

Perception of voice similarity and the results of a voice line-up

Jonas Lindh

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

Abstract

The perception of voice similarity is not the same as picking a speaker in a line-up. This study investigates the similarities and differences between a perception experiment where people judged voice similarity and the results from voice line-up experiment. Results give us an idea about what listeners do when they try to identify a voice and what parameters play an important role. The results show that there are similarities between the voice similarity judgments and the line-up results. They differ, however, in several respects when we look at speaking parameters. This finding has implications for how to consider the similarities between foils and suspects when setting up a line-up as well as how we perceive voice similarities in general.

Introduction

Aural/acoustic methods in forensic speaker comparison cases are common. It is possible to divide speaker comparison into 2 different branches depending on the listener. The 1st is the expert witness' aural examination of speech samples. In this case the expert tries to quantify and assess the similarities/dissimilarities between speakers based on linguistic, phonological and phonetic features and finally evaluate the distinctiveness of those features (French & Harrison, 2007). The 2nd branch is the speaker comparison made by naive listeners, for example victims of a crime where they heard a voice/speaker, but could not see the perpetrator. In both cases, some kind of voice quality would be used as a parameter. However, it is not thoroughly investigated if this parameter can be separated from so called articulation or speaking parameters such as articulation rate (AR) or pausing, which are parameters that have shown to be useful parameters when comparing speakers (Künzel, 1997). To be able to study this closer a web based perception experiment was set up where listeners were asked to judge voice similarity in a pairwise comparison test. The speech was played backwards to remove speaking characteristics and force listeners to

concentrate on voice quality similarity. 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 male and matched for general speaker characteristics like sex, age and dialect. The results from the ear witness study and the judgments of voice similarity were then compared. It was found, for example, that the occurrence of false acceptances (FA) 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? The results give implications on the existence of speech characteristics still present in backward speech. It can also be shown that speakers that are judged as wolves (term from speaker verification, where a voice is rather similar to many models) can be picked more easily in a line-up if they also possess speech characteristics that are similar to the target.

Method

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.

The ear witness study

To gain a better understanding of earwitness performance a study was designed in which children aged 7-8 and 11-12 and adults served as informants. A total of 240 participants were

equally distributed between the three age groups and exposed to an unfamiliar voice. Each participant was asked to come along with an experimenter to a clothes shop where they stopped outside a fitting cubicle. Behind the curtain they could here an unfamiliar voice planning of a crime (PoC). The recording they heard was played with a pair of high quality loudspeakers and was approximately 45 seconds long. 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 line-up was also played to the witness on loudspeakers from a computer and the voices presented on a power point slide. First an excerpt from a recording of a city walk of a bout 25 seconds was played. After that a shorter part of the excerpt of about 12-15 seconds was used. First they had to say whether they thought the voice was present in the line-up, and if so, they pointed the voice out. Secondly they were asked about their confidence and what they remembered from what the voice in the cubicle had said. This was done to see whether it was possible to predict identification accuracy by analyzing memory for content (Öhman, Eriksson & Granhag, 2009). To be able to quantify speaking parameters, pausing and articulation rate was measured. Articulation rate is here defined as produced syllables excluded pausing. Pauses are defined as clearly measurable silences longer than 150 ms.

The test material

The recordings 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) and age group (between 28–35). The speakers were 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.

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.

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, 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 and Discussion

The results will be presented separately in the first 2 paragraphs and then the comparison is done with a short discussion in the last section.

The overall results of the ear witness study

The original purpose of the study was to compare performance between the age groups. Here

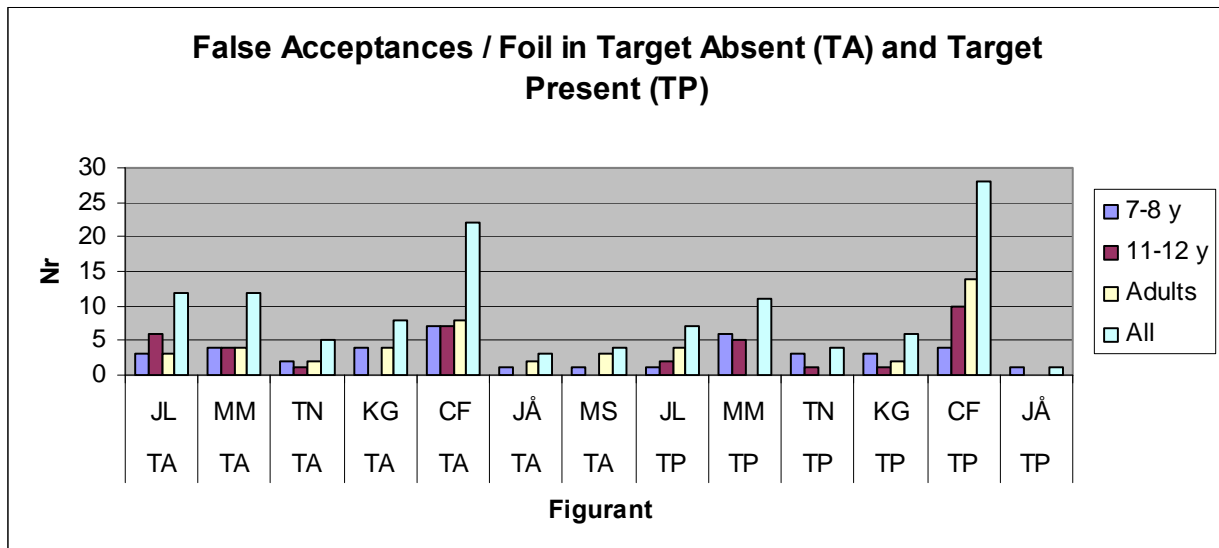


Figure 1. False acceptances for each figurant speaker in the 3 age groups and the sum (all) for both target absent (TA) and target present (TP).

we are only interested in the general tendencies of the false acceptances (the picking of the wrong voice) and the true, i.e. correct identifications. In Figure 1 we present the false acceptances given by the different age groups and all together.

In Figure 1 it is very clear that false acceptance is biased toward certain speakers such as speaker CF followed by MM and JL. It is noticeable that correct acceptances in TP was 27 and that can explain the decrease in FA for MM and JL, however, the degree of FA for speaker CF is even higher in TP (28).

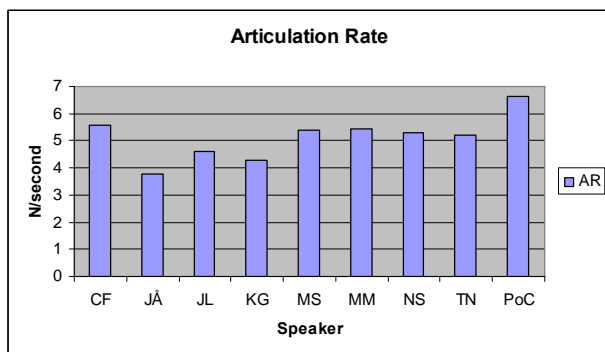


Figure 2. Articulation rate (produced syllables per second) for the speakers in the line-up.

In Figure 2 we can see that the target (PoC) was produced with a fast articulation rate. Several speakers follow with rather average values around 5 syllables per second. The speaker with the highest AR compared to PoC is CF. In Figure 3 we take a closer look at pausing.

Pauses tend to increase in duration with high articulation rate (Goldman-Eisler, 1961).

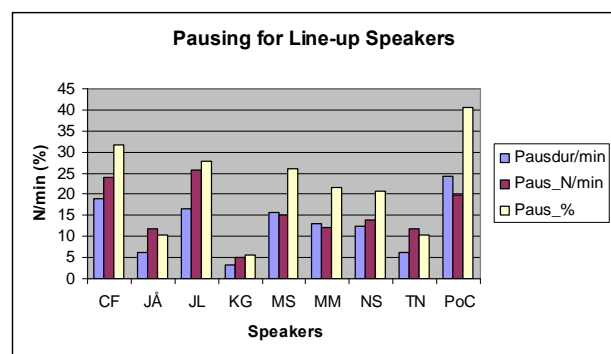


Figure 3. Pausing (pause duration per minute, number of pauses per minute and percentage pause from total utterance duration) for the speakers in the line-up.

The pausing measurement shows a bias towards speaker CF, which might explain some of the false acceptances.

The perception test results

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.

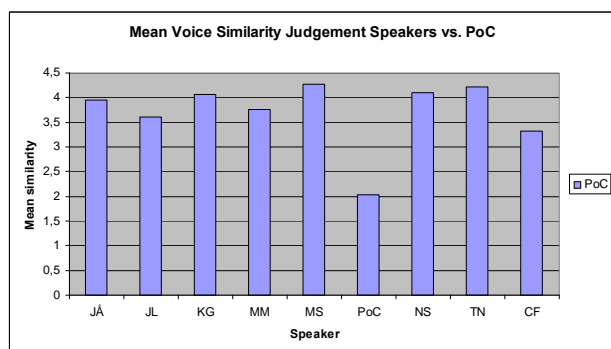


Figure 4. Mean voice similarity judgment by listeners comparing each speaker against target PoC. The closer to 1 the more similar voice according to judgments.

The voice similarity judgments indicate the same as the line-up regarding speaker CF, who is judged to be closest to the target followed by JL and MM. It is also noticeable that those speakers are among the speakers who get the highest mean overall similarity judgments compared to all the other speakers.

Table 1. The table shows speaker ranks based on mean similarity judgment for both listener groups pooled.

Speaker	JÅ	JL	KG	MM	MS	PoC	NS	TN	CF
JÅ	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

The mean rank in table 1 indicates how the speaker is ranked compared to the other voices in similarity judgment.

Comparison of results and discussion

The purpose of the study was to compare the general results from the line-up study and the results of the perception experiment presented here. A comparison between the results show

that CF is generally judged as most similar to the target speaker (even more than the actual target in the TP line-up). We have also found that the result can partly be explained by the similarity in speaking tempo parameters. However, since the result is also confirmed in the perception experiment it must mean either that the tempo parameters are still obvious in backward speech or that there is something else that make listeners choose certain speakers. Perhaps the indication that speakers are generally high ranked, or wolves (term from speaker verification, see Melin, 2006), in combination with similar aspects of tempo make judgments biased. More research to isolate voice quality is needed to answer these questions in more detail.

Acknowledgements

Many thanks to the participants in the listening test. My deepest gratitude to the AllEars project Lisa Öhman and Anders Eriksson for providing me with data from the line-up experiments before they have been published.

References

- French, P., and Harrison, P. (2007) Position Statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases, with a foreword by Peter French & Philip Harrison. *International Journal of Speech Language and the Law*. [Online] 14:1.
- Künzel, H. (1997) Some general phonetic and forensic aspects of speaking tempo. *Forensic Linguistics* 4, 48–83.
- Goldman-Eisler, F. (1961) The significance of changes in the rate of articulation. *Lang. and Speech* 4, 171-174.
- Öhman, L., Eriksson, A. and Granhag, P-A. (2009) Unpublished Abstract. Earwitness identification accuracy in children vs. adults.