section 5. Section 6 consists of concluding remarks and proposals for future research.

## 2. Reporting the outcome

Expressing the forensic outcome involves estimation of statistical distributions of the relevant features in question among some reference population. Concerning the DNA field for instance, once statistics are proved to converge, one can offer reliable figures for certain occurrences in a control population. ASR, on the other hand, suffers from intrinsic and extrinsic noise and, more significantly, noise cannot be objectively predicted and measured in the data. Speaker health, emotional state, background and channel noise are some of the factors challenging ASR robustness and are very common in typical forensic casework. All these factors introduce some inconsistency between the specific case and previously collected statistics. As a result, biased estimates might be offered to the court.

Typically, forensic ASR outcomes entail a previous estimation of an impostor (“innocent”) scores distribution and, possibly, a target (“criminal”) scores distribution (see Figure 1).



“Figure 1. Impostor and target score distributions”

In case just the impostor distribution is used, we can estimate the so known p-value for the match. The p-value is defined as the probability of observing a given

sample under the assumption that the null hypothesis is true. In our specific case, the sample under observation is the occurrence of a certain suspect/evidence match score (the ASR output) and the null hypothesis is that the suspect is innocent, i.e. the score obtained belongs to the impostor scores distribution. If the p-value is sufficiently low, then we reject the null hypothesis. In other words, the p-value can be interpreted as the estimated fraction of the reference population showing a greater similarity with the voice evidence than the suspect (The highlighted area in Figure 1). In this way, a very small p-value weakens the assumption of the null hypothesis.

One of the criticisms against the p-value formulation in ASR is that it cannot necessarily be viewed as a measure of strength of evidence, since no inference is made about the competitive hypothesis (the ASR score belongs to the target distribution). Therefore, an alternative likelihood ratio report relating both hypotheses could be seen as a more appropriate form of weighting the evidence strength. The likelihood ratio can be estimated as the ratio between the target and impostor distribution densities for a given ASR score (The two points in Figure 1). Nevertheless, the concept of likelihood ratio is often not properly understood by the courts. Moreover, it involves an additional match with the target score distribution, as compared with the p-value approach. This leads to another problem. Estimating this distribution is a much more elaborate and delicate process than the estimation of the impostor distribution. For the latter case, plenty of varied recordings are available, covering a vast range of scenarios. This data can be contrasted with the suspect and/or evidence samples in order to produce diverse potential impostor statistics. On the other hand, estimating a target scores distribution for a particular suspect [5] requires an exhaustive collection of time-spaced and diversified recordings of this person. In practice, this approach is not feasible. An alternative approach would be the employment of an average target distribution, obtained by several distinct within-speaker matches as proposed in [6]. Unfortunately, this will introduce an additional approximation to the inference process.

As can be seen, both approaches offer pros and cons. While the p-value bears perhaps a more simple intuition than the likelihood ratio and is already widely adopted by the forensic DNA community, it offers only a partial assessment in forensic voice comparison. On the other hand, the likelihood ratio approach, although a more proper inference tool, entails additional assumptions when modeling the alternative hypothesis and its interpretation by laymen is occasionally problematic. For this reason, we would like to propose the employment of a combined report in forensic ASR casework. Assessment would be in the form of p-value for the specific case. Nevertheless, this estimate would be complemented by an average measure of its implication strength.

Within this framework, the typical evaluation procedure would be as follows. In case the reported p-value is sufficiently high, the null hypothesis (the suspect's score belongs to the 'innocent' distribution) simply cannot be rejected. On the other hand, when the reported p-value is low the court must interpret the weight of the findings. This can be viewed as weighting the p-value absolute information, in view of system past performance in similar conditions. This parallel information is referred to here as the 'Identification Confidence', being the ratio between the system's rate of true identifications and rate of false identifications, for a specific p-value. The higher the ratio, the more confident is the weighting towards a true identification.

A key point in this framework is that the average p-value range should be as robust as possible to noise-related effects. Otherwise, the estimated confidence related to the findings could be exaggerated. To meet this requirement, proper system configuration is crucial.

### 3. Forensic nuances in ASR

In the previous section, we presented some reasons for expressing ASR output in terms of p-value, adding side information about its strength as a function of the particular system and case. Some desirable properties of this measure were discussed. In this section, we will discuss proper ASR configuration for implementing this idea, while adhering to forensic constraints.

In order to estimate the impostor scores distribution, two options are available. Either to keep the suspect model constant and obtain a sequence of scores, matching this model against reference data, or keeping the evidence (test) constant and obtaining scores matching it against reference speaker models. Such distributions are used in regular ASR as a means of output normalization, increasing scores stability. The former approach leads to the so-called Z-norm technique, while the latter is referred to as T-norm normalization. In both cases, first- and second-order statistics of the obtained distribution are used for score standardization. In our case, we would use those statistics, representing population facts, in the p-value estimation process. (ASR scores are supposed to follow a Gaussian distribution and thus can be reasonably parameterized by these statistics.) The common belief is that the T-norm version is the more efficient technique in terms of detection accuracy [7]. Does this mean that this should be the proper way for estimating impostor scores distributions in forensic ASR? In either case, what should be the impostor population employed? Again, the common belief is that it should be as close as possible to train and test data signal characteristics. (Another successful score normalization technique known as 'cohort' is based on this principle [7]). In an attempt to answer these questions, we performed a series of experiments simulating forensic cases, which will be presented in the next section.

## 4. Experiments

In this section, we describe the experiments performed in an attempt to define a suitable configuration for a forensic ASR system. The experiments were conducted following the male speakers part of the NIST-2001 evaluation plan [8]. Model speakers and tests served as the hypothesized suspects and evidences, and scores are expressed in terms of p-value as would be reported to the court. The original evaluation contains 76 enrolled speaker models, 850 target trials and 8500 impostor trials. This evaluation is almost entirely composed of cellular recordings. Speaker models are trained from 2 min. of speech. For consistency, test segments shorter than 30 sec. were discarded (about 35% of the tests) and the others were truncated to this length. In order to simulate fairly realistic forensic conditions, the experiments were duplicated in a noisy test mode. This was accomplished by introducing to the tests a random amount of additive noise (mean SNR in noise ~14 dB) followed by random band-pass filtering (300-440 Khz to 2000-3400 Khz). (Note that, in principle, in typical forensic casework the evidence is recorded under uncontrolled conditions of noise and channel, but the suspect model can, in theory, be recorded in controlled conditions. For this reason, we keep the suspect models as is).

The system used is a gender independent GMM-UBM (1000 clusters) similar to that described in [9]. The UBM was trained with part of the speakers contained in the NIST-2002 evaluation and with data from the OGI speaker verification database [10].

Regarding the impostor scores distribution estimation, four male background sets were adopted. The first set consisted of part of the enrolled speakers from the NIST 2002 evaluation (a disjoint set from those used for training the UBM). This set will be referred to as 'matched' (MTC), since the recordings are also mostly from cellular media and acquired and processed in a very similar fashion to the test data. The other three groups were extracted from the FBI speaker verification database (evaluated at the NIST-2002 campaign and currently available from LDC [11]): telephone recordings (TEL), high-quality microphone (MIC) and body-microphone (BOD). Each of the groups contained data of 48 speakers (limited to the maximum amount available from one of the groups). The distributions were estimated in the two forms previously described: fixed claimant speaker model ('suspect') matched against reference data (48 recordings from each of the four sets) or fixed test ('evidence') matched against reference speaker models (48 models for each of the reference sets). Appealing to forensic jargon, the former approach will be referred as suspect model pdf estimation ('s') and the latter will be denoted as evidence pdf estimation ('e').

In conclusion, for each processed voice comparison, eight different p-values will be estimated, according to reference set (MTC, TEL, MIC, BOD) and the way the

distribution was obtained ('s' or 'e'). In addition, this procedure will be duplicated using the artificially corrupted tests.

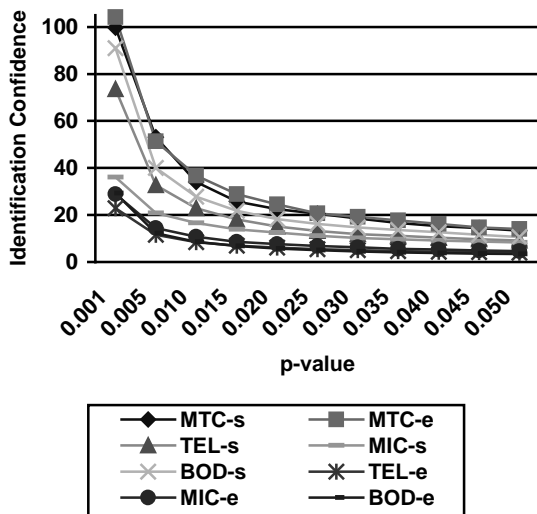
## 5. Results

Table 1 summarizes evaluation results according to the type of reference score distribution and noise in the test patterns. As expected, it can be observed that regarding overall error reduction, the reference set should be matched to evaluation data and reference scores pdf should be estimated keeping the test fixed ('evidence' pdf estimation, equivalent to 'T-norm'). However, the 'e'-type pdf's show an undesirable side effect: the estimated impostor distribution shifts across the p-value scale for all unmatched sets, as can be seen by the p-value at EER. In this sense, the 's'-type pdf's seem to be more stable. This phenomenon will undermine the defined confidence curves, meaning that curves estimated in certain circumstances would not be valid in more general environments. This is critical in forensic applications, since the p-value interpretation will be highly dependent on data mismatch issues. Recall that typical forensic casework is often characterized by noisy evidence and obtained in imprecise recording conditions.

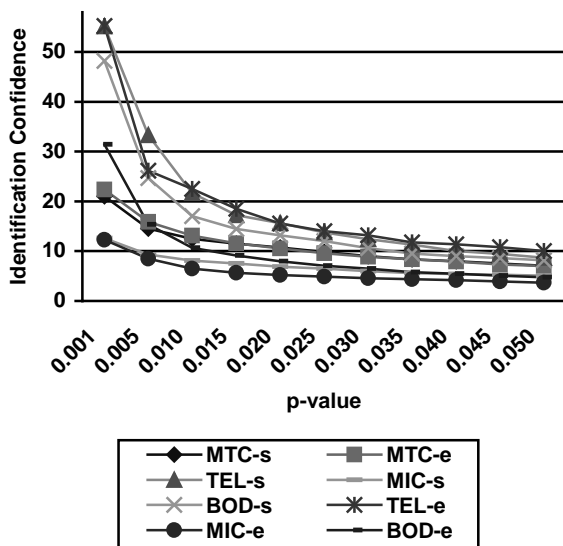
"Table 1: p-value average shift in clean, noise, matched and mismatched conditions"

Ref. Scores distribution	p-value @ EER (Clean/Noise)	EER (%) (Clean/Noise)
MTC-s	0.09 / 0.14	9.7 / 19.9
MTC-e	0.08 / 0.13	9.4 / 18.2
TEL-s	0.10 / 0.15	14.6 / 19.5
TEL-e	0.02 / 0.18	12.6 / 23.2
MIC-s	0.06 / 0.11	12.3 / 22.4
MIC-e	0.02 / 0.05	11.1 / 20.4
BOD-s	0.16 / 0.22	17.3 / 23.4
BOD-e	0.01 / 0.09	11.7 / 22.3

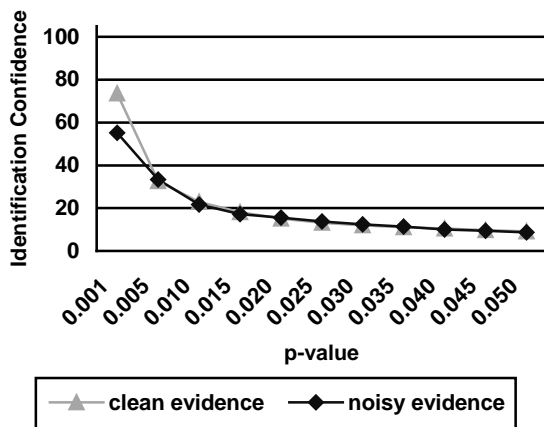
Figures 2 and 3 depict the previously defined "Identification Confidence" curve as a function of obtained p-value, for the several pdf types in clean and noisy tests. As explained, our motivation is to design an ASR system characterized by, as much as possible, a constant curve for clean and noisy test conditions, even at expenses of high "Confidence" values. In this sense, for the present evaluation, the preferable configuration would be TEL-s (normalizing the suspect model by means of a "mismatched" telephone reference set). Note that TEL can be reasonably considered the closest unmatched reference set for this evaluation. A close look for this specific configuration is seen in Figure 4.



“Figure 2: Identification Confidence - clean evidence”.

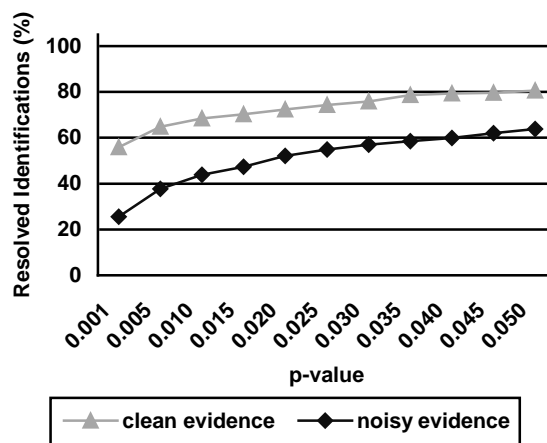


“Figure 3: Identification Confidence - noisy evidence”.



“Figure 4: Identification Confidence: TEL-s, clean and noise”.

Finally, Figure 5 depicts the proportion of true target identification as function of obtained p-value for the chosen TEL-s configuration. Note that, even in noisy conditions, about 50% of targets would receive a p-value as low as 0.015, with an attached confidence of around 20 (see Figure 4). This is a very reasonable conclusion, considering the strict forensic scenarios, which demand the sacrifice of sharp decisions on behalf of security.



“Figure 5: Resolved Identifications: TEL-s”.

## 6. Conclusions

Within the limited scope of the experiments performed, results suggest that ASR systems should be properly configured for forensic applications. Forensic idiosyncrasies require system settings not necessarily shared by current state-of-the-art systems. In particular, a forensic system should employ suspect model normalization for reference score distribution estimation. Furthermore, the reference data employed should be similar but not closely match the type of the questioned data. Such settings will in general decrease the strength of the reported outcome but will support consistency in reports for the typical irregular forensic environment. These conclusions are based on a series of evaluations attempting to simulate typical forensic environment and are presented in the context of a proposed framework for reporting the ASR outcome.

Future research should cover more simulations with diversified datasets in order to solidify the proposed methodology. In particular, special issues such as cross-language and speech under stress must be considered.

## 7. References

[1] Braun A., and Künzel H. J., “Is Forensic Speaker Identification Unethical - Or can it be Unethical Not to Do it?”, Proceedings of RLA2C Workshop on Speaker Recognition and its Commercial and Forensic Applications, Avignon, France, 145-148, 1998.

- [2] Broeders, A. P. A., "Forensic Speech and Audio Analysis, A Review", Proc. of 13th INTERPOL Forensic Science Symposium, Lyon, France, 16-19, 2001.
- [3] Champod C., and Meuwly D., "The Inference of Identity in forensic Speaker Recognition", Speech Communication 31, 193-203 (2000).
- [4] Rose, P., "Forensic Speaker Identification", London: Taylor and Francis, 2003.
- [5] Meuwly, D., and Drygajlo, A., "Forensic speaker recognition based on a Bayesian framework and Gaussian mixture modeling (GMM)", Proceedings of the Workshop ODYSSEY-2001, Crete, Greece 145-150, 2001.
- [6] Koolwaaij, J. W., and Boves, L.,J., "On decision making in forensic casework", Forensic Linguistics, the International Journal of Speech, Language and the Law, Vol. 6, Number 2, 242-264 (1999).
- [7] Auckenthaler R., Carey M., and Lloyd-Thomas H., "Score normalization for text-independent speaker verification systems", Digital Signal Processing, vol. 10, 42-54 (2000).
- [8] Przybocki, M., and A. Martin, "The NIST Year 2001 Speaker Recognition Evaluation Plan", <http://www.nist.gov/speech/tests/spk/2001/doc/>, 2001.
- [9] Reynolds, D., Quatieri, T., and Dunn, R., "Speaker verification using adapted gaussian mixture models", Digital Signal Processing 10, 19-41 (2000).
- [10] <http://www.cslu.ogi.edu/corpora/spkrec/>
- [11] <http://wave ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004S04>