# Voice Similarity - a Comparison Between Judgements by Human Listeners and Automatic Voice Comparison

*Jonas Lindh and Anders Eriksson*
*Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg*

## Abstract

*Comparison between the way human listeners judge voice similarity and how state-of-the art GMM-UBM systems for voice recognition compare voices is a little explored area of research. In this study groups of informants judged the similarity between voice samples taken from a set of fairly similar male voices that had previously been used in a voice line-up experiment. The result from the listening tests was then compared to the scores from a UBM-GMM automatic voice comparison system, built on the Mistral LIA_RAL open source platform. The results show a correlation between scores obtained from the automatic system and the judgements by the listeners. Listeners are, however, more sensitive to language dependent parameters or idiosyncratic phonetic features such as speaking tempo, while the system only bases its likelihood ratios on spectral similarities, i.e. timbre.*

## Introduction

Automatic methods, often referred to as Automatic Speaker Recognition systems, 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. However, a text independent system does (in most cases) not use information on how anything is spoken, so we will here refer to such a system as an Automatic Voice Comparison (AVC) system. Most importantly, we must know if and under what circumstances the outcome of the two types of analyses (automatic vs acoustic/auditory) may come into conflict and how to resolve the problem if they do. The present study is an attempt to shed some light on how human auditory voice similarity judgements compare with voice similarity scores obtained by automatic systems. We have only found a few studies where human perceptual evaluation of voice similarity and automatic methods have been directly compared. In her master's thesis, Kahn (2008) approached the problem in a manner very similar to the method applied in our study. Perception data was acquired by having subjects judge voice similarity in a pairwise comparison test using a five point scale. Likelihood ratio scores were obtained by a GMM-UBM system built on the Mistral/Alizé toolkit (http://mistral.univ-avignon.fr/). Speakers were related to some extent and read speech was recorded. Some classic acoustic measurements like mean F0 and formants were also done. The results show no significant correlation between system scores and perceptual evaluations or acoustic parameters. However there are correlations between listeners' judgments and acoustic speech parameters. The conclusion is that listeners base their judgements on acoustic information of the type represented by factors like F0 and formant values but also on speech style. Zetterholm et al. (2004) used an imitator as impostor to test a text dependant speaker verification system (Melin et al., 1998). The imitator was first trained by testing the system and receiving feedback both through listening to the target voice and being informed about the similarity (log likelihood ratio) scores. In an ABX perception test, 22 listeners rated voice similarity for all voices used in the system. The results of the listening test were then compared to the scores obtained by the system. Other studies have been conducted comparing aural similarity judgements with acoustic differences. Cleary et al. (2005) performed a study on voice similarity assessments by children with and without cochlear implants using a discrimination test. F0 needed to differ by at least 2–2.5 semitones for normal-hearing children to perceive the voices as belonging to different talkers. Several others have also studied speaker recognition correlation between aural judgements and classic acoustic parameters such as F0 and formants (Brown, 1981; Murry and Singh, 1980) or solely aural degree of similarity as a factor to be used to choose voices for a line-up (Rietveld et al., 1991).

The speech material used in the present study was originally produced for an ear witness study where 7 speaker line-ups were used to test voice recognition reliability in ear witnesses. The speakers in that study were matched for general speaker characteristics like sex, age and dialect. Some of the results obtained in the earwitness study served as inspiration for the present study. It was found, for example, that the occurrence of false identifications was not randomly distributed but systematically biased towards certain speakers. Such results raise obvious questions like: Why were these particular speakers chosen? Are their speaker characteristics particularly similar to those of the intended target? Would an aural voice comparison test single out the same speakers? And how would these voices be ranked by an automatic recognition system?

In the present study we have approached these questions by combining two experiments. In one of the experiments, listeners are asked to judge voice similarity in a pairwise comparison test. In another experiment the same stimuli are analyzed using a state-of-the-art GMM-UBM system. And in the final analysis we compare the outcomes of the two experiments and see to what extend they produce similar predictions and compare these predictions with some of the results obtained in the ear witness experiment mentioned above.

# Method

To correlate the two different kinds of measures, perceptual judgements on a five point scale and raw (not normalised) likelihood ratios, we first needed to choose compatible scales to represent both types of data. In the present case we decided that using ordinal scales (in this case rank order) for both results would suffice. There are possibilities to calculate a distance between models in the automatic system, for example cross likelihood ratio, suggested by Reynolds (1995) or normalised cross likelihood ratios (Le et al., 2007). But we choose instead to consider the informants' judgements as rankings of most to least alike since we cannot be sure that the subjects have judged the similarities on a more precise scale than a rank ordering.

To be able to collect sufficiently large amounts of data, two different web tests were designed. One of the web based forms was only released to people that could insure a controlled environment in which the test was to take place. Such a controlled environment could for example be a student lab or equivalent. A second form was created and published to as many people as possible throughout the web, a so-called uncontrolled test group. The two groups' results were treated separately and later correlated to see whether the data turned out to be similar enough for the results to be pooled.

## Mistral LIA_RAL - an open source toolkit for building a voice comparison system

The NIST speaker recognition evaluation campaign started already 1996 with the purpose of driving the technology of text-independent voice 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 about 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 Alizé SpkDet packages were developed and released as open source software under a so-called LGPL licence (Bonastre et al., 2005; Bonastre et al., 2008).

The very foundation of Mistral is Alizé, which is the umbrella for all developed packages you might include in your own application or framework. The feature extraction is handled by SPro, an open source signal processing toolkit (Guillaume, 2004). Using a background model trained using Maximum Likelihood Criterion and individual models trained to maximise the a posteriori probability that the claimed identity is the true identity given the data (MAP training) is called the GMM-UBM approach (Reynolds et al., 2000).

### *Description of the AVC system used for this study*

For the set up used in this study the so-called state-of-the-art GMM-UBM approach was adopted. Frame selection was made based on simple energy detection and the removal of silences longer than 100 milliseconds from each recording. 19 MFCCs were extracted together with delta and acceleration coefficients. 512 Gaussian mixture models were applied. The UBM was trained on 2 minutes of spontaneous speech (after frame selection) from 628 male speakers in the Swedia dialect database (Eriksson, 2004). The state-of-the-art performance of this kind of system for band limited (phone speech) is given in Fauve and Matrouf (2007). The recordings used here were sampled at 16 kHz/16 bits. The test recordings were between 13–15 seconds in duration.

## The web based listening tests

The listening tests had to be made interactive and with the results for the geographically dispersed listeners gathered in an automatic manner. Google docs provide a form to create web based question sheets collecting answers in a spreadsheet as you submit them and that was the form of data collection we chose to use for the perception part of the study. However, if one cannot provide a controlled environment, the results cannot be trusted completely. As an answer to this problem two equal web based listening tests were created, one intended for a guaranteed controlled environment and one openly published test, here referred to as uncontrolled. The two test groups are here treated separately and correlated before being merged in a final analysis.

### *The listening test material*

In the ear witness project mentioned above, the aim is to gain a better understanding of earwitness reliability. One study was designed in which children aged 7–9 and 11–13 and adults served as informants. A total of 240 participants, equally distributed between the three age groups, were exposed to an unfamiliar voice (the planning of a crime, PoC). After two weeks, the witnesses were asked to identify the target-voice in a line-up (7 voices). Half of the witnesses were exposed to a target-present line-up (TP), and the other half to a target-absent line-up (TA). The recordings used for the line-ups consisted of spontaneous speech elicited by asking the speakers to describe  a walk through the centre of Gothenburg based on a series of photos presented to them. The 9 (7 plus 1 in TA + target) speakers were all selected as a very homogeneous group, with the same dialectal background (Gothenburg area), age group (between 28–35). The speakers were also selected from a larger set of 24 speakers on the basis of a speaker similarity perception test using two groups of undergraduate students as subjects. The subjects had to make similarity judgments in a pairwise comparison test where the first item was always the target speaker intended for the line-up test. Subjects were also asked to estimate the age of the speakers. The recordings used for these tests were 16 kHz /16 bit wave files.

In the perception test for the present study, 9 voices were presented pair-wise on a web page and listeners were asked to judge the similarity on a scale from 1 to 5, where 1 was said to represent “Extremely similar or same” and 5 “Not very similar”. Since we wanted to minimize the influence of any particular language or speaking style influence the speech samples were played backwards. The listeners were also asked to submit information about their age, first language and dialectal background (if Swedish was their first language). There was also a space where they could leave comments after the completion of test and some participants used this opportunity. The speech samples used in the perception test were the first half of the 25 second samples used in the earwitness line-ups, except for the pairs where both samples were from the same speaker. In these cases the other item was the second half of the 25 second samples. Each test consisted of 45 comparisons and took approximately 25 minutes to complete. 32 (7 male, 25 female) listeners performed the controlled listening test and 20 (6 male, 14 female) the uncontrolled test.

# Results

The results are first outlined separately and then compared in the final subsection.

## System scores

Comparing all voices was done by training models for each voice as a target before testing. Models were also tested against themselves. The scores are presented as raw (not normalised) likelihood ratios.

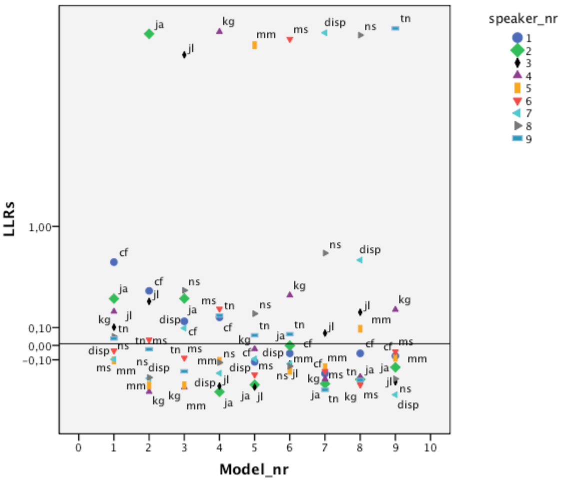

*Figure 1. Distribution of likelihood ratio scores on a logarithmic likelihood ratio scale (y-axis) and model number (x-axis).(The planning of a crime (PoC)=disp).*

For all subsequent comparisons we have converted the raw (not normalised) likelihood ratios to ranks. For easy comparison of the totality of the results we present the data as a rank matrix

*Table 1. The table shows speaker ranks when tested against models of each speaker including themselves. Mean values and standard deviation for each speaker's rankings are also shown.*

| Model | JA | JL | KG | MM | MS | PoC | NS | TN | CF |
|---|---|---|---|---|---|---|---|---|---|
| JA | 1 | 3 | 9 | 8 | 4 | 7 | 6 | 5 | 2 |
| JL | 3 | 1 | 9 | 8 | 6 | 5 | 2 | 7 | 4 |
| KG | 9 | 8 | 1 | 5 | 2 | 7 | 6 | 3 | 4 |
| MM | 8 | 9 | 4 | 1 | 7 | 5 | 2 | 3 | 6 |
| MS | 4 | 7 | 2 | 9 | 1 | 6 | 8 | 3 | 5 |
| PoC | 8 | 3 | 7 | 4 | 5 | 1 | 2 | 9 | 6 |
| NS | 7 | 3 | 6 | 4 | 9 | 2 | 1 | 8 | 5 |
| TN | 6 | 8 | 2 | 5 | 3 | 9 | 7 | 1 | 4 |
| CF | 2 | 4 | 3 | 9 | 7 | 8 | 5 | 6 | 1 |
| Mean rank | 5.33 | 5.10 | 4.77 | 5.88 | 4.88 | 5.55 | 4.33 | 5.00 | 4.10 |
| Std dev | 1.7 | 2.9 | 2.9 | 3.1 | 2.8 | 2.6 | 2.7 | 2.6 | 2.7 |

The results show that some speakers are generally higher ranked than others. For example speaker CF has a mean rank of 4.1, indicating that he is somewhat of a wolf. From the system scores (figure 1) it is also clear that the CF model suffers from a high degree of false acceptance, which indicates that the model is a lamb (Campbell, 1997; Doddington, 1985; Melin, 2006).

## Listening test result

Both listening tests separately (controlled and uncontrolled) show significant inter-rater agreement (Cronbach's alpha = 0.98 for the controlled and 0.959 for the uncontrolled test). When both datasets are pooled the inter-rater agreement remains at the same high level (alpha = 0.975) indicating that listeners in both subgroups have judged the voices the same way. This justifies using the pooled data from both groups (52 subjects altogether) for the further analysis of the perception test results. The results of the perception test are presented in Table 2. The rankings are based on the means of the similarity judgments.

*Table 2. The table shows speaker ranks based on mean similarity judgement for both listener groups pooled.*

| Speaker | JA | JL | KG | MM | MS | PoC | NS | TN | CF |
|---|---|---|---|---|---|---|---|---|---|
| JA | 1 | 4 | 5 | 3 | 6 | 8 | 9 | 7 | 2 |
| JL | 3 | 1 | 8 | 5 | 7 | 4 | 2 | 9 | 6 |
| KG | 5 | 9 | 1 | 2 | 3 | 7 | 8 | 6 | 4 |
| MM | 4 | 5 | 2 | 1 | 3 | 8 | 9 | 7 | 6 |
| MS | 7 | 8 | 6 | 5 | 2 | 9 | 3 | 1 | 4 |
| PoC | 5 | 3 | 6 | 4 | 9 | 1 | 7 | 8 | 2 |
| NS | 6 | 2 | 8 | 5 | 3 | 7 | 1 | 9 | 4 |
| TN | 6 | 9 | 5 | 4 | 1 | 7 | 8 | 2 | 3 |
| CF | 2 | 9 | 6 | 7 | 3 | 5 | 8 | 4 | 1 |
| Mean rank | 4.3 | 5.6 | 5.2 | 4.0 | 4.1 | 6.2 | 6.1 | 5.9 | 3.6 |
| Std dev | 2.0 | 3.2 | 2.4 | 1.8 | 2.6 | 2.5 | 3.2 | 2.9 | 1.7 |

Also in the perception test, speaker CF receives the highest mean rank with low variation. This indicates that speaker CF is also the most likely to be picked as the target if uncertain in a closed set line-up. This was indeed also the case in the earwitness study where speaker CF was the speaker most often confused with the target speaker resulting in a large number of false acceptances in both the TA and the TP conditions.

## Comparison between system scores and listening tests

In order to visualize the results presented in the matrices above we used Multidimensional Scaling to produce 2-dimensional Euclidean distance models (similar to Kahn, 2008).

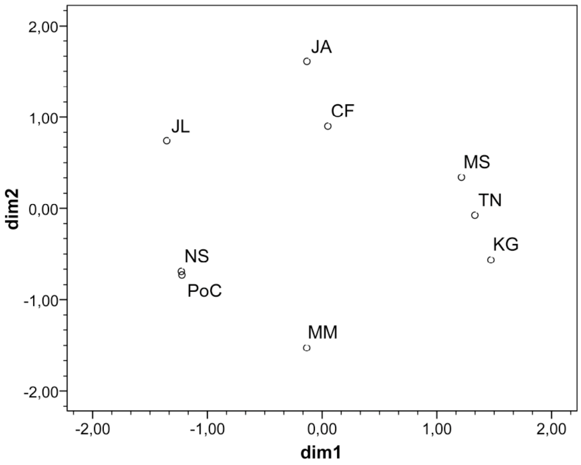

*Figure 2. Euclidean distance model representing the data rankings based on the log likelihood ratios from the ASC system.*

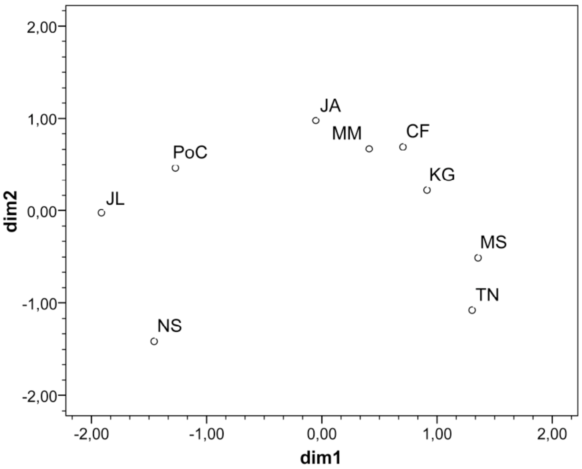

*Figure 3. Euclidean distance model representing the data rankings based on the perception results.*

We may observe several similarities between the representations of the AVC and the perception scores. Two groups of speakers cluster together in both diagrams. Speakers JL and NS group together with target voice in the earwitness study (PoC). Speakers MS, TN and KG who form a group in the AVC analysis also group together in the perception results. The latter speakers are also among those who are least often subject to false acceptance in the line-up experiment.

# Discussion and Conclusions

Even though it is difficult to draw any precise conclusions on whether humans and the system perform the same kind of comparisons, some conclusions may be drawn regarding the influence of linguistic factors. It seems as if humans include what is being said and how it is said in voice similarity judgements. This probably made the similarity judgment a rather difficult task using backward speech (also reflected in their comments). The system score ranking gives a clue to why the voice CF can be considered a 'wolf'. It is the voice with highest mean rank in both Table 1 and 2.

We also mean that many differences can be explained by the use of linguistic/phonetic cues still present in backward speech. Not all factors are eliminated this way, for example pausing and articulation rate. A separate analysis using rankings based on articulation rate shows MM, CF and PoC clustering together. This may explain why MM and CF cluster in Figure 2, in spite of the fact that they are quite dissimilar in the AVC analysis. These 2 voices are also subject to a high degree of the false acceptances made by participants in the voice line-up study. The false acceptances are most biased towards CF, which we suspect is a combination of speech feature similarities and his ‘wolfness’.

Speakers JL, NS and PoC also form a group in both analyses, but here we may see that whereas NS and PoC are regarded as identical in the ASC analysis they are quite widely separated in the perception data. This difference contains important information about the influence of speaking style for the perception results. Speaker NS is namely the speaker who was also used for the mock incriminating call. The voice is thus the same in both samples which is detected by the AVC system. The speaking style is, however, quite radically different. And as has been pointed out above such factors as articulation rate and pausing are at least partly present even if the speech samples are played backwards. It seems reasonable to suggest that the listeners observe this difference and therefore judge the speech samples as quite different in spite of the fact that the voice characteristics are very similar.

Generally we can conclude that identifying a speaker in the voice line-up study was also a difficult task. We suspect that listeners use at least two different strategies. They may pay a lot of attention to voice quality or concentrate on speaking parameters such as articulation rate and maybe to some extent pronunciation. In the first case they are much more likely to make correct identifications in a line-up task or judge the voices more in agreement with the automatic system in a voice similarity task.

One of the aims of the present study was, as pointed out in the introduction, to look for similarities between automatic and perceptual analyses, but also to detect possible conflicting differences. The present study does not contain any conflicting results, but several examples of how human listeners integrate factors which depend on speaking style even when the task is explicitly to judge voice similarity.

## References

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., Fredouille, 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.

Brown, B. (1981). An experimental study of the relative importance of acoustic parameters for auditory speaker recognition. Language and Speech 24: 295–310.

Campbell, J. (1997) Speaker recognition: A tutorial. Proceedings of the IEEE, 85(9):1437–1462.

Cleary, M., Pisoni, D. & Kirk, K. (2005) Influence of Voice Similarity on Talker Discrimination in Children With Normal Hearing and Children With Cochlear Implants. Journal of Speech, Language, and Hearing Research 48(2005/015): 204–223.

Doddington, G. (1985) Speaker recognition - identifying people by their voices. Proceedings of the IEEE, 73(11):1651–1664.

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.

Fauve, B., Matrouf, D. et al. (2007) State-of-the-Art Performance in Text-Independent Speaker Verification through Open-Source Software. IEEE Transactions on Audio, Speech and Language Processing 15, Issue 7: 1960–1968.

Guillaume, G. (2004) SPro: speech signal processing toolkit, Software available at http://gforge.inria.fr/projects/spro.

Kahn, J. (2008). Caractéristique propres au locuteur : Traitement automatique et distance perceptive. Université-Stendhal-Grenoble 3 : Unpublished Masters Thesis.

Le, V-B., Mella, O. & Fohr, D. (2007) "Speaker Diarization using Normalized Cross Likelihood Ratio". Proceedings of Interspeech 2007, Antwerp, Belgium, 2007.

Martin, A. F. and Przybocki, M. A. (1999) The NIST 1999 Speaker Recognition Evaluation-An Overview. Digital Signal Processing 10: 1–18.

Melin H., Koolwaaij J.W., Lindberg J., Bimbot F. (1998). A Comparative Evaluation of Variance Flooring Techniques in HMM-based Speaker Verification. Proc. of ICSLP ’98:1903–996.

Melin, H. (2006) Automatic speaker verification on site and by telephone: methods, applications and assessment, PhD thesis, KTH, Stockholm, December 2006.

Murry, T. and S. Singh (1980). "Multidimensional analysis of male and female voices." Journal of the Acoustical Society of America 68: 1294–1300.

Reynolds, D., (1995) Speaker identification and verification using Gaussian mixture speaker models, Speech Communication, vol. 17, issue 1–2, pp. 91–108, 1995.

Reynolds, D. A., Quatieri, T. F., Dunn, R. B., (2000) Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing, 2000.

Rietveld, A. C. M. and A. P. A. Broeders (1991). Testing the fairness of voice parades: the similarity criterion. Proceedings of the 12th International Congress of Phonetic Sciences. Aix-en-Provence, Université de Provence, Service des Publications. 5: 46–49.

Zetterholm, E., D. Elenius and M. Blomberg (2004) A comparison between human perception and a speaker verification system score of a voice imitation. Proceedings SST2004, Sydney, Australia, Dec 8–10 2004.