OSAC 2023-N-0023 Standard Guide to Forensic Speaker Recognition Landscape

Speaker Recognition Subcommittee
Digital/Multimedia Scientific Area Committee
Organization of Scientific Area Committees (OSAC) for Forensic Science

DRAFT OSAC Proposed Standard

OSAC 2023-N-0023
Standard Guide to the Forensic Speaker Recognition Landscape

Prepared by
Speaker Recognition Subcommittee
Version 1.0
September 2023

Disclaimer

This OSAC Proposed Standard was written by the [insert subcommittee or other unit name] of the Organization of Scientific Area Committees (OSAC) for Forensic Science following a process that includes an open comment period. This Proposed Standard will be submitted to a standards developing organization and is subject to change.

There may be references in an OSAC Proposed Standard to other publications under development by OSAC. The information in the Proposed Standard, and underlying concepts and methodologies, may be used by the forensic-science community before the completion of such companion publications.

Any identification of commercial equipment, instruments, or materials in the Proposed Standard is not a recommendation or endorsement by the U.S. Government and does not imply that the equipment, instruments, or materials are necessarily the best available for the purpose.

Acknowledgements

The Speaker Recognition Subcommittee thanks the following contributing authors who have provided substantial language to the document on multiple occasions: Sandra Ferrari Disner, Michael Jessen, and Finnian Kelly; as well as the following reviewers who provided valuable and feedback on content: Anil Alexander, John Hansen, Larry Kincaid, David Marks, Stephanie Domitrovich, and Jennifer Stathakis. The Subcommittee also thanks Erica Gold who has been the editor through the document's many revisions as well as lead author.

Table of Contents

1. Scope

Forensic speaker recognition, also referred to or covered by the terms forensic speech science, forensic phonetics, and speaker identification, aims to determine whether speakers are likely to be the same person or different people from at least two recordings (e.g., known and questioned recordings). This document provides a landscape of the methods used for analysis in the field of speaker recognition as well as the commonly used interpretation frameworks – Conclusion Frameworks. This document also establishes that the wider speaker recognition community has rejected previously held beliefs regarding the scientific validity of voiceprinting. This document is intended to serve as a general overview and reference (as it is currently practiced in the field) for forensic speaker recognition.

2. Referenced Documents

2.1 *OSAC Standards.*

2.2 Organization of Scientific Area Committees (OSAC) for Forensic Science Speaker Recognition Subcommittee, "Essential scientific literature for human-supervised automatic approaches to forensic speaker recognition."[1]

3. Terminology

For purposes of this document, the following definitions and acronyms apply.

3.1 Definition of terms specific to this standard.

3.2 **Automatic Speaker Recognition (ASR),** ***n.*** as used in this guideline, ASR requires the use of specialized software to compare speech samples, producing a numerical score that is evaluated from the perspective of the same-speaker origin as well as the different-speaker origin.

4. Summary of Practice

4.1 The main objective of forensic speaker recognition is to determine whether speakers are likely to be the same or different. Forensic practitioners are typically presented with a minimum of two audio recordings and asked to carry out an analysis of those audio recordings. The methods used for analysis in forensic speaker recognition have evolved far past previous traditions of voiceprinting, which has been rejected and discredited by the speaker recognition community. Methods of analysis that are currently in practice include: auditory phonetic analysis, acoustic phonetic analysis, semi-automatic acoustic analysis, automatic speaker recognition, human-assisted speaker recognition, and combined human and automatic speaker recognition analysis.

[1] Prepared by Scientific Literature Working Group, Forensic Speaker Recognition Subcommittee, "Essential scientific literature for human-supervised automatic approaches to forensic speaker recognition," Organization of Scientific Area Committees (OSAC), Online, Available: https://www.nist.gov/document/essentialscientificliteratureforhuman

Indeed, just as there are multiple methods for analysis being implemented in speaker recognition, there are also a number of different conclusion frameworks that have also been adopted. Interpretation frameworks (or Conclusion Frameworks) that are currently being utilized by the speaker recognition community include: the binary decision, probability scales, likelihood ratios (both verbal and numerical), the UK Position Statement, and support statements.

5. Significance of Use

5.1 Introduction: Forensic speaker recognition involves the comparison of at least two speech samples (typically from a questioned and known recording to determine whether speakers are likely to be the same or different. It is common for forensic speaker recognition to be referred to by a few other terms, largely dictated by the field in which the subject is researched and practiced. Within the forensic speech science and forensic phonetics communities, the task is often referred to as forensic speaker comparison or forensic voice comparison. Within the engineering communities, forensic speaker recognition is also sometimes referred to as forensic speaker identification. For clarification purposes, the task of comparing speech samples is referred to in this document as forensic speaker recognition. This document is intended to provide a general overview of forensic speaker recognition. For more detailed information about the topics discussed in this document, please see the OSAC Speaker Recognition Essential Literature document.[2]

5.2 The objective of the forensic practitioner carrying out forensic speaker recognition is to provide the trier(s) of fact with an informed opinion regarding the probability of obtaining the evidence (under the hypothesis that the samples came from the same person, versus under the hypothesis that two different speakers produced each sample). This objective can be reached by practitioners using a variety of methods (i.e., auditory phonetic and acoustic phonetic analysis, semi-automatic acoustic analysis, automatic speaker recognition, human-assisted speaker recognition, or combined human and automatic speaker recognition analysis). While questioned recordings often involve an array of different sounds and speech, the task of forensic speaker recognition is wholly concerned with the speech (and sounds) produced by individuals. Those sounds that cannot be attributed to a person are outside the scope of forensic speaker recognition, and fall more into the general area of audio or acoustic forensics. The aim of this document is not to promote or suggest any one method of analysis or interpretation framework over another, but rather to provide a general landscape of the methods used within the speaker recognition community.

5.3 *Voiceprinting:* While there are many ways to conduct forensic speaker recognition, it is important to note here that the method known as "voiceprinting" is not supported by the scientific community, and has been discredited. The term "voiceprint[2]" was coined by the author of an article (Kersta 1962) which appeared over a half-century ago. The name chosen for that methodology quite transparently implied parallels between a (never-proven) "theory of invariant speech" and the relative invariance of fingerprints.

[2] Gray and Kopp (1944) also used the term voiceprint with the same definition, however, they used the term with a space between the words voice and print. For all intents and purposes this document uses the term voiceprint without a space.

5.3.1 The so-called "voiceprints" were the product of sound spectrography, a technology carried forward even to the present day, which is still of great utility to speech scientists. The most notable scientific failing of the voiceprint method was that it did not provide examiners with the vocal output (i.e., the audio). This inevitably obscured the phonetic nature of the patterns of acoustic energy and reduced the analysis to a simple pattern-matching exercise. Nevertheless, that exercise was initially heralded with outsized claims of success. The article that introduced the voiceprinting method in 1962 reported that phonetically naïve examiners[3] were able to identify a target voice with 99% accuracy, even from a pool of a dozen speakers. Not surprisingly, this new methodology soon caught the attention of law enforcement, and was presented as evidence in a number of criminal prosecutions, in the US and elsewhere.

5.3.2 However, the scientific community remained skeptical. Well-known phoneticians such as Peter Ladefoged and Harry Hollien reported that mere pattern matching (which is all that the young voiceprint examiners were asked to do) was incapable of yielding the astonishing results reported in the 1962 article. Due to the variability present in speech productions from sample to sample, spectrographic template matching is not effective and it is inconsistent in speaker recognition work. In time, phoneticians began to provide expert testimony against the admissibility of voiceprint evidence, and in consequence, a number of lower-court convictions were eventually overturned. In response to these criticisms, another academic linguist, Oscar Tosi, initiated a more rigorous, and procedurally transparent, voiceprint study (Tosi et al. 1972). This yielded less vertiginous, but more scientifically reliable results – 6% false identification errors and 13% false elimination errors, under laboratory conditions. Still, given the high stakes of introducing a still largely unsupported procedure into courts of law, a report issued by the National Research Council[4] concluded that the voiceprint method lacked an adequate scientific basis for estimating reliability in many practical situations, pointing out in addition that laboratory evaluations of the voiceprint method showed increasing errors as the conditions for evaluation moved toward real-life situations, such as poor signal-to-noise ratios and dissimilar recording conditions.

5.3.3 The Federal Rules of Evidence, adopted in 1975, further challenged the voiceprint methodology by shifting the standards for admissibility in favor of practitioners whose "scientific, technical, and other specialized knowledge" can help the trier of fact "to understand the evidence or to determine a fact in issue"[5] Phonetically untrained voiceprint examiners, who sought to identify speakers simply by looking at pictures of their voice signals, were left at a marked disadvantage.

[3] To drive home his point, Kersta used high school students, who had been given only one week of training, as his examiners. The difficulty of the task was augmented by a forced-choice design; "not sure" was not an option.

[4] National Research Council. 1979. *On the Theory and Practice of Voice Identification*. Washington, DC: The National Academies Press. https://doi.org/10.17226/19814.

[5] *Federal Rules of Evidence* 702

5.3.4 For further overviews of the voiceprint approach one can consult (this is not intended to be an exhaustive list, but merely a few selected references):

5.3.4.1 Hollien (1990)

5.3.4.2 Hollien (2002)

5.3.4.3 Rose (2002)

5.4 While voiceprinting has been discredited from the speaker recognition community, it has also been declared inadmissible in court (for further information see U.S. v. Angleton 2003).

5.5 Additional methods that are properly applied and under certain circumstances are recognized as appropriate by the community are, in turn, described in detail below. It will become apparent that these methods have historically developed in parallel to one another, as they have grown out of different disciplines. However, it is not uncommon to see some cross-over between the various methods used in forensic speaker recognition. This will be explained further in the sections that follow.

6. Procedure

6.1 There is no one, single method that is used by all practitioners of forensic speaker recognition, and it is sometimes the case that some of these methods are combined when undertaking analysis. The methods most commonly employed in forensic speaker recognition are: auditory phonetic analysis, acoustic phonetic analysis, auditory phonetic + acoustic phonetic analysis, semi-automatic acoustic speaker recognition, automatic speaker recognition, human-assisted automatic speaker recognition, and a combination of auditory phonetic + acoustic phonetic analysis and (human-assisted) automatic speaker recognition. All seven approaches to speaker recognition are detailed below.

6.2 *Auditory Phonetic Analysis (AuPA):* AuPA is defined as the process by which “the expert listens analytically to the speech samples and attends to aspects of speech at the segmental and suprasegmental levels” (Gold and French 2011). AuPA is very important in the identification of language varieties, such as regional accent, foreign accent or in the detection of linguistic correlates of various social factors. Age, sex, and gender also fall into the category of characteristics most commonly judged auditorily. All of these can be classified as “group-level characteristics,” in contrast to “individual-level characteristics” (Hughes and Rhodes 2018). Group-level speaker characteristics are crucial in speaker profiling, but they also have their established place in speaker recognition. They can be important in defining the relevant population. They are also particularly powerful as evidence speaking against speaker identity: if, for example, the known and the unknown voices use two different regional varieties, it is likely that they are different individuals, given that bi-dialectalism is relatively rare. Group-level characteristics can also provide important information that supports inclusion, within the context of the case. They are also particularly helpful in excluding speakers. Part of the tradition of AuPA has been to narrow down dialect to a point that it achieves the status of a rare combination of linguistic parameters only used by a few individuals (and ideally, however, rather unrealistic, just one individual). This can occur when a speaker uses only some features of a dialect (or other language variety) to the exclusion of others, or if features from various language varieties are combined. Discussions of these aspects related to language variety are provided in Jessen (2010; 2021) and Hughes & Rhodes (2018).

6.2.1 AuPA is also used for the description and interpretation of various individual-level speaker characteristics. According to the survey by Gold and French (2011), voice quality is a particularly important one. Voice quality can be measured acoustically (Keating et al. 2015), but given the acoustic limitations typically encountered in forensic casework, most of these methods suffer from information loss or lack of applicability. Auditory analysis, instead, offers more robustness, though auditory analysis is also more subjective. Auditory voice quality assessment in forensics often builds upon the classificatory framework of the phonetician John Laver (1980), particularly in Europe. A description of that framework and how it is adapted to forensics is found in San Segundo et al. (2019), and a complete definition of voice quality can be found in McIntyre et al. (2021). Another classification framework for AuPA-based analysis has been developed for disfluency patterns, which include silent pauses, breathing pauses, filled pauses (utterance such as uh, and um) or sentence interruptions (McDougall and Duckworth 2018, de Boer and Heeren 2019, Hughes et al. 2016). Further speaker characteristics observed auditorily are listed in Gold and French (2011).

6.3 *Acoustic Phonetic Analysis (AcPA):* AcPA is the method by which "the expert analyzes and quantifies physical parameters of the speech signal using computer software. As with AuPA, this is labor intensive, involving a high degree of human input and judgment" (Gold and French 2011). AcPA traditionally has its strongest focus on speaker characteristics that have an anatomical motivation and that have been known since the 1950s to vary between women, men, and children, but also between individuals within these larger speaker categories. This applies to fundamental frequency and formant frequencies.

6.3.1 Fundamental frequency (f0), which is the frequency of the vibration patterns of the vocal folds, depends on the size of the larynx (especially vocal fold length), but it is to a degree controllable for linguistic purposes. As a way of disregarding locally determined linguistic factors, a common method is to average f0 (mean, mode, or median) across long utterances or the entire recording (Hudson et al. 2007). In this process, further irrelevant factors that have a strong influence on f0 must be controlled as much as possible; this is particularly important for the f0-raising effect of vocal effort (that is, speaking loudly) (Jessen et al. 2005). Speakers can also differ habitually in terms of how "melodically" they speak (scale from speaking monotonously to highly modulated). Standard deviation of f0 across long passages or the entire recording is a way of capturing these habitual speaker differences. Since mean and standard deviation of f0 are to some extent correlated, variability is sometimes expressed by the coefficient of variation (standard deviation divided by mean), by means of which the correlation almost disappears (Jessen et al. 2005).

6.3.2 Vowel formant frequencies, which are characteristic patterns of amplitude peaks in the speech spectrum, are associated with the length of the speaker's vocal tract and other anatomical features. However, formant frequencies – especially the first formant (F1) and the second formant (F2) – are also crucial carriers of linguistic information; they are the main correlates of vowel distinctions in a language, and transitions between successive vowels serve to distinguish any intervening consonants. There is thus a clear need to control for these linguistic factors. One way, which is analogous to the processing of f0, is to average all the formants across long stretches of speech. This method is referred to as long-term formant analysis (Nolan and Grigoras 2005). Another method is to measure formant frequencies separately for different vowels. This is the traditional way formants are measured in phonetics. Beyond anatomical restrictions, there are degrees of freedom in transitioning from one target sound to the next. Hence, measuring formant dynamics is a third way of capturing formant information in forensic speaker recognition. It has been shown in many studies that when formant measurements are not limited to targets but, rather, when one takes into account the entire dynamics of the formant movements, speaker recognition capability is improved (see McDougall 2006 and Morrison 2009 for some of the early studies).

6.3.3 There are other speaker characteristics that can be based upon acoustic phonetic analysis (Gold and French 2011). For example, it is possible to measure the spectral energy distribution in fricatives or nasals (Kavanagh 2012) or to make temporal measurements in the domain of rhythm and timing (Dellwo et al 2015; Plug et. al. 2021). But most actual AcPA casework utilizes f0 and formants.

6.4 *Auditory Phonetic + Acoustic Phonetic Analysis (AuPA+AcPA):* AuPA+AcPA is the combination of both auditory and acoustic analysis in speaker recognition as detailed in §6.2 and §6.3. The combination of AuPA and AcPA has also been referred to as an "auditory-acoustic-phonetic" method that is carried out by forensic practitioners (Morrison et al. 2016)." The term "auditory-acoustic-phonetic by forensic practitioners (qualitative opinion)" reflects how the phonetic data are traditionally interpreted by many practitioners of AuPA+AcPA: though there can be quantification on the feature level (e.g., formant frequencies in Hz; values on a scale of perceived voice qualities), the results are most commonly interpreted qualitatively, e.g., as distances of the values of the unknown and known speaker that is visible in a plot of formant values, or as an experience-based judgment of how frequently a certain speaker characteristic occurs in a relevant population (Morrison et al. 2016). Such a qualitative expression of AuPA+AcPA can approach a quantitative likelihood ratio-based method if there are data available of the relevant population, for example of the f0 values of male speakers (Gold and French 2019). But for full expression of quantitative likelihood ratios, the statistical methodology has to be present, as well as all the necessary data, such as non-contemporary same-speaker and different-speaker data.

Annex

Bibliography

This is not meant to be an all-inclusive list as the group recognizes other publications on this subject may exist. At the time this document was drafted, these were some of the publications available for reference. Additionally, any mention of a particular software tool or vendor as part of this bibliography is purely incidental, and any inclusion does not imply endorsement by the authors of this document.

American Nurses Association. (2021). Nursing: Scope and standards of practice (4th ed.). Nursesbooks.org.

Cambridge University Press. (n.d.). Meanings & definitions. Cambridge Dictionary. https://dictionary.cambridge.org/

National Commission on Forensic Science. (n.d.). Views document on definitions. https://www.justice.gov/archives/ncfs/page/file/477836/download

National Institute of Justice. (2020). National best practices for sexual assault kits: A multidisciplinary approach. Office of Justice Programs, U.S. Department of Justice.

Organization of Scientific Area Committees (OSAC) for Forensic Science, Crime Scene Investigation Subcommittee. (2021). Guiding Principles for Scene Investigation and Reconstruction (OSAC 2021-N-0015). OSAC, National Institute of Standards and Technology, US Department of Commerce. https://www.nist.gov/system/files/documents/2021/09/02/OSAC%202021-N0015%20Guiding%20Principles%20for%20CSI_FINAL%20OSAC%20PROPOSED%20FOR%20REGISTRY.pdf

Spellman, B. A., Eldridge, H., & Bieber, P. (2021). Challenges to reasoning in forensic science decisions. *Forensic Science International*. *Synergy*, *4*, 100200. https://doiorg.mutex.gmu.edu/10.1016/j.fsisyn.2021.100200