Evaluation of a foreign speaker in forensic phonetics: a report

Niels O. Schiller[1] *and Olaf Köster*[2]

[1] *Max Planck Institute for Psycholinguistics, Nijmegen*
[2] *Institute of Phonetics, University of Trier*

ABSTRACT Expert witnesses in phonetics find themselves more and more often in forensic situations in which they have to identify the voice of a speaker who does not speak their native language. Until recently, little has been known about the role which the native-language background of the listener plays in such speaker identification tasks. In this report, several aspects of an experimental investigation on the influence of native-language background on speaker identification are reviewed. Results of a first experiment are reported and some follow-up experiments currently being carried out are described within that context.

KEYWORDS Speaker identification; foreign language processing; forensic phonetics.

INTRODUCTION

Two tasks generally can be distinguished within the field of forensic speaker recognition. In one case, an expert witness often has to compare an anonymous voice sample with that of a known speaker ('speaker identification'). If no reference sample is available, the ear witness has to describe as many features of the incriminating voice sample as possible ('voice profiling'). On the other hand, phonetically naïve listeners sometimes have to judge the identity of speakers from their voices in a so-called 'voice lineup'. Further, it may be the case that lay or expert witnesses have to identify speakers of a foreign language. For example, a voice sample may have been produced by a foreigner (either in a completely different language or with a strong foreign accent) or an expert witness is called by a foreign court and has to work abroad.[1]

In such cases, a question arises concerning the degree to which the native-language background of the witness influences his or her ability to recognize a speaker auditorily. Very little empirical data are available on this topic to date. Goldstein *et al.* (1981) reported an experiment in which subjects (native American English listeners) were asked to identify voices with and

1350–1771

Forensic Linguistics 3 (1) 1996: 176–185

without a foreign accent. Results showed that they performed equally well in recognizing accented and non-accented voices. The authors concluded that 'voice recognition is just as good (or as poor) for foreign voices as it is for native voices' (page 220).

In contrast, other studies showed conflicting results. Thompson (1987) found that monolingual English listeners identified English speakers significantly better than they did either Spanish speakers or English speakers with a Spanish accent. In another study, Goggin *et al.* (1991) conclude from their data that 'voice identification is increased approximately twofold when the listener understands the language relative to when the message is in a foreign language' (page 456). A recent study by Köster *et al.* (1995) also supports these results. When they tested different groups of subjects, varying in the degree of their knowledge of the target language, they found that listeners with knowledge of the target language performed significantly better than those without such knowledge. The authors conclude that 'speaker recognition does not only involve purely phonetic features, but also incorporates linguistic information' (page 309). Their data also suggest that witnesses' level of knowledge of the target language (native versus non-native) seems to play little role in speaker identification. To further assess the results of the above-mentioned studies, additional experiments are necessary to test the effects of (1) the 'linguistic factor', i.e., the phonetic/linguistic distance between the native language of the listener and the target language, and (2) the 'listener factor', i.e., the dependency of the performance in speaker recognition on the degree of phonetic/linguistic knowledge of the listener.

The role of native-language background

Experimental manipulation is the most effective way to assess the effect native-language background has on speaker identification. Appropriate experiments allow one to control for those variables which might play a role in the speaker identification process, and to investigate them one at a time. Relevant variables include, for example, the degree of familiarity with the target language (native language versus non-native language), phonetic experience (expert witnesses versus naïve listeners), relatedness of the target and native languages (languages from the same language family versus from different language families), quality of the speech material (high fidelity versus telephone transmission), listeners' age (young versus elderly people), and listeners' sex (see Hollien 1990 for a list of features useful in speaker identification). Some of these factors already have been investigated by others; for instance, Künzel (1990), studies the familiarity of the listener with the target voice, the listener's age and the quality of the speech material. We deviate somewhat from this pattern and concentrate on factors such as the degree of familiarity with the target language and the relatedness of target and native language.

METHOD

Native-language background: first experimental evidence

In 1995, Köster, Schiller and Künzel evaluated the recognition ability of listeners' with different native-language backgrounds by means of a direct identification test in which three different groups of listeners were asked to identify the voice of one speaker from a set of six different speakers ('closed test'). This report involves an extension of this research.

Subjects

Subjects consisted of 53 female and 21 male listeners (N=74); they were divided into three different groups according to their knowledge of German. The first group consisted of native (American) English listeners without any knowledge of German (this group is further subdivided by age: subgroup 1a consisted of subjects 30 years and older and subgroup 1b of subjects under 30). The second group included native English listeners who had some knowledge of German, and the third group consisted of native German listeners who served as control subjects. All subjects took part in the experiment voluntarily. None of them reported any speech or hearing problems.

Speech material

The speech material used in the experiment was produced by six different male speakers. Speakers were of similar age (M=29.7 years, SD=5.45) and spoke Standard German with Hessian influences. Mean F_0 ranged from 86 Hz to 142 Hz (M=109.5, SD=18.7). All speakers read a German text of approximately one minute in length onto a DAT recorder. Then, three parts of the text, each between four and eight seconds in length, were spliced onto experimental tapes. To record exactly the same material under telephone transmission conditions, the speech samples were recorded again through a telephone line with each of the six re-recorded three times. In total, we obtained 108 speech samples.[2] All samples were randomized and re-recorded on DAT. One speaker was designated as speaker X, the target voice. From speaker X, the high-fidelity text was re-recorded on DAT five times to obtain a speech sample of approximately five minutes.

Procedure

Trials for all groups of listeners were conducted individually. First, subjects were familiarized with the voice of speaker X by listening to his five minute sample. While doing so, they were instructed to concentrate on his voice in order to try to memorize it. After this familiarization had been completed,

response sheets were handed out to the listeners; they then were administered a forced-choice test. Specifically, they were instructed to listen to the tape carefully, and after each sample to mark 'Yes' if they thought the voice was from speaker X and 'No' if it was not. Five second intervals were placed between each stimulus and a sinusoid of 300 Hz was placed after every tenth sample to help them keep track of the task.

RESULTS

The design of the experiment allowed differentiation between two error categories: subjects could either reject the target voice when it actually was produced by speaker X (false rejection; FR) or identify a speech sample as the target voice when it was in fact produced by one of the foil speakers (false identification; FI). Furthermore, FRs and FIs were split into errors made under the high fidelity versus telephone conditions to see if there was a transmission effect.

The false rejection versus false identification rates were contrasted by group. First, group 1a made 67 FRs (M = 4.4, SE = 2.5) and 256 FIs (M = 17.07, SD = 14.11) whereas Group 1b made 141 FRs (M = 5.88, SD = 5.18) and 163 FIs (M = 6.79, SD = 8.09). Moreover, there were 26 FRs (M = 1.44, SD = 2.43) and 39 FIs (M = 2.17, SD = 4.07) for Group 2 and 24 FRs (M = 1.41, SD = 1.97) and 37 FIs (M = 2.18, SD = 2.71) for Group 3. The respective error proportions are provided by Figure 1.

The performance in identification is expressed by the sensitivity measure d' and the response bias c as suggested by *Signal Detection Theory* (Macmillan and Creelman 1991). Hits and false alarms were pooled across participants in each group, and for each group d' was determined (Macmillan and Kaplan 1985). The respective d' values were 1.552 for group 1a (c = 0.102), 1.684 for group 1b (c = 0.563), and 3.459 for group 2 (c = 0.325), and 3.459 for group 3 (c = 0.325). Statistical comparisons between the groups (95 per cent confidence interval around the difference in sensitivity between two groups) revealed that the difference in identification sensitivity between group 1a and 1b was not significant. However, the response bias for the two groups was significantly different ($p < 0.05$). Group 1a and 1b were significantly different from Group 2 and 3 both in terms of identification sensitivity and response bias ($p < 0.05$). Group 2 and 3, however, were not significantly different from each other in identification sensitivity nor in response bias.

From the above-mentioned it follows that Groups 1a and 1b which had no knowledge of the target language showed a significantly worse sensitivity to identify the German target speaker than the two groups with knowledge of German (Groups 2 and 3). The sensitivity between Groups 1a and 1b and between 3 and 4, however, was not significantly different, suggesting that the degree of knowledge of the target language does not play an important role in identification.

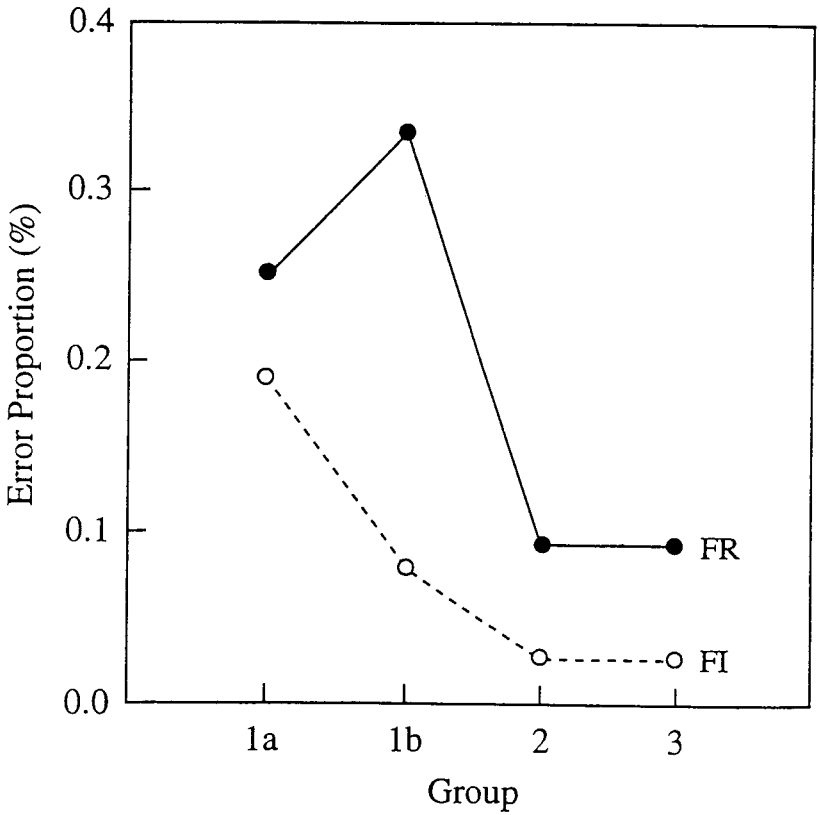

Figure 1 Error proportions for false rejections (FR) and false identifications (FI) respectively

Tables 1 and 2 show the distribution of FRs and FIs in the different transmission conditions. Most often participants made more errors in the telephone than in the high fidelity condition. The respective d' values in the high fidelity and in the telephone transmission condition, respectively, were 1.914 (c_{hifi} = 0.079) and 1.235 ($c_{telephone}$ = 0.1215) for group 1a, 2.182 (c_{hifi} = 0.385) and 1.241 ($c_{telephone}$ = 0.7215) for group 1b, 3.459 (c_{hifi} = 0.325) and 3,286 ($c_{telephone}$ = 0.238) for Group 2, and 3.501 (c_{hifi} = 0.576) and 3.632 ($c_{telephone}$ = −0.065) for Group 3. Statistical comparisons revealed that Groups 1a and 1b did not differ significantly from each other with respect to identification sensitivity in both transmission conditions, nor did Groups 2 and 3. However, the differences between the groups

Table 1: False rejections (FRs) per transmission condition

Group	*FR's (total)*	*Hi-Fi (H)*	*Telephone (T)*	*Ratio H:T*
1a	67	25	42	1:1.68
1b	141	44	97	1:2.21
2	26	13	13	1:1.00
3	24	19	5	1:0.26

Table 2: False identifications (FIs) per transmission condition

Group	*FI's (total)*	*Hi-fi (H)*	*Telephone (T)*	*Ratio H:T*
1a	256	99	157	1:1.59
1b	163	71	92	1:1.30
2	39	16	23	1:1.44
3	37	5	32	1:6.40

without any knowledge of the target language (Groups 1a and 1b) and the groups with knowledge of the target language (Groups 2 and 3) were significantly different from each other in both transmission conditions ($p < 0.05$).

These results show that with respect to the transmission conditions the identification sensitivity is generally higher in the hifi than in the telephone transmission conditions. Only for Group 3 sensitivity is slightly better in the telephone transmission condition. Furthermore, the main result obtained above showing that groups without any knowledge of the target language performed significantly worse than groups with knowledge of German also holds for the different transmission conditions.

DISCUSSION

The statistical analyses revealed that there was a main effect of group in the speaker recognition task. The results indicate 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. It seems that speaker recognition does not only involve purely phonetic features, but also incorporates linguistic information. The results further permit the interpretation that the degree of knowledge of the target language seems to be of but minor relevance because Group 2 and 3 performed equally well.

The influence of the listeners' age on the performance in speaker recognition remains rather unclear. Whereas the younger subjects without any

knowledge of German (Group 1b) made fewer FRs than the older ones (Group 1a), the situation is reversed with respect to the FIs; here, Group 1a made significantly more errors than Group 1b. This last result is in accord with Künzel (1990:54) who found that the amount of FIs rose with increasing age.

The effect of the acoustic quality of the speech samples was investigated by recording them both under high fidelity and telephone transmission conditions. The speech signal is reduced to the bandwidth interval between 300 and 3400 Hz when transmitted over German telephone lines and contains additional noise. On the whole, performance was worse when the speech sample was recorded via the telephone. The only exceptions were the ratios of Groups 2 and 3 for the FRs (see again Table 1). This finding leads to the interpretation that the acoustic quality of a speech sample is very important for speaker recognition purposes. It seems that speech samples recorded via the telephone lose some of the speaker specific features that aid in voice recognition. On the whole, these results suggest that there is an effect of native-language background in speaker identification.

To re-test these findings the authors are now developing a control experiment on the following bases. If there actually is an effect of native-language background in speaker identification, it must be due to the linguistic information in the speech material. This postulate would imply that listeners base their decisions about the identity/non-identity of two voices in part on linguistic information (if such information is available), i.e., if they have some knowledge of the language under consideration. If listeners relied on purely phonetic (acoustic) information, then a significant effect should not have been found between groups of listeners who knew German and those who knew no German. To test this suggestion, a control experiment was designed where native German speakers[3] are asked to read a 'text' that consisted only of combinations of the syllable 'ma', i.e., mono- and polysyllabic nonwords of the structure 'ma(ma)*' where the asterisk refers to the preceding expression in parentheses and means 'zero or more times'. All subjects were recorded on DAT, and one was designated to be the target speaker. The experimental procedure is identical to the one reported here. Again, two groups of native English listeners and one group of German controls are being familiarized with the voice of the target speaker and listeners are then asked to identify the target voice from a set of six different voices. Since the cues of the target language (German) were reduced to a minimum in the material, no effect of native-language background is expected in the performance among the three groups. Note, however, that some linguistic information, especially on the level of phonology and phonetics (e.g. the articulatory setting for German, prosodic features etc.), will remain and thus the experiment may not yield maximally clear results. This (control) experiment is currently being carried out.

A second issue has been identified by the current research; it also is being addressed. Specifically, while there seems to be evidence for the hypothesis

that performance in speaker identification is generally worse when listeners do not speak the same language as the target speaker, it remains unclear whether this effect is dependent on the relatedness between listeners' native-language background and the target language. In the experiments summarized above, native English and native German listeners were compared (target language: German) but both German and English are West Germanic languages. If a linguistic effect is operating, it could be hypothesized that *ceteris paribus*, listeners of other languages, typologically less related to German, would perform even worse on the above task than did the native English listeners. To test this hypothesis, the experiment reported above is being repeated with Spanish speaking listeners. Spanish can be considered to be less related to German than is English.

However, all three languages belong to the Indo-European language family. This relationship could be taken as an argument for predicting that the differences in performance between the native English and the native Spanish listeners should not in fact, be significant. To yet further test this relationship, the above experiment is being repeated with Chinese listeners; Chinese is a non-Indo-European language. If the association between target and native language plays no role in the process, then no significant performance differences will be found among the English, Spanish and Chinese listeners. If it plays a role, however, then significant differences should result either among all three groups or between the Chinese group and the two others. Significant differences between the English and Spanish groups plus no significant differences among any of the groups would be unexpected. These experiments currently are underway also.

A final question resulting from our research on the influence of native-language background to speaker identification accuracy concerns the phonetic experience of the listeners. In all the experiments reported above, care was taken that listeners were naïve with respect to the phonetic and linguistic aspects of the experiments. That is, all subjects were university students from either linguistic/phonetic undergraduate courses or courses of an unrelated discipline. This approach prevented any confounding of the main dependent variable 'native-language background' with the phonetic/linguistic experience of the listeners.

However, it might also be possible that the effects of native-language background are stable with respect to the variable of 'phonetic/linguistic experience'. To evaluate this possibility, a study has been constructed which will test English and German expert witnesses with the same speech material used in the experiments reported above. If the effect of native-language background is not affected by the amount of phonetic knowledge (and experience with speaker identification), then there still should be a (significant) difference in performance between these phonetic experts. If the difference observed between the two groups of naïve listeners disappears with the expert groups, it may be interpreted that evidence exists against the stability of the native-language background effect.

CONCLUSION

In conclusion, it can be stated that native-language background may play a role in speaker identification. This relationship may create a problem as an increasing number of cases are occurring where either naïve listeners or expert witnesses are called upon to identify the voice of a person speaking a foreign language.[4] Of course, much more research is needed to determine the full scope of the influence of this factor on speaker identification. First steps have been made to study these relationships (Goggin *et al.* 1991, Köster *et al.* 1995) and we hope our research will shed light on related issues.

ACKNOWLEDGEMENTS

The authors would like to thank Shanley Allen (Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands) for proof-reading the paper and Harry Hollien (University of Florida, Gainesville, USA) for useful comments on the paper. However, the authors take responsibility for all mistakes. The research in this paper was supported by a grant from the International Association for Forensic Phonetics.

NOTES

1 The International Association for Forensic Phonetics (IAFP) advises phonetic expert witnesses to be extremely cautious in such cases. In Paragraph 6 of the 'Code of Practice' of the IAFP it says:

   (a) Members should approach with particular caution forensic work on speech samples in languages of which they are not native speakers.
   (b) Members should approach with particular caution forensic work in cases where samples are in different languages.

   Nevertheless, sometimes non-native expert witnesses are required to judge the voice samples of a foreign speaker if no native speaking expert witness is available.

2 They consist of three parts of the text x 2 transmission conditions (high fidelity versus telephone) x 3 repetitions x 6 speakers = 108 speech samples.

3 The speakers in this case were different ones than those used in the experiment described in this report. It might be the case that the between-speaker variability will be different in the two groups of subjects and the speaker identification task in the first experiment could have been either easier or more difficult than in the second experiment. See Nolan (1983:11) for a discussion of between-speaker variability and within-speaker variability.

4 According to Künzel (p.c.) the percentage of cases at the German Bundeskriminalamt in which the voice of foreign speaker is involved amounts to about 30 per cent.

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.

Hollien, H. (1990) *The Acoustics of Crime: The New Science of Forensic Phonetics*, New York: Plenum Press.

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

Künzel, H. J. (1990) *Phonetische Untersuchungen zur Sprecher-Erkennung 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.

Nolan, F. (1983) *The Phonetic Bases of Speaker Recognition*, Cambridge: Cambridge University Press.

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