

# **A first step towards a text-independent speaker verification Praat plug-in using Mistral/Alize tools**

*Jonas Lindh*

*Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg*

## **Abstract**

*Text-independent speaker verification can be a useful tool as a substitute for passwords or increased security check. The tool can also be used in forensic phonetic casework. A text-independent speaker verification Praat plug-in was created using tools from the open source Mistral/Alize toolkit. A gate keeper setup was created for 13 department employees and tested for verification. 2 different universal background models were trained and the same set tested and evaluated. The results show promising results and give implications for the usefulness of such a tool in research on voice quality.*

## **Introduction**

Automatic methods are increasingly being used in forensic phonetic casework, but most often in combination with aural/acoustic methods. It is therefore important to get a better understanding of how the two systems compare. For several studies on voice quality judgement, but also as a tool for visualisation and demonstration, a text-independent speaker comparison was implemented as a plugin to the phonetic analysis program Praat (Boersma & Weenink, 2009). The purpose of this study was to make an as easy to use implementation as possible so that people with phonetic knowledge could use the system to demonstrate the technique or perform research. A state-of-art technique, the so called GMM-UBM (Reynolds, 2000), was applied with tools from the open source toolkit Mistral (former Alize) (Bonastre et al., 2005; 2008). This paper describes the surface of the implementation and the tools used without any deeper analysis to get an overview. A small test was then made on high quality recordings to see what difference the possession of training data for the universal background model makes. The results show that for demonstration purposes a very simple world model including the speakers you have trained as targets is sufficient. However, for research purposes a larger world model should be trained to be able to show more correct scores.

## **Mistral (Alize), an open source toolkit for building a text-independent speaker comparison system**

The NIST speaker recognition evaluation campaign started already 1996 with the purpose of driving the technology of text-independent speaker recognition forward as well as test the performance of the state-of-the-art approach and to discover the most promising algorithms and new technological advances (from <http://www.nist.gov/speech/tests/sre/> Jan 12, 2009). The aim is to have an evaluation at least every second year and some tools are provided to facilitate the presentation of the results and handling the data (Martin and Przybocki, 1999). A few labs have been evaluating their developments since the very start with increasing performances over the years. These labs generally have always performed best in the evaluation. However, an evaluation is a rather tedious task for a single lab and the question of some kind of coordination came up. This coordination could be just to share information, system scores or other to be able to improve the results. On the other hand, the more natural choice to be able to share and interpret results is open source. On the basis of this Mistral and more specifically the ALIZE SpkDet packages were developed and released as open source software under a so-called LGPL licence (Bonastre et al., 2005; 2008).

## **Method**

A standard setup was made for placing data within the plugin. On the top of the tree structure several scripts controlling executable binaries, configuration files, data etc. were created with basic button interfaces that show up in a given Praat configuration. The scripts were made according to the different necessary steps that have to be covered to create a test environment.

## Steps for a fully functional text-independent system in Praat

First of all some kind of parameterization has to be made of the recordings at hand. In this first implementation SPro (Guillaume, 2004) was chosen for parameter extraction as there was already support for this implemented in the Mistral programs. There are 2 ways to extract parameters, either you choose a folder with audio files (preferably wave format, however other formats are supported) or you record a sound in Praat directly. If the recording is supposed to be a user of the system (or a target) a scroll list with a first option "New User" can be chosen. This function will control the sampling frequency and resample if sample frequency is other than 16 kHz (currently default), perform a frame selection by excluding silent frames longer than 100 ms before 19 MFCCs are extracted and stored in parameter file. The parameters are then automatically energy normalized before storage. The name of the user is then also stored in a list of users for the system. If you want to add more users you go through the same procedure again. When you are done you can choose the next option in the scroll list called "Train Users". This procedure will control the list of users and then normalize and train the users using a background model (UBM) trained using Maximum Likelihood Criterion. The individual models are trained to maximise the a posteriori probability that the claimed identity is the true identity given the data (MAP training). This procedure requires that you already have a trained UBM. However, if you do not, you can choose the function "Train World" which will take your list of users (if you have not added others to be included in the world model solely) and train one with the default of 512 Gaussian mixture models (GMM). The last option on the scroll list is instead "Recognise User" which will test the recording against all the models trained by the system. A list of raw (not normalised) log likelihood ratio scores gives you feedback on how well the recording fitted any of the models. In a commercial or fully-fledged verification system you would also have to test and decide on threshold, as that is not the main purpose here we are only going to speculate on possible use of threshold for this demo system.

### Preliminary UBM performance test

To get first impression how well the implementation worked a small pilot study was made

using 2 different world models. For this purpose 13 colleagues (4 female and 9 males) at the department of linguistics were recorded using a headset microphone. To enroll them as users they had to read a short passage from a well known text (a comic about a boy ending up with his head in the mud). The recordings from the reading task were between 25-30 seconds. 3 of the speakers were later recorded to test the system using the same kind of headset. 1 male and 1 female speaker was then also recorded to be used as impostors. For the test utterances the subjects were told to produce an utterance close to "Hej, jag heter X, jag skulle vilja komma in, ett två tre fyra fem." ("Hi, I am X, I would like to enter, one two three four five."). The tests were run twice. In the first test only the enrolled speakers were used as UBM. In the second the UBM was trained on excerpts from interviews with 109 young male speakers from the Swedia dialect database (Eriksson, 2004). The enrolled speakers were not included in the second world model.

## Results and discussion

At the enrollment of speakers some mistakes in the original scripts were discovered such as how to handle clipping in recordings as well as feedback to the user while training models. The scripts were updated to take care of that and afterwards enrollment was done without problems. In the first test only the intended target speakers were used to train a UBM before they were enrolled.

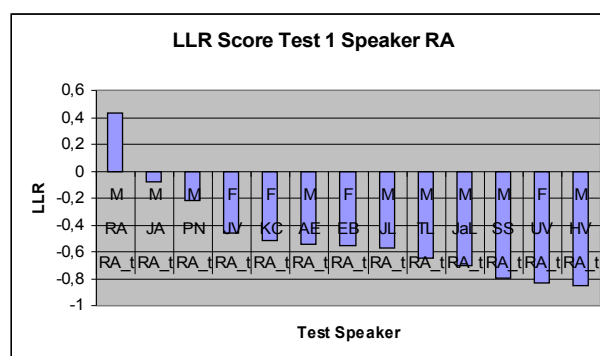


Figure 1. Result for test 1 speaker RA against all enrolled models. Row 1 shows male (M) or female (F) model, row 2 model name and row 3 the test speaker.

In Figure 1 we can observe that the speaker is correctly accepted with the only positive LLR (0.44). The closest following is then the model of speaker JA (-0.08).

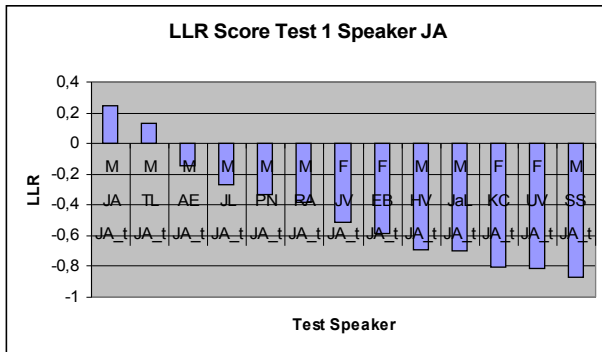


Figure 2. Result for test 1 speaker JA against all enrolled models.

In the 2<sup>nd</sup> test there is a lower acceptance score (0.25) for the correct model. However, the closest model (TL) also has a positive LLR (0.13).

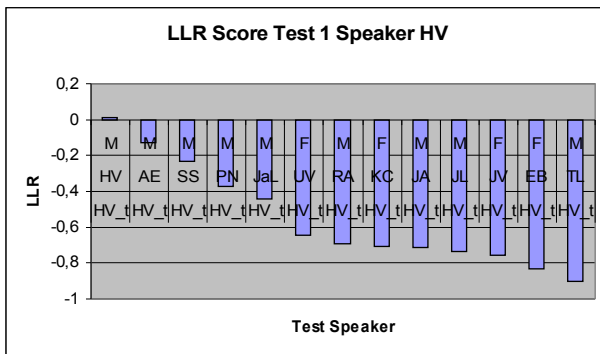


Figure 3. Result for test 1 speaker HV against all enrolled models.

In the 3<sup>rd</sup> test the correct model is highest ranked again, however, the LLR (0.009) is low.

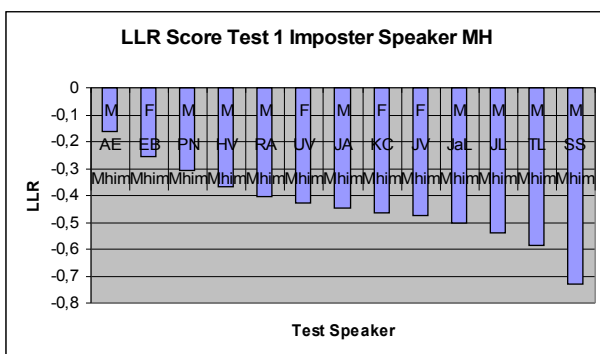


Figure 4. Result for test 1 imposter speaker MH against all enrolled models.

The 1<sup>st</sup> imposter speaker has no positive values and the system seems to successfully keep the door closed.

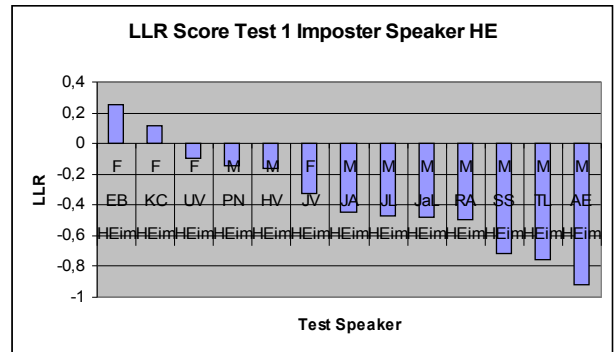


Figure 5. Result for test 1 female imposter speaker HE against all enrolled models.

The female imposter was more successful in test 1. She gained 2 positive LLRs for 2 models of enrolled speakers.

In test 2 the world model was exchanged and models retrained. This world model was trained on excerpts of spontaneous speech from 109 young male speakers recorded with a similar quality as the enrolled speakers.

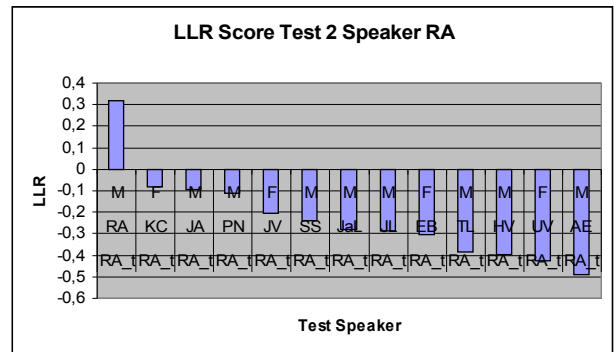


Figure 6. Result for test 2 speaker RA against all enrolled models.

The increase in data for world model training has had no significant effect in this case.

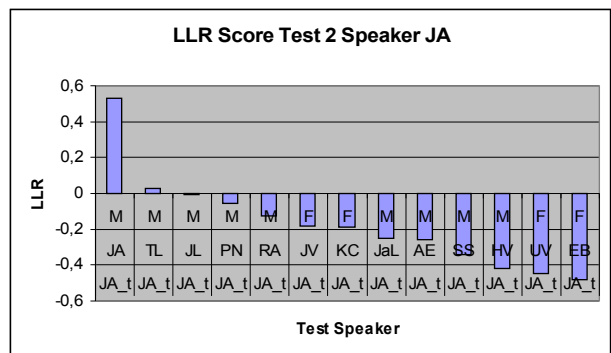


Figure 7. Result for test 2 speaker JA against all enrolled models.

For the test of speaker JA the new world model improved the test result significantly. The correct model now gets a very high score (0.53) and even though the second best has a positive LLR (0.03) it is very low.

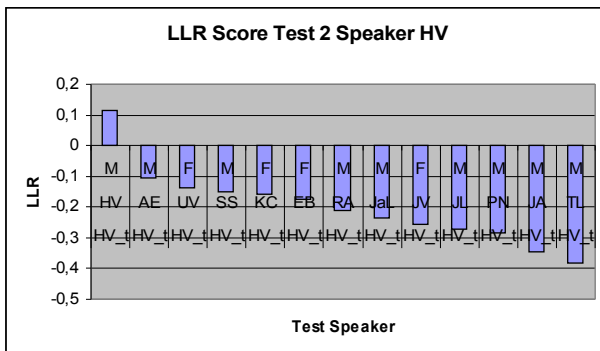


Figure 8. Result for test 2 speaker HV against all enrolled models.

Also for this test the new world model improves the correct LLR and creates a larger distance to the other models.

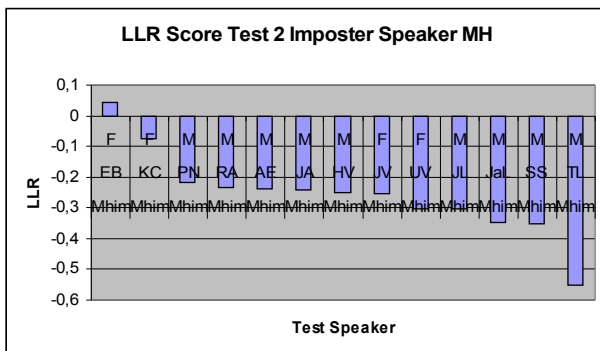


Figure 9. Result for test 2 imposter speaker MH against all enrolled models.

In the male imposter test for test 2 we obtained a rather peculiar result where the male imposter gets a positive LLR for a female target model. The lack of female training data in the world model is most probably the explanation for that.

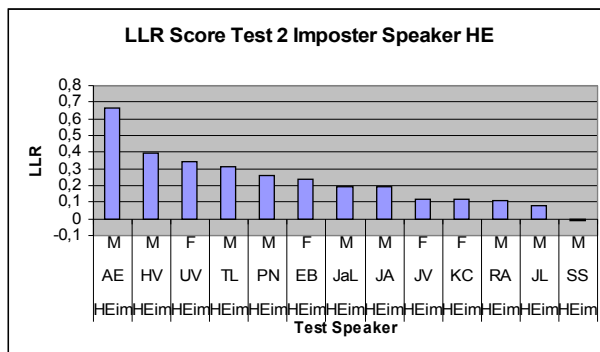


Figure 10. Result for test 2 female imposter speaker HE against all enrolled models.

When it comes to the female impostor the result it becomes even clearer that female training data is missing in the world model. All scores except 1 are positive and some of the scores very high.

## Conclusions

This first step included a successful implementation of open source tools, building a test framework and scripting procedures for text-independent speaker comparison. A small pilot study on performance of high quality recordings were made. We can conclude that it is not sufficient to train a UBM using only male speakers if you want the system to be able to handle any incoming voice. However, for demonstration purposes and comparison between small amounts of data it is sufficient to use the technique.

## References

- Boersma, Paul & Weenink, David (2009). Praat: doing phonetics by computer (Version 5.1.04) [Computer program]. Retrieved April 4, 2009, from <http://www.praat.org/>
- Bonastre, J-F, Wils, F. & Meigner, S. (2005) ALIZE, a free toolkit for speaker recognition, in Proceedings of ICASSP, 2005, pp. 737–740.
- Bonastre, J-F, Scheffer, N., Matrouf, C., Frouille, A., Larcher, A., Preti, A., Pouchoulin, G., Evans, B., Fauve, B. & Mason, J.S. (2008) ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition. In Odyssey 2008 - The Speaker and Language Recognition Workshop, 2008.
- Eriksson, A. (2004) SweDia 2000: A Swedish dialect database. In Babylonian Confusion Resolved. Proc. Nordic Symposium on the Comparison of Spoken Languages, ed. by P. J. Henrichsen, Copenhagen Working Papers in LSP 1 – 2004, 33–48
- Guillaume, G. (2004) SPro: speech signal processing toolkit, Software available at <http://gforge.inria.fr/projects/spro>.
- Martin, A. F. and Przybocki, M. A. (1999) The NIST 1999 Speaker Recognition Evaluation-An Overview. Digital Signal Processing 10: 1–18.
- Reynolds, D. A., Quatieri, T. F., Dunn, R. B., (2000) Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing, 2000.