The correlation between auditory speech sensitivity and speaker recognition ability

Olaf Köster,[*] *Markus M. Hess,*[†] *Niels O. Schiller*[‡] *and Hermann J. Künzel*[*]

[*] *Bundeskriminalamt (BKA), Wiesbaden, Germany*
[†] *Research Associate and Lecturer in Otolaryngology at the Massachussetts Eye and Ear Infirmary*
[‡] *Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands*

ABSTRACT In various applications of forensic phonetics the question arises as to how far aural-perceptual speaker recognition performance is reliable. Therefore, it is necessary to examine the relationship between speaker recognition results and human perception/production abilities like musicality or speech sensitivity. In this study, performance in a speaker recognition experiment and a speech sensitivity test are correlated. The results show a moderately significant positive correlation between the two tasks. Generally, performance in the speaker recognition task was better than in the speech sensitivity test. Professionals in speech and singing yielded a more homogeneous correlation than non-experts. Training in speech as well as choir-singing seems to have a positive effect on performance in speaker recognition. It may be concluded, firstly, that in cases where the reliability of voice line-up results or the credibility of a testimony have to be considered, the speech sensitivity test could be a useful indicator. Secondly, the speech sensitivity test might be integrated into the canon of possible procedures for the accreditation of forensic phoneticians. Both tests may also be used in combination.

KEYWORDS speaker recognition, auditory speech sensitivity, tests of musicality, computer speech sensitivity test

INTRODUCTION

One of the most important and basic problems in forensic phonetics is the reliability of voice recognition. In cases of auditory speaker identification by lay witnesses or experts (e.g. in a voice line-up) or when a voice profile has to be worked out, the question arises whether the courts can rely on the results. Additionally, recent demands for procedures for the accreditation of forensic phoneticians make it necessary to test the relationship between speaker recognition results and human perception/production abilities.

A common idea is that musicality correlates positively with performance in speaker identification. This would imply that if a person has musical talent or is trained in that area, s/he will also have a superior performance in speaker identification tasks. Another human perceptual/

1350–1771

productive ability that might be positively correlated with speaker recognition ability is 'auditory speech sensitivity' (Pahn and Pahn 1991).[1] This term refers to a test which is designed to investigate an individual's ability to perceive and produce different elements of speech, that is tonal movement, voice onset, pitch, rhythm, dynamics (intensity), nonsense syllables and all combinations of the preceding elements.

In the past, several tests which evaluate the musicality of a participant have been designed. Well known are the tests of musicality by Seashore (1967),[2] Gordon (1965) and Bentley [1966] (1983). In Germany, a new test has been developed by Arndt (1989). To our knowledge, no experiments have been carried out to test the relationship between musicality and perfomance in speaker recognition tasks. In contrast to the different tests of musicality, the test of speech sensitivity by Pahn and Pahn (1991) not only involves classical elements of music like pitch, rhythm and dynamics but also pure elements of speech like voice onset and nonsense syllables. Furthermore, in the Pahn test those elements adapted from musicality tests are produced by a human speaker and not by an instrument.

To test the hypothesis that auditory speech sensitivity and speaker recognition performance are correlated, thirty subjects were asked to participate (a) in the speech sensitivity test by Pahn and Pahn and (b) in a speaker recognition experiment (Köster *et al.* 1995; Schiller and Köster 1996).

EXPERIMENT

Test of speech sensitivity

The test of speech sensitivity was designed by Pahn and Pahn in 1991 to characterize a person's analytical sensitivity towards different elements of speech. This index helps the speech therapist to predict the success of voice treatment or speech therapy. Pahn and Pahn assume that if a person cannot *perceive* what is wrong with his/her voice or speech, s/he will not be able to change these habits either. In addition, the authors propose to use the speech sensitivity test to find out the 'aptitude for professions making high demands on voice and speech' (Pahn and Pahn 1991: 19). The Pahn test examines only formal aspects of speech. Semantic aspects (words and sentences) are excluded. Thus, according to the authors, the test is independent of the subject's native language or dialect. Furthermore, voice quality or difficult motor (articulatory) abilities are not tested.

In particular, the following parameters are examined: relative tonal movement (pitch contour), voice onset, relative pitch, rhythm (including pauses), dynamics (accent, intensity) and nonsense syllables.[3] Among the twelve tasks, five offer stimuli which include one of the tested parameters, six of the stimuli include two parameters and one stimulus includes three parameters in combination. The stimuli consist of either the sound [ɑ] or the syllable [dap] differing with respect to the tested parameters. The nonsense syllable sequence sounded like [pastʁeʃliɔχu] with a German spelling of <pastreschliochu>.

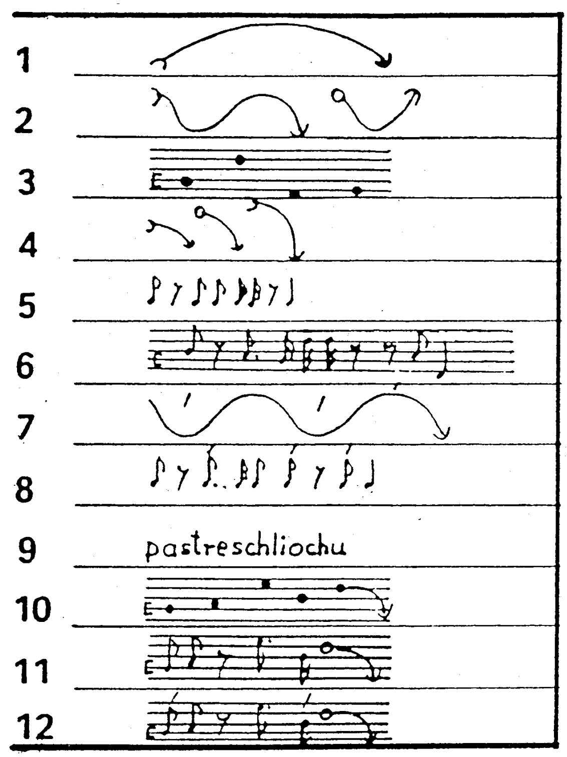


Figure 1 Schematic representation of the twelve tasks in the auditory speech sensitivity test

The test of speech sensitivity is divided into two parts. The first one evaluates a parameter which is called 'auditory speech output' ('auditory speech output analysis', ASOA). The subject is asked to imitate a speech sample produced by a speaker of his/her own sex. The subject's attention is drawn to the formal aspect that will be tested in the following task; then the stimulus is offered three times and the subject tries to imitate the sample. The stimulus is presented for a fourth time and the subject imitates again. The stimulus is presented a last time followed by an imitation. All three imitation trials are recorded. Each imitation is judged by an expert (the experimentor). If the imitation is identical to the stimulus with respect to the parameter that is tested the subject scores 3 points. If the imitation is 'almost correct' it scores 2 points. 1 point corresponds to the level beneath 'almost correct' and no points are scored if the tested parameter is not produced correctly at all. Altogether, the maximum score of a subject is 108 (twelve tasks, three trials each up to 3 points).

In the second part of the test, which is called the 'auditory speech input analysis' (ASIA), each subject is asked to evaluate his or her own imitation. The subject's task is to tell the experimentor if the imitation was identical to the original stimulus or if s/he can recognize any differences. The combination of the imitation and the original is offered three times to the subject. Each trial is judged a second time by the experimentor. If the subject recognizes all errors or recognizes that no errors have been made 3 points are scored. If not all errors have been recognized 2 points are scored. The subject obtains 1 point if no errors are detected but an 'impression of difference' exists. If the subject does not recognize any errors which have been made no points are scored. The maximum score for a subject is again 108.

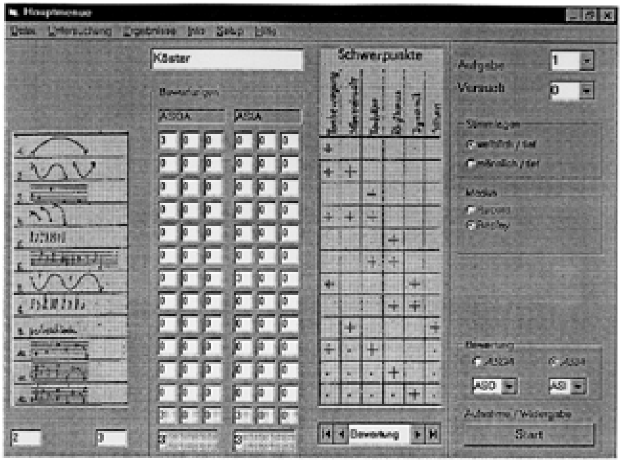
Due to the fact that in our experiment we wanted to correlate the parameter 'auditory speech sensitivity' with the perceptive performance in a speaker recognition experiment, only the ASIA values are taken into account. In a way, the 'auditory speech input' *includes* the 'auditory speech output'. The parameter 'auditory speech output' evaluates both perceptive *and* productive elements while the 'auditory speech input' only includes *perceptive* abilities. Experience shows that both the ASIA and the ASOA indices are similar for an individual subject with the ASIA values being usually 10 to 15 points higher.

In the test of speech sensitivity, a crucial element is the experimentor who evaluates the subject's performance. For this task, Pahn and Pahn demand a person with high speech analysis skills who has undergone at least two days' training in the test. According to Pahn and Pahn, trained and skilled persons maximally differ for 5 points in evaluating the speech sensitivity. In this study, two logopedists with experience in the Pahn test served as experimentors.

A computer version of the Pahn test has been developed (Grohmann *et al.* 1997). It replaces the old test which involved analogue tape recordings. The multimedia program which is available on a CD-ROM (in German or in English), runs on a PC (486) under Windows 3.1. It is easily operated by means of a user interface. (See Figure 2.) After the speech sensitivity test has been carried out, the program calculates the ASOA and ASIA values; the test record can be stored for further processing (which is planned for the future). In our study, this PC version was applied.

Speaker recognition experiment and correlation

In order to evaluate speaker recognition ability, we used the same experiment as in former studies (Köster *et al.* 1995; Schiller and Köster 1996). Subjects were familiarized with the voice of a male German speaker by listening to a five-minute sample of his voice. After a break of approxi-



Köster

Aufgabe 1

Versuch 0

Figure 2 User interface of the computer speech sensitivitiy test (German PC version)

mately five minutes subjects had to recognize the target speaker from a set of randomized voice samples. The test tape consisted of 108 voice samples from six different male German speakers[4] in high-fidelity and telephone transmission quality among which the target speaker appeared eighteen times. On a response sheet subjects had to mark whether they had recognized the target speaker or not (a forced-choice test). Subjects could either identify the target speaker correctly (hit), reject a dummy speaker correctly, reject the sample when in fact it was produced by the target speaker (false rejection), or identify a speech sample as the target voice when in fact it came from one of the dummy speakers (false alarm).

As in the earlier studies, the performance of identification, which takes into account the proportion of both the hit rate and the false-alarm rate and which is expressed by the sensitivity measure *d'*, was calculated using Signal Detection Theory (for a detailed description see MacMillan and Creelman, 1991; Schiller, Köster and Duckworth, 1997). *d'* was calculated for each subject taking part in the investigation. The maximum score of a subject was 6.18 corresponding to no errors in the recognition experiment.

Due to the fact that the duration of the speaker recognition experiment was about forty-five minutes and that of the speech sensitivity test was about one hour, subjects had a break between both tests of at least twenty minutes. All participants considered the time span between both tests sufficient for recovery. Generally, the speech sensitivity test was considered to be only moderately exhausting.

In our experiment, a correlation of the ASIA values (auditory speech sensitivity) and the *d'* values (speaker recognition performance) was carried out (Pearson's product–moment correlation).

Subjects

Altogether, thirty subjects participated in the test of speech sensitivity as well as in the speaker recognition experiment; their native language was German. The subjects' ages ranged from twenty to forty-three years (M: 29.8; SD: 5.5). There were ten male and twenty female participants. Eight were professionals in the field of speech and singing: they were either logopedists or undergraduate students of logopedics in their third year. One participant was a singing teacher at a university; he was the only male subject in the expert group. As far as 'speech' is concerned, all the other subjects were non-experts: they were either undergraduate students of logopedics in their first year (beginners; having no education in speech and voice) or other lay people. Two of the subjects who were not professionals in speech science had several years' experience in choir singing. All subjects took part in the investigation voluntarily; none of them reported any speech or hearing problems.

RESULTS

A Pearson product–moment correlation of the speech sensitivity index and the *d'* values of the speaker recognition performance revealed that the results of both experiments are correlated with a correlation coefficient of $r = 0.4$. This correlation is significant ($p<.05$). As can be seen from Figure 3 the different coordinates are distributed around the regression line. The figure reveals that the scores for the auditory speech input range between 55 and 108 points (M: 86.4; SD: 14.1). One subject (the university singing teacher) obtained the maximum score of 108 which means that he recognized all possible 'errors' of his own imitations in the production task. The *d'* values of the speaker recognition experiment are distributed between 0.35 and 6.18 (M: 4.27; SD: 1.73). Eight subjects reached the maximum value of 6.18 which means that they had recognized all target samples correctly and made no false identifications. Seven of these subjects had an auditory speech sensitivity which was over 80; one of these subjects had a score of 74 points. The singing teacher, with the maximum score in the auditory speech sensitivity, had a maximum *d'* value, too. Furthermore, it can be seen from the figure that subjects generally performed better in the speech sensitivity test than in the speaker recognition experiment.

Figure 4, which only displays the eight subjects who are professionals in speech or singing, shows that the 'experts' are distributed rather homogeneously (although the regression line indicates a slightly lower correlation between speech sensitivity and speaker recognition performance than in the complete group). The speech sensitivity scores range from 79 to 108 points (M: 93.9; SD: 10.3). The *d'* values range from 2.49 to 6.18 (M: 4.75; SD: 1.32). Revealing another aspect, Figure 5, which

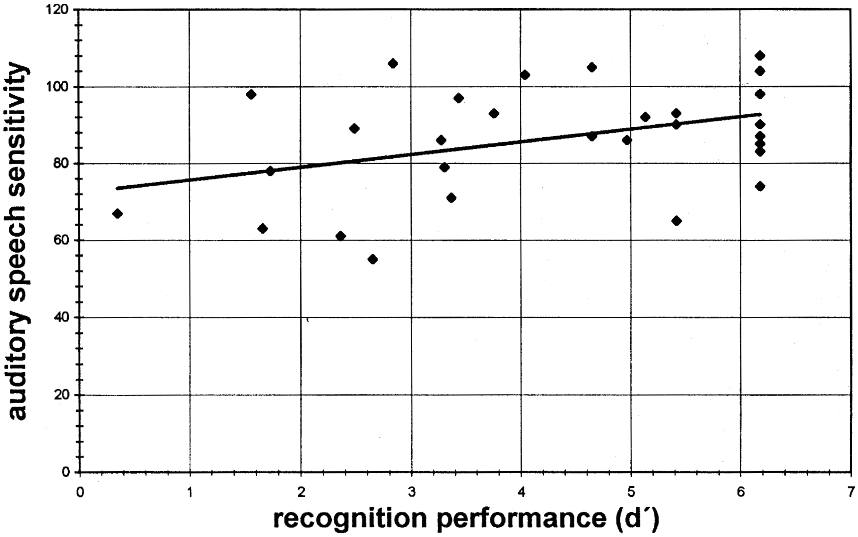


Figure 3 Correlation of speaker recognition performance and speech sensitivity

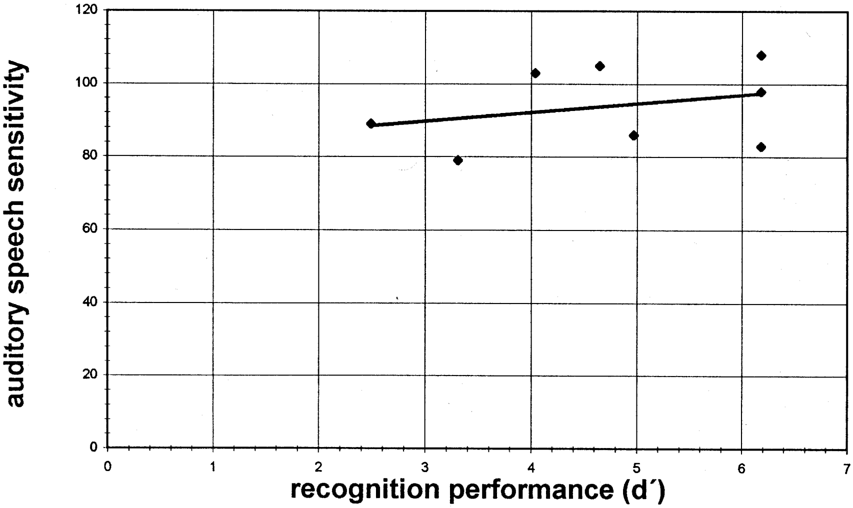


Figure 4 Correlation of speaker recognition performance and speech sensitivity (experts)

displays only the non-professionals in speech and singing, shows that subjects are distributed more heterogeneously. Speech sensitivity scores range from 55 to 108 points (M: 83.6; SD: 14.3). *d'* values range from

0.35 to 6.18 (M: 4.09; SD: 1.82). The two subjects of the non-professional group who had experience in choir singing (subjects 16 and 29, as indicated in Figure 5 by ↑) performed on or above average in both tests.

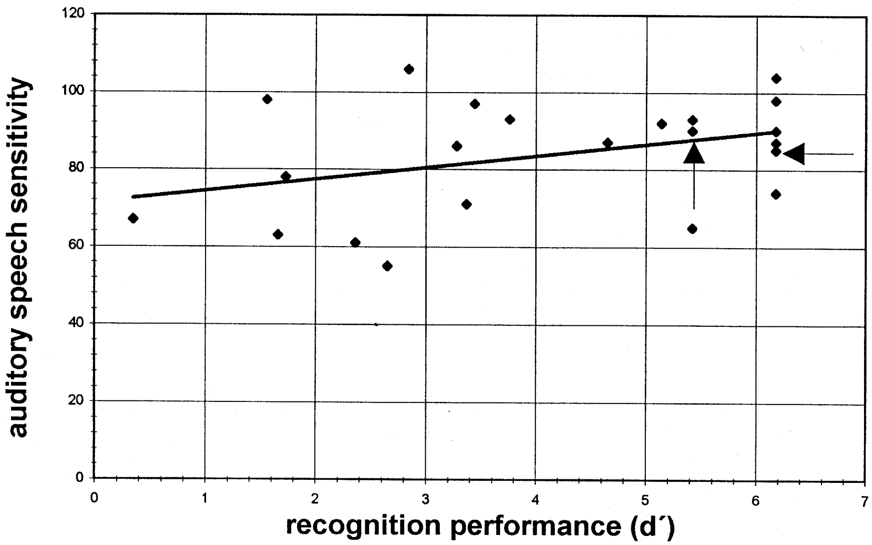


Figure 5 Correlation of speaker recognition performance and speech sensitivity (non-experts)

DISCUSSION

The statistical analysis has revealed that the parameter ‘auditory speech output’ of the speech sensibilty test (Pahn and Pahn 1991) and the performance in the speaker recognition experiment (Köster *et al.*, 1995; Schiller and Köster 1996) are correlated significantly. This leads us to the assumption that there is a general positive correlation between the ability to perceive and describe formal elements of speech and the ability to recognize a speaker aural-perceptually. Therefore, it can be predicted that a person who performs well in the speech sensitivity test will, in general, achieve good results in a speaker recognition task as well. In other words, a person with a high score of auditory speech sensitivity will probably produce more reliable speaker recognition results than a person with a poor score for auditory speech sensitivity. The results also show that, compared to the speaker recognition experiment for the subjects, it is easier to obtain good results in the speech sensitivity test (correlation coefficient: 0.4). This must be considered when speech sensitivity scores are applied to assess reliability in speaker recognition tasks.

There is a tendency for professionals in speech or singing to produce good results in both tests, especially in the speech sensitivity test. Both their ability to recognize a speaker and their sensitivity to formal elements of speech seems to be distributed more homogeneously than the results of the non-expert group (further statistical analyses were not carried out because of the small size of this test group; further investigations are needed). Choir singing might be good training to render a person sensitive to formal speech elements and speaker specific features, as both subjects with experience in choir singing performed exceptionally well in both tests (however, there is no statistical basis for this assumption). This would also confirm the hypothesis that in fact musicality or musical training correlate positively with speaker recognition ability. If we assume that the perception and processing of speech and music are connected in some way this makes it interesting for future research to correlate speaker recognition experiments with different tests of musicality (e.g. Seashore 1967; Arndt 1989). It must be considered that the test of speech sensitivity already incorporates musical elements.

Although subjects generally performed well in the auditory speech sensitivity test (the average score was 86.4 out of 108), two aspects stand out which seem to create unequal conditions for different subjects. Firstly, only *one* speaker demonstrates the stimuli which have to be imitated by the subjects. As it seems to be easier to imitate a voice which is similar to one's own voice (e.g. in pitch or sound quality) those subjects with a voice similar to the speaker's might have the lead. Secondly, the nonsense syllable sequence is not independent from the dialect or the native language of the subjects. Some of the syllables or sounds do exist in standard German but not in certain dialects. Syllables, sounds or phonotactic combinations of sounds might not occur in foreign languages. For example, in English no [ʁ], [eː], or [χ] exist.

In contrast to established tests of musicality, the test of auditory speech sensitivity is not yet a scrutinized standard diagnostic tool in the area of speech pathology or voice treatment. Nevertheless, our experiment provides first evidence that this method is a useful indicator for the ability to recognize a speaker as both tests are correlated significantly. It is unlikely that a person with a very high score in 'auditory speech input' will perform very poorly and unreliably in speaker recognition tasks. However, there is no certainty that a subject will perform equally well in both tests as the correlation between both tests is only 0.4.

In the practice of forensic phonetics the results may lead to different applications of the speech sensitivity test. First, in cases where the reliability of voice line-up results or the credibility of a testimony have to be considered, the test might be a moderate but useful indicator. Second, a test like the speech sensitivity test might be integrated into the canon of possible procedures for the accreditation of forensic phoneticians. Both tests, the test of auditory speech sensitivity and the speaker recognition experiment used in this study, may also be used in combination.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Herbert Masthoff, Department of Phonetics, University of Trier, for helpful comments on the paper. Furthermore, we appreciate the support of the members of the 'Deutsches Tonarchiv für Stimm- und Sprachstörungen' (FU Berlin), especially of Günter Grohmann (knowledge engineer).

NOTES

1 We replaced the original translation 'auditive speech sensibility' (Pahn and Pahn 1991) by the term '*auditory* speech *sensitivity*' which might be more appropriate.

2 The Seashore test evaluates the parameters of pitch, intensity, rhythm, duration of a tone, quality of a tone and memory of tones.

3 Experience shows that the elements pitch, rhythm, and dynamics are especially difficult to perceive and to produce (Pahn and Pahn 1991).

4 All speakers including the target speaker were of similar age (M: 29.7 years; SD: 5.45 years) and spoke standard German with Hessian influences. Their mean F0 ranged from 86 Hz to 142 Hz (M: 109.5 Hz; SD: 18.7 Hz). All speakers read a text from which three different parts were selected as test samples (always the same parts for each speaker).

REFERENCES

Arndt, J. (1989) 'Test zur Diagnostik musikalischer Wahrnehmungsfähigkeiten (TMW)', unpublished Ph.D thesis, University of Halle.

Bentley, A. [1966] (1983) *Musikalische Begabung bei Kindern und ihre Meßbarkeit* (3rd edn), Frankfurt–Berlin–München: Diesterweg (first published in English, 1966).

Gordon, E. (1965) *Musical Aptitude Profile*, Boston: Houghton Mifflin.

Grohmann, G., Pahn, J., Gross, M. and Hess, M. (1997) 'Sprachsensibilitätstest nach Professor Pahn: Multimediale PC Version', in M. Gross and U. Eysholdt (eds), *Aktuelle phoniatrisch-pädaudiologische Aspekte* 1996 (4), Göttingen: Verlag Phoniatrie, 224.

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 XIIIth International Con-*

gress of Phonetic Sciences, vol. 4, Stockholm: KTH and Stockholm University, 306–9.

Pahn, J. and Pahn, E. (1991) 'Formblatt, Eigenschaften, Ablauf und Bedeutung des Tests der Sensibilität formaler sprachlicher Elemente im Hinblick auf Perzeption und Produktion', *Sprache–Stimme–Gehör*, 15: 19–23.

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

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', *Forensic Linguistics*, 4 (1): 1–17.

Seashore, C. E. (1967) *Psychology of Music*, New York: Dover (first published in 1938 by McGraw-Hill).