CON(gruence)-plots for assessing agreement between voice comparison systems

Michael Jessen[1], *Anil Alexander*[2], *Thomas Coy*[2], *Oscar Forth*[2] *and Finnian Kelly*[2]
[1]*Department of Text, Speech and Audio, Bundeskriminalamt, Germany*
michael.jessen@bka.bund.de
[2]*Oxford Wave Research Ltd., Oxford, U.K.*
{anil|tom.coy|oscar|finnian}@oxfordwaveresearch.com

In forensic voice comparison, a practitioner may use multiple systems to compare recordings. Traditional performance measures such as EER and Cllr, or representations such as Tippett plots, can be used to select the 'best' system in the conditions of the case. However, these measures do not inform of the agreement, or congruence, between the different systems for each single comparison; for example, they do not inform about how often System-1 and System-2 both output LRs that support speaker identity. In order to capture such information in an accessible manner, and exploit benefits from multiple systems, a representation called the CON(gruence)-plot has been developed, and implemented as a component of the software BIO-METRICS (https://oxfordwaveresearch.com/products/bio-metrics/).

A CON-plot (Fig. 1) consists of the LR scores output by two voice comparison systems for the same set of comparisons, with same-speaker (H0) and different-speaker (H1) scores clearly differentiated. The plot is divided into four quadrants based on either a log LR value of zero or the Equal Error Rate (EER) score threshold. The relative occupancy of each of the quadrants is indicative of the potential errors introduced by each of the systems; in an ideal scenario, all same-speaker scores would exist in the upper right quadrant (high scores on both methods), and all different-speaker scores in the lower left (low scores on both methods). Entries in the remaining quadrants indicate disagreement, or incongruence, between the two systems, i.e. one of them supporting speaker identity the other non-identity. The level of congruence between systems is also expressed numerically in terms of a (Spearman) rank correlation metric.

CON-plots were applied here to a scenario in which System-1 is high in speaker discrimination but limited in explainability, whereas System-2 is lower in discrimination but higher in explainability (based on phonetic theory). Specifically, an automatic speaker recognition system using x-vector technology was used as System-1 and a semiautomatic system based on long-term formant analysis (LTF) was used as System-2. These systems were applied to a test set called GFS (German Forensic Speech; Solewicz et al. 2017). Further details about the x-vector system applied to this set are presented in Klug et al. (2021) and the LTF system in Jessen (2021).

The resulting CON-plot is presented in Fig. 1. As shown, the level of correlation between the two systems is relatively high. This is expected because both the MFCC features (Mel Frequency Cepstral Coefficients) of the x-vector system and the formant frequencies of the LTF system are strongly or entirely influenced by vocal tract shape. It can be seen that the x-vector system performs much better in terms of speaker discrimination than the LTF system. For example, if LogLR=0 is used as a decision threshold, there are many more false acceptances for the LTF system (red dots above line y=0) than the x-vector system (red dots right of line x=0). The exact performance indices are: x-vector EER 3.0%, Cllr 0.13; LTF EER 17.3%, Cllr 0.66; fused EER 3.3%, Cllr 0.14.

Interestingly, despite the performance difference between the systems in terms of EER and Cllr, there is a high level of congruence among the same-speaker comparisons. In only 2 of 23 comparisons is there disagreement, and in 20 comparisons both systems correctly support speaker identity (in one they both incorrectly support nonidentity, but barely). This prompts the idea of an “explainability-enhanced” mode, where LRs are taken from the more discriminant system, and any incongruent results are declared inconclusive. This and further casework implications will be discussed at the presentation.

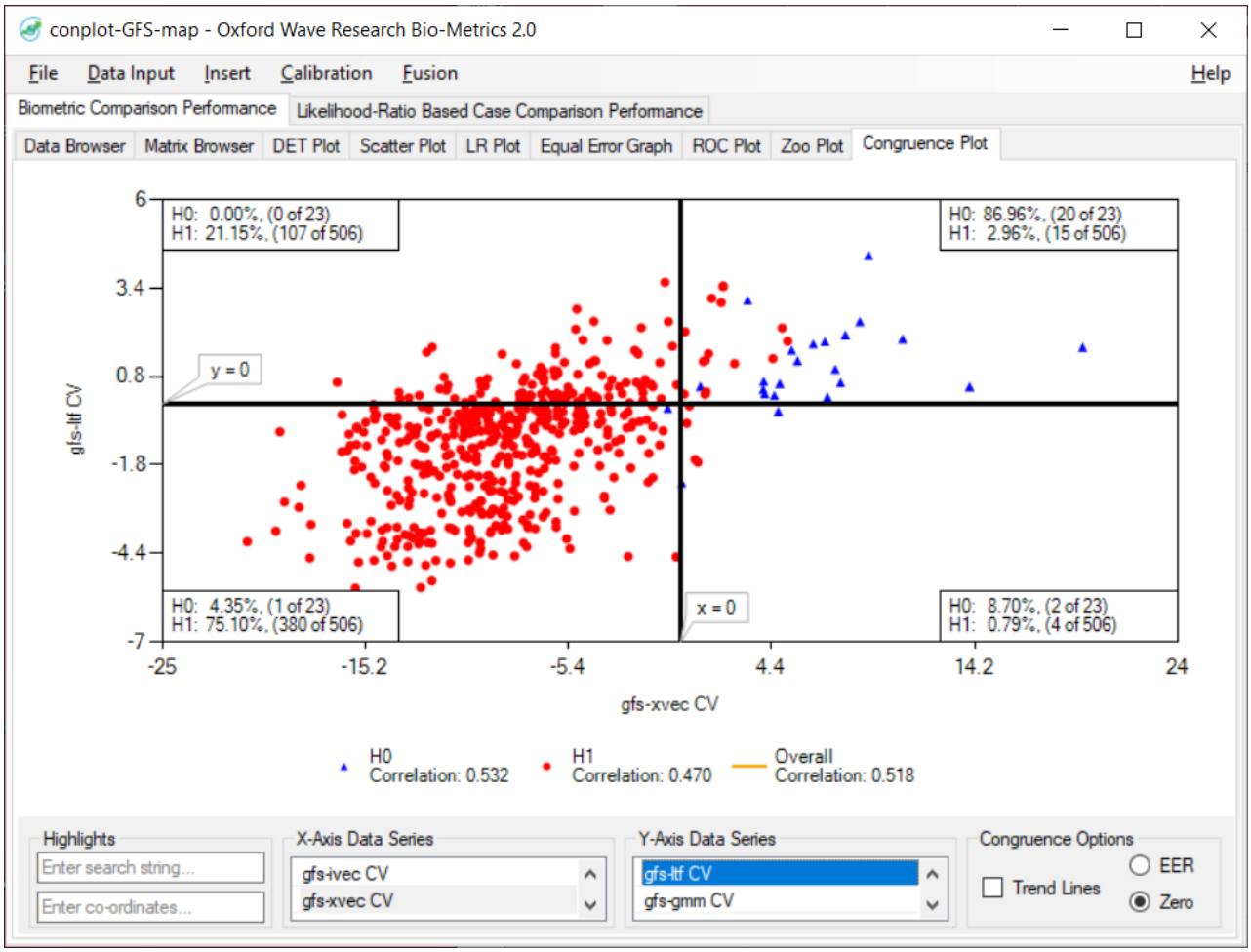

Figure 1. CON-plot for comparison of an x-vector system (X-axis) with an LTF system (Y-axis). Each point on the plot represents the output scores (in terms of Log_eLR after logistic regression cross-validation calibration) of both systems for a single comparison. The blue triangles represent the 23 same-speaker comparisons (H0) and the red dots the 506 different-speaker comparisons (H1). The horizontal and vertical lines represent LogLR=0. Each quadrant shows the number of H0 and H1 comparisons it contains, both in absolute terms and as a percentage of the total number of H0 and H1 comparisons. Also shown in the plot are the Spearman's rank correlation coefficients for H0 and H1 comparisons, and for all comparisons. Further features of the CON-plot will be explained in the presentation.

References

Jessen, M. (2021). MAP adaptation characteristics in forensic long-term formant analysis. In *Proc. Interspeech*, Brno, 411–415.

Klug, K., Jessen, M., Solewicz, Y.A. & Wagner, I. (2021). Collection and analysis of multi-condition audio recordings for forensic automatic speaker recognition. In C. Bernardasci et al. (Eds.), *Speaker individuality in phonetics and speech sciences: Speech technology and forensic applications* (pp. 57–76). Publ. by Associazione Italiana Scienze della Voce, Series Studi AISV, Vol. 8.

Solewicz, Y.A., Jessen, M. & van der Vloed, D. (2017). Null-Hypothesis LLR: A proposal for forensic automatic speaker recognition. In *Proc. Interspeech*, Stockholm, 2849–2853.