AUDITORY-PERCEPTUAL IDENTIFICATION OF VOICE QUALITY BY EXPERT AND NON-EXPERT LISTENERS*

Olaf Köster[1], Michael Jessen[1], Freshta Khairi[2], and Hartwig Eckert[3]

[1]Department of Speaker Identification and Audio Analysis, Bundeskriminalamt, Germany
[2]Institute of Communication Research and Phonetics, University of Bonn, Germany
[3]Department of English, University of Flensburg, Germany
olaf.koester@bka.bund.de, michael.jessen@bka.bund.de, freshta-khairi@gmx.de, eckert@uni-flensburg.de

ABSTRACT

In a perception task 13 types of voice quality were to be identified by two listener groups. Expert listeners with a professional background in forensic phonetics performed significantly better than the non-expert group. Furthermore, the non-experts produced more heterogeneous types of error. For prominent types of voice quality and stimuli with a strong scalar degree low error rates were observed for the experts.

Keywords: forensic phonetics, speaker identification, voice quality.

1. INTRODUCTION

Voice quality is not only used as a tool in voice pathology diagnosis [2] but also frequently applied as a feature in forensic speaker comparison reports.

Voice quality can be understood in a narrower sense equivalent to the term “phonation type” referring to laryngeal characteristics only. It can also be understood in the broader sense of including both laryngeal and supralaryngeal characteristics, the latter of which can be captured by the term “articulatory settings”. It is this broader understanding of voice quality according to Laver [4] that will be applied in the present study. Voice quality can be investigated on various levels, ranging from production/physiology over acoustics to audition/perception. This paper focuses on the auditory-perceptual level.

In forensic phonetics, unfortunately the ability of expert listeners to identify voice quality has not been a matter of experimental research in the past (for an overview of the current situation cf. [3, 5]). Yet, the question whether or not a listener´s judgment is reliable regarding the quality of a speaker´s voice is a crucial issue. Two main reasons should motivate more investigation in this area:

First, since crime has generally become more globalized, since more immigrants from disadvantaged countries have become criminal in economically advantaged societies, and since extremism/terrorism involving the Arabic language has become a growing problem, more forensic cases occur in which the language involved is not the native language of an expert. Thus, aspects of the linguistic system of these languages cannot be judged by the forensic expert her/himself and have to be evaluated with the help of an external expert of the language in question. The only aspects in foreign languages that the expert can still process are from domains frequently referred to as para- or extralinguistic, including voice quality.

Also, as in many other fields *quality assurance* is becoming more and more important in forensic expert reports. This means that forensic experts are expected not only to describe their methods and analysis tools properly but also to validate their skills. In this case, it needs to be shown that a specialist is able to identify and consistently recognize an individual´s voice quality.

Within the framework of quality management, the standard procedure of evaluating identification performance is a proficiency test. Accordingly, in the following experiment the performance of forensic phonetic experts and lay persons regarding the identification of different types of voice quality was evaluated in a blind collaborative exercise. The types of voice quality in question were adapted from Laver´s well established framework [4] and represent some of the most frequently occurring vocal features in forensic expert reports [3].

2. EXPERIMENT

2.1. Experimental design

In order to make the test accessible to all participants at the same time, a web page was set up. Each anonymous subject was able to enter the URL through a user ID and password. After activating the first stimulus the listener had to select from a list the particular voice quality that s/he had perceived. The stimulus could be repeated an unlimited number of times and no time limit was set for the answer, but once selected, no retrospective changes were possible. After having selected the respective voice quality, the next stimulus was presented (random order). If a break was needed, the experiment could be interrupted for any length of time. Via the modal-voice-button participants were able to listen to the unmarked neutral setting of the speaker (see 2.2.) at any time. By comparing a stimulus of a certain voice quality to the neutral anchor stimulus a judgment can be made easier and with more confidence. In order to test the ability of listeners to judge the *degree* of a voice quality, subjects were provided with the actual type of voice quality of a sample in a second part of the experiment, when the task was to mark the degree (see 2.2.). The current paper focuses on the first part of the experiment, where selection of an incorrect voice quality but not of an incorrect voice quality degree was counted as an error. The experiment was repeated after 4 and after 8 weeks to measure the consistency of the listeners´ judgments. These re-tests together with the second part of the experiment will be evaluated in a follow-up study.

2.2. Speech material

The speech material came from a 64 year old male native speaker of German, who produced - in addition to modal voice - the following voice qualities: breathiness, roughness (harshness), creak, pressed voice, tremolo, falsetto, nasality (both closed and open), labial protrusion, labial spreading, open jaw and closed jaw. Most of the qualities were produced in three different scalar degrees, that is: weak, moderate, and strong. The classification into only three different degrees modifies the suggestions by Köster and Köster [3] and reflects an easier applicability in everyday casework. Falsetto voice as a categorical type of phonation was offered only in one grade, as well as pressed voice, which was presented in a stimulus produced with strong vocal fold adduction and a stimulus produced with raised larynx. Tremolo and creak appeared only in a weak and moderate degree. As a result, the subjects listened to a total amount of 32 voice quality stimuli.

Each speech sample consisted of two German sentences from the standard text “The North Wind and the Sun” with a duration of approx. 16 seconds. For the experiment, the speech material had been digitally recorded in HiFi-quality (for a discussion of voice quality recognition of telephone transmitted speech, see [5]). In order to limit identification factors other than voice quality alone, only one single speaker was chosen to produce the speech samples. This speaker was able to produce all the types of voice quality as illustrated on the accompanying tape of Laver [4]. In previous work with Laver he had demonstrated his skills in relevant German publications including self-produced sound samples [1].

2.3. Participants

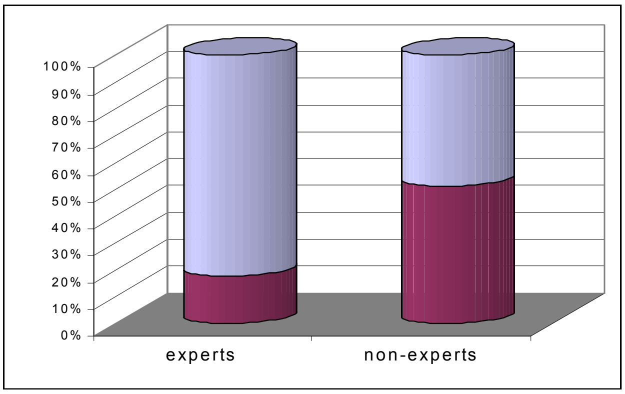
The group of experts consisted of 11 participants; all of them currently work as experts in the field of forensic speaker identification either for German and Austrian public authorities or as private experts. They have been in the field for at least several years and have an educational background in speech science, phonetics as well as speech signal analysis. All experts had previously attended a workshop on voice quality and had been provided with a multimedia CD comprising prototypes of the tested voice qualities (description, sound, videos).

The non-expert group consisted of 20 participants; 18 of them were students, 2 had no academic background. No one had any educational background in phonetics or linguistics. Each participant was familiarized individually with the concept of voice quality by one of the authors (detailed powerpoint presentation). In the course of a one-hour training session the non-experts listened to all relevant types of voice quality. After the training, all of them were provided with a CD comprising the same audio samples the expert group received during their workshop.

3. RESULTS

The raw data produced by the participants consisted of binary answers which were either correct or incorrect. When a type of voice quality was identified, the answer was registered as correct, when a voice quality was misidentified, the answer was counted as incorrect. The type of voice quality of each presented stimulus, the corresponding answer as well as age, gender and response time of the participants were imported into excel-data-sheets. Correct and incorrect answers were counted for each participant and also pooled across the two groups. The expert listeners produced 62 incorrect answers altogether, with a total amount of 348 registered answers; this resulted in a share of 17.8 % of errors. In contrast, the group of non-expert listeners was unable to identify the correct voice quality 325 times with a total number of 634 answers; here the share of false identifications was 51.2 % (see fig. 1).

Figure 1: Overall identification performance; correct identifications: light color, false identifications: dark color.

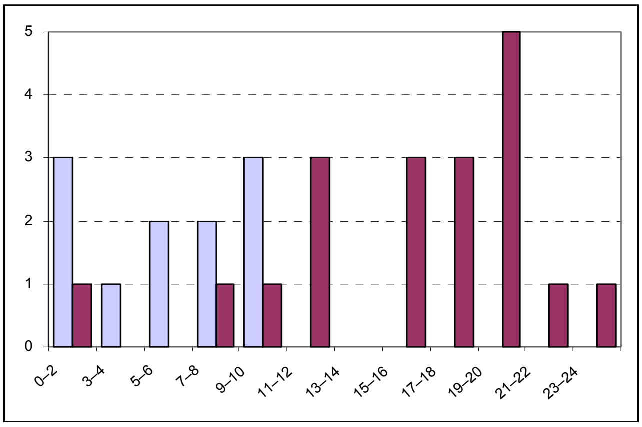


A comparison between the two listener groups showed that the overall difference between speaker identification experts and non-expert listeners was highly significant on a 1%-significance level (*p* < 0.01). Since the variance for the non-experts was significantly higher than for the experts a t-test with separate variances (Welch modification) was applied.

None of the participants could identify the 32 stimuli without any error. One of the experts incorrectly marked only one sample, two other experts made two incorrect choices. One of the non-experts showed an outstanding performance with only two incorrect answers, while another non-expert incorrectly identified all stimuli (32 errors). All other naïve listeners ranged between 7 and 23 false identifications out of 32 answers; experts ranged between merely 1 and 10 mistakes. Between the two groups there was only a small overlap with three non-experts (2, 7, and 9 false identifications respectively) reaching the results of the expert group (see fig. 2; outstanding non-expert with 32 false identifications excluded from fig.).

A more detailed analysis of the incorrect answers shows that none of the 13 types of voice quality was identified without any error at all. The group of lay listeners produced between 13 % (for breathy voice quality) and 81 % false identifications (for closed jaw voice quality) per type of voice quality. The experts spread between 3 % (for breathy voice) or 5% errors (tremolo voice) respectively and 39% (for creaky voice) or 42% errors respectively (for close jaw voice quality) per type of quality.

Figure 2: Number of subjects (y-axis) ordered by number of errors in two-error-intervals (x-axis); experts: light column, non-experts: dark column.



When all 32 stimuli including the degree of the different voice qualities were taken into account, it turned out that the experts either showed no false identifications for 10 stimuli or were performing well (that is only one or two errors per type of voice quality for the whole group of experts) for most of the other samples. High negative scores were reached for *weak* closed jaw voice quality (64 % confusions with modal voice) and *weak* creaky voice (45 % confusions with modal voice). For the non-experts, results showed that not a single stimulus escaped a false identification; instead, most voice samples were confused at a wide range. In fact, while on average experts confused each stimulus with 1.25 of other types, the non-expert listeners confused a stimulus with 5.1 of differing voice types. For the lay listener group, worst performance was observed for modal voice which was confused with creaky voice in 50 % of the cases. Confusions on a large scale (between 30 and 40 %) were also found for (weak and moderate) closed nasality with open nasality, (moderate and strong) open nasality with closed nasality, moderate rough voice with creaky voice, and weak creaky voice with rough voice.

4. DISCUSSION

The overall comparison of the identification performance of expert and non-expert listeners regarding 13 different voice quality types has revealed clear differences between the two groups. It has become evident that forensic phoneticians are (highly) significantly more successful in this task than lay persons. As could have been expected, professional skills, extended training, and experience generally make the speaker identification specialist superior to the lay listener even in cases in which the lay listener has acquired some basic knowledge of voice quality.

Nevertheless, factors such as intrinsic talent or perceptive skills may also play a role in the identification of voice quality, as can be seen from the good performance of three out of twenty non-experts: two of the participants with no phonetic background who performed within the range of the expert listeners were either involved in music production and played instruments, or had some training in singing.

The performance of the experts was not only better from the overall perspective but also for each type of voice quality: For each of the 13 voice qualities as well as for each of the 32 graded stimuli the error rate was higher for the non-experts. While the lay listeners produced incorrect answers for each stimulus, ten out of 32 graded types of voice quality were identified correctly by all expert listeners.

Another important difference between the two groups concerns both the distribution of the individual error rates and the spreading of incorrect answers. First, as can be seen from fig. 2, the group of non-experts produced much more heterogeneous results as compared to the experts: Lay listeners performed between 2 and 32 false identifications while the specialists only spread between 1 and 10 mistakes. Secondly, if a type of voice quality was identified incorrectly, the phonetic experts varied only between one and two (1.25) alternative qualities on average while the non-experts picked more then 5 (5.1) alternatives on average. This suggests that the naïve listeners perceived the concept of voice quality in a more confused, less conscious and less straightforward way than the speaker identification specialists did.

For both groups the best results were obtained for breathy voice quality; probably, breathy voice is a comprehensible and known concept of phonation. In addition, a closer look at all three respective voice samples revealed a strong coloring with breathiness. For the experts, other voice qualities which were also identified almost without error were the following: tremolo, falsetto, lip rounding and closed nasality; obviously here we are dealing with rather prominent voice qualities in terms of perception.

The total number of 17.8 % error rate in the identification of voice quality seems quite high for the professionals in speaker identification at first glance. But a closer look revealed that the most frequent incorrect answers (see 3.) were due to mistakes that can easily be explained: The confusion of weak closed jaw quality with modal voice was understandable as in the stimulus in question the closing of the jaw was indeed hardly marked and therefore difficult to recognize. The confusion of weak creaky voice with modal voice could be expected as the speaker of the speech samples indeed showed a slight creak in his normal phonation. It was striking that for the expert group only 5.1 % of the answers were incorrect when a strong scalar degree of a voice quality was involved.

It needs to be kept in mind that the identification labels in this experiment were specialized phonetic terms (voice qualities according to [4]). It cannot be ruled out, therefore, that some lay persons - even after some training as described above - might have perceptually identified a given voice quality correctly but were unsuccessful in handling the terminology.

5. REFERENCES

[1] Eckert, H., Laver, J. 1994. *Menschen und ihre Stimmen*. Weinheim: Beltz Psychologie Verlags Union.

[2] Hirano, M. 1981. *Clinical Examination of Voice*. Wien, New York: Springer.

[3] Köster, O., Köster, J.-P. 2004. The auditory-perceptual evaluation of voice quality in forensic speaker recognition. *The Phonetician* 89, 9–37.

[4] Laver, J. 1980. *The Phonetic Description of Voice Quality*. Cambridge etc.: Cambridge University Press.

[5] Nolan, F. 2005. Forensic speaker identification and the phonetic description of voice quality. In: Hardcastle, W.J., Mackenzie Beck, J. (eds), *A Figure of Speech*. London: Lawrence Erlbaum Associates, 385–411.