THE INFLUENCE OF NATIVE-LANGUAGE BACKGROUND ON SPEAKER RECOGNITION

O. Köster, N. O. Schiller and H. J. Künzel
Trier University, Trier, Germany, Max-Planck-Institute for Psycholinguistics, Nijmegen, The Netherlands, and Bundeskriminalamt, Wiesbaden, Germany

ABSTRACT

The influence of native-language background on the ability of speaker recognition was tested with different groups of subjects: group 1 had no knowledge of the target language (i.e. German), group 2 had some knowledge, and group 3 spoke the target language as its native language (control group). In a direct identification task, subjects had to recognize a speaker's voice with which they were familiarized before. The differences in performance between the groups were significant.

1. INTRODUCTION

In forensic speaker recognition, it sometimes occurs that the voice from a speaker of a foreign language has to be evaluated in a voice line-up or by an expert witness. The question arises in how far this process is influenced by the native-language background of the listener. Human listeners may make use of linguistic information when remembering voices (in addition to purely acoustic information) (cf. [1], [2]). Therefore, it may be the case that the performance in auditory speaker recognition is related to a listener's familiarity with the language under consideration.

Few studies have focussed on the effect of native-language background on speaker recognition. Goldstein et al. [3] found that native American English listeners showed no differences in recognizing speakers with and without a foreign accent and concluded that "[...] voice recognition is just as good (or as poor) for foreign voices as it is for native voices" (Goldstein et al. [3]: 220). Thompson [4] investigated monolingual English natives listening to speech samples from Spanish speakers, native English speakers and English speakers with Spanish accents, and found that the monolingual English listeners identified speakers of their own language best. Goggin et al. [5] tried to quantify the relationship between language familiarity and performance in speaker recognition. They concluded that "[...] voice identification is increased approximately twofold when the listener understands the language relative to when the message is in a foreign language" (Goggin et al. [5]: 456).

In the experiment reported here, we examined the performance of different groups of subjects in a speaker recognition task, with the groups differing in the degree of familiarity with the target language. Additionally, the influence of the voice transmission condition (hifi vs telephone) was tested. This is of primary interest in the forensic situation where most of the recorded speech material is transmitted via the telephone.

2. EXPERIMENT

To test the ability of listeners with a different native-language background in speaker recognition, a direct identification test was designed, in which four different groups of listeners had to recognize the voice of one German speaker in a set of six different German speakers.

2.1. Subjects

Subjects consisted of 53 female and 21 male listeners (n = 74). The age of the subjects was between 16 and 56 years (m = 26.28, SD = 11.85). Subjects were divided into three groups with respect to their knowledge of German. The first group consisted of native English speakers with no knowledge of German at all. The second group consisted of native speakers who had some knowledge of German.[1] The last group included native speakers of German (control group).

The first group of English speakers was further divided into two categories of age: group 1 (n = 15) included all subjects ≥ 30 years of age (m = 47.4, SD = 8.23), group 2 (n = 24) consisted of subjects under 30 years of age (m = 18.42, SD = 3.32). Subjects in both group 3 (n = 18; some knowledge of German) and group 4 (n = 17; German controls) were all under 30 years of age (group 3: m = 21.22, SD = 1.32; group 4: m = 26.28, SD = 3.38).

All subjects took part in the investigation voluntarily. None of them reported any hearing problems.

2.2. Speech material

The speech material used in the experiment was produced by six different male speakers. Speakers were of similar age (m = 29.67, SD = 5.45) and spoke Standard German with Hessian influences. The F_0 of the six speakers ranged from 86 Hz to 142 Hz (m = 109.5, SD = 18.7). All speakers had to read a small German text of approximately one minute in length onto a DAT recorder. Then three parts of the text between four and eight seconds in length were spliced out of the recordings of every speaker. To record exactly the same material under telephone transmission conditions, the speech samples were recorded again through a telephone line. Each of the six speech samples was re-recorded three times. In total, we obtained 108 speech samples[2]. All of the speech samples were randomized and re-recorded on DAT.

One speaker was designated as speaker X, the target voice. From speaker X, the hifi text was re-recorded on DAT five times to obtain a speech sample of approximately five minutes.

2.3. Method

All four groups of listeners were tested individually. Firstly, subjects were familiarized with the voice of speaker X by listening to speaker X's five minutes speech sample. Subjects were instructed to concentrate on the voice in order to try to memorize it. After this familiarization, response sheets were handed out to the subjects. After a short break of approximately five minutes, the subjects were given a forced-choice test. They were instructed to listen to the tape with the randomized speech samples carefully. After each sample the sujects marked "Yes" if they thought the voice was from speaker X and "No" if it was not. There were five seconds between each stimulus which the subjects considered to be enough time to make a decision. After every tenth speech sample, there was a sine tone of 300 Hz to help subjects to keep track of the task.

3. RESULTS

The design of the experiment allows to differentiate between two error categories: subjects could either reject the target voice speech sample when it actually came from speaker X (false rejection; FR) or identify a speech sample as the target voice when it was in fact produced by one of the dummy speakers (false identification; FI). Furthermore, FRs and FIs were split into the errors made under the hifi vs telephone transmission conditions to see whether there was a difference.

3.1. False rejections vs false identifications

If subjects were randomly identifying the speaker, we would expect an FRs to FIs error ratio of 1:5 (18 target voice samples compared to 90 dummy samples). The observed error ratios fall below the expected value in all four groups: group 1 made 67 FRs (m = 4.4, SD = 2.5) and 256 FIs (m = 17.07, SD = 14.11) (ratio = 1:3.82), in group 2 there were 141 FRs (m = 5.88, SD = 5.18) and 163 FIs (m = 6.79, SD = 8.09) (ratio = 1:1.16), in group 3 there were 26 FIs (m = 1.44, SD = 2.43) and 39 FIs (m = 2.17, SD = 4.07) (ratio = 1:1.5), and group 4 made 24 FRs (m = 1.41, SD = 1.97) and 37 FIs (m = 2.18, SD = 2.71) (ratio = 1:1.54).

[1] Subjects of group 3 were students of German; they took part in a university exchange program and had already been in Germany for several months when the experiment was run.

[2] 3 parts of the text x 2 transmission conditions (hifi vs telephone) x 3 repetitions x 6 speakers = 108 speech samples.

χ^2-tests revealed that the FR to FI error ratios fall significantly below the expected value of 1:5 in all four groups (group 1: χ^2 value of 1:5 in all four groups (group 1: $\chi^2 = 18.16$, $df = 1$, $p < .001$, group 2: $\chi^2 = 416.69$, $df = 1$, $p < .001$, group 3: $\chi^2 = 63.7$, $df = 1$, $p < .001$ and group 4: $\chi^2 = 57.41$, $df = 1$, $p < .001$). The respective error proportions are given in figure 1.

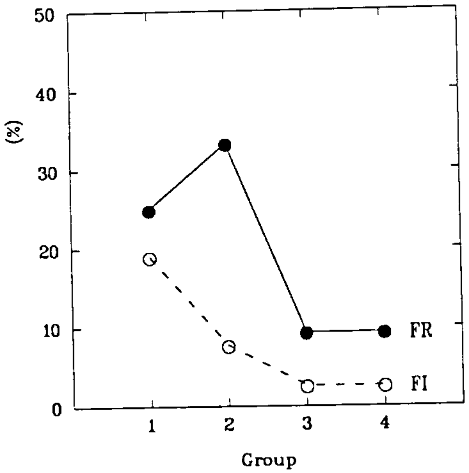

Figure 1. Error proportions for FRs and FIs, respectively.

To see whether there were differences in the amount of errors (FRs and FIs) between the four groups ANOVAs were carried out with the degree of knowledge of German as the dependent variable. The data therefore were arcsin transformed (cf. Winer [6]: 400). The results were highly significant for both the FRs, $F(3, 70) = 7.85$, $p < .001$, and the FIs, $F(3, 70) = 11.897$, $p < .001$.

Post-hoc analyses (pairwise comparisons; *Scheffé* tests) revealed that with respect to the FRs, group 2 made significantly more errors than either group 3 ($p = .003$) or group 4 ($p = .004$). Concerning the FIs, group 1 made significantly more errors than any of the other groups (group 2 $p = .007$, group 3 $p < .001$ and group 4 $p < .001$). None of the other differences between the four groups were significant.

3.2. Hifi vs telephone transmission condition

The ratio of speech samples recorded under hifi vs telephone transmission conditions was 1:1 (54:54). Within the 18 target voice samples (9:9) and the 90 dummy samples (45:45), the respective ratios were also 1:1. The expected error ratio both for FRs and FIs was therefore 1:1. Again, the observed ratios deviated from this *a priori* value in different ways (see tables 1 and 2).

group	FRs	hifi	teleph.	ratio h:t
1	**67**	**25**	**42**	**1:1.68**
2	**141**	**44**	**97**	**1:2.21**
3	**26**	**13**	**13**	**1:1**
4	**24**	**9**	**5**	**1:0.56**

Table 1. FRs in the two transmission conditions for the groups.

group	FIs	hifi	teleph.	ratio h:t
1	**256**	**99**	**157**	**1:1.59**
2	**163**	**71**	**92**	**1:1.3**
3	**39**	**16**	**23**	**1:1.44**
4	**37**	**5**	**32**	**1:6.4**

Table 2. FIs in the two transmission conditions for the groups.

With respect to FRs, group 2 made significantly more errors when the speech sample was recorded over the telephone ($\chi^2 = 9.96$, $df = 1$, $p < .005$). The difference between the number of errors for group 4 reached only marginally significance ($\chi^2 = 4.08$, $df = 1$, $p < .05$). But note that in this case there were fewer errors for the telephone transmission condition.

Concerning FIs, all four groups made fewer mistakes in the hifi condition. Significance was reached for group 1 ($\chi^2 = 6.57$, $df = 1$, $p < .025$) and for group 4 ($\chi^2 = 9.85$, $df = 1$, $p < .005$).

4. DISCUSSION

All four groups made significantly fewer FIs relative to FRs than could be theoretically expected. This means that, on the average, subjects were inclined not to identify a speech sample as coming from speaker X. This leads to the interpretation

that subjects were in general quite prudent with identifying a voice as the one from the target speaker. This result is in contrast to the result obtained by Künzel [7]. Künzel tested the speaker recognition abilities of linguistically naive listeners and found that in his groups on the whole subjects showed the tendency to identify two speech samples as coming from the same speaker even when this was not the case (cf . Künzel [7]: 35).

As the statistical analyses revealed, there were significant differences in performance in the speaker recognition experiment between the four groups. 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 less relevance because group 3 and 4 performed equally well.

The influence of the listeners' age on the performance in speaker recognition remains rather unclear. Whereas the younger subjects of group 2 made fewer FRs than the older ones of group 1, the situation is reversed with respect to the FIs; here, group 1 made significantly more errors than group 2. This last result is in accord with Künzel ([7]: 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 the speech samples under hifi vs 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 3 and 4 for the FRs (see table 1). This leads to the interpretation that the acoustic quality of the speech sample is very important for speaker recognition purposes. In accord with what Künzel ([7]: 26) found, it seems that in the speech samples recorded via the telephone some speaker specific features that help in voice recognition are missing or obscured.

Acknowledgements

The authors would like to thank Bernadette Schmitt and Jan-Peter de Ruiter (both Max-Planck-Institute for Psycholinguistics, Nijmegen, The Netherlands) for helpful comments on the statistical analyses and Julie Christiansen (Max-Planck-Institute for Psycholinguistics, Nijmegen, The Netherlands) for proof-reading the paper. The authors take responsibility for all remaining mistakes.

5. REFERENCES

[1] Ladefoged, P., Ladefoged, J. (1980), "The ability of listeners to identify voices", *UCLA Working Papers in Phonetics*, vol. 49, pp. 43-51.

[2] Lorch, M. P., Meara, P. (1989), "How people listen to languages they don't know", *Language Science*, vol. 11, pp. 243-253.

[3] Goldstein, A. G., Knight, P., Bailis, K., Conover, J. (1981), "Recognition memory for accented and unaccented voices" *Bulletin of the Psychonomic Society*, vol. 17, pp. 217-220.

[4] Thompson, C. P. (1987), "A language effect in voice identification", *Applied Cognitive Psychology*, vol. 1, pp. 121-131.

[5] Goggin, J. P., Thompson, C. P., Strube, G., Simental, L. R. (1991), "The role of language familiarity in voice identification", *Memory & Cognition*, vol. 19, pp. 448-458.

[6] Winer, B. J. (1971), *Statistical principles in experimental design*, second edition. New York et al.: McGraw-Hill, 1971.

[7] Künzel, H. J. (1990), *Phonetische Untersuchungen zur Sprecher-Erkennung durch linguistisch naive Personen*, Stuttgart: Steiner (Zeitschrift für Dialektologie und Linguistik, Beihefte; 69).