Evaluation of Batvox under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01)

David van der Vloed[1], Geoffrey Stewart Morrison[2,3], and Ewald Enzinger[2]
[1]*Netherlands Forensic Institute, The Hague, The Netherlands*
d.van.der.vloed@nfi.minvenj.nl
[2]*Morrison & Enzinger, Forensic Consultants, Vancouver, BC, Canada & Corvallis, OR, USA*
{geoff-morrison|ewald-enzinger}@forensic-evaluation.net
[3]*Department of Linguistics, University of Alberta, Edmonton, AB, Canada*

In order to validate a system intended for use in casework, a forensic laboratory needs to evaluate the degree of validity and reliability of the system under forensically realistic conditions. To contribute to this, Morrison & Enzinger have released a set of training and test data representative of the relevant population and reflecting the conditions of an actual forensic voice comparison case. The evaluation data set is named *forensic_eval_01*. For more information see: http://databases.forensic-voice-comparison.net/#forensic_eval_01

The case involved a substantial mismatch in recording conditions. The speech signal in the offender recording was transmitted via a landline telephone to a call center, there was background office noise (babble and typing noises) at the call center, and the recording was saved in a compressed format. The call included verbal exchange of information, and the duration of speech from the offender was 46 seconds. The suspect recording was of a police interview recorded in a room with substantial reverberation, there was ventilation-system noise, and the recording was saved in a different compressed format. The speaker on each recording was male and spoke English with an Australian accent. The procedures used to select data representative of the relevant population and speaking styles, and the signal processing procedures used to simulate the technical recording conditions of the suspect and offender recordings are described in Enzinger et al (2016). The training data consist of a total of 423 recordings from 105 speakers, and the test data consist of a total of 222 recordings from 61 speakers.

The *forensic_eval_01* data were used to train and test a commercially marketed forensic voice comparison system, Batvox 4.1. Users can optionally enter two sets of case-specific data: "reference population" data, which should reflect the suspect conditions, and "imposter recordings" which should reflect the offender conditions. Both should also represent the relevant population. Another user option is to use all the reference population data, or use a subset which is automatically selected by Batvox. Four variants were tested, constituting a factorial combination of the user options: reference data + imposter data versus reference data only, and all reference data versus a selected subset of 30 recordings. Results in the form of metrics of validity and reliability are shown in Figure 1 (see Morrison, 2011, for explanations of the metrics). Tippett plots are shown in Figure 2.

For this specific data set, better performance resulted when imposter data were used and when all the reference data were used.

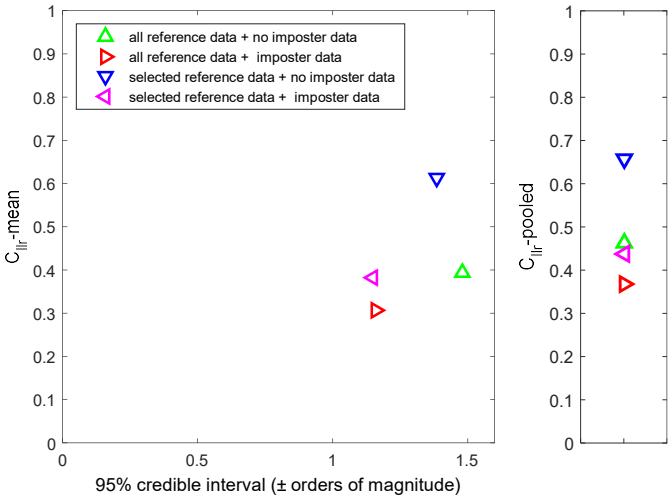

Figure 1. Validity and reliability metrics.

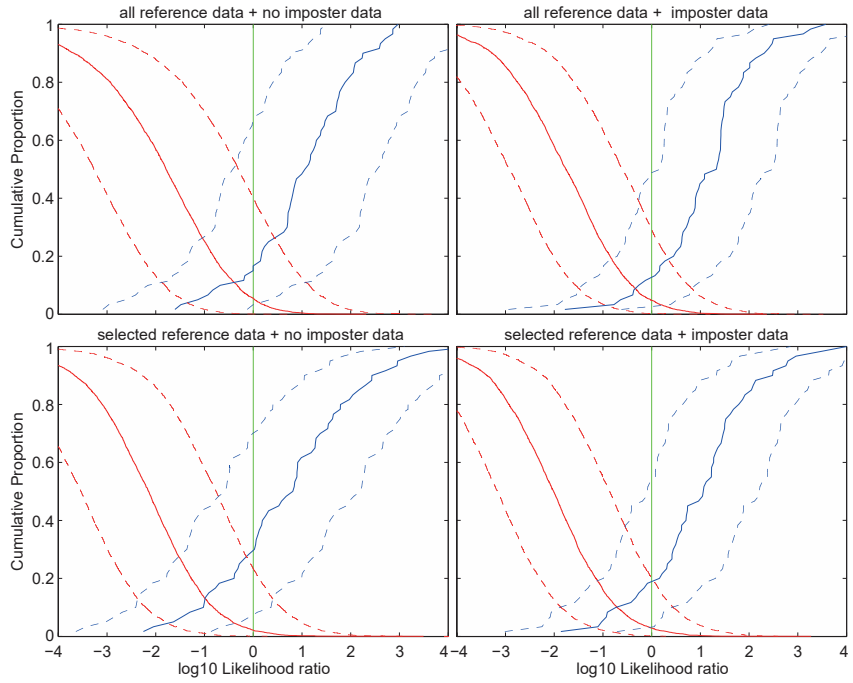

Figure 2. Tippett plots. Results of different source comparisons are in red, same source comparisons are in blue, the dotted lines denote 95% credible interval.

References

Enzinger, E., G.S. Morrison, and F. Ochoa. (2016). A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case. *Science & Justice*, **56**, 42–57. doi:10.1016/j.scijus.2015.06.005

Morrison, G.S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, **51**, 91–98. doi:10.1016/j.scijus.2011.03.002