iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 11 **Issue:** IX **Month of publication:** September 2023

**DOI:** https://doi.org/10.22214/ijraset.2023.55714

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 11 Issue IX Sep 2023- Available at www.ijraset.com*

# Forensic Speaker Recognition: Review of Methods and Evaluations

Vinay Mishra
*Regional Forensic Science Laboratory, Bhopal (MP) India*

***Abstract:*** ***Forensic voice comparison is quite challenging because the quality of speech recordings in casework is often poor and there is often mismatch between the recording conditions of the known- and questioned-speaker recordings. Further, the field of forensic voice comparison is characterized by a large diversity of methods, procedures and evaluation of comparison results. The present paper aims to provide a concise review of various comparison methods and assessment of comparison results. The need of an extensive amount of experimental studies and statistical evaluation on a large database is highlighted to use the likelihood-ratio framework for reporting conclusions that has been now firmly established as a theoretical framework for any forensic discipline.***

***Keywords: Forensic Speaker Recognition, Auditory, Spectrographic voice identification, Acoustic-Phonetic voice identification***

## I. INTRODUCTION

The term speaker recognition is used as a cover term for the wide variety of situations in which people are identified, or strictly speaking individualized, on the basis of the sound of their voices. The term speaker identification is used for comparing a test speaker against all the voices in a particular database to determine his or her identity; and the term speaker verification is used for comparing a test sample with a reference sample of the speaker who is claimed to have produced the test sample. In a court of law, to establish the identity of a speaker on an audio recording, the court may call for an opinion of forensic practitioner on forensic speaker recognition. The task of the forensic practitioner is to compare the recording of the speaker of questioned identity with a recording of the speaker of known identity to help the court decide whether the voices on the questioned recording were produced by the same speaker or by different speakers [1]. In this sense, the forensic application typically amounts to a verification task, in that the question that needs to be answered tends to be whether the recorded voice is that of a particular speaker (i.e., the suspect) [1]. Forensic voice comparison is challenging because the quality of speech recordings in real-world forensic casework is often poor and there is often mismatch between the speaking style in and the recording conditions of the known- and questioned-speaker recordings. The recordings may contain only a few seconds of speech, they may include background noise of various sorts (e.g., babble, ventilation system noise, vehicle noise) at varying intensities, they may have been recorded in reverberant environments, they may have been recorded using microphones that are distant from the speaker of interest (e.g., covert microphones, telephone microphones picking up speakers other than the caller), they may have been transmitted through different transmission channels (e.g., landline telephone, mobile telephone, voice over internet protocol) that distort the signal, and they may have been saved in compressed formats (e.g., MP3) that also distort the signal [1].

## II. METHODS AND PROCEDURES

There are probably few forensic disciplines that are characterized by such a diversity of methods and procedures as the field of forensic speaker identification by experts [2]. The “auditory” or “aural-perceptual” method of voice identification is based on critical listening by a trained practitioner. The practitioner listens to the questioned-speaker recordings and known -speaker recordings in search of-*similarities in speech properties that they would expect if the two recordings were produced by the same speaker but not if they were produced by different speakers*; and, in search of *-differences in speech properties that they would expect if the two recordings were produced by different speakers but not if they were produced by the same speaker*.

The “spectrographic identification” method of voice identification is based on visually comparing spectrograms of words or phrases that occur in both the questioned-speaker recordings and known- speaker recordings. The practitioner visually compares the questioned-speaker recordings and known -speaker recordings in search of-*similarities in voice spectrograms that they would expect if the two recordings were produced by the same speaker but not if they were produced by different speakers*; and, in search of -*differences in voice spectrograms that they would expect if the two recordings were produced by different speakers but not if they were produced by the same speaker.*

The basis for the use of spectrograms in speaker identification in forensic settings is derived from the work of Lawrence Kersta [6] and Oscar Tosi [7]. Interestingly, Kersta, who was the first person to testify as a voice identification "expert "in 1966, did not utilize aural comparisons in his work and, incorrectly, claimed that the accuracy of voiceprints was comparable to that of fingerprints. In the forensic setting, spectrographic identification is different from fingerprint identification. This is because, in spectrographic identification, comparisons of spectrograms involve finding sufficient similarities, while fingerprints always have exactly the same pattern. The individual differences in the spectrograms, caused by intra-speaker variation, make it the examiners task to subjectively determine if a match exists. There has been substantial controversy surrounding the use of the spectrographic approach. Morrison [8] and Morrison &Thompson [9] have critically reviewed the controversy around its use and admissibility.

The "aural-spectrographic "method of voice identification involves both aural and spectrographic comparisons in forming an opinion based on the similarity or dissimilarity of the totality of the observed patterns [10]. As to the aural comparison, the examiner listens to each sample uttered by the known and unknown played in close temporal juxtaposition, thus enabling him to gauge the similarity of their aural patterns. The visual component of the comparison involves an assessment of the pattern similarity between the known and unknown samples' spectrograms. The examiner, rather than trying to somehow rate individual characteristics, allows knowledge of acoustic phonetics and experience in spectrographic pattern matching to guide him in evaluating the aggregate of appropriate patterns at his disposal. Such pattern matching is known in cognitive psychology as a "gestalt".

The "gestalt" pattern matching is only effective if the patterns being matched are comparable, it is crucial that these patterns correspond to the same phonetic utterance and simply having a known speaker utter the text of a transcription of an unknown's words does not assure that all of the utterances thus obtained will be usable in carrying out the comparison [10].

The "acoustic- phonetic" method of voice identification involves both aural and spectrographic comparisons and also makes quantitative measurements of acoustic- phonetic parameters, e.g. fundamental frequency and formant frequencies, voice onset time (VOT), fricative spectra, nasal spectra, etc. [11,12]. The examiner using acoustic-phonetic parameter measurements have to take into account the mismatch in recording channel between the known- and questioned-speaker recordings. Since questioned-speaker recordings are often telephone recordings and traditional telephone systems have band passes of around 300 Hz – 3.4 kHz, first formant (F1) frequencies are often distorted, and high frequency spectral information in bursts and fricatives is often missing; the effect of landline and mobile phone transmissions on voice comparisons have been reported by Zhang et al. [13]. From the measurements done for acoustic-phonetic parameters, some practitioners make tables or plots of the values, and use their training and experience to assess those tables or plots and use those as input to a subjective judgment process, which may also include consideration of the results of an auditory analysis. The measurements of "acoustic- phonetic" parameters may seem somewhat objective and thus obviously superior to the subjective "aural-spectrographic" method, but the author has observed sufficient real life instances of forensic comparisons to believe that grounding judgment on measurements of parameters alone represents a significant potential for error. Though the subjective method of "aural-spectrographic" identification is very tedious and time consuming, under the real life instances of forensic comparisons where the voice disguise and the mismatch in recording conditions is routinely observed, the subjective method provides better results. This has also been concluded by Wojciech Majewski [14] while comparing subjective and objective speaker recognition under voice disguise conditions.

The automatic speaker recognition techniques developed for non-forensic applications have also been adapted for forensic application though; the output of the system is not a binary decision but a quantification of strength of evidence [2]. In forensic application, in contrast to most security applications, the questioned speaker is generally not cooperative, the recording conditions and speaking styles are variable and the quality of the recordings are often much poorer. Hansen & Hasan [15] have reviewed human versus machine speaker recognition and attempted to point out strengths and weaknesses of each.

Morrison et al in INTERPOL survey of the use of speaker identification by law enforcement agencies [16] have reported use of a variety of approaches to speaker identification: the human-supervised-automatic approach was the most popular in North America, the auditory-acoustic-phonetic approach was the most popular in Europe, and the spectrographic/auditory-spectrographic approach was the most popular in Africa, Asia, the Middle East, and South and Central America.

## III. ASSESSMENT OF RESULTS

Morrison et al in 2016, in INTERPOL survey of the use of speaker identification by law enforcement agencies [16] reported that perhaps the clearest pattern to emerge from this survey was that of diversity. There was substantial variation in comparison methods and assessment procedures or frameworks,*-the term they used to refer to ways of reasoning or ways of drawing inferences, for reporting conclusions*, used for speaker identification by law enforcement agencies in different parts of the world.

In the identification/*exclusion/inconclusive* framework for reporting conclusions the practitioner only reports either "identification", i.e., 100% probability for same speaker, or "exclusion", i.e., 100% probability for different speaker, or declines to express an opinion "inconclusive". In making an "identification" or "exclusion" the forensic practitioner has made the decision as to same speaker or different speaker, which is properly a decision to be made by the trier of fact who also takes other evidence into consideration [1],[4]. Morrison and Thompson [9] referred to the PCAST report that opined "the expert should not make claims or implications that go beyond the empirical evidence and the applications of valid statistical principles" and considered the definitive statements of 100% certainty beyond what can be empirically supported.

In the *posterior-probability* framework for reporting conclusions the expressions such as "identification," "probable identification," "possible identification," "inconclusive," "possible elimination," "probable elimination," and "elimination" are verbal expressions of posterior probabilities. Logically, posterior probabilities cannot be derived solely via comparison of the properties of the known- and questioned-speaker recordings. The only logical way to derive a posterior probability is to combine the likelihood ratio with a prior probability and the logically correct framework for reporting conclusions for the evaluation of forensic evidence is the likelihood ratio framework. It may be the case that practitioners who present posterior probabilities without combining likelihood ratio with a prior probability may not be aware of the logical problems [1]. The mere consideration of the degree of similarity, without considering the typicality, of the voices on the known- and questioned-speaker recordings can not quantify strength of the evidence; may it be a subjective or a statistical mode. The proper way to evaluate forensic speech samples, and thus to evaluate the weight of the forensic-phonetic evidence, is by estimating the probability of observing the differences between them assuming that the same speaker is involved; and the probability of observing the differences between them assuming that different speakers are involved. This method is thus inherently probabilistic, and as such will not yield an absolute identification or exclusion of the suspect [4].

The *likelihood-ratio* framework for reporting conclusions was considered as the logically correct framework for the evaluation of forensic evidence by many forensic statisticians, forensic scientists, and legal scholars [1],[9],[17]. A forensic likelihood ratio is the probability of the evidence if the same-origin hypothesis were true divided by the probability of the evidence if the different-origin hypothesis were true. The forensic practitioner must estimate both the degree of similarity of the properties of the voice on the questioned-speaker recording with respect to the known speaker, and the degree of typicality of the properties of the voice on the questioned-speaker recording with respect to the relevant population. This, however, requires a large database to recognize the statistical distribution of the relevant parameters in the relevant populations. The relevant population is the population of people who could plausibly have produced the voice on the questioned-speaker recording if it were not produced by the known speaker. The relevant population and the conditions of questioned-speaker and known-speaker recordings can vary from case to case [19].

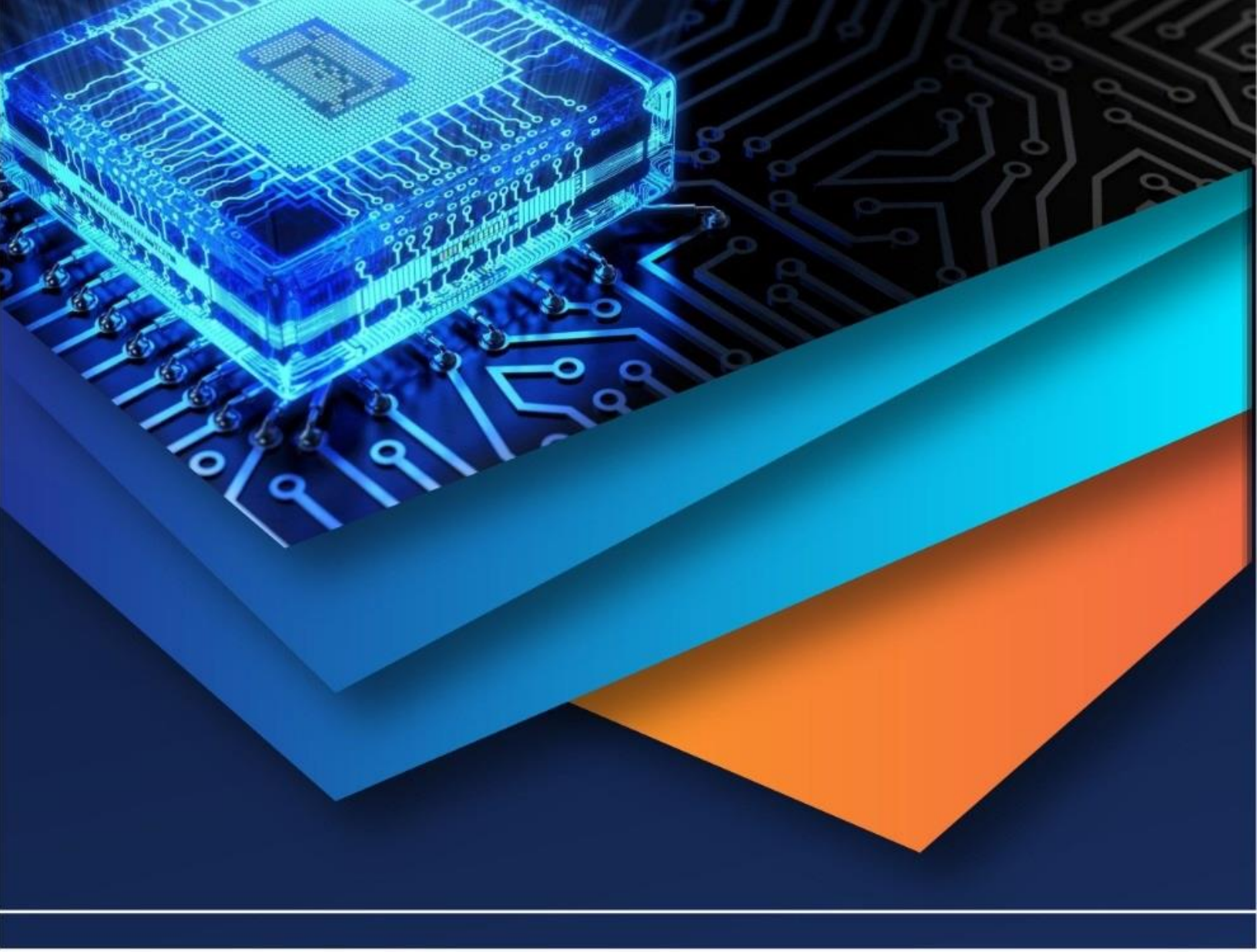
## IV. CONCLUSION

Gold, E. and French J.P. [11], in an international survey on forensic speaker comparison practices pointed out lack of consensus over fundamental matters in forensic voice comparisons, such as how speech samples are to be analyzed and compared, which aspects of the samples are to be assigned greatest importance during the analytic process, and how conclusions are to be expressed; though majority of experts participating in the survey were expressing their conclusions in terms of a classical probability scale: the probability or likelihood assessment being a verbal rather than a numerical one. This type of pattern evidence relied on the subjective judgments of trained examiners, and not on rigorous statistical analysis. Recently Thomas Busey and Meredith Coon [18] have observed that in the pattern comparison disciplines there are no widespread quantitative approaches and therefore most conclusions rely on subjective human evaluations; this indeed makes use of several subjective thresholds. In order to facilitate statistical analysis examiners need large database and a set of rules for feature selection that could then feed a statistical model that describes how unusual the set of similarities between two samples really is, relative to similarities between two randomly selected samples from the population.

The likelihood-ratio framework for reporting conclusions is now firmly established as a theoretical framework for any forensic discipline. The expert, lacking the knowledge of the prior probability (background information relative to the case: concern to the court), cannot logically combine it with their LR (likelihood-ratio) to estimate the posterior [20]. As pointed out by Gonzalez-Rodriguez et al [21], referring to the identification in forensic cases in general, an agreement must be achieved in every identification area, especially in the process of selection of the involved populations, and what characteristics to be used from this population. This requires an extensive amount of experimental studies and statistical evaluation of the different parameters with uniform conclusions based upon a large database; this, however, is not the case with forensic voice identification as it is being performed by many forensic practitioners till the present moment.

## REFERENCES

[1] Morrison, G.S., Enzinger, E. (2019), "Introduction to forensic voice comparison". In: The Routledge Handbook of Phonetics (ch. 21, pp. 599–634). Publisher: Abingdon, UK: Taylor &Francis, Editors: Katz W.F., Assmann P.F.
[2] A.P.A. Broeders (2008), "Speaker Identification in Forensic Arena", In : Law and Language: Theory and Society (pp.59-85), Publisher: Duesseldorf University Press, Editors: F. Olsen, A. Lorz, D. Stein.
[3] Nolan, F. (1997), "Speaker recognition and forensic phonetics", In: The handbook of phonetic sciences (pp. 744–767), Publisher: Oxford, UK: Blackwell, Editors: Hardcastle, W.J. and Laver, J.
[4] Rose, P. (2002), Forensic speaker identification. Publisher: London, UK: Taylor and Francis.
[5] Hollien, H. (2002), Forensic voice identification. Publisher: San Diego, CA: Academic Press.
[6] L. G. Kersta(1962), "Voiceprint Identification", Nature, 196.
[7] O. Tosi, H. J. Oyer, W. Lashbrook, C. Pedney, J.Nichol, W. Nash, (1972), "Experiment on voice identification", Journal of the Acoustical Society of America, 51:2030-43.
[8] Morrison, G.S., (2014), "Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison", Science & Justice, 54, 245–256.
[9] Morrison, G.S. and Thompson W.C. (2017), "Assessing the admissibility of a new generation of forensic voice comparison testimony", Columbia Science and Technology Law Review, 18, 326–434.
[10] Poza, F. and Begault, D.R. (2005), "Voice identification and elimination using aural-spectrographic protocols", In: Proceedings of the Audio Engineering Society 26th International Conference: Audio Forensics in the Digital Age, paper number 1-1.
[11] Gold, E. and French J.P. (2011), "International practices in forensic speaker comparison", International Journal of Speech, Language and the Law, 18, 143–152.
[12] French, J.P. and Stevens L. (2013), "Forensic speech science", In: The Bloomsbury Companion to Phonetics. pp. 183–197, Publisher: London, UK: Bloomsbury, Editors: Jones M.J. and Knight R.A.
[13] Zhang, C., Morrison, G.S., Enzinger, E. and Ochoa F. (2013), "Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – female voices", Speech Communication, 55, 796–813.
[14] Wojciech Majewski (2007), "Comparison Of Subjective And Objective Speaker Recognition Under Voice Disguise Conditions", Archives Of Acoustics, 32, 4 (Supplement), 173–178.
[15] J. H. L. Hansen and T. Hasan (2015), "Speaker Recognition by Machines and Humans: A tutorial review," IEEE Signal Processing Magazine, 32, 6: 74-99.
[16] Morrison G.S., Sahito F.H., Jardine G, Djokic D, Clavet S, Berghs S, Goemans Dorny C.(2016), "INTERPOL survey of the use of speaker identification by law enforcement agencies", Forensic Sci Int., 263:92-100.
[17] Morrison G.S., Enzinger E. and Zhang C. (2018), "Forensic speech science", In: Expert Evidence, ch. 99, Publisher: Sydney, Australia, Editors: Thomson Reuters Freckelton I. and Selby H.
[18] Thomas Busey and Meredith Coon (2023), "Not all identification conclusions are equal: Quantifying the strength of fingerprint decisions", Forensic Science International, 343, 111543
[19] Geoffrey Stewart Morrison, Ewald Enzinger, Vincent Hughes, Michael Jessen, Didier Meuwly, Cedric Neumann, S. Planting, William C. Thompson, David van der Vloed, Rolf J.F. Ypma, Cuiling Zhang, A. Anonymous, B. Anonymous (2021), "Consensus on validation of forensic voice comparison", Science & Justice, 61, 3: 299-309.
[20] Rose P. (2022), "Likelihood Ratio-based Forensic Semi-automatic Speaker Identification with Alveolar Fricative Spectra in a Real-world Case", Proceedings of the 18th Australasian Int'l conf. on Speech Science and Technology, Canberra, Australia.
[21] Gonzalez-Rodriguez, Joaquin & Fierrez, Julian & Ortega-Garcia, Javier & Lucena-Molina, Jose. (2002), "Biometric Identification in Forensic Cases According to the Bayesian Approach", In: Biometric Authentication, 177-185, Publisher: Springer, Berlin, Heidelberg, Editors: Tistarelli, M., Bigun, J., Jain, A.K.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY