Different influences of the native language of a listener on speaker recognition

Olaf Köster and Niels O. Schiller†*

** University of Trier, Germany; †Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands*

ABSTRACT In forensic phonetics, lay or expert witnesses might be confronted with voice samples for auditory evaluation from a language they do not understand. In speaker identification experiments, it has been shown that knowledge of the target language affects recognition results. Köster *et al.* (1995) showed that German listeners and English listeners with a knowledge of German identified a German voice better than English listeners without knowledge of German. Replicating the same experiment with Spanish and Chinese listeners, the results of this study show that (a) Spanish and Chinese listeners with knowledge of German obtain significantly better recognition results than their compatriots with no knowledge of the target language, and that (b) Spanish and Chinese listeners with knowledge of German perform significantly worse than native Germans and English listeners with a knowledge of German. No clear evidence was found that the typological difference between the native language of the listener and the target language influenced recognition performance.

KEYWORDS forensic phonetics, speaker identification, voice line-ups, recognition memory for foreign voices, language familiarity, typological influences

INTRODUCTION

In aural-perceptual voice identification, a voice line-up can be important if an 'earwitness' is available. In these cases, a phonetically untrained listener has to recognize a voice he or she is familiar with from a set of different speakers. Increasingly, cases occur in which a lay earwitness or an expert witness is confronted with speech material from a language (target language) that he or she does not speak or only masters as a second language (Künzel, Huntley Bahr, personal communications). A reason for this situation is the disproportionately large percentage of crimes committed by foreigners. Furthermore, an expert witness might be called by a foreign court if no local forensic phoneticians are available in that particular country.

If lay or expert witnesses are confronted with voice samples for auditory evaluation from a language they do not understand, the question of the

1350–1771

Forensic Linguistics 4(1) 1997

reliability of their speaker recognition abilities arises. Irrespective of the fact that a listener may or may not know the target language, two further problems come up. Firstly, is there a difference in recognition performance between native-speaker listeners and those listeners who are only second or foreign-language speakers of the target language? Secondly, if the target language is unknown to the listener, does the extent to which the listener's native language and the target language are related influence recognition results?

Only a few experiments, with partly conflicting results, have been published about the interrelation between the listener's language and the target language in speaker recognition. Goldstein *et al.* (1981), who investigated the recognition of voices both with and without foreign accents, came to the conclusion that 'voice recognition is just as good (or as poor) for foreign voices as it is for native voices' (220). In contrast, three other studies show different results. According to Thompson (1987), monolingual English listeners identified English speakers significantly better than either Spanish speakers or English speakers with a Spanish accent. The data presented by Goggin *et al.* (1991) suggest that 'voice identification is increased approximately twofold when the listener understands the language relative to when the message is in a foreign language' (456). In another study, Köster *et al.* (1995) investigated what level of competence in the target language would have an effect on speaker recognition performance. In their experiment, three groups of listeners differing in their knowledge of the target language (German) had to recognize a speaker (whom they had been familiarized with before) from a set of 108 utterances. The comparison of German listeners (students at the University of Trier), English listeners with knowledge of German (exchange students of German at the University of Trier) and English listeners without knowledge of German (Canadian college students and teachers) revealed that 'subjects with knowledge of German performed generally better than subjects without any knowledge of German' (309). There was no difference between German native speakers and English listeners who had learned German as a second language.

These results lead to the hypothesis that in general a listener can recognize a speaker more reliably if s/he has command of the speaker's language. On the other hand, it does not seem to be relevant whether the listener speaks the target language as his or her native language or as a second language. In order to test these assumptions, more experiments using groups of listeners with different native languages (both with and without knowledge of the target language) need to be carried out.

In addition, whenever a listener has to recognize a speaker of a language s/he does not know at all, the typological distance[1] between the two languages might play a role in the forensic setting. The question is to what extent the typological distance between the listener's native language and the target language influences proficiency in speaker recognition. Is

it easier to recognize a foreign voice if the language is more similar to the native language of the listener?

EXPERIMENT

In order to extend the pilot study by Köster *et al.* (1995) and to check the hypotheses mentioned above, a new experiment was carried out. The design of this test was exactly the same as that described by Köster *et al.* (1995) so that a comparison of the results is possible. In a direct identification task, Spanish and Chinese listeners with and without knowledge of German were asked to identify a German target speaker among five foils.

Subjects

A group of seventy-two subjects participated in the investigation. All participants were aged between seventeen and twenty-six years (m = 20.48, SD = 1.98). Among the subjects, there were fifty-two female and twenty male listeners. In the evaluation, no further differentiation between the gender groups was made.

Subjects were divided into four different groups according to their nationality and their knowledge of the target language, German. Since not only the effect of the difference between subjects with/without knowledge of the target language but also the effect of typological differences of the listeners' native languages was to be tested, the following groups were selected: (1) Spanish speakers with no knowledge of German at all (n = 16); (2) Spanish speakers with a knowledge of German (n = 10); (3) Chinese speakers with no knowledge of German (n = 23); (4) Chinese speakers with a knowledge of German (n = 23). To extend the testing of the effect of typological difference, English speakers with no knowledge of German were added from another experiment (see Köster *et al.* 1995). Consequently, one group of subjects (English) consisted of listeners of the same language family (West Germanic languages) as the reference language (German, see Speech material, below). Spanish is generally considered to be less related to German than English is, as it is a Romance language. Therefore, in a recognition experiment the group of Spanish speakers (with no knowledge of German) might be expected to obtain (slightly) worse recognition results than the English group (with no knowledge of German). Furthermore, as German, English and Spanish are all Indo-European languages, the subjects of the third language, Chinese, which is typologically very different (a Sino-Tibetan as well as a tone language) could be expected to perform significantly worse than any other group.

Most of the Spanish speakers with a knowledge of German were students of German at the University of Santiago de Compostela, Spain; others

were exchange students at the University of Trier. The Spanish speakers with no knowledge of German were students in Santiago de Compostela. The Chinese speakers with a knowledge of German were students of this language at the II. University for Foreign Languages, Beijing. Native Chinese listeners without knowledge of German were students at the same university. All participants belonged to a very homogeneous age-group. Group 1 subjects were between nineteen and twenty-four years old (m = 21.81, SD = 1.22); members of group 2 were between twenty-two and twenty-six years of age (m = 23.74, SD = 1.49). Chinese participants (groups 3 and 4) were between seventeen and twenty (m = 18.74, SD = 0.72) and between eighteen and twenty-one years of age (m = 19.68, SD = 0.76) respectively. All subjects were phonetically naive listeners and took part in the experiment voluntarily. None of them reported any hearing problems.

Speech material

The speech material used in this investigation was exactly the same as in the above-mentioned study by Köster *et al.* (1995). In order to obtain a homogeneous set of utterances, six different male speakers of standard German from the same geographical and dialectal area (having a slight Hessian accent) read a German text of approximately one minute's duration. The readings were tape recorded (for details of the recording procedure see Künzel 1989). All speakers were of a similar age (m = 29.67, SD = 5.45). Their average fundamental frequency for the passage used in the experiment ranged from 86 Hz to 142 Hz (m = 109.5, SD = 18.7). The test tape was structured as follows: three parts of the text between four and eight seconds in length were spliced out of each speaker's recording giving three utterances for each of the six speakers. In the next step, the eighteen digital speech samples were reproduced under telephone transmission conditions by re-recording exactly the same material through a telephone line. These thirty-six parts were then copied three times giving, in total, 108 utterances (6 speakers x 3 parts of the text x 2 transmission conditions [hifi vs. telephone] x 3 repetitions). Finally, the samples were randomized and copied to the test tape.

One speaker whose verbal behaviour was not in any way 'marked' was designated as the target voice (speaker X). Speaker X's complete original hifi text was recorded five times on DAT to obtain a speech sample which had a duration of approximately five minutes.

Method

The method used was the same as in the Köster *et al.* (1995) study. The four listener groups (Spaniards with and without a knowledge of German, Chinese with and without a knowledge of German) were tested in separate

sessions. At the beginning, subjects were familiarized with the target voice by listening to the five-minute sample of speaker X. Subjects were instructed to listen to the utterances carefully and to memorize the voice in order to recognize it in a identification task later on. After the familiarization stage response sheets were handed out and the subjects read the instructions; they were told to listen carefully to the tape with the 108 speech samples and to mark 'yes' after any sample they thought had come from speaker X and 'no' after any other (a forced-choice test). The time lag between the familiarization stage and the test was approximately five minutes.

In order to keep demands on the subjects within reasonable limits, the experiment did not exceed forty minutes (including the familiarization process). The stimuli were offered with a response-interval of five seconds, which the subjects considered long enough to make a decision. After every tenth speech sample, there was a 300 Hz pure-tone signal to help subjects to keep track of the stimuli.

All experiments were carried out by the authors themselves, except for the experiment in China. In Bejing, Prof. Tang Lunyi followed detailed written instructions in order to provide the same experimental conditions as described above. In all cases, it was made sure that the participants perceived the speech samples well.

RESULTS

In general, two different error categories as well as correct answers can be distinguished. Subjects can (a) identify the target speaker among the samples (hit), (b) they can reject speaker X when it actually was the target voice (miss), (c) they can identify a speech sample which in fact came from one of the dummy speakers (false alarm) and (d) they can correctly reject a sample which in fact was not produced by speaker X (correct rejection). Overall performance of identification was calculated using Signal Detection Theory (Macmillan and Creelman 1991; McNicol 1972).

This performance of identification can be expressed by the sensitivity measure d' and the response bias c. Hits (H) and false alarms (F) were pooled across participants in each of the four groups, and for each group d' was determined (Macmillan and Kaplan 1985) (for a more detailed description of the statistical procedure see Schiller *et al.* 1997). This analysis resulted in d' values of 1.191 for group 1 (H = 0.65; F = 0.21), 2.681 for group 2 (H = 0.87; F = 0.06), 2.147 for group 3 (H = 0.79; F = 0.09), and 2.425 for group 4 (H = 0.75; F = 0.04) (see Figure 1). Since positive d' values indicate that subjects are sensitive regarding the discrimination between target voice samples and foils all groups generally performed better than chance level ($d' = 0$).

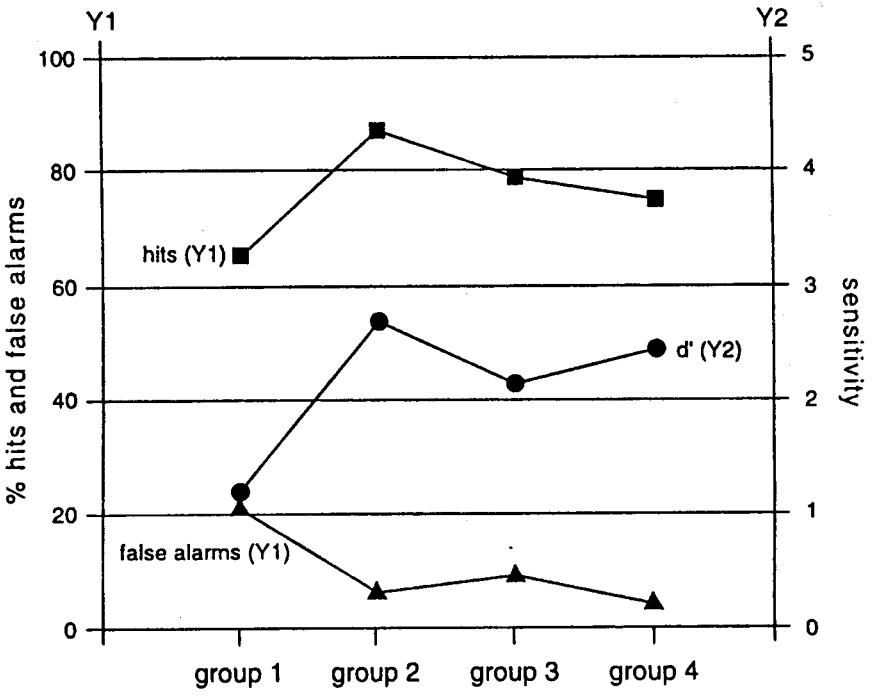

Figure 1 Hit rate (H), false-alarm rate (F), and sensitivity (*d'*) for the four listener groups

Pairwise comparisons between the respective groups showed that, in terms of identification sensitivity, the difference between group 1 (Spanish, without German) and group 2 (Spanish, with German) as well as between group 3 (Chinese, without German) and group 4 (Chinese, with German), was significant ($p < 0.05$). This means that both Spanish and Chinese listeners recognized the German target speaker better if they had knowledge of German than if they had no knowledge of the target language. On the other hand, although group 2 (Spanish, with knowledge of German) and group 4 (Chinese, with knowledge of German) did not differ significantly from each other, both groups performed significantly worse than English listeners with knowledge of German and native German listeners (see Schiller and Köster 1996).

A comparison between groups incorporating listeners who had no knowledge of the target language revealed that group 3 (Chinese listeners without German) performed significantly better than group 1 (Spanish listeners without German). Group 3 also performed better than the English listeners without knowledge of German (see Schiller and Köster 1996). Spanish speakers without a knowledge of the target language obtained worse recognition results than both the Chinese and the English subjects.

According to signal detection theory, the response bias of the different groups is expressed by the *c* values. This measure indicates whether subjects disproportionately tended to mark 'yes' (yes-bias) or to mark 'no' (no-bias) on the answer sheet. The *c* values were 0.2105 for group 1, 0.2145 for group 2, 0.2675 for group 3, and 0.5385 for group 4. The *c* value for native English listeners without knowledge of German was 0.563; for native English listeners with a knowledge of German it was 0.3245, and similarly for native Germans it was 0.3245 (Schiller and Köster 1996). Since a positive *c* value indicates a no-bias all groups disproportionately tended to mark 'no'. One possible explanation for the general no-bias in our experiment could be the fact that the test tape incorporated five times as many foils as target voice samples. Therefore, subjects might have been influenced by the predominance of foils.

Statistical comparisons between the groups revealed that all groups differed significantly from each other with the exception of Spaniards with versus without knowledge of German; Spanish versus Chinese listeners without knowledge of German; and Spanish versus English listeners with no knowledge of German.

As in the pilot study (Köster *et al.* 1995), the difference of the recognition results between high-fidelity speech samples and telephone-transmitted speech samples was also tested. The respective *d'* values were (hifi vs. telephone): Spaniards with no knowledge of German 1.379 vs. 1.028; Spaniards with a knowledge of German 2.666 vs. 2.746; Chinese with no knowledge of German 2.441 vs. 1.865; Chinese with a knowledge of German 3.008 vs. 0.964. Further measurements of German and English groups (Schiller and Köster 1996) showed the following *d'* values: English listeners with no knowledge of German 2.182 vs. 1.241; English listeners with a knowledge of German 3.459 vs. 3.286; native German listeners 3.501 vs. 3.632. Statistical analyses revealed that almost all groups performed significantly better when good quality speech samples were judged (with the exception of Spanish and English listeners with knowledge of German where there was no significant difference).

DISCUSSION

In contrasting English listeners without knowledge of German with compatriots having knowledge of German and native German listeners, Köster *et al.* (1995) found that 'unfamiliarity with the target language affects the ability to recognize a speaker, as subjects with knowledge of German performed generally better than subjects without any knowledge of German' (309). This study, replicating that first experiment with Spanish and Chinese listeners, clearly confirms these results. In all cases (English, Spanish, Chinese), subjects with a knowledge of the target language (German) were able to identify a German speaker better than subjects without any

knowledge of the target language (pairwise comparisons). It can generally be concluded that in speaker recognition tasks knowledge of the target language increases the reliability of recognition results. Accordingly, if no linguistic information on the target language is understood, recognition results are poorer. Therefore, results of voice line-ups involving speech samples in a language which the witness does not understand should be handled with caution.

As far as the *degree* of knowledge of the target language[2] is concerned, Köster *et al.* (1995) came to the conclusion that 'the degree of knowledge of the target language seems to be of less relevance because group 3 [English listeners with knowledge of German] and 4 [German listeners] performed equally well' (309). In the present experiment, both Spanish and Chinese listeners speaking German were less successful at recognizing the German target speaker than Germans and English listeners with knowledge of German. Therefore, the correlation between the degree of competence in a language and performance in voice line-up experiments remains unclear. The question of whether a listener has to be a native speaker of the target language or if competence in the target language as a second language is sufficient cannot be answered clearly.

While, in general, a clear positive correlation between knowledge of the target language and speaker recognition performance exists, statements about situations in which the target language is *un*known seem to be more difficult. For an unknown target language it might be hypothesized that the closer it is related to the listener's native language typologically the better s/he will recognize a speaker of that language. One reason for this assumption is that the more similar the target language and the listener's language are (i.e., the more segmental and suprasegmental similiarities which exist), the more parameters the listener might be able to extract in order to remember the speaker. Thus, in a voice line-up subjects should perform better the more their native language is related to the target language even if no semantic information is understood. According to the typological relatedness of the languages involved in our experiment, English subjects could be expected to perform better than Spanish and Chinese subjects; Spanish subjects could be expected to perform better than the Chinese group.

However, the experimental results did not support these hypotheses. It is true that Spanish listeners recognized the target speaker less successfully than English listeners did, but Chinese participants performed significantly better than both Spanish and English subjects. From a linguistic point of view, there seems to be practically no explanation for this result. While the fact that English listeners without knowledge of German showed better recognition results than Spaniards without knowledge of German supports the hypothesis, the superior proficiency of the Chinese group remains striking. All groups were homogeneous as far as gender and age[3] are concerned. As far as the response bias of the subjects is concerned, this

measure cannot explain the unexpectedly good recognition results of the Chinese listeners either. The *c* values of the Chinese group do not differ from Spanish listeners with no knowledge of German, but they are significantly different from English listeners with no knowledge of German. We speculate that the results of the Chinese participants can possibly be explained by a special ability of Chinese speakers to recognize intonation phenomena. However, from a linguistic point of view, further investigations about language processing in tone languages are necessary.

In summary, both the experiment by Köster *et al.* (1995) and this study showed that, in forensic speaker recognition tasks such as voice line-ups, more reliable results can be expected if a listener ('earwitness') who is a native speaker of the target language or who knows it as a second language is involved. If, in a voice line-up, the listener does not understand the speaker's language, recognition results should be treated with care. At this stage, there does not seem to be evidence that recognition performance is correlated with typological difference. This is at least true for the taxonomically selected languages English, Spanish, Chinese and the reference language German. Furthermore, this experiment as well as the pilot study by Köster *et al.* (1995) shows that recognition results are degraded if the quality of the speech samples is poor (e.g., recorded over a telephone line).

It must be pointed out that the experimental results only refer to linguistically and phonetically naive listeners. Experts trained in forensic phonetics perform much better than lay witnesses. In another experiment by Köster and Schiller (in preparation) the same speech material and the same method were used to test German experts working in the field of forensic speaker identification. When compared to the German control group consisting of phonetically naive subjects (Köster *et al.* 1995) the experts recognized the target speaker significantly better. Nevertheless, the 'Code of Practice' of the International Society for Forensic Phonetics (IAFP) also advises expert witnesses to be extremely cautious in cases where speech samples from a language which is different from the expert´s native language have to be evaluated.

ACKNOWLEDGEMENTS

The authors wish to thank Prof. W. Barry (Universität des Saarlandes, Saarbrücken) for helpful comments on the paper and Prof. Tang Lunyi (II. University for Foreign Languages, Beijing) for his help in carrying out the experiments with the Chinese listeners. The authors take responsibility for all errors remaining.

This study was supported by a grant from the International Association for Forensic Phonetics.

NOTES

1 The typology of a language incorporates different grammatical features. The initial definition of language typology emphasizes morphological criteria. To characterize a language more completely, nowadays phonological and syntactic aspects are also taken into account. Language typology generally describes how meaning is encoded on different grammatical levels. According to typological criteria, languages are divided into language families.

2 By the term 'degree' we want to differentiate between the knowledge of a language as a native language and the knowledge of this language as a second language. Generally, a person can be assumed to have command of his/her native language at a higher degree than s/he will have of a second language.

3 'The influence of the listeners' age on performance in speaker recognition remains rather unclear.' (Köster *et al.* 1995: 309; see also Künzel 1990: 54)

REFERENCES

Goggin, J. P., Thompson, C. P., Strube, G. and Simental, L. R. (1991) 'The role of language familiarity in voice identification', *Memory and Cognition*, 19: 448–58.

Goldstein, A. G., Knight, P., Bailis, K. and Conover J. (1981) 'Recognition memory for accented and unaccented voices', *Bulletin of the Psychonomic Society*, 17: 217–20.

Köster, O., Schiller, N. O. and Künzel, H. J. (1995) 'The influence of native-language background on speaker recognition', in K. Elenius and P. Branderud (eds) *Proceedings of the Thirteenth International Congress of Phonetic Sciences*, vol. 4, Stockholm: 306–9.

Köster, O. and Schiller, N. O. (in preparation) 'Expert witnesses and their ability to identify foreign voices'.

Künzel, H. J. (1989) 'How well does average fundamental frequency correlate with speaker height and weight?', *Phonetica*, 46: 117–25.

Künzel, H. J. (1990) *Phonetische Untersuchungen zur Sprechererkennung durch linguistisch naive Personen*, Stuttgart: Steiner.

Macmillan, N. A. and Kaplan, H. L. (1985) 'Detection theory analysis of group data: estimating sensitivity from average hit and false alarm rates', *Psychological Bulletin*, 98: 185–99.

Macmillan, N. A. and Creelman, C. D. (1991) *Detection Theory: A User's Guide*, Cambridge: Cambridge University Press.

McNicol, D. (1972) *A Primer of Signal Detection Theory*, London: Allen and Unwin Ltd.

Schiller, N. O. and Köster, O. (1996) 'Evaluation of a foreign speaker in forensic phonetics: a report', *Forensic Linguistics*, 3(1): 176–85.

Schiller. N.O., Köster, O. and Duckworth, M. (1997) 'The effect of removing linguistic information upon identifying speakers of a foreign language', present volume, 1–17.

Thompson, C.P. (1987) 'A language effect in voice identification', *Applied Cognitive Psychology*, 1: 121–31.