University of Zurich[UZH]

PHONEXIA

Forensisches Institut Zürich
Eine Organisation der Kantonspolizei und Stadtpolizei Zürich

Testing the Newest Phonexia Speaker Recognition System With Forensic Eval 01 and Further Corpora

Andrea Fröhlich

Zurich Forensic Science Institute

Department of Computational Linguistics UZH

Introduction

The performance of automatic speaker comparison software has increased dramatically with the introduction of artificial neural networks. To include the automatic approach in a forensic phonetic analysis, forensic practitioners need to be aware of the system's performance and potential pitfalls. As a continuation of previous testing efforts performed and described by Enzinger & Morrison (2016, 2019) and Jessen (2019), Phonexia's newest neural network models were tested and results are described hereafter. A real-case dataset, designed specifically for system testing, was used (Morrison & Enzinger 2016) and widely accepted performance metrics such as EER and various types of Cllr's were obtained (described in (Morrison & Enzinger 2016).

Methods

The system test followed the testing guidelines set by Morrison & Enzinger (2016). The dataset contains files with two different recording conditions that vary in their quality and content (details provided in Morrison & Enzinger (2016)). Additional training data was further provided alongside the test dataset, which was used in the current test to further enhance the system. MFCC's were calculated as features and embedded with deep neural networks into so-called x-vectors (Snyder et al. 2018) for comparison. The tested models are similar to the SID-Beta4 version as described in detail in Jessen et al. (2019). Two enhancement methods were used, namely Mean Normalization and FAR (False Acceptance Rate)–Calibration. The calculation of the Cllr's and EER's was done in Matlab using a script provided by Geoffrey Morrison.

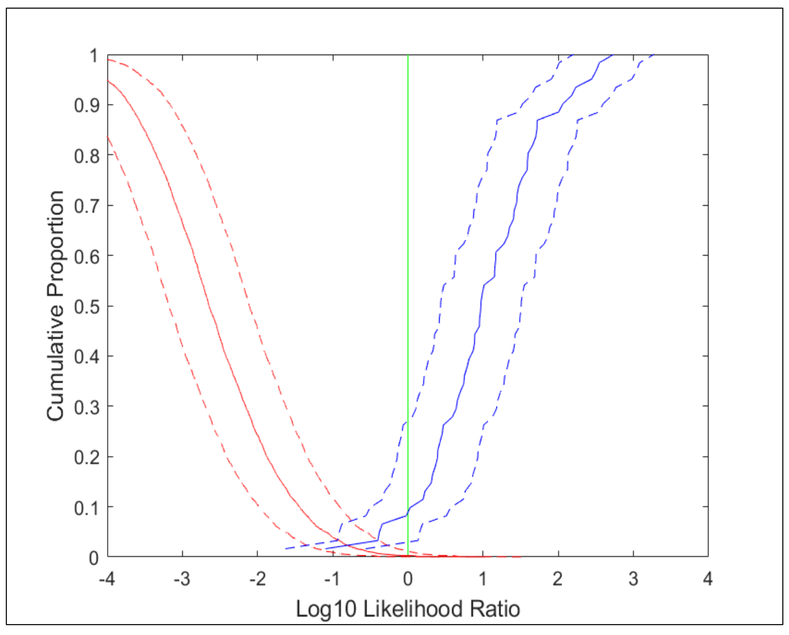

Figure 1: Tippett plot (with precision) xl4 (baseline)

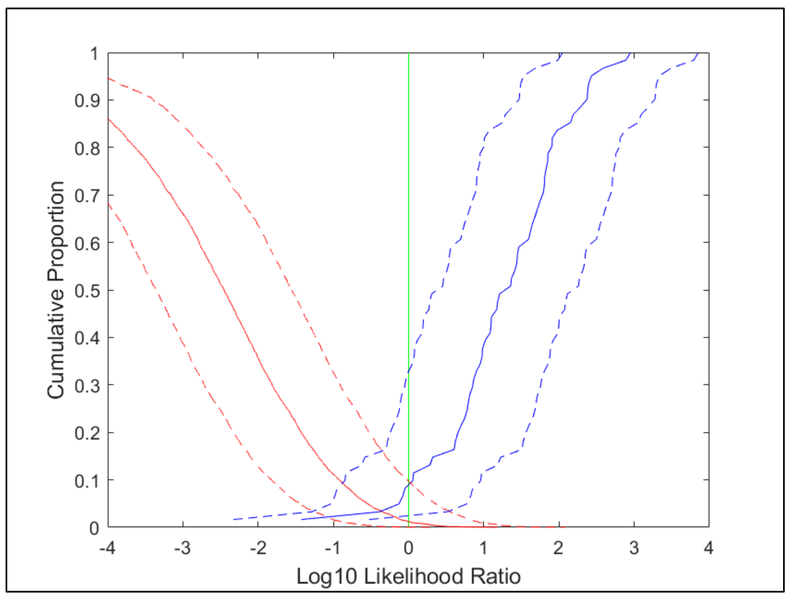

Figure 2: Tippett plot (with precision) l4 (baseline)

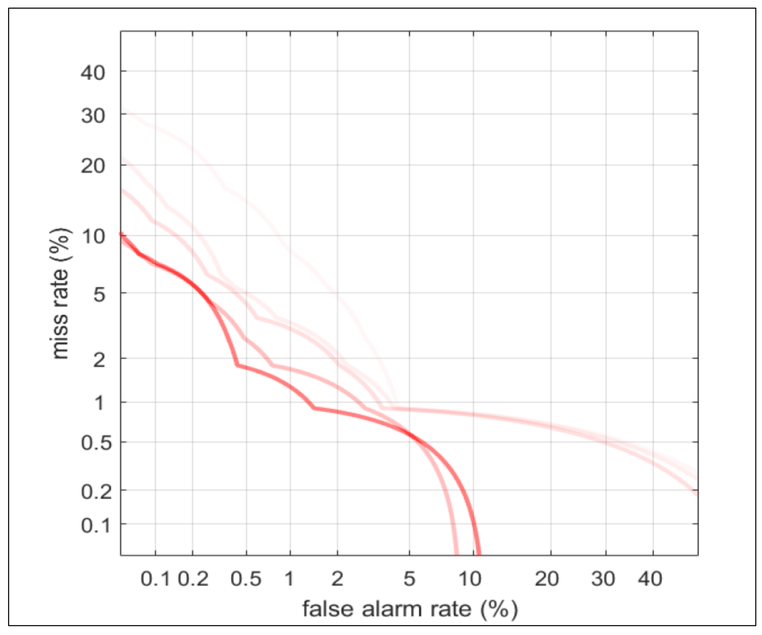

Figure 3: DET plots for all models (model xl4 (mean normalised) shows the best results

Results

	$Cllr_{pooled}$	$Cllr_{mean}$	95% CI	$Cllr_{min}$	$Cllr_{cal}$	EER
Xl4 Baseline	0.219	0.184	0.535	0.085	0.134	0.021
Xl4 10 % FAR	0.184	0.149	0.559	0.0784	0.106	0.019
Xl4 10% FAR + Mean Normalisation	0.127	0.107	0.686	0.050	0.077	0.015
Xl4 Mean Normalisation	**0.089**	**0.055**	**0.0659**	**0.045**	**0.043**	**0.012**
L4 Baseline	0.194	0.179	0.902	0.111	0.083	0.028

Conclusion

In comparison to the older versions of Phonexia (tested in Jessen 2019), the system performance has increased. Comparing the obtained test results to previous evaluations of other systems (Morrison & Enzinger, 2019) shows that Phonexia's newest Xl4 model is scoring the best, respectively is showing the lowest error rates and log likelihood ratios. However, further testing should be conducted, e.g. investigating the performance on low-resource languages such as Swiss-German.

Further Work & Outlook

- Further testing should be conducted, e.g. investigating the performance of the systems on low-resource language such as Swiss-German and synthetic voices
  - Preliminary tests with Swiss-German were very promising
- Testing of systems should occur on a regular basis, since the models and hence the performance of the systems change rapidly
- Further testing should be done with speech messages of e.g. WhatsApp and Snapchat

Contact

andrea.froehlich@for-zh.ch

Acknowledgements:

The author wants to thank Daniel Ramos, Niko Brümmer, Geoffrey Stewart Morrison and Lutz Duembgen for providing forensic_eval_01 data and corresponding analysis scripts.

References

Jessen, M., Bortlík, J., Schwarz, P., & Solewicz, Y. A. (2019). Evaluation of Phonexia automatic speaker recognition software under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01). Speech Communication, 111(March), 22–28.

Morrison, G. S., & Enzinger, E. (2016). Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Introduction. Speech Communication, 85, 119–126.

Morrison, G. S., & Enzinger, E. (2019). Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Conclusion. Speech Communication, 112(June), 37–39.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S., (2018): X-vectors: robust DNN-embeddings for speaker recognition. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)