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

Elisabeth Zetterholm[1], Farhan Sarwar[2] and Carl Martin Allwood[3]
[1]School of Language and Literature, Linnaeus University, Växjö
[2]Department of Psychology, Lund University
[3]Department of Psychology, University of Gothenburg

Abstract

This contribution is a partial report from a study of the identification accuracy and realism in the confidence judgments of the correctness in the identification reports in two kinds of target-present voice lineup. 24 men and 54 women were asked to identify a voice that they had heard previously in a dialogue context that simulated the planning of a burglary by two males 22 and 27 years old. The voice lineup either consisted of recordings of each of six male speakers reading a text from a book (text-lineup condition) or each of the same six speakers having a spontaneous dialogue with another male speaker (dialogue-lineup condition). Each recording lasted 30 seconds. The results showed a tendency ($p<.06$) for better accuracy and better ability to separate correct from incorrect identification responses by means of ones' confidence judgments for the text-lineup condition compared with the dialogue-lineup condition. The text-lineup condition also showed a tendency for lower overconfidence. These results deviate from expections following from the encoding specificity principle in memory psychology (Tulving & Thomson, 1973), maybe because text reading provides a more varied representation of the features of the human voice compared to dialogues.

Introduction

A speaker with a distinctive and characteristic voice and speech is easier to recognize compared to a speaker with less distinctive features referring to different studies presented in Hollien (2002). An earlier study, focusing on characteristic features in the speech and voice, reported that a foil with characteristic voice features, close to the target speaker, is confusing for listeners in a voice lineup (Zetterholm et al, 2009). This paper focuses on the effect of different speaking styles, namely read text and spontaneous speech, in a voice lineup. It is known that speakers with an unfamiliar dialect or accent sound more similar to a listener than speakers with a familiar dialect or accent (Yarmey, 1995). In the present study, as in the previous study (Zetterholm et al, 2009), all speakers, both the target speaker and the foils, had the same Swedish dialect in order to avoid confusion about the dialect.

When preparing an earwitness lineup the samples should be equal in length and Hollien (2002) points out that it might be good to use different types of voice samples, e.g. natural speech, text-dependent words, phrases and sentences as well as read text. Hollien also recommends between five to eight foils of approximately the same age as the suspect, or target voice, as well as speaking the same accent or dialect in general.

Previous research about speaking styles have used different data and the material has been classified in two main groups. *Connected speech* consisting of read speech and *spontaneous speech* consisting of speech produced in more or less unprepared situations with both professional and non-professional speakers (Llisterri, 1992). Some of the studies show that average F0 is higher in reading than in spontaneous speech. This is not confirmed in present study, see Table 1 and 2. However, the speech samples used in this study is not longer than 30 s. All speakers are non-professional.

The task presented to the participants in this study involves retrieval from memory. In this context it is relevant that Tulving and Thomson (1973) formulated a central conclusion from psychological research on memory: the *encodning specificity principle*. In this context retrieval from memory is seen to be occasioned by cues to (or in) memory which activate similar memory elements as the content in the cues. For

example, the content in a question will act as a cue to memory and will help to activate a possible answer to the question posed.

According to the *encodning specificity principle* we tend to remember better the more similar the cues provided are at the time of recall are to the encoded material that we want to retrieve. This means, for example, that the more similar the context of the occasion when the memory was encoded is to the context of retrieval, the more successful the retrieval is expected to be. The context is here taken to be the total situation that the memory retriever is in, especially aspects that are at the focus of the person's attention.

For example, in a recognition task situation such as a lineup, the more similar the features of one (or more) of the persons in a target present lineup are to the voice features of the culprit the greater is the chance for recognition. For this reason, in the present study we speculated that if the people in the lineup are presented in dialogue form when the originally encoded experience took the form of a dialogue then the chances for correct identification would be better compared with when the people in the lineup are presented in some other format, such as reading from a book.

In brief, the aim of this study thus was to analyze if there is a difference in the ability to recognize a voice in a voice lineup depending on the speaking style of the people being presented in the lineup.

Method

Participants

Seventy-eight students (54, females, 24 males) from Lund University participated in the study. The mean age was 26.5 years (range 20 - 65 years). Each participant received a lottery ticket worth 25 SEK. Two participants were removed from the data because they choose more then one option in the voice lineup.

Design

A between-subjects design with two conditions was used. The conditions differed with respet to the format of the presentations of the voices in the lineup. In the first condition, the *Text-lineup condition* (n = 38), each of the people in the lineup read the same passage from a book. In the second condition, the *Dialogue-lineup condition* (n = 43), each of the voices in the lineup was engaged in a dialogue.

Material

Original event dialogue. A 2 minutes long dialogue between two male speakers was recorded. The dialogue simulated the planning to break into a house by two burglars. The speakers were 27 and 22 year old respectively and both spoke with a Scanian dialect.

Voice lineups. Two kinds of voice lineups were used. A "text" voice lineup which consisted of six recordings of six male speakers reading a text, the target speaker and five foils. They all read the same text. A "dialogue" voice lineup which consisted of six recordings of the same six male speakers having a spontaneous dialogue with another male speaker. The discussion partner was the same for all the six speakers. The voice of the conversation partner was not audible in the lineups. They all engaged in a dialogue about the same news article. The duration of each recording, in both conditions, was about 30 seconds. All the speakers in both lineups had a Scanian dialect (i.e., the dialect of the southern-most part of Sweden) and were of almost the same age. The target speaker was present in the test. The results of the acoustic measurements of the mean fundamental frequency (F0) and the standard deviations (SD.) for both lineups are shown in Table 1 and 2 respectively. It is obvious that the there is almost no difference in mean F0 or SD. between the recordings with the different speaking styles.

Table 1. Age, F0 mean and standard deviations (SDs) for the target speaker and the five foils, in the text-lineup condition.

	Age	F0, mean	SD.
Target	22	125 Hz	21 Hz
Foil 1	19	99 Hz	15 Hz
Foil 2	21	121 Hz	21 Hz
Foil 3	23	86 Hz	12 Hz
Foil 4	23	123 Hz	26 Hz
Foil 5	22	90 Hz	16 Hz

Table 2. Age, F0 mean and standard deviations (SDs) for the target speaker and the five foils, in the dialogue-lineup condition.

	Age	F0, mean	SD.
Target	22	132 Hz	23 Hz
Foil 1	19	100 Hz	21 Hz
Foil 2	21	126 Hz	30 Hz
Foil 3	23	87 Hz	16 Hz
Foil 4	23	114 Hz	23 Hz
Foil 5	22	87 Hz	17 Hz

Confidence judgments. Participants rated their confidence in their lineup decision on an 11-point scale beginning at 0% ("Completely sure that I remember wrong") and ending at 100% ("Completely sure that I remember correct").

Procedure

The experiment was run in small groups with 2 to 5 participants in each group. Each group was randomized into one of the two conditions. The participants were received in the lab where they first signed the consent form. Next, the participants were told that they were going to listen to a dialouge between two men. They were instructed to just listen to the dialouge and that they would recieve further instructions afterwards. Participants listened to the 2 minutes long dialogue between two men planning to brake into somebody's home. After this, the participants participated in another experiment for about 15 minutes as a filler task.

The participants were then told that their task was to identify the speaker that that they had heard most in the original dialogue. Then the participants listened to the voice lineup relevant for their condition and each lineup was played twice. After listening to the voice lineup the listeners answered the question: "Do you recognize if any of these six voices is from the person from the dialouge that you listened to earlier?" If they could not identify anyone in the lineup as the culprit they could choose the option "do not recognize any of the voices". The participants were also told to be aware of that the voice they had heard in original dialogue might not be present at all in the lineup. After this, the participants gave their confidence judgement about the correctness of their idenfication decision on the 11-point confidence scale described above. Finally, the participants were given a lottery ticket, debriefed and thanked before leaving.

Measures

Apart from the accurcy in the identification responses (proportion correct of all identification responses) and the participants' confidence in the correctness of their identification responses, we also calculated three measures of the realism in the participants' confidence judgments. By the *realism* in the participants' confidence judgments we mean how veridical the confidence judgments were with respect to the correctness of the identification responses (this is sometimes called the participants' metamemory realism).

Two of these measures, *calibration* and *over-/underconfidence*, concerned the relation between the level of the participants' confidence judgments and the proportion correct identification responses.

Calibration is calculated by first dividing a person's confidence judgments into different confidence classes based on the level of confidence (11 confidence classes were used, since the participants were allowed to use 11 confidence levels, that is, 0%, 10%, 20%, etc). Calibration is computed by the following formula:

$$\text{Calibration} = 1/n \sum_{t=1}^{T} n_t \, (r_{tm} - c_t)^2$$

Here n is the total number of responses rated, T is the number of confidence classes used and n_t is the number of responses within confidence class r_t , r_{tm} is the confidence level of the confidence class r_t, and c_t is the percent of correct responses within the confidence class t. For each confidence class the percent of correct answers within that class was thus subtracted from the mean level of confidence within that class. This difference was squared and multiplied with the number of times this confidence class was used by the listners. The resulting product was then summed over the corresponding products for the other confidence classes and finally this sum was divided by the total number of responses (for further details see e.g., Yates, 1994).

Over-/underconfidence is computed by subtracting the listners' average proportion correct responses (in percentage) from their average confidence level for all responses (also

in percent). Just as for calibration, the result zero expresses perfect realism, in the sense that there is no over-/underconfidence. A negative value indicates underconfidence and a positive value indicates overconfidence.

We also used a measure of the listners' ability to discriminate correct from incorrect identifications by means of their confidence judgments. For this purpose we used a measure called *resolution* which is computed as:

$$\text{Resolution} = 1/n \sum_{t=1}^{T} n_t (c_t - c)^2$$

Here, c is the proportion of all items for which the correct identification response was given. A higher value reflects better resolution than a lower.

Results

Calibration curves

Figure 1 shows that calibration curves for the text-lineup condition and the dialogue-lineup condition. The x-axis shows the eleven different confidence levels (from 0 to 100%) and the y-axis shows the percent of correct answers, but in the graph the data-points have been reduced to five (0 %, 10 - 40 %, 50 %, 60 - 90% and 100 %). This reduction was done in order to smooth the calibration curves since the number of listners was small. The numbers inside the graph give the number of answers for each of the five reported confidence level in each condition. The diagonal shows perfect calibration. As can be seen in Figure 1, the calibration curves for the text-lineup and dialogue-lineup conditions show a difference in that the text-lineup condition evidence less overconfidence compared with the dialogue condition, However, at the 50% level there was no difference between the conditions and at the confidence levels below 50 % there is even a hint of underconfidence for the text-lineup condition.

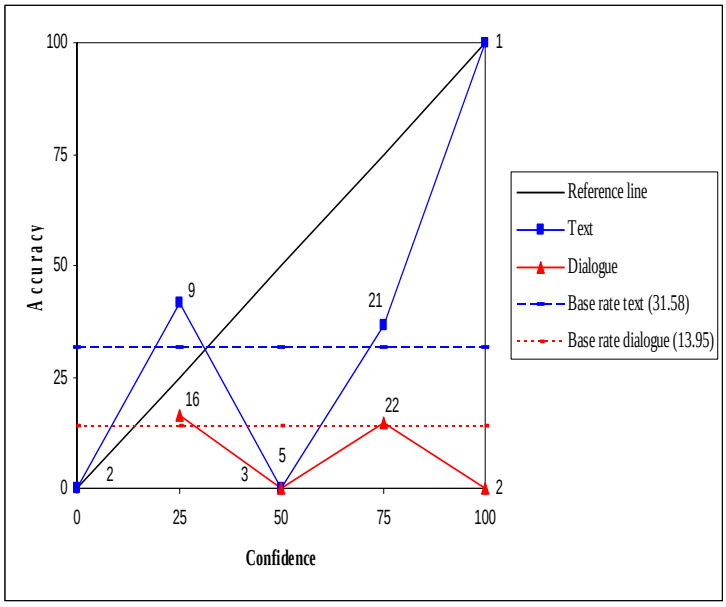

Figure 1. Calibration curves for text-lineup condition (squares) and dialogue-lineup condition (triangles). Digits at each point show the number of listners in the condition that used this confidence level.

Only 17 listners selected the target in the lineup, which is 21 % of all listeners. The average confidence level of these 17 listners was 64 %, which can be compared with the average confidence level 53 % for the remaining 79 % listners who selected the foils in the lineup. There is no significant difference between the mean confidence level of the listners who selected the target and those listners who selected the foil.

To study the relationship between age and accuracy a point-biserial correlation was computed. There was a very weak correlation between the two variables (r = .12, n.s, n = 76). In order to study the relationship between age and confidence a Pearson correlation was computed that similarly showed a very weak relationship between age and confidence (r = -.03, n.s., n = 76).

To test the importance of gender in voice identification we used t-tests to compare the males and females on their accuracy, confidence, calibration, over-/underconfidence and resolution scores but no differences were found.

The listners' scores in the two conditions were also submitted to independent-sample t-tests. The results are shown in Table 3. It can be seen that there was a trend towards a significant

difference between the conditions for accuracy $t(79) = 1.92$, $p < .06$ and for resolution $t(79) = 1.92$, $p < .06$, meaning that the participants in the text-lineup condition showed better performance than the participants in the dialogue-lineup condition for accuracy and resolution. Furthermore, a trend towards a significant difference between the two conditions was also found for over-/underconfidence. The participants in the dialogue-lineup condition showed a tendency towards greater overconfidence compared with the participants in the text-lineup condition. No significant differences were found between the text-lineup condition and the dialogue-lineup condition for confidence and calibration. For accuracy and resolution the Levene's test showed that equal variance for these two variables cannot be assumed, but the t-value for accuracy and resolution was same both with and without assuming equal variance and both t-values were significant at $p < .06$ level, as shown in Table 3.

Table 3. Accuracy, confidence, calibration, over/underconfidence, and resolution for the text-lineup condition and the dialogue-lineup condition.

	Text-Lineup	**Dialogue-lineup**	**t**	**p**
Acc.	31.58(47.11)	13.95(35.06)	1.92	.06
Conf.	55.26(27.09)	54.88(26.58)	.06	.95
Calib.	.28(.25)	.36(.28)	-1.42	.16
O-/u.	.24(.48)	.41(.44)	-1.68	.09
Res.	.32(.47)	.14(.35)	1.92	.06

Note. Acc. = Accuracy, Conf. = Confidence, Calib. = Calibration, O-/u. = Over-/under-confidence, Res. = Resolution.

Discussion and conclusions

In the present study we tested whether voice-lineup recordings resulted in better identification performance when the lineups were presented in the form of the voices reading from a book (the text-line condition) or voices participating in a dialogue (the dialogue-lineup condition). Our expectation was that the lineup-condition using the dialogue presentations whould result in the better performance. This expectation was based on the encoding specificity principle presented by Tulving and Thomson (1973) which says that the accuracy in memory retrieval should be better when the there is a better match between the features of the encoding situation (including the focussed "object") and the features of the retrieval situation (included the attended-to "object").

In contrast to our expectation, the results showed that it was the listners in the text-lineup condition that showed the best performance of the two conditions. However, this is not the first time that the encoding specificity principle has been challenged by empirical data (e.g., Bower & Mayer, 1989; Higham 2002). For example, Bower and Mayer reported six experiments that failed to show stable evidence for mood-dependent retrieval.

In line with previous research on voice lineups (e.g., Olsson, Juslin, & Winman, 1998; Yarmey, 2007) the level of overconfidence evidenced by the listners in this study was quite high (especially in the dialogue condition) if compared to what is reported in research on eyewitness lineups. It remains a task for future research to provide information as to why this is the case.

As noted above, the present contribution reports the results from the first 78 listners of the at least 200 listners that we plan will participate in the study when it is completed. A further limitation of the results in the present report is that it is not clear what the effect was of the fact that the recordings in the dialogue-lineup condition did not include the conversation partner, only the person being part of the lineup. This was done so that the listner should be able to concentrate on the voice of the person in the lineup and so that the listner should not be confused about which voice we wanted them to respond to. Moreover, our study only used target-present voice lineups. In future research also target-absent voice lineups should be investigated with the two conditions used in the present study.

Given that our results hold up in our remaining data-collection and in future research it is of great interest to investigate why the presentation of text recordings in voice lineups lead to better indentification performance, including better meta-memory performance. One speculations is that this effect, if it is real, is due to that text reading provides more varied and representative information about a speaker's

voice compared with a voice participating in a dialogue.

Acknowledgements

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

References

Bower GH, Mayer JD (1989). In search of mood-dependent retrieval. In D. Kuiken (Ed.), Mood and memory: Theory research and applications. *Special issue of Journal of Social Behavior and Personality*, *4*, 121-156.

Higham PA (2002)/. /Strong cues are not necessarily weak: Thomson and Tulving (1970) and the encoding specificity principle revisited. *Memory & Cognition*, *30*(1), 67-80.

Hollien H (2002). *Forensic voice identification*. San Diego, CA: Academic Press.

Llisterri J (1992). Speaking styles in speech research. *ELSNET/ESCA/SALT Workshop on Integrating Speech and Natural Language*, Dublin, Ireland, 15-17 July 1992. http://liceu.uab.es/~joaquim/publications/SpeakingStyles_92.pdf.

Olsson N, Juslin P, Winman A (1998) Realism of confidence in earwitness versus eyewitness identification. *Journal of Experimental Psychology: Applied*, *4*, 101–118.

Tulving E, & Thomson DM (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*, 352-373.

Yarmey AD (1995) Earwitness speaker identification. Psychology, Public Policy, and Law 1, 792-816.

Yarmey AD (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 JF (1994). Subjective probability accuracy analysis. In G Wright & P Ayton (Eds.), *Subjective probability* (pp. 381-410). New York: John Wiley & Sons.

Zetterholm E, Sarwar F, Allwood CM (2009). Earwitnesses: The effect of voice differences in identification accuracy and the realism in confidence judgments. *Proceedings, Fonetik 2009*, Dept. of Linguistics, Stockholm University.