University of Twente Students Journal Of Biometrics and Computer Vision 2017

Language Dependency of Speaker Recognition Systems

Rianne van Dommelen, Dewi Jhanjhan, and Anne Schrader

Abstract—One of many possible biometric identification methods is voice recognition. A voice recording can be compared to a suspect sample to determine whether suspect and perpetrator are the same. Often the necessary materials are not all available in the same language. This aspect called cross-language speaker recognition can make identification much more challenging. It is therefore fundamental to ensure that the systems used, can correctly perform the assessment in cross-language situations. This paper compares two systems used for biometric speaker recognition that both support cross language identification. The first tool is the BatVOX system, made by Agnitio corp. The second system is Nuance Forensics, made by Nuance. Several contradictions concerning language dependence have been seen. In some situations the language match test performs better than the language mismatch test, while this has been seen the other way around as well. In general the Nuance system seems to be slightly better, however nothing can be said about the language dependency of both systems. It is recommended to obtain more data from both languages in order to make a proper comparison.

Index Terms—forensics, speaker recognition, language mismatch, BatVOX, Nuance, likelihood ratio,

1 Introduction

Biometrics refers to establishing the identity of individuals based on their behavioral and biological characteristics. [1] Nowadays biometric systems are used in a multitude of situations like airport identification, access to buildings, and crime scene investigations. Especially in the latter situation not all identifiers are available, so methods like speaker recognition need to be used.

Each individual speaker sounds different because the larynx sizes, vocal tract shapes and other parts of the voice production organs that are different for each person. Speaker recognition uses the vocal characteristics of speakers to deduce information about their identities. In other words, persons are recognized based on their voice. [2]

First, the vocal tract characteristics of a person are modeled by a speaker recognition system. Once the speaker recognition system has established a model and the model has been associated with an individual, new instances of speech may be evaluated. These new instances of speech are used to determine the likelihood of them having been generated by the concerned model in contrast with different observed models. This methodology is used for all speaker recognition applications. [3]

Methods of automatic speaker recognition mainly extract speaker dependent characteristics of 3 types of specimens. These 3 are recordings of a reference population, suspect and perpetrator. In figure 1 the generic processing chain of biometric speaker recognition is given. The speaker recognition model is trained by means of the reference population. A comparative analysis of the extracted features of suspect and perpetrator is performed with the trained model. This gives a statistical distance measure resulting in a similarity score. This score is converted to a likelihood ratio (LR). After calibration the LR can be evaluated to determine if the comparison should be accepted or rejected.

Forensics is an important application of speaker recognition technology. Forensics is the use of science or technology to investigate and establish evidence or facts in the court of law. The process to determine if a specific individual (suspect) is the source of a voice recording (perpetrator), is called forensic speaker recognition (FSR). This process compares unknown recordings of a perpetrator with one or more known suspect voice recordings. The term forensic automatic speaker recognition (FASR) is used when forensic applications are adapted by automatic speaker recognition methods. [1] As described before, a biometric system is trained by means of a reference population, also called a training set. When this biometric system is trained properly, a test can be performed to check whether suspect and perpetrator are the same. However the language spoken in the reference population is not always the same as the language of the suspect and perpetrator. Due to this language mismatch, the test performance may suffer. Therefore, the NFI is interested in finding out whether the reference population should be in the same language as the perpetrator. Two hypotheses are computed and tested, H_p: the perpetrator and the suspect are from the same speaker and H_d: the perpetrator and the suspect are from 2 different speakers. Thus, the influence of a language match or mismatch in reference population is measured. In this paper a short literature study is performed on the language match and mismatch in the relevant population dataset in forensic automatic speaker recognition evaluation. Furthermore, the performance of 2 speaker recognition methods producing likelihood ratios (BatVOX and Nuance) are compared. There is a language match or mismatch in the relevant population dataset in these experiments. The 2 languages used in the experiments are Dutch and Turkish. Matlab LR toolbox (created by Dr. Rudolf Haraksim) is used to measure and compare the performance of the 2 methods.

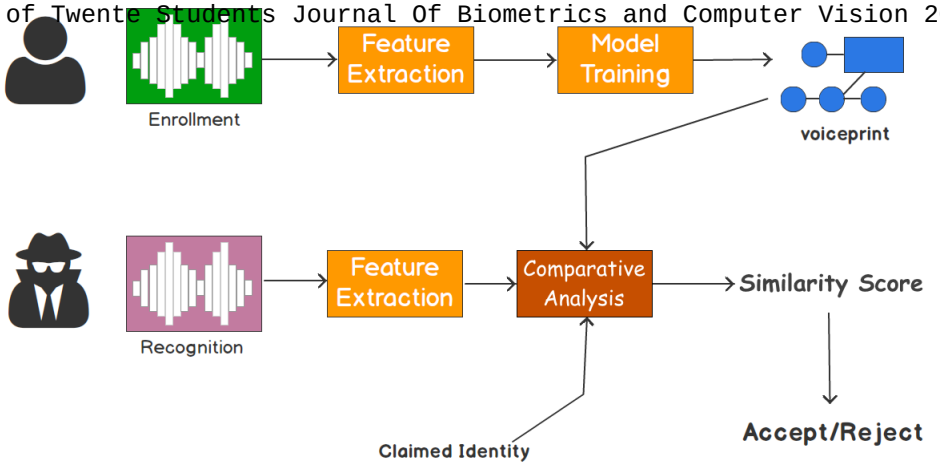


Fig. 1: The processing chain of biometric speaker recognition [12]

2 Literature Study

2.1 Language match/mismatch

In speaker recognition systems differences may occur due to language variations among the group of speakers used as input data. Often a system is originally designed to analyze one specific language and difficulties may occur in using it for other languages. These language variations can be seen in two ways. Firstly, it is possible to have the perpetrators and suspects speak a different language, called language mix. [9] For example, a Dutch speaking suspect is being compared with Turkish speaking perpetrators. In this case, speaker recognition should be performed independently of language differences. Secondly, it is possible to have a different language spoken in the test (suspects and perpetrators) group compared to the reference group by which the system was trained, called cross language effects. [9] An example of this is a system that is trained by using the Dutch language, and is then tested by using the Turkish language. The latter case of language variations is the situation that is analyzed in this paper. In literature several studies on language dependency of speaker recognition systems have been performed. In this short literature study we have focused on the cross language effects only, since this is tested in our own dataset as well. In doing so, some interesting results have been found.

In a small part of the study by van der Vloed et al. [9] the Turkish language was tested against a Dutch training set and vice versa. Both Dutch and Turkish spoken segments were used as test population, with the other language being the reference language. In this study it was seen that the discrimination performance is quite similar compared analyses in which test and reference language were the same. The calibration however performed a lot less. Another study performed by van Leeuwen and Bouten [10] focused on discrimination only, described by the EER. It was found that the discrimination performed less in a language mismatch situation compared to the language match case. Within these cross language effects a lower EER was seen in systems trained by Dutch spoken segments and tested by non-Dutch spoken segments, compared to if the system was trained in non-Dutch and tested by Dutch segments. This means the discrimination seems to be easier in the first situation, but it is not known if this difference is statistically significant. The study of Brmmer et al. [11] is contradictory to the study of van der Vloed et al. described above. In the cross-language analyses between English and non-English, the discrimination is performing a lot less , whilst the calibration is performing quite similar to the language match trials.

These three studies show unclarity in the effects of language matches or mismatches. The second and third study [10], [11] give totally different results than the first one described. [9] However both van der Vloed et al. and van Leeuwen and Bouten point out the low number of cross language trials. Besides the influence of the non-native effect may have influenced the results of these studies. The non-native effect means that most people are not raised bilingually, resulting in the non-native language being spoken with a native accent. When the reference and test populations consist of two different groups of people, the non-native effect might not play a role.

2.2 Nuance/BatVOX

Two types of speaker recognition systems are used in this research to create the Likelihood Ratios (LR). BatVOX and Nuance are both speaker recognition methods for experiments in which there is a language match or mismatch in the relevant population dataset.

BatVOX is according to AGNITIO an expert 1:1 voice biometric tool which is designed to perform speaker verification and compile expert reports that can be used in court [5]. The system provides speaker verifications through the computation of Likelihood Ratios (LR) and is considered text-independent, channel independent and language independent. To perform an identification only 7 seconds of speech are needed [6].

The second system is Nuance Forensics [7]. Nuance is a voice biometric software solution designed to give forensic specialists the ability to match an individuals voice with sound captured through any type of audio channel. The system is not language independent, but supports more than 20 languages, including Dutch and Turkish. [7]

3 METHODS

3.1 Evaluation approach

The used speaker recognition methods are BatVOX and Nuance. The performance of these methods will be compared in several situations where there is a language match or mismatch with the population dataset. To measure the performance of these methods the LR Toolbox by Dr. Rudolf Haraksim is used. To evaluate the quality of the likelihood ratios, performance characteristics will be used. Table 1 shows an overview of the characteristics, their performance metric and the graphical representation.

TABLE 1: Performance characteristics for likelihood ratios. [4]

Performance characteristics	Performance metric	Graphical representation
Accuracy	Cllr	ECE plot
Discrimination power	EER, Cllr-min	ECE-min plot DET plot
Calibration	Cllr-cal	Tippett plot ECE plot
Robustness	Cllr, EER, range of LR values	ECE plot DET plot Tippett plot
Coherence	Cllr, EER	ECE plot DET plot Tippett plot
Generalization	Cllr, EER	ECE plot DET plot Tippett plot

3.2 Parameters

A detection error trade-off (DET) plot represents a trade of between the False Acceptance Rate (FAR) and the False Rejection Rate (FRR). The FAR is plotted against the FRR, with the value where the FAR and FRR are equal being the Equal Error Rate (EER). The closer the plotted curve lies to the origin, the higher the discrimination power, as that means a lower FAR and lower FRR, which in turn means less error. In the same way is the EER a measure of the discrimination power, as a lower EER value implies a higher discrimination power [8].

A Tippett plot shows the complement of the empirical cumulative distribution of the LR values. Basically this can be used to see for how many values of logLR the hypothesis H is true, so how often the log-likelihood-ratio is above the threshold [8].

Lastly there is the log-likelihood-ratio cost (Cllr), which can in turn be split up into the Cllr.min and Cllr.cal. Cllr is a measure of discrimination and calibration, whereas Cllr.min is a measure of only discrimination and Cllr.cal a measure of calibration. For all values is true that the lower the value the better the system [8].

3.3 Toolbox

To evaluate the data the Performance Evaluation Toolbox by Dr. Rudolf Haraksim [4] is used. This toolbox is meant to enable users to measure the quality of their log-likelihood-ratio resulting from their experiments. The toolbox itself is made for MATLAB and can thus does not have to be installed. It can just be put in a folder, which is part of the matlab paths. Running the 'Perf_Ev_Tool.m' will then start up the toolbox. Figure 2 shows the main window of the toolbox, with a normalized histogram already made.

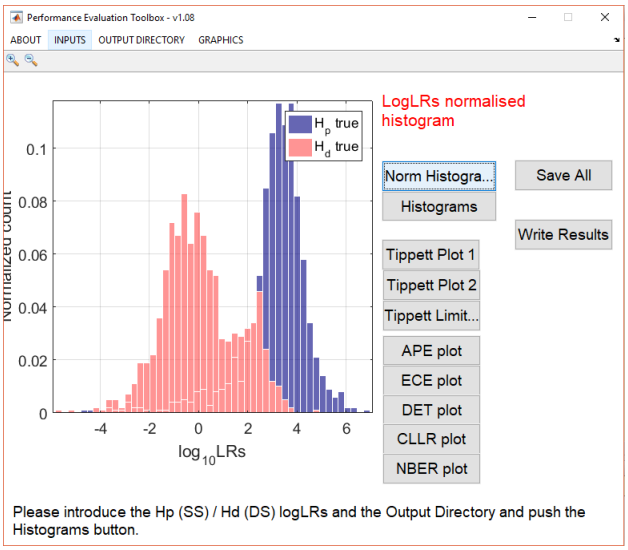


Fig. 2: Main window of Performance Evaluation Toolbox

Data can be inputted with one excel file containing both the ds and ss data. Another option is two text files, where each file contains one source type. After the input and output is decided, it becomes possible to create several figures, including the Tippett and DET plot. A summary of the results is also given, which includes the cllr, cllr.min, cllr.cal and the eer.

4 DATA

Datasets with likelihood ratios are provided by the NFI - FRITS. Each dataset is analyzed by the BatVOX and Nuance system, thus doubling the amount of results.

TABLE 2: Available data from NFI-FRITS

RP Language	Suspect/Perpetrator language	# sets
Dutch	Turkish	4
Turkish	Dutch	3
Turkish	Turkish	1
Dutch	Dutch	11

5 RESULTS

5.1 Cllr plots

In figures 3 and 4 the Cllr plots of both systems are given. The data that was used to create these plots can be found in Appendix B, together with the EER values. In these plots the Cllr.cal (measure of calibration) is plotted against the Cllr.min (measure of discrimination). Ideally, both values are very low (close to zero), however this is hardly seen in practice. From this graph the performance of the speaker recognition test can be extracted, which can be used to balance the discrimination and calibration performances. In the BatVOX system it is seen that the Turkish-Dutch

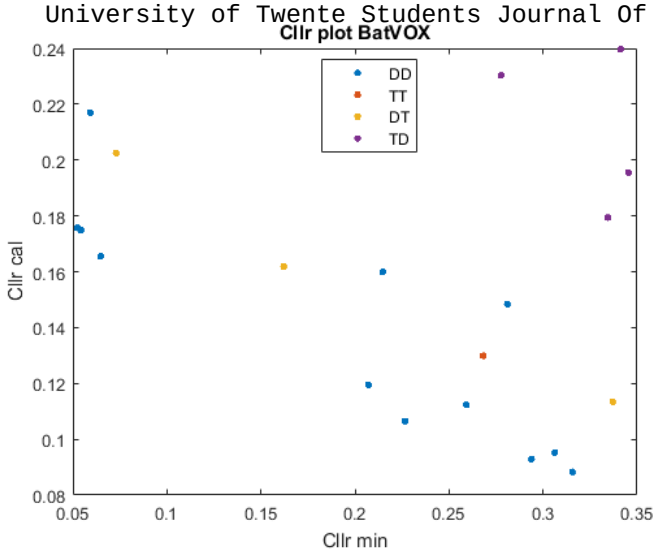


Fig. 3: Cllr spread of the BatVOX system

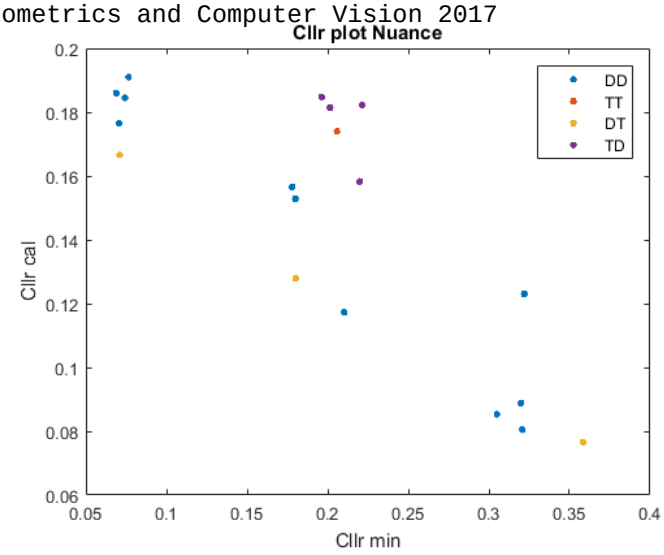


Fig. 4: Cllr spread of the Nuance system

mismatch test has high Cllr values, which means it is performing not that well. Dutch match and Dutch-Turkish mismatch tests are spread a lot across the graph, although some clustering is visible in the Dutch test populations (for both Dutch and Turkish reference populations). The Turkish match test seems to have a somewhat better discrimination than its calibration, but this is based on only a very small dataset. Using the Nuance system again the DD and DT tests are spread, the TD test is performing slightly better, with especially an improved discrimination performance. The TT test is performing worse compared to the BatVOX analysis, especially in the calibration performance.

5.2 EER plots

In figures 5 and 6 the averages of the EER values of each type of language match/mismatch are plotted against the corresponding language match/mismatch type. This is done for both systems, BatVOX and Nuance. In the graph of BatVOX, DD1 has an average EER of 0.019. This is in contrary with the average EER value of about 0.026 for DT1. In the graph of Nuance, the average EER value of DD1 is about 0.022. However, DT1 has a value of about 0.023. A lower EER value implies a higher discrimination power. In Nuance, almost all values of EER for the language match/mismatch are lower than BatVOX, except for DD1.

5.3 Tippett plots

In figures 7, 8, 9 and 10, 4 Tippett plots are given. All plots are made with test group 1 and reference population 4. In these graphs the proportion of cases is plotted against the logarithm of the likelihood ratio (LR).

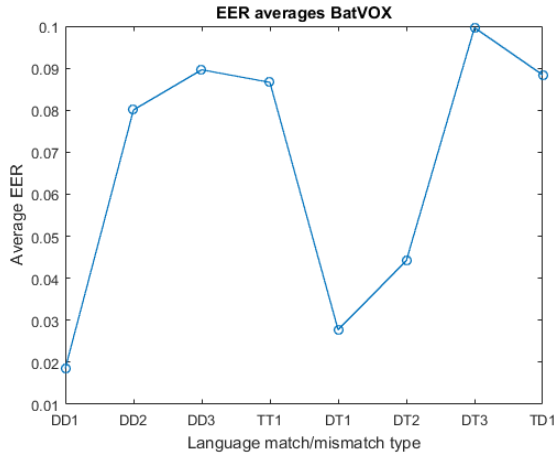


Fig. 5: EER values of the batVox system

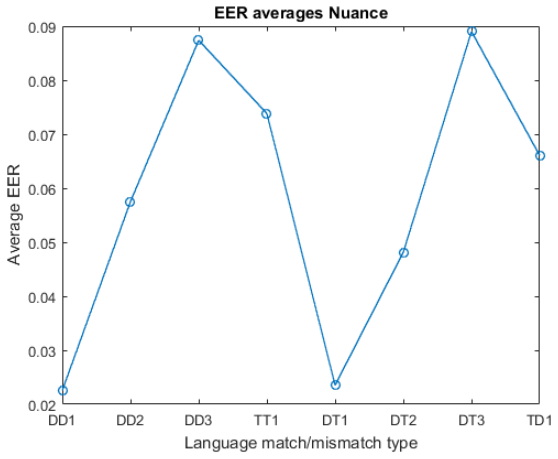


Fig. 6: EER values of the Nuance system

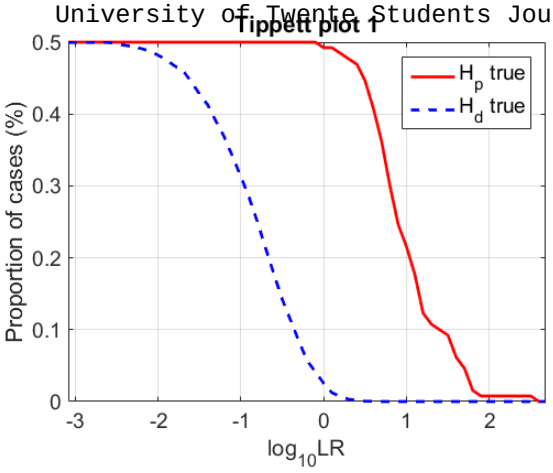


Fig. 7: Dutch-Dutch BatVOX plot

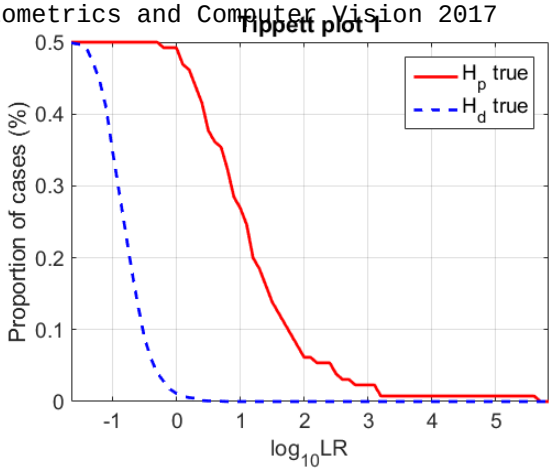


Fig. 8: Dutch-Dutch Nuance plot

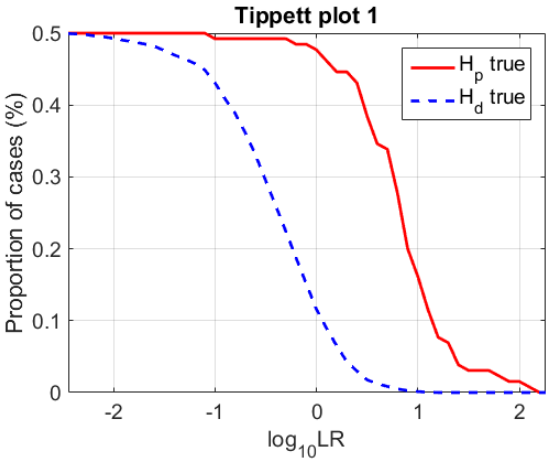


Fig. 9: Turkish-Dutch BatVOX plot

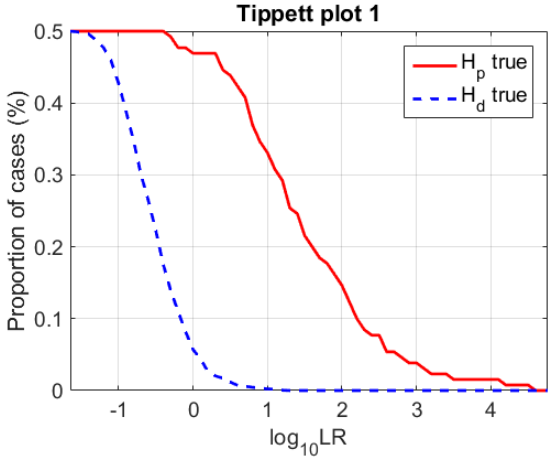


Fig. 10: Turkish-Dutch Nuance plot

In the Tippett plots two hypotheses are visualised; H_p: the suspect and perpetrator are the same, H_d: the suspect and perpetrator are different. In these plots one should take notice of the y-axis, which is wrongly defined due to an error in the toolbox. This should state 0 to 100%.

To evaluate the Tippett plots one should look at the relative amount of false positives and false negatives. These numbers can be found by looking at the intersection between the red/blue lines and the vertical line $log_{10}(LR) = 0$, corresponding to a LR = 1. [13] The amount of false positives is defined by the area under the blue curve on the right hand side of the vertical line, so y-value of intersection. The amount of false negatives is defined by the area above the red curve on the left hand side of the vertical line, so one minus the y-value of the intersection. Because, the y-axis is not correct the evaluation is performed qualitatively only.

The number of false positives and negatives is larger in case of the TD1 compared to DD1 in both BatVOX and Nuance, resulting in a easier speaker discrimination in the language match situation. In comparing the BatVOX system with Nuance in DD1, no significant differences can be found. However when comparing both systems in TD1, a higher number of false positives is seen in BatVOX, while the number of false negatives is slightly lower in BatVOX compared to Nuance. This would suggest speaker discrimination is more difficult with the Nuance system in the situation of language mismatch.

If we look however to the Tippett plots in Appendix A no significant difference in performance between DD and DT can be seen in the Nuance system, while again DD gives better results than DT in the BatVOX system. If we also take into consideration the EER values, we see that this supports the above theory. The EER values are very similiar between the DD and DT systems, but in general the DD values are slightly lower. As said before a lower EER implies better discrimination power, which is in line with the tippet plots. Similar as in the figures above, Nuance generally performs better than BatVOX.

6 Discussion & Conclusion

When looking at the Cllr values not much can be said, as the data is spread out widely and only few results are available per test. What can be said is that the results of Cllr in the Turkish-Dutch test of Nuance are slightly better than that of BatVOX. Furthermore, the graphs of EER are quite similar in shape, but Nuance has in general somewhat lower EER values. Only in the first Dutch match test BatVOX gives a better EER value. Considering the Tippett plots the results

are very contradictory. It differs per test type which system, Nuance or BatVOX, seems to perform better. From these plots it is also difficult to say anything about the language dependency since no consistent results are found. All in all, it is not possible to draw hard conclusions from this study. Too little data is available on the Turkish language, with only one reference population being available. In general the Nuance system seems to perform slightly better than the BatVOX system. Language dependency, however, cannot really be compared between both systems. There is too much variation amongst all populations, even same language populations. Differences between language match and mismatch tests cannot be attributed to language dependencies of speaker recognitions systems based on this analysis. This is due to the fact that there is too much variation between populations of each language. It is therefore recommended to obtain more data of each language being compared. In this way a larger analysis can be performed and conclusions regarding language dependencies may be drawn.

ACKNOWLEDGEMENT

We want to thank Prof. Dr. D. Meuwly for his time and guidance for this paper and David van der Vloed for his time and useful feedback.

REFERENCES

[1] Andrzej Drygajlo *Automatic Speaker Recognition for Forensic Case Assessment and Interpretation* Book: Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism Year: 2012

[2] Tomi Kinnunen, Haizhou Li*An overview of text-independent speaker recognition: From features to supervectors* Book: Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism Year: 2010

[3] Homayoon Beigi *Speaker Recognition, Biometrics* InTech, DOI: 10.5772/17058. Available from: https://www.intechopen.com/books/biometrics/speaker-recognition Year: 2011

[4] Rudolf Haraksim, Daniel Ramos, Didier Meuwly, Andrzej Drygajlo *Performance Evaluation Toolbox for Likelihood Ratio Methods used in Forensic Evidence Evaluation*

[5] Agnitio corp. *BatVOX Product Data Sheet* Available from: http://www.agnitio-corp.com/sites/default/files/BATVOX_Datasheet.pdf Year: 2015

[6] Agnitio corp. *BatVOX: Your Voice Biometrics Partner for Homeland Security* Available from: https://www.pegasus.cl/descargas/BATVOX_Brochure_November_2014.pdf Year: 2014

[7] Nuance Communications. Inc *Nuance Forensics: Public Security for a Safer World* Available from: https://www.nuance.com/content/dam/nuance/en_us/collateral/enterprise/data-sheet/Nuance_PublicSecurity_DS.pdf Year: 2014

[8] Didier Meuwly, Daniel Ramos, Rudolf Haraksim *A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation* Available from: Forensic Science International Year: 2016

[9] David van der Vloed, Jos Bouten and David A. van Leeuwen *NFI-FRITS: A forensic speaker recognition database and some first experiments* Netherlands Forensic Institute, Radboud University Nijmegen Year: 2014

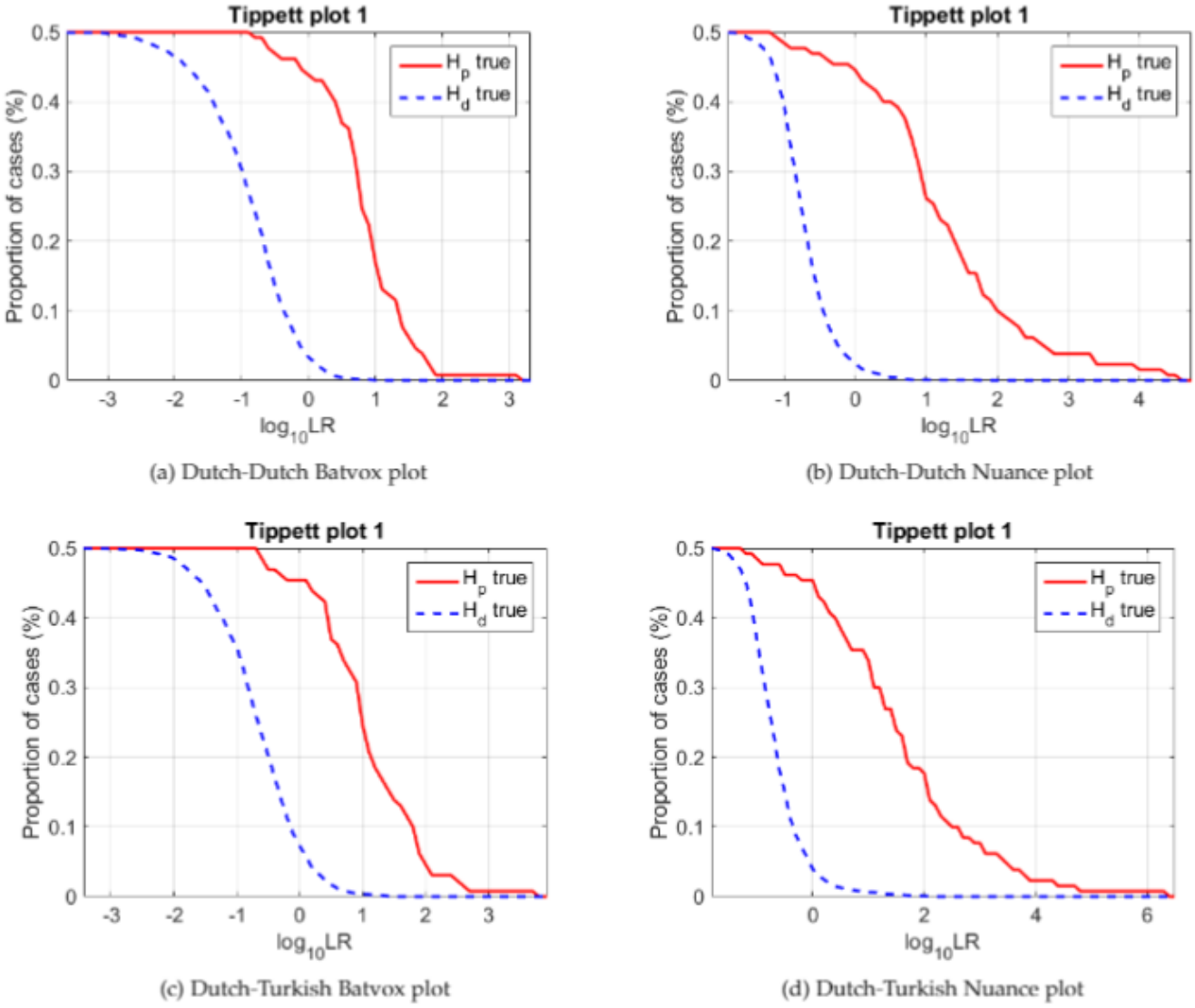
[10] Jos S. Bouten and David A. van Leeuwen *Results of the 2003 NFI-TNO Forensic Speaker Recognition Evaluation* Netherlands Forensic Institute, TNO Human Factors Year: 2004

[11] Niko Brmmer, Luk Burget, Jan "Honza" ernock, Ondej Glembekd, Frantiek Grzl, Martin Karafit, David A. van Leeuwen, Pavel Matjka, Petr Schwarz, Albert Strasheim *Fusion of hetero geneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006* Year: 2006

[12] Ryan Galgon *Now Available: Speaker & Video APIs from Microsoft Project Oxford* Available from: https://blogs.technet.microsoft.com/machinelearning/2015/12/14/now-available-speaker-video-apis-from-microsoft-project-oxford/ Year: 2015

[13] Daniel Ramos, Joaquin Gonzalez-Rodriguez, Grzegorz Zadora, Colin Aitken *Information-Theoretical Assessment of the Performance of Likelihood Ratio Computation Methods* Year: 2013

APPENDIX A
IMAGES



(a) Dutch-Dutch Batvox plot

(b) Dutch-Dutch Nuance plot

(c) Dutch-Turkish Batvox plot

(d) Dutch-Turkish Nuance plot

Fig. 11: Tippett plots made with test group 3 and reference population 1 of both systems.

APPENDIX B
DATA VALUES OF BOTH SYSTEMS

TABLE 3: Cllr and EER values of both systems

(a) Values of the Nuance system

Nuance	cllr	min	eer	cal
DD1	0.2472	0.07066	0.02434	0.1764
	0.2548	0.06877	0.02118	0.186
	0.2587	0.07421	0.02421	0.1845
	0.2674	0.07647	0.02047	0.191
DD2	0.3275	0.2103	0.06289	0.1173
	0.3345	0.1779	0.05096	0.1566
	0.3328	0.1799	0.05873	0.1528
DD3	0.3903	0.3049	0.08516	0.08534
	0.4013	0.3208	0.08730	0.08047
	0.4145	0.3219	0.08801	0.1231
	0.4089	0.32	0.08912	0.08860
TT1	0.3797	0.2057	0.07382	0.1739
DT1	0.2375	0.07095	0.02353	0.1666
DT2	0.308	0.1801	0.04814	0.1279
DT3	0.4304	0.3583	0.08901	0.07660
TD1	0.3781	0.2199	0.07011	0.1583
	0.4036	0.2214	0.07307	0.1822
	0.3828	0.2015	0.06381	0.1814
	0.3811	0.1963	0.05724	0.1848

(b) Values of the BatVOX systems

BatVOX	cllr	min	eer	cal
DD1	0.2303	0.06474	0.01955	0.1655
	0.228	0.05232	0.01694	0.1757
	0.2292	0.0543	0.01885	0.1749
	0.2763	0.05929	0.01871	0.217
DD2	0.3331	0.2266	0.07272	0.1065
	0.3267	0.2072	0.07114	0.1195
	0.3716	0.2592	0.09677	0.1123
	0.3746	0.2147	0.07965	0.1599
DD3	0.3868	0.2939	0.09271	0.09293
	0.404	0.3156	0.08446	0.08833
	0.4045	0.3063	0.09605	0.0952
	0.4294	0.281	0.08504	0.1484
TT1	0.3981	0.2682	0.08663	0.1299
DT1	0.2755	0.07293	0.02774	0.2025
DT2	0.324	0.1621	0.04417	0.1619
DT3	0.4504	0.337	0.09964	0.1134
TD1	0.5079	0.2276	0.07392	0.2303
	0.5809	0.3412	0.08547	0.2397
	0.5138	0.3345	0.09914	0.1794
	0.5409	0.3454	0.09507	0.1955