

Earwitnesses: The effect of voice differences in identification accuracy and the realism in confidence judgments

Elisabeth Zetterholm¹, Farhan Sarwar² and Carl Martin Allwood³

¹Centre for Languages and Literature, Lund University

²Department of Psychology, Lund University

³Department of Psychology, University of Gothenburg

Abstract

Individual characteristic features in voice and speech are important in earwitness identification. A target-absent lineup with six foils was used to analyze the influence of voice and speech features on recognition. The participants' response for two voice foils were particularly successful in the sense that they were most often rejected. These voice foils were characterized by the features' articulation rate and pitch in relation to the target voice. For the same two foils the participants as a collective also showed marked underconfidence and especially good ability to separate correct and incorrect identifications by means of their confidence judgments for their answers to the identification question. For the other four foils the participants showed very poor ability to separate correct from incorrect identification answers by means of their confidence judgments.

Introduction

This study focuses on the effect of some voice and speech features on the accuracy and the realism of confidence in earwitnesses' identifications. More specifically, the study analyzes the influence of characteristic features in the speech and voices of the target speaker and the foils in a target-absent lineup on identification responses and the realism in the confidence that the participants feel for these responses. This theme has obvious relevance for forensic contexts.

Previous research with voice parades has often consisted of speech samples from laboratory speech, which is not spontaneous (Cook & Wilding, 1997; Nolan, 2003). In spontaneous speech in interaction with others, the assumption is that the speakers might use another speaking style compared with laboratory speech. In forensic research spontaneous speech is of more interest since that is a more realistic situation.

Sex, age and dialect seem to be strong and dominant features in earwitness identification (Clopper et al, 2004; Eriksson et al., 2008; Lass et al., 1976; Walden et al., 1978). In these studies, there is nothing about how the witness' confidence and its' realism is influenced by these features.

The study presented in this paper focuses on the influence of differences and similarities in voice and speech between a target voice and six foils in a lineup. A week passed between the original presentation of the target speaker (at for example the crime event) and the lineup, which means that there is also a memory effect for the listeners participating in the lineup. Spontaneous speech is used in all recordings and only male native Swedish speakers.

Confidence and realism in confidence

In this study, a participant's confidence in his or her response to a specific voice in a voice parade with respect to if the voice belongs to the target or not, relates to whether this response is correct or not. Confidence judgments are said to be realistic when they match the correctness (accuracy) of the identification responses. Various aspects of realism can be measured (Yates, 1994). For example, *the over-/underconfidence measure* indicates whether the participant's (or group's) level of confidence matches the level of the accuracy of the responses made. It is more concretely computed as: Over-/underconfidence = (The mean confidence) minus (The mean accuracy).

Another aspect of the realism is measured by *the slope measure*. This measure concerns a participant's (or group's) ability, by means of one's confidence judgments, to, as clearly as possible, separate correct from incorrect judgments. This measure is computed as: Slope = (The mean confidence for correct judgments) minus (The mean confidence for incorrect judgments). The relation between a participant's level of confidence for a voice with a

specific characteristic compared with the participant's confidence for the other voices in the lineup might indicate how important that characteristic is for the participant's judgment. In a forensic context the level of realism indicates how useful a participant's confidence judgments are in relation to targets' and foils' voices with specific features.

Method

Participants

100 participants took part in the experiment. The mean age was 27 years. There were 42 males and 58 females. 15 participants had another mother tongue than Swedish, but they all speak and understand Swedish. Four of them arrived in Sweden as teenagers or later. Five participants reported minor impaired hearing and four participants reported minor speech impediment, one of them stuttering.

Materials

The dialogue of two male speakers was recorded. They played the role of two burglars planning to break into a house. This recording was about 2 minutes long and was used as the familiarization passage, that is, as the original experienced event later witnessed about. The speakers were 27 and 22 year old respectively when recorded and both speak with a Scanian dialect. The 22 years old speaker is the target and he speaks most of the time in the presented passage.

The lineup in this study used recordings of six male speakers. It was spontaneous speech recorded as a dialogue with another male speaker. This male speaker was the same in all recordings and that was an advantage since he was able to direct the conversation. They all talked about the same topic, and they all had some kind of relation to it since it was an ordinary situation. As a starting point, to get different points of view on the subject talked about and as a basis for their discussion, they all read an article from a newspaper. It had nothing to do with forensics. The recordings used in the lineups were each about 25 sec long and only a part of the original recordings, that is, the male conversation partner is not audible in the lineups.

All the six male speakers have a Scanian dialect with a characteristic uvular /r/ and a slightly diphthongization. They were chosen

for this study because (as described in more detail below) they share, or do not share, different features with the target speaker. It is features such as pitch, articulation rate, speaking style and overall tempo and voice quality.

The target speaker has a mean F0 (mean fundamental frequency) of 107 Hz, see Table 1. The speech tempo is high overall and he has an almost forced speaking style with a lot of hesitation sounds and repetition of syllables when he is excited in the familiarization passage. The acoustic analysis confirms a high articulation rate.

Foil 1 and 2 are quite close in their speech and voices in the auditory analysis. Both speak with a slightly creaky voice quality, although foil 2 has a higher mean F0. Their articulation rate is quite high and close to the target speaker. Foil 3 and 6 speak with a slower speech tempo and a low and a high pitch respectively is audible. In the acoustic analysis it is also obvious that both foil 3 and 6 have an articulation rate which is lower than the target speaker. Foil 4 is the speaker who is closest to the target speaker concerning pitch and speaking style. He speaks with a forced, almost stuttering voice when he is keen to explain something. His articulation rate is high and he also uses a lot of hesitation sounds and filled pauses. Foil 5 has quite a high articulation rate, in similarity to the target speaker, but he has a higher pitch and his dialect is not as close to the target speaker as the other foils.

All the speakers, including the target speaker, are almost the same age, see Table 1. The results of the acoustic measurements of mean fundamental frequency (F0) and std.dev. are also shown in the table. The perceptual auditory impression concerning the pitch is confirmed in the acoustic analysis.

Table 1. Age, F0 mean and standard deviations (SDs) for the target speaker and the six foils

	Age	F0, mean	SDs.
target	22	107 Hz	16 Hz
foil 1	23	101 Hz	26 Hz
foil 2	21	124 Hz	28 Hz
foil 3	23	88 Hz	15 Hz
foil 4	23	109 Hz	19 Hz
foil 5	23	126 Hz	21 Hz
foil 6	25	121 Hz	17 Hz

Procedure

The experimental sessions took place in classes at the University of Lund and in classes with

final year students at a high school in a small town in Northern Scania. The experiment conductor visited the classes twice. The first time, the participants listened to the 2 minute dialogue (the original event). The only instruction they got was that they had to listen to the dialogue. Nothing was said about that they should focus on the voices or the linguistic content. The second time, one week later, they listened to the six male voices (the foils), in randomized order for each listener group. Each voice was played twice. The target voice was absent in the test. The participants were told it was six male voices in the lineup. This was also obvious when looking at the answer sheets. There were six different listener groups for the 100 participants presented in this paper. The number of participants in each group differed between seven and 27 people.

For each voice, the participants had to make a decision if the voice was the same as the one who talked mostly in the dialogue last week. There were two choices on the answer sheet for each voice; 'I do recognize the voice' or 'I do not recognize the voice'. They were not told if the target voice was absent or not, nor were they initially told if a voice would be played more than once. There was no training session. The participants were told that they could listen to each voice twice before answering, but not recommended to do that.

Directly after their judgment of whether a specific voice was the target or not, the participants also had to estimate the confidence in their answer. The confidence judgment was made on scale ranging from 0% (explained as "absolutely sure this voice sample is not the target") via 50% ("guessing") to 100% ("absolutely sure this voice sample is the target").

Results and Discussion

Since this is an ongoing project the results presented here are only results from the first 100 participants. We expect 300 participants at the end of this study.

Figure 1 shows the number of the 'I do recognize the voice' answers, or 'yes' answers. Since the voice lineups were target-absent lineups, a "yes" answer equals an *incorrect* answer, that is, an incorrect identification.

When looking at the answers focusing on the presentation order, which, as noted above, was randomized for each group, it is obvious that there is a tendency not to choose the first voice. There were no training sessions and that

might have had an influence. Each voice was played twice, which means that the participants had a reasonable amount of time to listen to it. The voices they heard in the middle of the test as well as the last played voice were chosen most often.

Only 10 participants had no 'yes' answers at all, which is 10% of all listeners, that is, these participants had all answers correct. The result for confidence for these 10 participants showed that their average confidence level was 76 %, which can be compared with the average confidence level for the remaining 90 participants, 69 %. No listener had more than 4 'yes' answers, which means that no one answered 'yes' all over without exception.

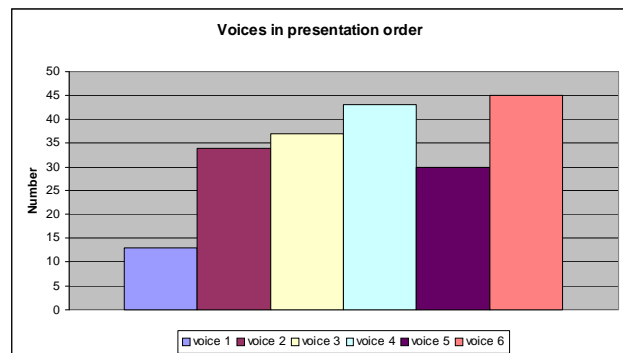


Figure 1. Numbers of 'yes' answers (e.g., errors) focusing on the presentation order.

In Figure 2 the results are shown focusing on each foil 1-6. Most of the participants selected foil 4 as the target speaker. The same results are shown in Table 2. Foil 4 is closest to the target speaker and they share more than one feature in voice and speech. The speaking style with the forced, almost stuttering voice and hesitation sounds is obvious and might remind the listeners of the target speaker when planning the burglary. Foil 3 and 6 were chosen least often, and these foils are different from the target speaker both concerning pitch, articulation rate and overall speaking style. It is not surprising that there were almost no difference in results between foil 1 and 2. These male speakers have very similar voices and speech. They are also quite close to the target speaker in the auditory analysis (i.e. according to an expert holistic judgment). Foil 5 received many 'yes' answers as well. He reminds of the target speaker concerning articulation rate, but not as obviously as foil 4. He also has a higher pitch and a slightly different dialect compared to the target speaker.

The results indicate that the participants were confused about foil 4 and this is an expected result. The confusion might be explained both in the auditory and the acoustic analyses. The overall speaking style, the articulation rate as well as the pitch, of the target speaker and foil 4 are striking.

The results show that the mean accuracy of all the foils was 66.17 with a standard deviation (SD) of 47.35. The results for each of the foils are shown in Table 2. These results were analyzed with a one-way ANOVA and the outcome shows that the difference in how often the six foils were (correctly) rejected was significant, $F(5, 594) = 12.69, p < .000$. Further *post hoc Tukey* test revealed

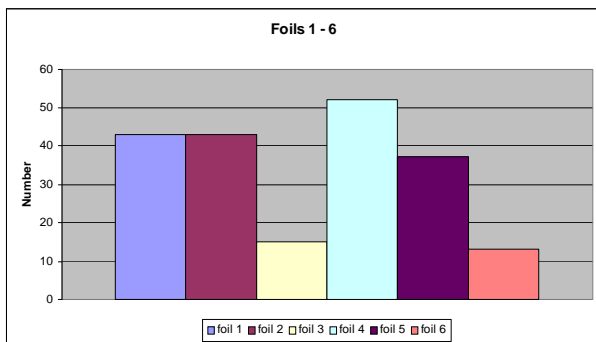


Figure 2. Numbers of 'yes' answers (e.g., errors) focusing on the foils 1-6.

that participant rejected the foil 3 ($M=85$) and foil 6 ($M=87$), significantly more often as compared with the other foils.

We next look at the results for the confidence judgments and their realism. When analyzing the confidence values we first reversed the confidence scale for all participants who gave a "no" answer to the identification question. This means that a participant who answered "no" to the identification question and then gave "0 %" ("absolutely sure that this voice sample is not the target") received a 100% score in confidence when the scale was reversed. Similarly, a participant who gave 10 as a confidence value received 90 and a participant who gave 70 received a confidence value of 30 after the transformation, etc. 50 in confidence remained 50 after the transformation. In this way the meaning of the confidence ratings could be interpreted in the same way for all participants, irrespective of their answers to the identification question.

The mean confidence for all the foils was 69.93 ($SD = 25.00$). Table 2 shows that there was no great difference in the level of the confidence judgments for the given identification answers for the respective foils. A one-way ANOVA showed no significant difference between the foils with respect to their confidence ($F = .313$).

Turning next to the over/underconfidence measure (O/U-confidence), the average O/U-confidence computed over all the foils was 3.77 ($SD = 52.62$), that is, a modest level of overconfidence. Table 2 shows the means and SDs for O/U-confidence for each foil. It can be noted that the participants showed quite good realism with respect to their level of over/underconfidence for item 5 and an especially high level of overconfidence for item 4. Moreover, the participants' showed underconfidence for items 3 and 6, that is, the same items showing the highest level of correctness.

A one-way ANOVA showed that there was a significant difference between the foils with respect to the participants' O/U-confidence, $F(5, 394) = 9.47, p < .000$. Further *post hoc Tukey* tests revealed that the confidence of the participants who rejected foil 3 and foil 6 showed significantly lower over/underconfidence as compared to the confidence of participants for foil 1, foil 2 and foil 4.

Table 2. Means (SDs) for accuracy (correctness), confidence, over-/underconfidence and slope for the six foils

	Foil 1	Foil 2	Foil 3	Foil 4	Foil 5	Foil 6
Accuracy	57.00 (49.76)	57.00 (49.76)	85.00 (35.89)	48.00 (50.21)	63.00 (48.52)	87.00 (33.79)
Confidence	69.90 (22.18)	70.00 (24.82)	70.20 (28.92)	70.40 (21.36)	67.49 (23.81)	71.70 (28.85)
Over-/under conf	12.9%	13.0%	-14.8%	22.4%	4.4%	-15.3%
Slope	-5.07	-2.04	18.27	0.43	1.88	12.57

Table 2 also shows the means for the slope measure (ability to separate correct from incorrect answers to the identification question by means of the confidence judgments) for the 6 foils. The overall slope for all data was 2.21. That is, the participants on average showed a very poor ability to separate correct from incorrect answers by means of their confidence judgments. However, it is of great interest to note that the only items for which the participants showed a clear ability to separate correct from incorrect answers was for the two foils (3 and

6) for which they showed the highest level of correct answers to the identification question.

Summary and conclusions

In this study the original event consisted of a dialogue between two persons and, similarly, the recordings for the foils were a dialogue. This is an important feature of this study and something that contributes to increasing the ecological validity in this research area since previous research often has used monologue readings of text both as the original events and as the recognition stimuli. To what extent this feature of the study influenced the results is not clear since we did not have a comparison condition in this context.

Characteristic voice features had an impact upon the listeners in this study. The results, so far, are expected since the participants seem to be confused and thought that the voice of foil 4 was the target speaker, compared with the other foils in the lineup. Foil 4 was the most alike concerning pitch and speaking style. It might be that the speaking style and the forced voice was a kind of hang-up for the listeners. Even though all male speakers had almost the same dialect and the same age as the target speaker, there were obvious differences in their voices and speech behavior. The listeners were not told what to focus on when listening the first time. As noted above, we don't know if the use of a dialogue with a forensic content had an effect upon the result. The recordings in the lineup were completely different in their content.

In brief, the results in this study suggest that prominent characteristic features in voice and speech are important in an earwitness identification situation. In a forensic situation it would be important to be aware of characteristic features in the voice and speech.

Turning next to the realism in the participants' confidence judgments it is of interest that the participants in this study over all, in contrast to some other studies on earwitnesses (e.g., Olsson et al, 1998), showed only a modest level of overconfidence. However, a recent review of this area shows that the level of realism found depends on the specific measure used and various specific features of the voices involved. For example, more familiar voices are associated with better realism in the confidence judgments (Yarmey, 2007). Had a different mixture of voices been used in the present

study, the general level of realism in the O/U-confidence measure might have been different.

We next discuss the variation between the foils with respect to their level of overconfidence. It can be discerned from Table 2 that the level of overconfidence follows the respective foil's level of accuracy. When the level of identification accuracy is high, the level of O/U-confidence is lower, or even turns into underconfidence. Thus, a contributing reason to the variation in overconfidence between the foils (in addition to the similarity of the foils' voices to that of the target), may be that the participants expected to be able to identify a foil as the target and when they could not do so this resulted in less confidence in their answer. Another speculation is that the participants' general confidence level may have been the most important factor. In practice, if these speculations are correct it is possible that the different speech features of the foils' voices did not contribute very much to the participants' level of confidence or degree of overconfidence. Instead the participants' confidence may be regulated by other factors as speculated above.

The results for the slope measure showed that the participants evidenced some ability to separate correct from incorrect answers by means of their confidence judgments for the two foils 3 and 6, that is, the foils for which the participants showed the highest level of accuracy in their identifications. These two foils were also the foils that may be argued to be perceptually (i.e., "experientially") most separate from the target voice. For the other four foils the participants did not evidence any ability at all to separate correct from incorrect identification answers by means of their confidence judgments.

Finally, given that they hold in future research, the results showed that earwitnesses' confidence judgments do not appear to be a very reliable cue as to the correctness of the their identifications, at least not in the situation investigated in this study, namely the context of target-absent lineups when the target voice occur in dialogues both in the original event and in the foils' voice sample. The results showed that although the average level of overconfidence was fairly modest when computed over all foils, the level of over-underconfidence varied a lot between the different foils. Still it should be noted that for those two foils where the participants had the best accuracy level they also tended to give higher confidence judgments.

ments for correct answers as compared with incorrect answers. However, more research is obviously needed to confirm the reported results.

Acknowledgements

This work was supported by a grant from Crafoordska stiftelsen, Lund.

References

- Clopper C.G. and Pisoni D.B. (2004) Effects of talker variability on perceptual learning of dialects. *Language and Speech*, 47 (3), 207-239.
- Cook S. and Wilding J. (1997) Earwitness testimony: Never mind the variety, hear the length. *Applied Cognitive Psychology*, 11, 95-111.
- Eriksson J.E., Schaeffler F., Sjöström M., Sullivan K.P.H. and Zetterholm E. (submitted 2008) On the perceptual dominance of dialect. *Perception & Psychophysics*.
- Lass N.J., Hughes K.R., Bowyer M.D., Waters L.T. and Bourne V.T. (1976) Speaker sex identification from voice, whispered, and filtered isolated vowels. *Journal of the Acoustical Society of America*, 59 (3), 675-678.
- Nolan F. (2003) A recent voice parade. *Forensic Linguistics*, 10, 277-291.
- Olsson N., Juslin P., & Winman A. (1998) Realism of confidence in earwitness versus eyewitness identification. *Journal of Experimental Psychology: Applied*, 4, 101-118.
- Walden B.E., Montgomery A.A., Gibeily G.J., Prosek R.A. and Schwartz D.M. (1978) Correlates of psychological dimensions in talker similarity. *Journal of Speech and hearing Research*, 21, 265-275.
- Yarmey A.D. (2007) The psychology of speaker identification and earwitness memory. In R.C. Lindsay, D.F. Ross, J. Don Read & M.P. Toglia (Eds.), *Handbook of eyewitness psychology, Volume 2, Memory for people* (pp. 101-136). Mahwah, N.J.: Lawrence Erlbaum Associates.
- Yates J.F. (1994) Subjective probability accuracy analysis. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 381-410). New York: John Wiley & Sons.