# HUMANS VERSUS MACHINE: FORENSIC VOICE COMPARISON ON A SMALL DATABASE OF SWEDISH VOICE RECORDINGS

*Jonas Lindh*[a,b] *& Geoffrey Stewart Morrison*[c]

[a]Division of Speech and Language Pathology, Dept. of Clinical Neuroscience and Rehabilitation, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Sweden;
[b]The Swedish Language Bank (Språkbanken), Dept. of Swedish, University of Gothenburg, Sweden;
[c]Forensic Voice Comparison Laboratory, School of Electrical Engineering & Telecommunications, University of New South Wales, Sydney, Australia
jonas.lindh@gu.se; geoff-morrison@forensic-voice-comparison.net

## ABSTRACT

A procedure for comparing the performance of humans and machines on speaker recognition and on forensic voice comparison is proposed and demonstrated. The procedure is consistent with the new paradigm for forensic-comparison science (use of the likelihood-ratio framework and testing of the validity and reliability of the results). The use of the procedure is demonstrated using a small database of Swedish voice recordings.

**Keywords:** forensic voice comparison, human listeners, automatic speaker recognition

## 1. INTRODUCTION

There has been considerable interest recently on comparing the performance of humans and machines on speaker recognition, due to the human assisted speaker recognition (HASR) test introduced as part of the 2010 National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE) [7]. Most of the participants in the HASR test did not produce systems including human supervision of acoustic measurement as is common in acoustic-phonetic forensic voice comparison [13], but rather had panels of listeners (naïve with respect to auditory phonetics and psychoacoustics) attempt to decide whether pairs of recordings were produced by the same or different speakers.[1] Use of panels of listeners, even naïve listeners, is a procedure which has been applied to forensic voice comparison in the past [8] p. 204-205]. An iterative procedure could be adopted whereby the listeners who perform the best in each round of testing are retained in the panel and new listeners replace those who perform worst. This leads to the need for procedures in speaker recognition and in forensic voice comparison to compare the performance of humans against humans, and the performance humans against machines. As argued in [13] if a panel of listeners is found to outperform some other forensic-voice-comparison system in terms of validity and reliability, then for casework the panel of listeners should be preferred over the other system.

We propose a procedure for comparing panels of human listeners with other forensic-voice-comparison systems in a manner which is consistent with the new paradigm for forensic-comparison science (use of the likelihood-ratio framework and testing of the validity and reliability of the results, see [11, 13, 14, 15, 16]). We demonstrate the procedure by comparing a panel of listeners and a generic automatic forensic-voice-comparison system; both tested on the same set of pairs of Swedish voice recordings.

## 2. METHODOLOGY

### 2.1. Data

The data consisted of 45 pairs of recordings, 9 same-speaker pairs and 36 different-speaker pairs. The pairs were constructed from a total of 18 recordings, 2 recordings from each of 7 speakers, and 4 recordings of 1 speaker. The first recording from each speaker was paired with their own second recording and with the first recording of every other speaker, and for the speaker with 4 recordings the third recording was paired with his own fourth recording and with the first recording of every other speaker. Each recording was 13-15 seconds long. The first two recordings per speaker were actually different portions of an originally longer recording, for the speaker with 4 recordings the third and fourth recordings were from different original recordings. For most pairs of recordings, the within speaker variability did not therefore

include inter-session variability and was not forensically realistic in this respect.

The 8 speakers were a homogeneous group of male speakers of Swedish. All speakers spoke the same dialect (Gothenburg area) and ranged in age from 21 to 40. Initially 17 speakers were recorded, and the final 8 were selected on the basis of being most similar on a preliminary test in which 37 Swedish listeners (undergraduate students) gave similarity judgments on different-speaker pairs of recordings, each recording being 11-12 seconds long (see [17] for details of selection criteria).

The recordings were made in a quiet room using the built-in microphone of a Zoom H2 solid-state recorder, and saved as 16 kHz 16 bit raw wave files. The recorder was placed on a table about 60 cm in front of the speaker. The aim was to obtain good-quality recordings but not studio quality. The recordings consisted of spontaneous speech elicited by asking the speakers to describe a walk through the center of Gothenburg, based on a series of photos presented to them. The fourth recording of the speaker with 4 recordings was part of one side of a conversation.

### 2.2. Automatic system

The automatic forensic-voice-comparison system was of generic design, built using the MISTRAL platform [1]. 19 mel-frequency-cepstral-coefficient (MFCC) values were extracted every 10 ms over the entire speech-active portion of each recording (a simple energy detector removed silences of longer than 100 ms). Delta and double-delta coefficient values were also calculated and included in the subsequent statistical modeling [6]. A Gaussian mixture model - universal background model (GMM-UBM) [19] was built using 2 minutes net spontaneous speech from each of 628 male speakers in the SweDia dialect database [5] as data to train the background model. The model used 512 Gaussians. For each comparison pair, the first recording was used to build a suspect model and the second as offender probe data to calculate a score. The scores were calibrated and converted to likelihood ratios using linear logistic regression [4, 9] implemented using the FOCAL TOOLKIT [2] with a robust version of the training function [12]. Calibration was conducted using a cross-validated procedure in which the calibration weights were calculated using all the scores except those which were calculated from comparison pairs which included recordings of the same speaker (or speakers) as in the comparison pair corresponding to the score which was being calibrated (see, for example [14]).

### 2.3. Panel-of-human-listeners system

A panel of listeners judged the similarity of the pairs of recordings. There were 52 listeners, 13 males and 39 females, with ages ranging between 20 and 60. The listeners had a mixture of different first languages but most were first-language Swedish speakers. The experiment was presented using an online interface.[2] On each trial a listener was presented with a pair of recordings, they could listen to each recording as many times as they liked, and then gave their judgment as to the similarity of the speakers on a 5-point scale where 1 represented "extremely similar or same" and 5 represented "not very similar". It took approximately 25 minutes for a listener to judge the similarity of the 45 comparison pairs.

Each listener completed two versions of the similarity-judgment task. In one version the recordings were played forwards and in the other version the recordings were played backwards. The idea was that playing the recordings backwards would remove much of the phonetic and linguistic information which the listeners might otherwise rely on in making their similarity judgments, a situation which is more comparable to the generic MFCC automatic system.

The listeners' similarity judgments were converted to log-likelihood-ratio type scores using the procedure given in Eq. 1-3:

$$y = \frac{5 - x}{4} \quad (1)$$

$$z = 0.9999(y - 0.5) + 0.5 \quad (2)$$

$$s = \ln\left(\frac{z}{1 - z}\right) \quad (3)$$

where $x$ is the mean of all the listeners' similarity judgments for a given comparison pair. Eq. 1 converts $x$, with a range of 1 to 5 where lower numbers indicate greater similarity, to $y$ with a range of 0 to 1 where higher numbers indicate greater similarity. Eq. 2 shrinks $y$ to $z$ with a range of 0.00005 to 0.99995. This prevents the generation of infinitely valued scores at the next step, Eq. 3, which converts $z$ to a score $s$ that has the form of a log likelihood ratio.

The scores were then converted to likelihood ratios using the same calibration procedure as was applied in the case of the automatic system.

## 3. RESULTS

The validity (accuracy) of each system was assessed using the log-likelihood-ratio cost ($C_{llr}$) [4] and Tippett plots [10] (Descriptions of both of these can be found in [13]). Since we only had one pair of recordings for most speaker-comparison pairs no attempt was made to separately assess the reliability (precision) of each system (see [15, 16]).

The Tippett plots are shown in Figs. 1–3. The humans in the backwards condition had the worst performance, $C_{llr}$ of 0.687, the humans in the forwards condition had better performance, $C_{llr}$ of 0.359, and the automatic system had the best performance, $C_{llr}$ of 0.033 (smaller $C_{llr}$ values indicate greater validity).

## 4. DISCUSSION AND CONCLUSION

The panel of human listeners clearly performed much better when the recordings were played to them forwards than when they were played backwards. This indicates that in the normal forwards condition the human listeners were exploiting phonetic and/or linguistic information, information which was obscured in the backwards condition. In backward speech there are still features such as rate of speech and pausing, which could guide the listeners' judgments of voice-quality similarities.

The automatic system outperformed the human panel of listeners even in the forward condition. In fact the automatic system achieved complete separation. Given the size of this test, however, one should be cautious about generalizing the results and drawing the conclusion that there is nothing to be gained from employing panels of listeners. A possible way to improve the performance of an automatic system could be to fuse its scores with the scores from a panel of human listeners using logistic regression [3, 18], a very basic way of combining phonetic and/or linguistic information with an automatic system. Since the automatic system alone achieved complete separation, it was not possible to test this hypothesis in the present study. The vast improvement in the human listeners' results when they were able to exploit phonetic and/or linguistic information indicates that directly incorporating this type of information into an automatic system might also lead to improvements in performance [20].

**Figure 1:** Tippett plot of test results from human panel of listeners in backwards condition.

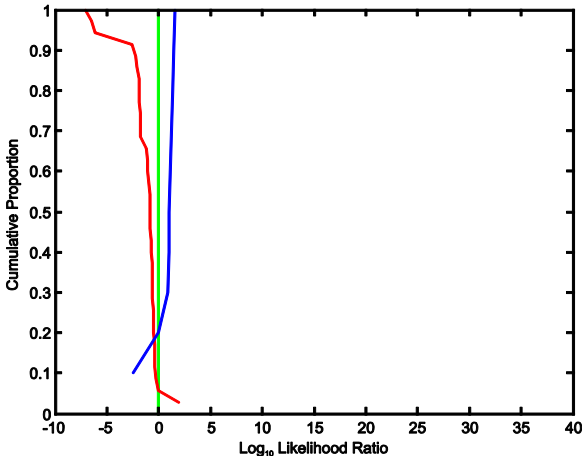



**Figure 2:** Tippett plot of test results from human panel of listeners in forwards condition.

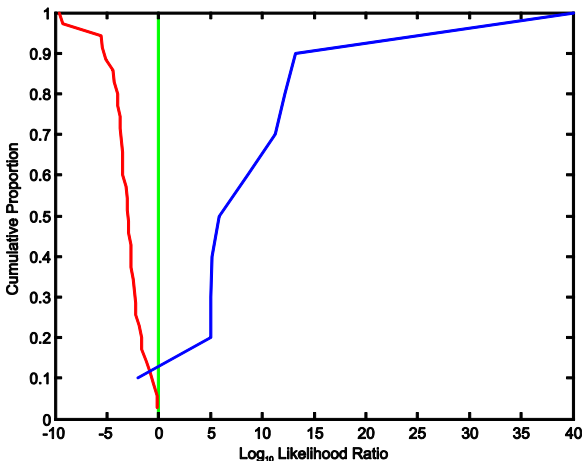
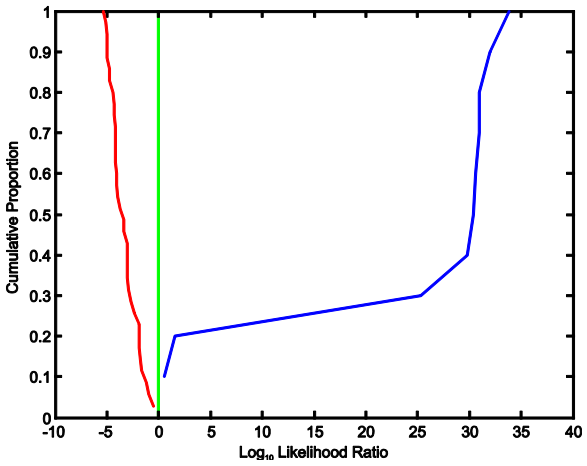



**Figure 3:** Tippett plot of test results from automatic system.



## 5. ACKNOWLEDGMENTS

This research was supported by the Swedish Crime Victim Fund (grant 03347/2007 awarded to Anders Eriksson). We thank the AllEars project, and Anders Eriksson and Lisa Öhman for their assistance with audio data. The AllEars project is funded by the Swedish Crime Victim Compensation and Support Authority.

## 6. REFERENCES

[1] Bonastre, J.F. (project coordinator), 2009. Mistral: Open source platform for biometrics authentification, version 1.3. *http://mistral.univ-avignon.fr/*

[2] Brümmer, N. 2005. Tools for Fusion and Calibration of automatic speaker detection systems. *http://niko.brummer.googlepages.com/focal*

[3] Brümmer, N., Burget, L., Cernocký, J.H., Glembek, O., Grézl, F., Karafiát, M., van Leeuwen, D.A., Matejka, P., Schwarz, P., Strasheim, A. 2007. Fusion of heterogenous speaker recognition systems in the STBU submission for the NIST SRE 2006. *IEEE Trans. Audio, Speech, Lang. Process.* 15, 2072-2084. doi:10.1109/TASL.2007.902870

[4] Brümmer, N., du Preez, J. 2006. Application independent evaluation of speaker detection. Comp. *Speech Lang.* 20, 230-275. doi:10.1016/j.csl.2005.08.001

[5] Eriksson, A. 2004. SweDia 2000: A Swedish dialect database. *Copenhagen Working Papers in LSP* 1, 33-48.

[6] Furui, S. 1986. Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum. *IEEE Trans. Acoust., Speech and Sig. Proc.* 34, 52-59. doi:10.1109/TASSP.1986.1164788

[7] Greenberg, C., Martin, A., Brandschain, L., Campbell, J., Cieri, C., Doddington, G., Godfrey, J. 2010. Human assisted speaker recognition in NIST SRE10. *Proceedings of Odyssey 2010 The Speaker and Language Recognition Workshop* Brno, 180-185.

[8] Hollien, H. 1990. *The Acoustics of Crime: The New Science of Forensic Phonetics*. New York: Plenum.

[9] van Leeuwen, D.A., Brümmer, N. 2007. An introduction to application-independent evaluation of speaker recognition systems. In Müller, C. (ed.), *Speaker Classification I: Fundamentals, Features, and Methods*, Heidelberg, Germany: Springer-Verlag, 330-353. doi:10.1007/978-3-540-74200-5_1

[10] Meuwly, D. 2001. *Reconnaissance de Locuteurs en Sciences Forensiques: l'apport d'une Approche Automatique*. PhD diss., U. Lausanne.

[11] Morrison, G.S. 2009. Forensic voice comparison and the paradigm shift. *Sci. & Justice* 49, 298-308. doi:10.1016/j.scijus.2009.09.002

[12] Morrison, G.S. 2009. Robust version of train_llr_fusion.m from Niko Brümmer's FoCal Toolbox, release 2009-07-02. *http://geoff-morrison.net/*

[13] Morrison, G.S. 2010. Forensic voice comparison. In Freckelton, I., Selby, H. (eds.), *Expert Evidence*. Sydney. Australia: Thomson Reuters, 99.

[14] Morrison, G.S. 2011. A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multvariate kernel density (MVKD) versus Gaussian mixture model – universal background model (GMM-UBM). *Speech Commun.* 53, 242-256. doi:10.1016/j.specom.2010.09.005

[15] Morrison, G.S. In press. Measuring the validity and reliability of forensic likelihood-ratio systems. *Sci. & Justice*. doi:10.1016/j.scijus.2011.03.002

[16] Morrison, G.S., Thiruvaran, T., Epps, J. 2010. Estimating the precision of the likelihood-ratio output of a forensic-voice-comparison system. *Proceedings of Odyssey 2010 The Speaker and Language Recognition Workshop* Brno, 63-70.

[17] Öhman, L., Eriksson, A., Granhag, P.A. 2010. Mobile phone quality vs. direct quality: How the presentation format affects earwitness identification accuracy. *European J. of Psychology Applied to Legal Context* 2, 161-182.

[18] Pigeon, S., Druyts, P., Verlinde, P. 2000. Applying logistic regression to the fusion of the NIST'99 1-speaker submissions. *Digit. Signal Process.* 10, 237-248. doi:10.1006/dspr.1999.0358

[19] Reynolds, D.A., Quatieri, T.F., Dunn, R.B. 2000. Speaker verification using adapted Gaussian mixture models. *Digit. Signal Process.* 10, 19-41. doi:10.1006/dspr.1999.0361

[20] Shriberg, E, Stolke, A. 2008. The case for automatic higher-level features in forensic speaker recognition. *Proceedings of Interspeech* Brisbane, 1509-1512.

[1] It is our view that the NIST SRE10 HASR was not designed in such a way as to facilitate participation by members of the acoustic-phonetic forensic-voice-comparison community, the people with existing expertise in this area, and thus the opportunity to promote meaningful research was missed. The participants were generally the same signal-processing engineering groups as participate in the regular fully-automatic SRE, with no expertise in phonetics or the psychoacoustics of speech/speaker perception.

[2] *http://www.ling.gu.se/~jonas/webb_test/*