DOI: 10.17469/O2108AISV000010

ANGELIKA BRAUN

The notion of speaker individuality and the reporting of conclusions in forensic voice comparison

This contribution addresses some principal issues in forensic voice comparison, reflecting on some of the topics which have dominated the discussion among experts in the past two decades. The issue of speaker individuality is linked to the way in which conclusions in forensic voice comparison cases are expressed. The recent discussion about expressing conclusions in terms of likelihood ratios in forensic voice comparison is critically reviewed here. It is argued that likelihood ratios are not as unequivocal as they are said to be, neither are they popular with the triers of fact. Results of a survey among members of the judiciary on this topic are presented. It demonstrates once again that verbal probabilities are preferred even though, strictly speaking, they are logically flawed.

Keywords: voice comparison, idiosyncrasy, likelihood ratio, speaker individuality, phrasing of conclusions.

1. *Is speaking individual?*

Intuitively, there is little doubt that speaking is highly individual and that it is among the phenomena which lend themselves to be used as a biometric. Indeed, there is a large number of publications indicating that listeners are very good at recognizing familiar voices (Hollien, Majewski & Doherty, 1982, Skuk, Schweinberger, 2013, Braun, Kraft, 2013, Maguinness, Roswandowitz & Kriegstein, 2018) or famous voices (van Lancker, Kreiman & Emmorey, 1985, Schweinberger, Herholz & Sommer, 1997). The terms "idiosyncratic" or "idiosyncrasy" have been used by a number of authors to underline the individual character or voices (cf. Baldwin, French, 1990: 80; Kienast, Glitza, 2003, Dellwo, Leeman & Kolly, 2012 etc.). These terms, however, are somewhat misleading because they are prone to creating the misconception of speaker identification being comparable to fingerprint or DNA evidence. This view culminated in Lawrence Kersta's *voiceprint* analogy (Kersta, 1962), which essentially stated that one could positively identify speakers by visually comparing spectrograms of ten[1] words frequently used when talking over the telephone (Kersta, 1962). This voiceprint analogy – in combination with the emergence of color spectrograms which can be interpreted as bearing some similarity to fingerprints – has been deeply engraved in people's minds and has proven very difficult to eradicate. Major films, e.g. *Clear and Present Danger* and

[1] It is quite obvious that this number was deliberately chosen in order to emphasize the analogy to fingerprinting. *Honi soit qui mal y pense*.

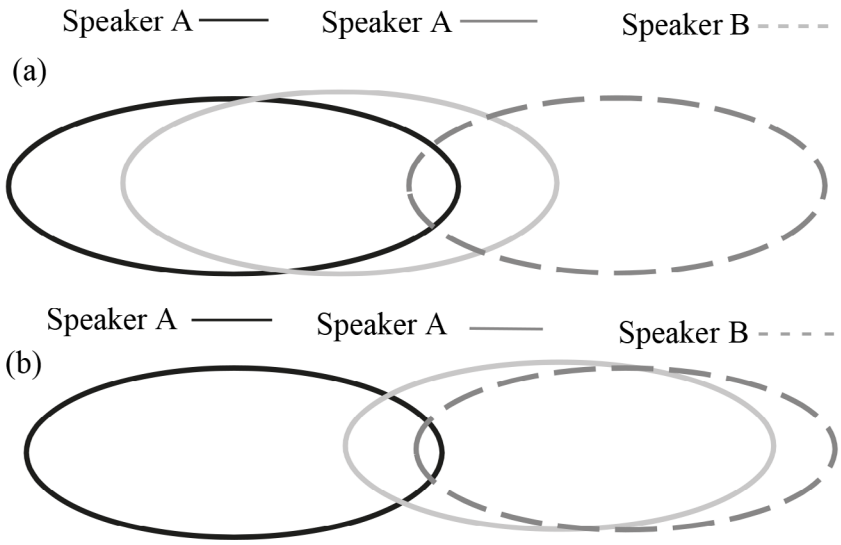
TV series such as *NCIS* have contributed to the misconception that voices work like fingerprints: smart young scientists take two-time signals or spectrograms, superimpose them and conclude "It's a match." The impact of these movies and TV shows has generated expectations with which the expert is then confronted in court, and it is sometimes no easy task to convince the triers of fact that this rendition is entirely fictional.

This is because the voiceprint analogy is simply untenable for two main reasons: Neither the larynx nor the vocal tract is an anatomical constant as, e.g., the fingertips are, and speaking goes far beyond implementing the anatomical bases. Francis Nolan (1983: 27-28) has coined the term *plasticity of the vocal tract* in order to describe the fact that the vocal tract configuration may be altered by, e.g., pulling up the larynx in a stressful situation and thereby reducing the size of the pharyngeal cavity or by protruding one's lips and thereby enlarging the oral cavity. In a similar way, the source function is to a certain extent subject to change according to mood, situation, and individual preference. Once again, Francis Nolan provides us with a fitting description: "In the real world, speakers *communicate* rather than merely exercise their vocal apparatus" (Nolan, 1983: 73; emphasis mine, AB). Since communicating is part of human behavior, it is inherently variable. Varying emotional states will alter the phonetic form of an utterance, and indirect speech (e.g. in verbal irony) will influence phonetic realizations (Braun, Heilmann, 2012; Braun, Schmiedel, 2018). Consuming alcohol or other neurotoxic agents have been demonstrated to induce changes to voice and articulation (cf. e.g. Künzel, Braun & Eysholdt, 1992). Beyond these short-term behavioral aspects, long-term changes to the human vocal apparatus are induced by the aging process. This will cause the human voice to sound different with advancing age (cf. e.g. Linville, 2001). This list does not even take into account any illnesses which involve the speech organs, among them common colds, sinusitis or – as an extreme example – laryngeal or pulmonic cancer, or, for that matter, deliberate disguise. In other words, while speaking gives the impression of being highly individual, even those familiar with a speaker will, under certain circumstances, fail to recognize him/her. This is why speech has been called a *performance biometric* (Hansen, Hasan, 2015: 76, emphasis mine, AB).

We can therefore draw the interim conclusion that voices are definitely not individual in the same way as fingerprints or DNA are. This has consequences for the expression of conclusions in voice comparison reports (see below). Voices are, however, individual in the sense that they reflect a combination of a speaker's anatomy, physiology and learned behavior. "Reflect" means that there are anatomical limitations within which a speaker can produce "sound"[2]. It is within that range that all of his/her speaking behavior takes place. Exceeding the anatomical / physiological limits is not possible.

[2] For reasons of brevity, I take this to include respiratory, phonatory, articulatory and linguistic features alike.

Figure 1 - *Different scenarios in voice comparison (see text for explanations)*



In effect, the question of voice individuality implies the question of whether *intra*speaker variability is always smaller than *inter*speaker variability. At present, there is no proof that this is actually the case – it may well not be (cf. Lee, Keating & Kreiman, 2019). Fig. 1 exemplifies the problem. Let us assume that there are two recordings of speaker A and one recording of speaker B. Scenario (a) is the default case: speaker A's speech behavior shows a much closer resemblance to his own voice on a different occasion than to that of speaker B. Yet there is likely to be some overlap[3] with speaker B as well, e.g. the fact that they are both male[4]. If we are willing to accept that scenario (b) is also conceivable (i.e., speaker A showing more overlap with speaker B than with himself on a different occasion) then speaking is not as individual as intuition suggests. Stretching articulation to the limits may create the kind of overlap with other speakers that is shown in Fig. 1 (b). For instance, if a person screams in panic, his or her voice is likely to be more similar to another person screaming in panic than to him/ or herself when talking quietly. There is speaker identity between the two recordings of (A), but that does not show in the samples provided. Studies by Lavan and colleagues (Lavan, Burston, & Garrido 2019; Lavan, Burton, Scott & McGettigan 2019) confirm this theoretical assumption. They found that listeners unfamiliar with the voices in question had

[3] "Overlap" is used here with reference to phonetic properties used in the auditory-acoustic framework, but the same principle would apply to the distribution of parameters such as MFCCs in (semi-) automatic approaches.

[4] It is only natural that the overlap between the different recordings of speaker A and speaker B may vary slightly.

considerable difficulties in "telling voices together". Unlike fingerprints or DNA, no one-to-one relationship between anatomy and speech output can be expected.

This brings us to the second question, i.e. the robustness of speaker specific features to the circumstances which characterize the forensic setting. They often exhibit mismatch conditions of various origins: there may be a situational mismatch (shouted vs. normally articulated speech; joyous vs. fearful speech, etc.) or a technical mismatch (landline vs. GSM transmission; HiFi vs. low mp3 coding, etc.) or even volitional changes to a given individual's speech behavior with the intention of disguising his or her identity (cf. Jessen, 2008: 677). Thus the features to rely on in the forensic setting not only have to be speaker specific but also resistant to technical issues such as transmission and coding (Wolf, 1972).

It has been suggested in the past that automatic speaker recognition systems cannot deal very well with mismatches caused by e.g. emotional states because the mel frequency cepstral coefficients (MFCCs) which most of them rely on are said to represent vocal tract geometry (Becker, 2012). If, however, vocal tract configuration cannot be assumed to be invariable, systems relying on it should be expected not to perform very well if e.g. the speaking situation changes. This actually seems to be the case: Automatic systems are extremely susceptible not only to channel mismatch (cf., e.g., Becker, 2012; Ajili, 2017) but also to behavioral mismatch and cannot deal with disguise at all (González Hautamäki, Sahidullah, Hautamäki & Kinnunen, 2017; González Hautamäki, Hautamäki & Kinnunen, 2019).

2. *The phrasing of conclusions*

2.1 Verbal scales vs. likelihood ratios

Since speech is a performance biometric with all the ramifications described above, there is no straightforward way of expressing the conclusions in voice comparison reports. There is evidently no easy way of arriving at some kind of numerical format. A similar problem occurs in handwriting reports. That is why forensic phoneticians have been applying a verbal probability scale resembling that used by handwriting experts (cf. Köller, Nissen, Rieß & Sadorf, 2004). Traditionally, i.e. in reports using the auditory-acoustic approach, verbal probability ratings are common which are based in part on background data and in part on the expert's individual assessment (Künzel, 1987; Wagner, 2019). This practice, however, has met with harsh criticism for about the past two decades. Verbal probabilities have mainly been criticized for

- being logically and statistically flawed and prone to the so-called prosecutor's fallacy, i.e. transposing the conditional or assessing the probability of the hypothesis given the evidence rather than the other way round (cf., e.g., Champod, Meuwly, 2000; Rose, Morrison, 2009; Morrison, Enzinger, 2016);
- being logically flawed because they state posterior probabilities without having a valid basis for determining the prior probability (Morrison, Enzinger, 2019);
- being subjective (i.e. not comparable between experts) (Hansen, Hasan, 2015);

– addressing similarity only and ignoring the question of typicality (Rose, 2006: 168).

Instead, Bayesian statistics has been invoked as the logically correct alternative. Specifically, stating likelihood ratios as opposed to verbal scales has been proposed as the method of choice. Two publications with reference to voice comparison initiated major steps in this direction: a paper by Christophe Champod and Didier Meuwly (2000) and Phil Rose's 2002 monograph. Ever since then, a likelihood ratio-based approach has been considered "modern" (cf. e.g. Rose, Morrison, 2009: 142) as opposed to verbal probabilities which, by way of implication, are seen by those authors as old-fashioned at best but essentially as untenable.

To the advocates of Bayesian statistics, calculating likelihood ratios as opposed to estimating probabilities represents a "paradigm shift" (Morrison, 2009) to "modern thinking" (Rose, 2006). They did not shy away from using grand words:

> We are in the midst of a paradigm shift in the forensic comparison sciences. The new paradigm can be characterised as quantitative data-based implementation of the likelihood-ratio framework with quantitative evaluation of the reliability of results" (Morrison, 2009: 298).

While criticism of the traditional verbal scales is certainly justified to some extent, there are practical considerations which raise questions about the use of likelihood ratios as well. Some of them will be addressed in the following sections[5].

2.2 Similarity and typicality

Likelihood ratios can be explained in terms of the notions of similarity and typicality (cf., e.g., Rose, 2006: 168). In this framework, the numerator captures the degree of similarity between suspect and offender, i.e. the probability of the evidence given that the suspect is the offender whereas the denominator reflects the degree of typicality of both within a reference population.

At first glance, it looks as if verbal probabilities are concerned with similarity only. Phil Rose deserves credit for relentlessly pointing to the fact that similarity is only one element in voice comparison. The other element is typicality, i.e. an assessment of the frequency with which such similarity is encountered in the relevant population without the questioned and the known samples originating from the same speaker. This puts the similarity ratings into perspective.

While it was very important to point out that typicality must not be neglected, it would be wrong to allege that traditional verbal scales address only the similarity aspect while ignoring typicality. In fact, typicality has always been considered, explicitly or implicitly. Formulating a conclusion in traditional terms never did and still does not preclude attention to the typicality side of the medal (cf. also Broeders, 1999: 237; French, Harrison, 2007). In fact, it has actually been addressed in court for

[5] I am fully aware of the fact that this contribution contains some highly controversial ideas which will meet with harsh criticism from parts of the forensic community. However, I am convinced that a principled discussion about the use of likelihood ratios in forensic speaker comparison is long overdue.

decades. The only difference is that typicality is assessed by the expert, either by way of background statistics if they are available (F0 and some disfluencies) or by way of forensic experience. It is not stated as a numerical value, but in court testimony it is always pointed out that similarity is a necessary condition for a positive conclusion, but by no means a sufficient one for a high rating on the probability scale. The position on the scale depends on the typicality of the matching results. Instead, this view of establishing similarity and typicality in a two-stage process is also reflected in the UK Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases (French, Harrison, 2007).

While voice similarity may easily be established at the surface level, the decision of whether the differences (and, as a matter of fact, also the similarities) found are consistent with[6] speaker identity under the given circumstances is a task which can only be undertaken by the trained expert. There may, e.g., be cases in which a considerable degree of similarity is established due to situational or technical mismatch (cf. scenario [b] in Fig. 1) without speaker identity.

2.3 My likelihood ratio – your likelihood ratio

One of the questions that seems to be addressed only rarely in this context is whether likelihood ratios are really so unambiguous and "objective" as purported. In this section, a number of issues which raise doubt as to their objectivity are discussed.

A key issue in likelihood ratio computation is the choice of what is called the relevant population, i.e. the population which is used to calculate typicality. Much of the literature on likelihood ratios is devoted to that. Only if the databases used to train the various models meet the necessities of the case at hand will the likelihood ratio calculation produce adequate results. Therefore, the advocates of likelihood ratio solutions go to great lengths to explain the compilation of the appropriate background data (Morrison, 2018: 5-6.). That author develops complex scenarios about which data to use for training the models of the numerator and the denominator of the likelihood ratio under mismatch conditions of various sorts. It is quite clear from his wording that the outcome will vary according to the relevant population used for the model of the denominator.

Behavioral mismatch conditions in general are a problem in a likelihood ratio environment because, strictly speaking, the relevant population would have to be tailored to the individual case. For example, a specific reference population would have to be used which matches the emotional and physiological states displayed in the questioned and/or the reference recording. However, it is completely unrealistic to record a separate reference population for each and every case, because neither the time nor the necessary means exist. Depending on the recordings used to represent

[6] It should be noted that the British "Position Statement" (French, Harrison, 2007) states just this fact in those same words. This author is in complete accord with the spirit of that statement in this respect despite the criticism by Rose, Morrison (2009). The concerns about the semantics of "consistent with" apply just as well to the "support" wording as proposed in Rose (2002: 62).

the relevant population, there is clearly more than one likelihood ratio per case. This can be expected to be very difficult to communicate in court.

In the course of the voice comparison, there are still more – and possibly more critical – ways for the expert to influence the outcome of the likelihood ratio. The brief list which will be discussed here is expanded by advocates of the use of likelihood ratios (cf. Rose, 2003: 167-172). To start with, different statistical models will obviously produce different likelihood ratios, and unless the ground truth is known, there is no way of determining which one works best in a given case. Obviously, likelihood ratios will turn out differently depending on the parameters employed, whether they be MFCCs, formants, F0 or something different still.

On a different strand, selection and preprocessing of the materials by the expert will affect the likelihood ratio. For instance, either the questioned and/or the reference recordings may be edited in order to work with speech which is "representative" of the respective speaker, i.e., shouted passages, crying or laughing may or may not be removed. In rare cases, the expert may even choose to filter a recording before analyzing it. More examples could easily be added to this list.

On a final note, likelihood ratios do not allow for attaching weight to certain findings. For instance, the forensic practitioner will easily identify findings which will effectively rule out identity. An example would be a case in which one sample shows severe disfluencies throughout whereas the other consists of perfectly fluent speech only. If both samples may be assumed to represent the respective speaker's natural behavior it is safe to conclude based on this feature alone that they come from different speakers. This kind of finding, which could be termed a "killer criterion", will thus override similarities between the samples with respect to other criteria. Scenarios like this one are not adequately reflected in the likelihood ratio framework.

In conclusion, likelihood ratios are not the single objective measure which uniquely describes two competing probabilities – they are highly "negotiable" instead. And yet – by virtue of being reported as specific numbers – they suggest a degree of precision which is not tenable after a closer look. An opposing expert may arrive at a different likelihood ratio depending on the reference population, the Universal Background Model (UBM) and the fine detail of calculation (Morrison, 2017, Rose, 2006: 170ff.). While it is difficult enough to communicate different verbal probabilities to the trier of fact, it would seem close to impossible to make it clear in court that the "exact" numbers may differ even though none of the experts has committed an outright error.

Morrison and Enzinger (2019) actually address this issue briefly:

> It should be noted that procedures based on relevant data, quantitative measurements, and statistical models do require subjective judgments, but these are judgments about relevant populations and relevant data which are far removed from the output of the system. The appropriateness of such judgments should be debated before the judge at an admissibility hearing and/or the trier of fact at trial. After these initial judgments, the remainder of the system is objective (p. 31).

This procedure seems problematic in a number of ways. First of all, it is realistic in case-law jurisdictions only. There is nothing like an admissibility hearing in common-law contexts. Secondly, it seems as if the task to decide about the appropriateness of the relevant populations is passed on to the judge. It is argued here that instead it is incumbent on the expert to make such a decision, because that constitutes part of his or her expertise. On the other hand, there is abundant evidence that the triers of fact are not statistic-savvy and lack comprehension of the concepts underlying Bayesian statistics (cf. Sjerps, Bliesheuvel, 1999; de Keijser, Elffers, 2012, and §4 below) and are thus much less than the expert in a position to judge the appropriateness of the methodology employed. Finally, the remark about the objectivity of the "remainder of the system" seems a bit naive, because if the initial steps, on which all the following steps are based, are subjective, the output can hardly be described as objective. It should be noted that this is not an argument against subjective elements in forensic voice comparison. It is, however, an argument against the rash claim that likelihood ratios are objective.

One of the most ardent advocates of likelihood ratios, Geoff Morrison, adds an interesting facet to the discussion:

> The discussion [...] raises a distinction to be made between the likelihood ratio reported by the forensic practitioner and the likelihood ratio actually used by the trier of fact. These will not necessarily have the same value. The effective likelihood ratio that the trier of fact employs, i.e., the extent by which they update their beliefs with respect to the relative probabilities of the competing prosecution and defence hypotheses, will likely depend on the trier of fact's assessment of how much they trust what the forensic practitioner reports. For example, if the practitioner reports a likelihood ratio of 1000 and their appearance and manner instil confidence, the trier of fact might use an effective likelihood of 1000, but if the practitioner's appearance and manner do not instil confidence the trier of fact might be less trustful of what the practitioner reports and use an effective likelihood ratio of 100 instead (Morrison, Enzinger, 2016: 375).

Morrison and Enzinger argue that therefore the practitioner should "supply the trier of fact with empirical information about the precision of the system used to calculate the likelihood ratio" (ibid.). However, the trier of fact may judge the expert's report about the precision of his methodology just as subjectively as the likelihood ratio reported. Still, this argument brings us to the receiving end of the forensic report: the trier of fact.

3. *The role of the judicial system – adversarial vs. inquisitorial*

The belief that likelihood ratios will solve essential problems of evidence weighing in court is closely tied to what is called an adversarial jurisdiction. It does not come as a surprise that the most adamant advocates of the likelihood ratio framework are working in jurisdictions with a case law tradition, such as Great Britain, the U.S.A., and Australia. It has largely been ignored that the principles of evidence presentation

and evidence weighting are entirely different in continental Europe. For instance, the German judicial system goes back to Napoleonic law (code law) and is inquisitorial by nature, i.e. the court does not function as an arbiter but has to investigate the case. This implies that any evidence has to be laid out in court, and that the judges need to establish any facts which constitute the basis for the verdict (for a summary cf. Margot, 1998 Braun, Köster, Künzel & Odenthal, 2005). There is usually just one court-appointed expert, which entails the danger that the court will not realize the relativity of numerical likelihood ratios and fall for numbers instead.

The court decides on the question of guilt as well as the sentence. There is no jury proper, but instead a panel of three or five judges (depending on the severity of the offence), two of whom are lay judges. Decisions are made by majority vote, which means that in a panel of three, the two lay judges may outvote the professional judge. They act according to the principle of free assessment of evidence (*liberté d'appréciation*) instead. The verdict is based on their subjective conviction (*l'intime conviction*); in order to convict, they do not need certainty or even probability bordering on certainty, but "a degree of certainty which is viable in real life and which silences doubts without ruling them out completely" (BGH, rulings of 14 January 1993 and of 11 December 2012; translation mine, AB)[7].

Strictly speaking, there is thus no need for calculating the weight of the evidence – the only prerequisite for a conviction is that the judges are convinced that the defendant is guilty. If that is not the case, the principle of *in dubio pro reo* applies. In this system, the triers of fact are seeking an informed opinion by an experienced expert about the individual case at hand, not necessarily a numerical value involving complex statistical considerations such as a discussion of the appropriateness of the relevant population as is suggested by Morrison, Enzinger (2019: 17). Or, as Broeders (1999, 238) put it: "The crucial question is not whether a conclusion arrived at by an expert is subjective or objective but whether it can be relied upon to be correct." This may help explain their preference for statistically incorrect conclusions over likelihood ratios (Sjerps, Biesheuvel, 1999; de Keijser, Elffers, 2012).

4. *What our "clients" expect – another survey*

Since Sjerps and Biesheuvel's results date back to the 1990s, the present author conducted her own survey among members of the judiciary. The survey was carried out in spring of 2021. It was distributed through the German Judges' Academy, Trier. A total of 41 judges and prosecutors participated in it. They were presented with a total of nine ways of expressing conclusions, among them two versions each of verbal probabilities and likelihood ratios. A short version and a long version were given of both verbal probabilities (VP) and likelihood ratios (LR). In the former case this means "There is a high probability in favor of voice identity" vs. "The

[7] IX ZR 238/91 NJW 1993, 935 under II 3a, and VI ZR 314/10, NJW 2013, 790 Rn. 16f.

examination of the materials revealed a high degree of matching elements between the questioned and the known recordings and at the same time an absence of relevant differences. Based on experience, the matching elements are judged to be rare". The same principle applied to likelihood ratios: the short version was "The likelihood ratio amounts to 1000/1 in this case", the long one was "The probability of the evidence given that questioned and reference materials originate from the same speaker is 1000 times greater than if they do not". The participants were asked to rate each phrase on its own merit on a five-point scale. (1) meant "unacceptable", (3) was neutral, and (5) meant "this is what I would really like to see in court." Only a selection of the results is presented here for reasons of space. They cover likelihood ratios and verbal probabilities only. Tab. 1 and Fig. 2 show the results.

Figure 2 - *Conclusion ratings by type and length*

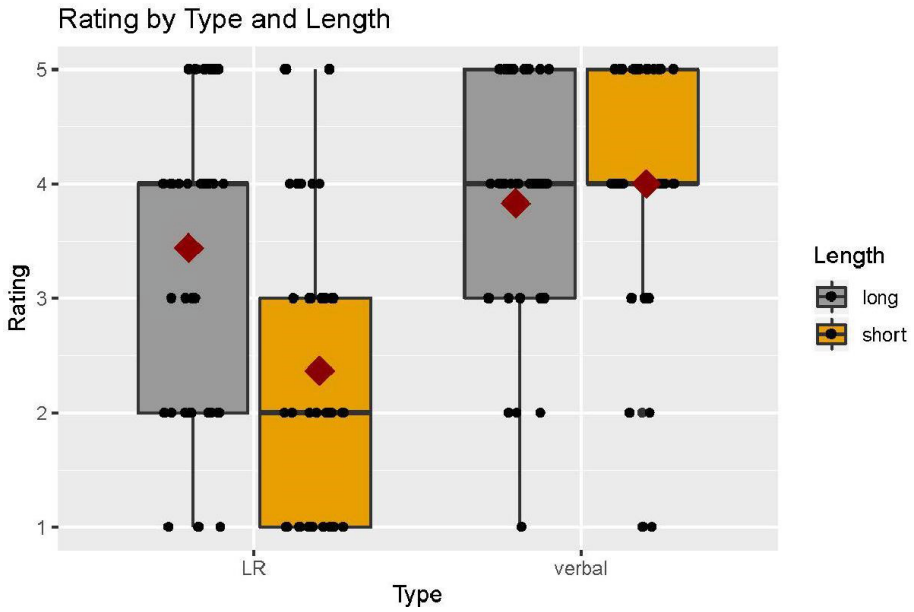


Table 1 - *Results of a survey on the expression of conclusions (N = 41)*

Wording No of votes	VP short	VP long	LR short	LR long
Score (mean)	4.0	3.8	2.3	3.4
1 & 2 combined	4	5	24	13
4 & 5 combined	33	29	9	24
(4+5)/(1+2)	8.3	5.8	0.4	1.8
5/1	7.0	11.0	0.2	3.3

It is quite obvious that verbal probabilities are the preferred way of phrasing conclusions. They obtain by far the highest mean score of 4.0 (cf. Tab. 1, row 1)[8]. The picture becomes even clearer if we ignore the "neutral" judgements and calculate the ratio between the positive ratings (4&5; cf. Tab. 1, row 3) on the one hand and the negative ones (1&2; cf. Tab. 1, row 2) on the other. Results are shown in Tab. 1, row 4. Higher numbers imply more client satisfaction. It is quite clear that both variants of the verbal probabilities are judged to be superior to likelihood ratios. A similar result emerges if only the extremes are considered, i.e., 1 and 5 (cf. Tab. 1, row 5).

Statistical analyses were carried out with respect to verbal probabilities and likelihood ratios. A Friedman test plus post-hoc pairwise comparisons was run. The rating score was significantly different for the four conditions (LR_short, LR_long, verbal_short, and verbal_long), $X2(3) = 29.1$, $p < .0001$. Post-hoc pairwise comparisons were carried out using a paired Wilcoxon signed-rank test. P-values were adjusted using the Bonferroni procedure for multiple testing correction. Only three of the pairs showed highly significant results: the two versions of likelihood ratios (long vs. short; $p = .000417$), the short version of likelihood ratios vs. the long version of verbal probabilities ($p = .000239$), and the short version of likelihood ratios vs. the short version of verbal probabilities ($p = .000116$). The difference between verbal probabilities and the long version of likelihood ratios did not reach significance. This finding could be interpreted to imply a growing acceptance of likelihood ratios provided that they are verbalized and not given as sheer numbers. However, it cannot be taken to imply a growing understanding of the concept of likelihood ratios, because their acceptance rate does not differ significantly from that of the "dummies" (cf. footnote 7).

The results of the present survey confirm those obtained by Sjerps, Biesheuvel (1999) and de Keijser, Elffers (2012) to a large extent. Verbal probability ratings are preferred over likelihood ratio formats, and the understanding of statistical correctness seems to be limited. All things considered, the preference for verbal probabilities as currently used is evident.

5. *Reactions to the surveys*

There have been two principal reactions by the advocates of likelihood ratios to results like these: the resolve (a) to educate the members of the judiciary (cf., e.g., Morrison, Enzinger, 2016), and (b) to reconvert likelihood ratios to verbal probabilities (cf., e.g., Rose, 2002). Neither of them sounds really convincing. What is overlooked in (a) is that it is not the expert's role to educate the members of the court, and any attempt do to so may lead to that expert not being commissioned again. Furthermore, even if the expert were to get her or his point across and actually

[8] Likelihood ratios are even surpassed in popularity by the two "dummies" which were introduced because they are common in fictional contexts: "per cent agreement" and "per cent identical".

use likelihood ratios, it is very probable that they will not be interpreted correctly (de Keijser, Elffers, 2012).

In recognition of this situation and in order to comply with the needs of the "clients", there have been several attempts to reconvert likelihood ratios to verbal probabilities. This way forward was chosen by Evett and colleagues at the British Forensic Science Service (Champod, Evett, 2000). Similar scales are cited by Rose (2002, 62); and were adopted in the European Network of Forensic Science Institutes (ENFSI) guideline for evaluative reporting in forensic science (Willis, McKenna, McDermott, O'Donell, Barrett, Rasmusson, Nordgaard, Berger, Sjerps, Lucena-Molina, Zadora, Aitken, Lunt, Champod, Biedermann, Hicks & Taroni, 2015: 17).

The latter expressly states that "[...] likelihood ratios can be informed by subjective probabilities using expert knowledge" (p. 16). While this can be regarded as a compromise among the many laboratories organized within the ENFSI framework, it carries the label of likelihood ratios without implementing the idea behind them.

Reconverting likelihood ratios to verbal scales is probably the worst of all choices because it combines the problems of verbal scales with those of the likelihood ratios. In addition, it would potentially be very confusing if different experts were to use different conversion scales, which might imply that one and the same numerical value would correspond to different verbal probabilities depending on who is being asked. Furthermore, there is the problem of cliff edge effects. This describes the fact that at certain points a small change in one parameter will induce a vast change in the results while it does not have this effect elsewhere. For example, according to Rose's verbal likelihood ratios (Rose 2002: 62), a difference in likelihood ratio of a magnitude of 2 between 999 and 1,001 amounts to one full step on the verbal scale, whereas the difference of 8,999 between 1,001 and 10,000 does not. Accordingly, these attempts at translating likelihood ratios into verbal scales have been pointed out as equally flawed statistically as are the verbal scales.

Two major surveys on voice comparison methodology in the past ten years have included the reporting of conclusions. They reveal a deplorable proliferation of conclusion frameworks: In Europe alone there are five different ones listed by Morrison et al. (2016: 96), while Gold, French (2019: 11) even list six. In Britain, there have been several changes (traditional probabilities, UK Position Statement, Support Statement, and verbal likelihood ratios) within less than ten years. This does not convey the impression that the scientific community has a unified strategy, which is, however, a prerequisite for rendering a method acceptable in court.

6. *Summary and conclusions*

"Expert opinions in the forensic sciences are always uncertain" (Edmond, Towler, Growns, Ribeiro, Found, White, Ballantyne, Searston, Thompson, Tangen, Kemp & Martire, 2017:148). This applies to voice comparison in particular because speaking is part of human behavior and thus inherently variable. This variability is one factor introducing mismatches between questioned and reference materials.

Another source of mismatch is constituted by differing technical specifications of the materials. In the context of (semi-)automatic speaker recognition, where the reference recording is used to model the numerator and a relevant population is used to model the denominator of the likelihood ratio, a mismatch of mismatches may occur on top (Morrison, 2018), if, e.g. there is a behavioral mismatch between questioned and known samples and a technical mismatch between questioned sample and the relevant population. This is, of course, a scenario which is frequently encountered in the forensic setting.

Furthermore, there are a number of ways in which the processing of the materials by the expert will affect likelihood ratios. There are various preparatory steps to be taken before the analysis proper, which involves decisions on the part of the expert. These decisions include the selection of materials to be analyzed in the first place – coughs, laughter, throat clearing, shouting should be edited out before making the comparison in order not to distort the results. There may be a need for preprocessing the data if the data format or the recording quality differs between the questioned and the reference materials. All these steps will invariably affect the likelihood ratio.

For these and other reasons as spelled out by Rose (2006: 167-172), the conclusions expressed in terms of likelihood ratios are by no means carved in stone but depend to some degree on the handling of the materials by the expert. It is sometimes suggested that likelihood ratios are essentially "objective" in contrast to the traditional probability scales (Morrison, Enzinger, 2019: 31). The present author would recommend rethinking this notion. Specifically, the assumption that they are objective simply because likelihood ratios are derived from using an automatic speaker ID system is untenable.

There can be absolutely no doubt that both similarity and typicality have to be assessed in forensic voice comparison. However, it would be wrong to say that typicality is not included in the traditional verbal probabilities. This is inherent in the expert's task, comparable to assessing the regional dialect or speech impairments. In the conviction of this author, this "subjective" element is something the courts know about and can deal with if they seek the assistance of an expert. The courts rely on the expert's personal (and therefore to some extent subjective) judgment, and this is what they are entitled to receive. What is important, though, is that the experts point to this subjective element in their reports.

The rephrasing of likelihood ratios in terms of traditional wordings constitutes no real advantage over the traditional assessment by the expert. It rather makes things worse by purporting to have an "objective" basis, which it really does not. Another argument against this way forward is that it would not work within common-law jurisdictions, where it would be entirely inappropriate for an expert to refer to a "prosecutor's or defense attorney's hypothesis"[9].

In both case-law and code-law jurisdictions, expert evidence is supposed to rely on established procedures which are widely recognized by the scientific community. Even though one individual author, by way of a multitude of publications, is currently

[9] This, by the way, was expressly stated by one of the participants in the present survey.

trying to dominate the discussion, this does not mean that his assertions are established knowledge which is generally accepted by the forensic phonetics community. This is simply not the case, as is documented by the Interpol survey (Morrison, Sahito, Jardine, Djokic, Clavet, Berghs & Dorny, 2016) as well as the one by Gold, French (2019). It is one intention of the present paper to make that very clear.

We may after all have to accept the notion that even after 20 years of debate and countless publications, the "objectivity" of likelihood ratios can be challenged on some counts and that the much criticized verbal probability scales, which admittedly involve subjective judgment on the part of the expert, do have some forensic merit, not least that they conform to the expectations of the triers of fact. It would be a potentially interesting endeavor to compare verbal likelihood ratios with traditional verbal probabilities based on identical material and determine whether they are really that far apart.

Acknowledgement

The author would like to thank Katharina Zahner-Ritter for her assistance with the statistical analysis.

Bibliography

Ajili, M. (2017). Reliability of voice comparison for forensic applications. PhD Dissertation, University of Avignon.

Baldwin, J.R., French, P. (1990). *Forensic phonetics*. London, UK: Pinter Publishers.

Becker, T. (2012). *Automatischer Forensischer Stimmenvergleich*. Verlag Book on Demand.

Braun, A., Köster, J.-P., Künzel H.J. & Odenthal, H.-J. (2005). Speaker Recognition in Germany. In Wolf, D. (Ed.), *Beiträge zur Geschichte und neueren Entwicklung der Sprachakustik und Informationsverarbeitung, Werner Endres zum 90. Geburtstag*. Dresden: w.e.b. Universitätsverlag, 78-86.

Braun, A., Heilmann, C.M. (2012). *SynchronEmotion*. Frankfurt etc.: Peter Lang.

Braun, A., Kraft, L. (2013). Die Erkennbarkeit vertrauter Stimmen bei Verstellung. In Mehnert, D., Kordon, U. & Wolff, M. (Eds.). *Systemtheorie. Signalverarbeitung. Sprachtechnologie. Rüdiger Hoffmann zum 65. Geburtstag.* Dresden: TUDpress, 226-233.

Braun, A., Schmiedel, A. (2018). The phonetics of ambiguity: a study on verbal irony. In Winter-Froemel, E., Thaler, V. (Eds.). *Cultures and traditions of wordplay and wordplay research*. Berlin, New York: De Gruyter, 111-136. https://doi.org/10.1515/9783110586374-006

Broeders, A.P.A. (1999). Some observations on the use of probability scales in forensic identification. In *The International Journal of Speech, Language, and the Law* 6(2), 228-241. https://doi.org/10.1558/sll.1999.6.2.228

Champod, C., Evett, I.W. (2000). Commentary on APA Broeders (1999) 'Some observations on the use of probability scales in forensic identification', Forensic Linguistics 6 (2): 228–241. In *The International Journal of Speech, Language and the Law*, *7*(2), 239-243.

Champod, C., Meuwly, D. (2000). The inference of identity in forensic speaker recognition. In *Speech communication,* 31(2-3), 193-203. https://doi.org/10.1016/S0167-6393(99)00078-3.

Dellwo, V., Leemann, A., & Kolly, M.J. (2012). Speaker idiosyncratic rhythmic features in the speech signal. In *Proceedings of Interspeech 2012*, Portland, Oregon, 9-13 September 2012. 10.5167/uzh-68554.

De Keijser, J., Elffers, H. (2012). Understanding of forensic expert reports by judges, defense lawyers and forensic professionals. In *Psychology, Crime & Law,* 18(2), 191-207. https://doi.org/10.1080/10683161003736744

Edmond, G., Towler, A., Growns, B., Ribeiro, G., Found, B., White, D., Ballantyne, K., Searston, R.A., Thompson, M.B., Tangen, J.M., Kemp, R.I. & Martire, K. (2017). Thinking forensics: Cognitive science for forensic practitioners. In *Science & Justice,* 57(2), 144-154. https://doi.org/10.1016/j.scijus.2016.11.005

French, J.P., Harrison, P. (2007). Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases. In *International Journal of Speech Language and the Law,* 14(1), 137-14. 10.1558/ijsll.v14i1.137

Gold, E., French, J.P. (2019). International practices in forensic speaker comparison: second survey. In *International Journal of Speech, Language and the Law,* 26(1), 1-20. 10.1558/ijsll.38028

González Hautamäki, R., Sahidullah, M., Hautamäki, V. & Kinnunen, T. (2017). Acoustical and perceptual study of voice disguise by age modification in speaker verification. In *Speech Communication,* 95, 1-15. https://doi.org/10.1016/j.specom.2017.10.002

González Hautamäki, R. Hautamäki, V. & T. Kinnunen (2019). On the limits of automatic speaker verification: Explaining degraded recognizer scores through acoustic changes resulting from voice disguise. In *The Journal of the Acoustical Society of America,* 146, 693-704. https://doi.org/10.1121/1.5119240

Hansen, J.H., Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. In *IEEE Signal processing magazine,* 32(6), 74-99. 10.1109/MSP.2015.2462851

Hollien, H., Majewski, W. & Doherty, E.T. (1982). Perceptual identification of voices under normal, stress and disguise speaking conditions. In *Journal of Phonetics*, 10(2), 139-148.

Jessen, M. (2008). Forensic Phonetics. In *Language and Linguistics Compass,* 2, 1-41. https://doi.org/10.1111/j.1749-818X.2008.00066.x

Kersta, L.G. (1962). Voiceprint identification. In *Nature,* 196, 1253-1257. https://doi.org/10.1038/1961253a0

Köller, N., Nissen, K., Riess, M. & Sadorf, E. (2004). *Probabilistische Schlussfolgerungen im Schriftgutachten. Zur Begründung und Vereinheitlichung von Wahrscheinlichkeitsaussagen im Sachverständigengutachten.* München: Luchterhand.

Nolan, *F.* (1983). *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press.

Kienast, M. & Glitza, F. (2003). Respiratory sounds as an idiosyncratic feature in speaker recognition. In *Proceedings of 15th ICPhS,* Barcelona, Spain, 3-9 August 2003, 1607-1610.

Künzel, H.J. (1987). *Sprechererkennung. Grundzüge forensischer Sprachverarbeitung*. Heidelberg: Kriminalistik-Verlag.

Künzel, H.J., Braun, A. & Eysholdt, U. (1992). *Einfluss von Alkohol auf Sprache und Stimme.* Heidelberg: Kriminalistik-Verlag.

Lavan, N., Burston, L.F. & Garrido, L. (2019). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. In *British Journal of Psychology*, 110(3), 576-593. https://doi.org/10.1111/bjop.12348

Lavan, N., Burton, A.M., Scott, S.K., & McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. In *Psychonomic Bulletin & Review*, 26(1), 90-102. https://doi.org/10.3758/s13423-018-1497-7

Lee, Y., Keating, P., & Kreiman, J. (2019). Acoustic voice variation within and between speakers. In *The Journal of the Acoustical Society of America*, 146(3), 1568-1579. https://doi.org/10.1121/1.5125134

Linville, S.E. (2001). *Vocal aging.* San Diego: Singular.

Maguinness, C., Roswandowitz, C., & von Kriegstein, K. (2018). Understanding the mechanisms of familiar voice-identity recognition in the human brain. In *Neuropsychologia,* 116, 179-193. 10.1016/j.neuropsychologia.2018.03.039

Margot, P. (1998). The role of the forensic scientist in an inquisitorial system of justice. In *Science & Justice* 38(2), 71-73. 10.1016/S1355-0306(98)72080-5

Morrison, G.S. (2009). Forensic voice comparison and the paradigm shift. In *Science & Justice*, 49(4), 298-308. https://doi.org/10.1016/j.scijus.2009.09.002

Morrison, G.S. (2017). The impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings. In *Forensic Science International*, https://doi.org/10.1016/j.forsciint.2017.12.024

Morrison, G.S. (2018). The impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings. In *Forensic Science International*, 283, e1-e7. 10.1016/j.forsciint.2017.12.024

Morrison, G.S., Enzinger, E. (2016). What should a forensic practitioner's likelihood ratio be? In *Science & Justice*, 56(5), 374-379. 10.1016/j.scijus.2016.05.00

Morrison, G.S., Enzinger, E. (2018). Score based procedures for the calculation of forensic likelihood ratios – Scores should take account of both similarity and typicality. In *Science & Justice*, 58(1), 47-58. https://doi.org/10.1016/j.scijus.2017.06.005

Morrison, G.S., Enzinger, E. (2019). Introduction to forensic voice comparison. In Katz, W.F., Assman, P.F. (Eds.), *The Routledge Handbook of Phonetics*. London: Routledge, 599-634. https://doi.org/10.4324/9780429056253-22

Morrison, G.S., Sahito, F.H., Jardine, G., Djokic, D., Clavet, S., Berghs, S. & Dorny, C.G. (2016). INTERPOL survey of the use of speaker identification by law enforcement agencies. In *Forensic Science International*, 263, 92-100. 10.1016/j.forsciint.2016.03.044

Rose, P. (2002). *Forensic speaker identification*. London, New York: Taylor & Francis.

Rose, P. (2003). The Technical Comparison of Forensic Voice Samples. In Selby, H., Freckelton, I. (eds). *Expert Evidence*. Sydney: Thompson Lawbook Co..

Rose, P (2006). Technical forensic speaker recognition. Evaluation, types and testing of evidence. In *Computer Speech and Language,* 20(2-3), 159-191. 10.1016/j.csl.2005.07.003

Rose, P., Morrison, G. (2009). A response to the UK position statement on forensic speaker comparison. In *The International Journal of Speech, Language and the Law*, 16(1), 139-163. 10.1558/iisll.v16i1.l 39

Schweinberger, S.R., Herholz, A. & Sommer, W. (1997). Recognizing famous voices: Influence of stimulus duration and different types of retrieval cues. In *Journal of Speech, Language, and Hearing Research*, 40(2), 453-463. 10.1044/jslhr.4002.453

Sjerps, M. & Biesheuvel, D.B. (1999). The interpretation of conventional and 'Bayesian' verbal scales for expressing expert opinion: a small experiment among jurists. In *Forensic Linguistics* 6(2), 214-227. 10.1558/ijsll.v6i2.214

Skuk, V.G. & Schweinberger, S.R. (2013). Gender differences in familiar voice identification. In *Hearing Research*, 296, 131-140. 10.1016/j.heares.2012.11.004

Van Lancker, D., Kreiman, J. & Emmorey, K. (1985). Familiar voice recognition: patterns and parameters Part I: Recognition of backward voices. In *Journal of Phonetics*, 13(1), 19-38. https://doi.org/10.1016/S0095-4470(19)30723-5

Wagner, I. (2019). Examples of casework in forensic speaker comparison. In Calhoun, S., Escudero, P., Tabain, M. & Warren, P. (2019). *Proceedings of the 19th International Congress of Phonetic Sciences,* Melbourne*,* Australia, 5-9 August 2019, 721-725.

Willis, S.M., McKenna, L., McDermott, S., O'Donell, G., Barrett, A., Rasmusson, B., Nordgaard, A., Berger, C.E.H., Sjerps, M.J., Lucena-Molina, J.J., Zadora, G., Aitken, C.G.G., Lunt, L., Champod, C., Biedermann, A., Hicks, T.N. & Taroni, F. (2015). *ENFSI guideline for evaluative reporting in forensic science.* European Network of Forensic Science Institutes.

Wolf, J.J. (1972). Efficient acoustic parameters for speaker recognition. In *The Journal of the Acoustical Society of America*, 51(6B), 2044-2056. https://doi.org/10.1121/1.1913065