Effect of Voice Disguise on the Performance of a Forensic Automatic Speaker Recognition System

Hermann J. Künzel, Joaquín Gonzalez-Rodriguez** and Javier Ortega-García***

*Department of Phonetics, University of Marburg, Marburg, Germany
**ATVS, Universidad Politecnica de Madrid, Madrid, Spain
kuenzelh@staff.uni-marburg.de jgonzalez@diac.upm.es

Abstract

This paper presents first results of an ongoing study on the effects of common types of voice disguise, including increased voice pitch (even falsetto speech), lowered voice pitch and pinching the nose while speaking, on forensic speaker recognition (FSR) techniques. Natural and disguised speech data from 100 German speakers recorded 5 times over a period of 7 to 9 months were used in a series of speaker recognition experiments, using the LR-based forensic automatic speaker recognition system developed by ATVS at Universidad Politécnica de Madrid. In this paper, experiments are limited to estimate the performance degradation when the suspect is known to be the author of the disguised test speech (no impostor trials are reported). Results indicate that the three types of voice disguise selected affect the performance of the system only marginally if reference populations contain speech data which exhibit the same type of disguise. If, however, the reference population is assembled with normal speech only, effects are generally more severe and also different for the three types of disguise under evaluation.

1. Introduction

Forensic speaker recognition has long been a subject of scientific and methodological debate. Widely different techniques have been used in the past for the purpose of identifying the author of an anonymous recording in various scenarios, ranging from purely auditory to automatic [1]; see also the excellent review in [2].

Some instance of voice disguise can be observed in a significant portion of offences. In a survey paper of 1975, the author [1] mentioned a general average of 15 % of the cases supplied to the Speaker Recognition Department of the German Federal Police Office. Internal statistics covering a range of more than 20 years reveal, however, that disguise may be found in the majority of cases for certain types of offences, particularly blackmail and kidnapping. Interestingly, electronic devices such as the so-called voice changers have been used extremely rarely at least in Germany (less than a dozen cases) whereas the three modes of disguise used also in this study may be called "classics".

The authors have recently started a research project [3][4] that aims at an improvement of forensic expert reports by combining the strongholds of the classical phonetic-acoustic method with those of a NIST- and NFI-assessed automatic speaker recognition system that has been adapted to the forensic environment by using the Bayesian approach to the analysis of evidence [5][6]. The Bayesian principle was originally applied to speaker recognition in [7] and has been successfully employed with phonetic-acoustic techniques [8] and automatic systems [9][11]. The current ATVS-UPM LR-based system [11] provides a meaningful and valid LR score for every test-speech/suspect-model that can be used directly in an expert report, as the numerical value of the LR has meaning in itself, as it is a factor of reduction of the potential population of sources, and it is directly used in Court measuring the strength of the support to the prosecution hypothesis.

Last fall the system has obtained excellent results in the NFI Forensic Speaker Recognition Evaluation [10][11][13], as it has been the first system to obtain robust and meaningful likelihood ratios for every questioned-speech/suspect-speech pair in a public blind evaluation.

The purpose of the present study is to shed light on the effect of three of the most frequent (on the author experience) modes of voice disguise which may severely handicap FSR - regardless of the method employed. Literature on the forensic implications of disguise is scarce, particularly in connection with automatic identification procedures. In this sense, the present study may be regarded as a complement to a previous phonetic-acoustic study for which a part of the same data base was used [12].

2. The ATVS-UPM FSR System

The ATVS-UPM FSR system is a two layer system:

(i) a NIST-eval type speaker recognition (SR) system, allowing for speech parameterisation, model training and obtention of likelihood scores relating test speech files with trained models.

(ii) a likelihood ratio (LR) estimator, fully compliant with the bayesian framework for the analysis of forensic evidences. This subsystem, making use of the above speaker recognizer for the obtention of "raw" scores, deals with reference populations, within- and between-source pdf estimations, and finally, robust LR estimation.

Note that hundreds/thousands of "raw" scores are computed to obtain each LR score for any speech evidence (test file) as between-source distributions are evidence-dependent and within-source ones are suspect dependent. Then, the LR estimator is in the upper layer fed with lots of required "raw" scores from the lower layer SR system.

The ATVS-UPM SR system is basically a UBM-MAP adapted GMM system, fully implemented in our own classes library (C++), from basic signal processing to ML training or MAP adaptation (no HTK or similar software have been used). Speech is parameterised with 20 ms. window length, 10 ms. overlapping, Hamming windows, 20 mel-spaced (300-3200 kHz) magnitude filters, 38 coefficients per frame (19 MFCC + delta). A 1024 ML UBM is used in the trials and adapted when possible to the conditions (channel, language, sex ...) of the corresponding task. Speaker models are 1024 mixtures GMM, adapted with MAP (only means) from the

1024 UBM described above. Only five gaussians per frame are used for likelihood computations. In these experiments, a simple channel compensation is performed via cepstral mean subtraction, where channel-dependent Z-Norm likelihood normalization is used to obtain the final "raw-score" ("raw" with regard to the Bayesian LR score as shown below).

The ATVS-UPM LR estimator performs the following actions (details in [9][11]):

- Between-source pdf estimation using EM-ML criteria to obtain a 16 mixtures 1D-GMM from scores obtained when the test utterance is compared with the reference population.
- Within-source pdf (single gaussian) estimation in two different cases depending on the available suspect speech:
  - Suspect speech from one recording session: estimation is obtained by bootstrapping training data, fixing the within-source maximum adaptation values and predicting user variability degradation
  - Suspect speech from more than one session: estimation is obtained by bootstrapping training data, estimating the average value of such distribution and predicting user variability degradation.

In both cases, the algorithm/strategy is determined with databases out of the experiment, and specific values for each LR are real-time blindly obtained from case/test data.

- In this paper, results for each trial are presented in the form of log(LR) (Log10 of LR-Scores), which turns the classical 10^{-5} to 10^{5} (higher reduction factors are unlikely with speech evidences , which is not the case of e.g. DNA) in a range from -5 to +5.

With this Bayesian forensic system, ATVS-UPM has recently participated in NFI-TNO evaluation [10], obtaining excellent results as meaningful LR scores were blindly reported for every suspect-speech/questioned-speech pair of the evaluation (see details in [11]), which moreover was prepared with real forensic case data. This means that each one of the thousands of computed LRs in this NFI-eval could directly have been used in Court to support (LR>1) or not (LR<1), and with the LR-value strength, the prosecution hypothesis (the test speech was uttered by the suspect –the author of the model-).

3. The speech data base

Three modes of voice disguise were investigated: increased voice pitch (including the use of falsetto speech), lowered voice pitch, and pinching the nose (blocking the nasal cavity) while speaking. High-quality analogue speech recordings of a set of 100 adult male speakers of German were used. At this stage of the investigation (January 2004), the first 50 were used for Z-normalization and also to build a reference population. The remaining 50 speakers were used as test speakers. All 100 speakers had produced the same faked kidnapper's call in 5 recording sessions at intervals of ca. 6 weeks, so that their speech behaviour was covered over a period of 7 to 9 months. In view of the forensic perspective of the study this period was considered sufficiently long to provide a truly representative account of an individual's vocal behaviour, including potential short-term and long-term imponderables such as diseases, general physical condition, daytime (vocal fatigue!) and others. The duration of the text read at normal speed is 50 - 55 seconds. In each of the 5 recording sessions a speaker had to read the text three times: first, with undisguised voice and then in two of the three proposed modes of voice disguise that he was encouraged to choose in the first session. Thus, a total of 1500 recordings was obtained (100 speakers, 5 sessions, 3 recordings).

4. Experiments

Previous acoustic phonetic experiments with the disguised and undisguised speech material that was also used for the present study have shown [12] that all three modes of disguise involve complex modifications of a speaker's vocal behaviour und thus may cause severe problems to his recognition. Some of the most salient "supplementary" features are a slow-down of speaking tempo, monotonization (reduced variability of Fo), change of voice quality, particularly when register changes are involved, reduction of amplitude, lowering of Fo due to the transglottal air pressure changes induced by blocking the nasal passage (pinched-nose condition). Furthermore, there are sex-related differences in the strategies employed for disguise by men and women[12].

In this paper, using the ATVS-UPM FSI system described above, "anonymous" voices are related to "suspect" models by means of Likelihood Ratios. Within the Bayesian framework for the analysis of evidence the composition of the reference population is an important issue. In this context, regional dialect, age group, and speaking condition are among the most frequently mentioned categories. It may be anticipated, however, that features such as different modes of voice disguise will be at least as important, considering their sometimes dramatic changes of the acoustic structure of speech. In the present study this aspect was investigated by running all tests with two different reference populations, one consisting of the undisguised-speech models of the first 50 speakers of the above-mentioned set, and the other consisting of the 3 subsets of the first 50 speakers who had opted for the respective modes of voice disguise. Thus the following distributions were obtained:

1. Ref. Pop. "normal" (undisguised speech): 50 speakers
2. Ref. Pop. "high" (increased pitch): 27 speakers (out of 50)
3. Ref. Pop. "low" (lowered pitch): 36 speakers (out of 50)
4. Ref. Pop. "nose pinched": 37 speakers (out of 50)

The models for all four reference populations were trained with the (undisguised / disguised) speech samples produced in the first recording session. This means that hitherto only 1/5 of the material has been exploited. Suspect models are obtained with the undisguised section of first recording session.

Furthermore it is evident that if a test sample in a forensic case is actually affected by a certain mode of disguise, if detected by the forensic expert, the Z-norm should also be made with impostor speech samples that exhibit that same disguise. Conversely, if no disguise is involved, the Z-norm is calculated with normal speech. For the purposes of this study it seemed adequate to use data from the speakers of the four reference populations also for these normalizations.

5. Results

At this juncture we will briefly present the first results of this study with known guilty disguiser suspects. Table 1 contains the proportion of LRs obtained for the speakers in the three modes of disguise and corresponding reference populations consisting of (a) undisguised and (b) disguised samples.

In order to ease understanding, the author have grouped the results data in three levels of degradation: no degradation (logLR=5), moderate degradation (logLRs between 5 and 2), and significant degradation (logLR<2).

		Degradation level		
Disguise mode	No. of speakers	Undegraded (logLR > 5)	Moderate (2<logLR<5)	Significant (logLR< 2)
undisguised speech	*50*	*100%*	*0%*	*0%*
(a) Undisguised reference population				
High	44	43%	20%	36%
Low	29	79%	7%	14%
Nose p.	27	22%	7%	70%
(b) Disguised reference population				
High	44	82%	2%	16%
Low	29	93%	3%	3%
Nose p.	27	92%	7%	0%

Table 1. Test results split into 3 degradation levels of LRs and three modes of disguise with undisguised (a) and disguised (b) ref. population with the same mode of disguise as test sample.

The first row contains the results of control tests made for every speaker with his own undisguised speech samples used for testing, model training and speaker controls, as well as undisguised speech material for both reference population and Z-normalization.

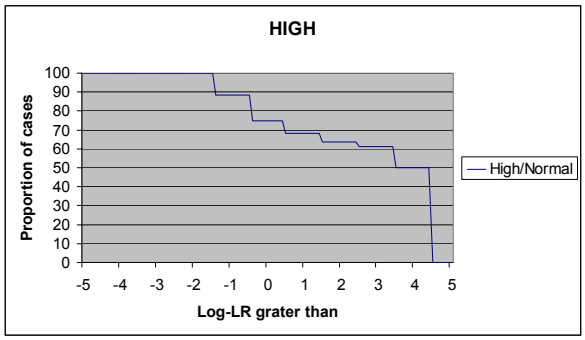

Figure 1a. Distribution of Log LRs of 44 sorted speakers with disguise mode "high". Ref.Pop.: normal speech

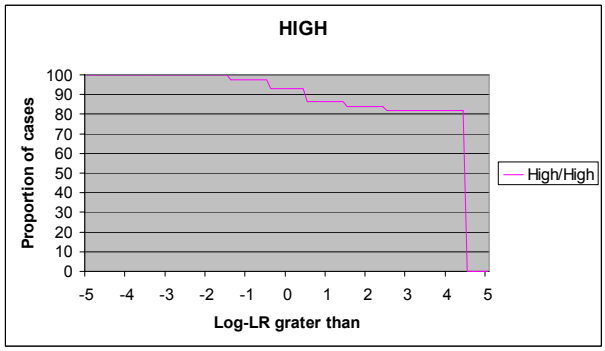

Figure 1b. Distribution of Log LRs of 44 speakers with disguise mode "high". Ref.Pop.: disguised speech .

Under these optimal circumstances the maximum level of logLR was obtained for all 50 speakers. If we regard this result as a - rather theoretical - benchmark for tests involving disguised speech, some conclusions may be drawn from the results:

1. The performance of the system is degraded by all three modes of disguise, where the smallest degradation is present for the mode “lowered pitch”.

2. The size of the effect depends to a large degree on the fact whether or not the reference population accounts for the particular disguise mode. If it does not, significant degradation is present with 36%, 14% and 70% of the speakers for the disguise modes "high", "low" and "pinched-nose", respectively. If, however, the particular reference population consists of models trained with the same type of disguised speech by which the test sample is affected, significant degradation is then present just with 16% and 3% of the speakers for the disguise modes "high" and "low". Most interestingly, no degradation at all evolves for the pinched-nose mode of disguise, although it poses by far the most serious problem to the system if the reference population consists only of undisguised speech.

Figures 1 to 3 illustrate these findings. LogLRs for the three modes of disguise are shown in ascending order. Parts a and b of each figure contain the results on the basis of undisguised and disguised reference populations.

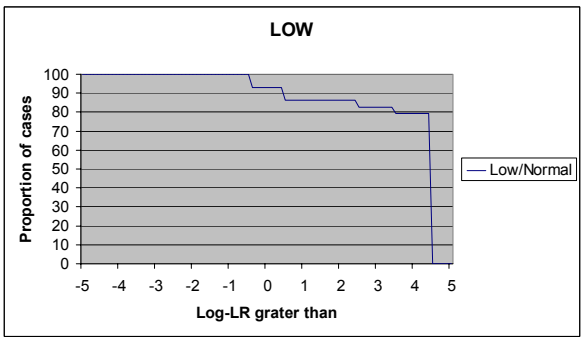

Figure 2a. Distribution of Log LRs of 29 speakers with disguise mode "low". Ref.Pop.: normal speech.

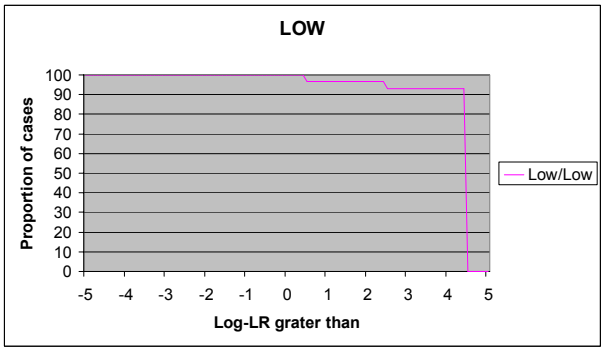

Figure 2b. Distribution of Log LRs of 29 speakers with disguise mode "low". Ref.Pop.: disguised speech.

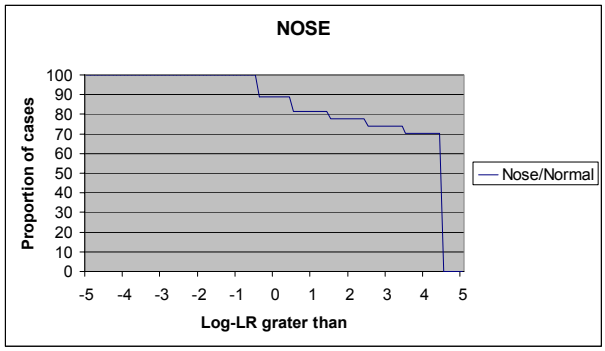

Figure 3a. Distribution of Log LRs of 27 speakers with disguise mode "pinched-nose". Ref.Pop.: normal speech.

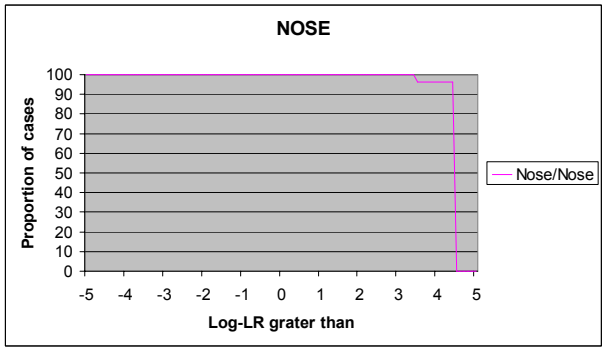

Figure 3b. Distribution of Log LRs of 27 speakers with disguise mode "pinched-nose". Ref.Pop.: disguised speech.

6. Discussion

Automatic systems for forensic speaker recognition have become of age. However, before systems can be used for forensic fieldwork and ultimately for expert reports - preferably together with "conventional" phonetic-acoustic methods - their performance under forensic real-world conditions has to be tested carefully. One of the most detrimental features of the real world is the occurrence of voice disguise. In this context it is extremely important to decide in every single case whether or not an automatic FSR system is so robust towards the type of disguise involved that it may still be applied. The present study has given a first assessment of the influence of three frequent kinds of voice disguise on the performance of a particular system but many more experiments are needed. First, non-target (impostor) trials have to be performed, even though no degradation is expected because of voice disguise. The next stages of this project will be dedicated to an increase of the data base. First, swapping the two sets of 50 speakers used for reference population and LR-tests, respectively, will double the number of individual tests. Second, the remaining disguised speech samples of each speaker (from recordings sessions 2 - 5) will be used as tests samples, which will again increase the amount of data by a factor of four.

As soon as this enlarged material basis is available, the most demanding phonetic task will follow, i.e. an analysis of the phonetic features of the different modes of disguise and their implications on the automatic system. As an example, auditory analysis during the preparation of the speech files for the tests has revealed that almost all of those speakers who had chosen "high" as a mode of disguise and were not properly recognized by the system had not just increased their fundamental frequency more or less, but changed the voice register from modal to falsetto. From a phonetic and physiological point of view it goes without saying that such a drastic alteration of the vocal apparatus as a whole will also affect the resonance characteristics of the vocal cavities which in turn are the basis for the extraction of the MFCCs used by the automatic system. Knowledge about these interrelations is crucial for the understanding of the possibilities and limitations of automatic FSR systems in these conditions.

7. Acknowledgement

Hermann J. Künzel wants to thank Marta Garcia-Gomar for her advice and tutorial on the use of the system.

8. References

[1] Künzel, H.J., "Current Approaches to Forensic Speaker Recognition", Tutorial paper, Proc. of ESCA workshop on Automatic Speaker Recognition, pp. 135-141, Martigny (Switzerland), April 1994.

[2] Meuwly, D., *Reconnaissance de Locuteurs en Sciences Forensiques: L'apport dúne Approche Automatique*, PhD Thesis, IPSC-Université de Lausanne, 2001 (available at http://homepage.mac.com/dmeuwly/ Publications/)

[3] Künzel, H.J. and Gonzalez-Rodriguez, J., "Combining automatic and phonetic-acoustic speaker recognition techniques for forensic applications", Proceedings 15th Intern. Congress of Phonetic Sciences (ICPhS), Barcelona (Spain), 2003.

[4] Künzel, H.J. and Gonzalez-Rodriguez, J., "Testing IdentiVox with German Speech data", Proc. of International Association of Forensic Phonetics (IAFP) 2002 Annual Meeting, Moscow (Russia), July 2002.

[5] Evett, I.W., "Towards a Uniform Framework for Reporting Opinions in Forensic Science Casework", Science & Justice 1998: 38(3), pp. 198-202.

[6] Aitken, C.G.C., "Statistical Interpretation of Evidence /Bayesian Analysis", Encyclopedia of Forensic Sciences, pp. 717-724, Academic Press, 2000.

[7] Champod, C. and Meuwly, D., "The Inference of Identity in Forensic Speaker Recognition", Speech Comm., 31, 193-203, 2000.

[8] Rose, P., *Forensic Speaker Identification*, Taylor & Francis Forensic Science Series, London, 2002.

[9] J. Gonzalez-Rodriguez, D. Garcia-Romero, M. Garcia-Gomar, D. Ramos-Castro and J. Ortega-Garcia, "Robust Likelihood Ratio Estimation in Bayesian Forensic Speaker Recognition", Proceedings of EuroSpeech 2003, pp. 693-696, Geneva (Suitzerland), September 2003.

[10] Bouten, J.S. and van Leeuwen, D.A., "Results of the 2003 NFI-TNO Forensic Speaker Recognition Evaluation", Proceedings of Odyssey 2004.

[11] J. Gonzalez-Rodriguez, D. Ramos-Castro, M. Garcia-Gomar, and J. Ortega-Garcia, "On Estimation of Robust Likelihood Ratios: the ATVS-UPM system at 2003 NFI/TNO Forensic Evaluation", Proceedings of Odyssey 2004.

[12] Künzel, H.J, Effects of voice disguise on speaking fundamental frequency, *Forensic Linguistics 7(2),*149-179, 2000.

[13] http://speech.tm.tno.nl/aso/