OSAC Technical Guidance Document 0006

Issues in Data Processing and Relevant Population Selection

https://doi.org/10.29325/OSAC.TG.0006

OSAC Speaker Recognition Subcommittee

OSAC
Organization of Scientific Area Committees for Forensic Science

OSAC Technical Guidance Document 0006

Issues in Data Processing and Relevant Population Selection

1st edition

Prepared for
The Organization of Scientific Area Committees (OSAC) for Forensic Science

By
The OSAC Speaker Recognition Subcommittee

September 23, 2022

Document Disclaimer:

This OSAC Technical Guidance Document was produced using a consensus process within the Organization of Scientific Area Committees (OSAC) for Forensic Science and is made available by the U.S. Government. All OSAC members had an opportunity to comment on the document and provide suggestions for revisions on this OSAC Technical Guidance Document. The views expressed in the document do not necessarily reflect the views or policies of the U.S. Government. The document is provided "as is" as a public service, and the U.S. Government is not responsible for its contents.

Any mention of commercial equipment, instruments, or materials in this Technical Guidance Document does not imply recommendation or endorsement by the U.S. Government; neither does it necessarily imply that the materials or equipment identified are the best available.

Copyright Status:

Abstract

In Forensic Automatic Speaker Recognition (FASR), forensic examiners typically compare audio recordings of a speaker whose identity is in question with recordings of known speakers to assist investigators and triers of fact in a legal proceeding. The performance of automated speaker recognition (SR) systems used for this purpose depends largely on the characteristics of the speech samples being compared. Examiners must understand the requirements of specific systems in use as well as the audio characteristics that impact system performance. Mismatch conditions between the known and questioned data samples are of particular importance, but the need for, and impact of, audio pre-processing must also be understood. The data selected for use in a relevant population can also be critical to the performance of the system. This document describes issues that arise in the processing of case data and in the selections of a relevant population for purposes of conducting an examination using a human supervised automatic speaker recognition approach in a forensic context. The document is intended to comply with the Organization of Scientific Area Committees (OSAC) for Forensic Science Technical Guidance Document.

Keywords

speaker recognition; forensic speaker recognition; automatic speaker recognition; forensic speaker identification; forensic speaker comparison; forensic voice comparison; corpora; data sets; data processing; relevant population selection; intrinsic speaker variability; extrinsic speaker variability; context speaker mismatch

Scope

This document describes issues that arise in the processing of data and in the selection of a relevant population data for purposes of conducting an examination using a human supervised automatic speaker recognition approach in a forensic context.

Approaches to forensic speaker recognition include spectrographic, auditory, acoustic phonetic, and automatic. While practitioners often employ combinations of these approaches, only automatic speaker recognition, and specifically, only human-supervised automatic speaker recognition, fall within the scope of this document.

Methods of speaker recognition that do not take into account the specific challenges of forensically realistic material are not covered by this document. Also, out of the scope of this document are practices such as familiar voice recognition, which exploits a person's natural ability to recognize the voice of a speaker known to them, and ear witnessing, when a person is asked to determine if the voice of a speaker whom they heard previously but who is otherwise unknown to them is present in a voice lineup.

This document covers issues in a general way and may be used by different audiences for different purposes. As the document offers recommendations on data processing and, when possible, on handling of mismatched conditions, practitioners should use it in conjunction with the OSAC-SR document titled "Forensic speaker recognition: Evaluation of evidence to inform

legal decision making“, which addresses additional considerations with respect to relevant population selection; they may wish to consult both documents as part of their training and before beginning an investigation that contains conditions unfamiliar to them. As the document cites literature wherever possible on the impact of mismatched conditions, early career researchers may consult the document before beginning new investigations. As the document mentions the frequency and importance of conditions in casework and notes when there is insufficient scientific evidence to assess the impact of mismatched conditions on system performance, funding body program officers may wish to consult the document when planning new research or resource development programs.

Table of Contents

1. Executive Summary

Before forensic examiners can begin comparing the speech samples that comprise case data, they must preprocess the case data into the form required by the investigation and select relevant population data. The quality and quantity of both case data and relevant population data can have a significant impact on the performance of speaker recognition systems and thus the reliability of their outputs. In comparing speech samples, the forensic examiner must exercise great care informed by an understanding not only of the requirements of the specific systems in use but also of the potential impact of speech pre-processing and the potential impact of mismatches in the conditions under which speech samples are acquired. A 2017 OSAC survey of forensic speaker recognition experts, including practitioners, researchers and legal experts, ranked twenty-seven factors according to their importance in forensic investigations, how well the factors were understood in the scientific literature and how much guidance was available for dealing with the factors. The scientific literature, sponsored predominantly by agencies outside the forensic community, nonetheless offers experimental results that partially address forensic concerns. This literature has shown for example that systems built and tested with relevant population data involving normal speech perform particularly poorly when comparing speech samples that are whispered, shouted or sung.

2. Acknowledgments

The Speaker Recognition Subcommittee thanks the following contributing authors who have provided substantial language to the document on multiple occasions: Ewald Enzinger, Kevin Farrell, John Hansen, Alysha Hiller, Michael Jessen, Colleen Kavanagh, Finnian Kelly, Larry Kincaid, Aaron Lawson, Ken Marr, Mitchel McLaren, Peter Milne, and David van der Vloed; as well as the following reviewers who provided valuable and extensive feedback on content: Dana Delger, Sandra Disner, Stephen Gibbs, David Marks, Johanna Morley, Geoffrey Stewart Morrison, Omid Sadjadi, William Thompson and Emily Whitmarsh. The Subcommittee also thanks Christopher Cieri who has been the editor through the document's many revisions as well as a contributing author.

3. Background

Speaker Recognition (SR) addresses the question, among others, of whether two speech samples, typically in the form of digital audio recordings, were uttered by the same speaker. The OSAC Speaker Recognition (OSAC-SR) subcommittee has adopted this term instead of other commonly used terms such as voice comparison and voice recognition. Elsewhere, the terms speaker detection or speaker verification have been used to refer to this specific comparison while speaker recognition has been used an umbrella for this and other comparisons. Other potential use cases, such as speaker clustering, voice search or comparison with multiple known samples, can be expressed in terms of this more basic question.

The terms audio recording and speech sample, sometimes abbreviated as audio, recording and sample are used interchangeably in this document.

In Forensic Automatic Speaker Recognition (FASR) an examiner typically compares an audio recording of a speaker whose identity is in question, henceforth the **questioned** speaker or sample, with a recording of a **known** speaker to assist investigators and triers of fact (judge or jury) in a legal proceeding. Together the questioned and known samples are called the **case data**. Forensic speaker recognition may also involve two questioned recordings. The questioned recordings may have been made at locations and under conditions that are uncontrolled and may or may not be fully documented or even known.

Human-supervised Automatic Speaker Recognition (HASR) compares speech samples using automated methods that employ signal processing and machine learning to create and compare statistical models of the features of the known and questioned speaker recordings.

HASR often requires the use of recordings in which the speaker identity is known and that are representative of the case data. This is called the relevant population data and it can be used as a tool for optimizing an ASR system, or it can be used to build a model with which the strength of evidence is calculated. More precisely, relevant population data are recordings of a sample of the population that is relevant to the characteristics of the case data. Relevant population data should resemble as much as possible the case data in terms of the speaker intrinsic, contextual and speaker extrinsic factors described in detail in §6.3 through §6.5. Finally, relevant population data selection should be informed by the conditions that influence the ASR system being used, since mismatch greatly impacts ASR system performance.

Prerequisites for applying any of the data selection, processing, analysis or comparison procedure described herein are that the procedure has been validated either in the scientific literature or via independent tests in the laboratory, and that the examiner possesses the expertise necessary to apply the procedure correctly and understands its impact upon the examination. It is also necessary that the examiner document the characteristics of the data used in any investigation, the processing applied, and the decisions made to assess whether the analysis can continue.

As the National Academy of Science report on Strengthening the Forensic Sciences noted: *"The findings of forensic science experts are vulnerable to cognitive and contextual bias."* and *"Unfortunately, at least to date, there is no good evidence to indicate that the forensic science community has made a sufficient effort to address the bias issue; thus, it is impossible for the committee to fully assess the magnitude of the problem."* [1] Although the OSAC Speaker Recognition subcommittee is developing a document concerning the "Reduction of the potential for cognitive bias in evaluation of evidence" specifically for forensic speaker recognition, that document is not yet available at the time of writing. The current document identifies subjective decisions in the processing of data and relevant population data selection that may be susceptible to unconscious bias. For a fuller discussion of bias and recommendations on reducing it, readers are directed to the National Commission on Forensic Science's *"Ensuring That Forensic Analysis Is Based Upon Task-Relevant Information"* [2].

4. Problem Statement

The performance of automated SR systems depends largely on the characteristics of the speech samples being compared. It is important for the forensic practitioner to understand the conditions that impact system performance in order to avoid the misuse of ASR technology that can lead to flawed or misleading analyses. Case data that is of insufficient quantity or quality can adversely affect a SR system. Additionally, for a forensic speaker recognition examination to produce interpretable and relevant results, the system must be trained, tested and validated using relevant population data with acoustic, demographic and contextual characteristics similar to those found in the case data.

The dependence of automatic speaker recognition system performance upon case data attracted broad attention during the US legal case *State of Florida v. George Zimmerman* [3]. In that case, a 911 emergency call captured a scream for help in the background. The State sought "to introduce expert opinion testimony that the screams heard in the 911 call belong to Martin." The Defense responded by seeking "to exclude the opinions of the State's experts on the basis that the techniques applied are not generally accepted in the scientific community." [4] When practitioners, hired by the State, attempted to use semi-automatic methods to compare the original scream and a simulated scream obtained from Zimmerman, controversy arose over the comparison of actual and simulated screaming as two independent experts, including one from the US Federal Bureau of Investigation (FBI), testified that the methods used were unreliable. A brief probe analysis of scream and speaker recognition technology confirmed the limitations of current technology [5] [6]. A subsequent and more extensive study analyzed acoustic properties of audio from screaming vs. speech and benchmarked current SR technologies to show their complete inability when applied to non-speech vocalizations such as screaming [7]. Many SR systems are trained on, and assume speakers produce, speech consisting of the core units of language (i.e., the phonemes of a language). Non-speech vocalizations, such as screams, without lexical content are void of such linguistic structure [8]. Non-speech vocalizations could prove as informative as speech vocalizations with some modern technologies that do not rely on the core units of language for constructing the models. However, irrespective of the technology used, the mismatch issue still remains relevant.

The case data raised a number of issues other than the mismatch between screaming and normal speech including the concatenation of copies of the signal to bypass the duration requirements of the system used, and mismatches between the questioned and known samples in terms of distance from the microphone and level of stress experienced by Zimmerman in the original and simulated samples that were compared.

Most forensic scenarios are not so complicated. When there is sufficient speech of acceptable quality available in the questioned and known samples, acquired under well matched conditions, a comparison between the samples is possible. One outcome of *Florida v. George Zimmerman*, however, was the redoubling of effort toward establishing best practices in forensic examinations employing automatic SR technology.

Our knowledge of the factors affecting automated SR performance is connected to the history of the field. Early efforts focused more on the telecommunications domain, where telephone hand-

set and communication channel variation were the primary concerns. Over time, as telephone technology has evolved, the nature of the challenge has changed, and its complexity has grown. With mobile cellular phone technology now dominating the world's telecommunications market, the diversity of phone and network types has expanded considerably along with their associated data transmission compression tradeoffs. In addition, virtually all cell phones have a speaker option which allows voice interaction at a distance greater than was previously possible introducing a wider range of channel variability. As research has progressed on mitigating the effects of microphone and channel mismatch, focus has begun to shift to other factors with the result that our knowledge of some factors is much better than of others.

One goal of this document is to gather what we currently know of the factors in audio sample processing and the selection of relevant population data in order to reduce mismatches in case data of the kinds shown to negatively impact performance of automatic SR systems. As there is active research on many of factors and mismatch conditions discussed below the state of the art will change over time.

5. Data Processing

The speech sample(s) submitted to a practitioner may need to be processed to be suitable for evaluation. Audio characteristics and processing that may be considered during sample preparation include the following. Such processing is most often applied to questioned-speaker samples that are acquired under uncontrolled circumstances. However, it may be necessary to apply such processes to known-speaker samples and even relevant population data samples if they too were not acquired under careful control. Data processing includes several points at which subjective decisions are required. Because subjective decisions are susceptible to cognitive bias, examiners should be aware of potential bias and guard against it.

5.1 Diarization

Diarization is the segmentation of an audio stream, along the time domain, according to who is speaking. In short, diarization answers the question "who spoke when". In forensic SR, diarization is used to isolate the areas of speech belonging to the speaker of interest. The speaker of interest should be clearly identified using information provided by the individual or entity requesting the examination. Only utterances from the speaker of interest should be included in the sample and should be selected by using multiple types of information, for example the content of the speech and turn-taking behavior and, in the case of wiretaps, the audio channel on which speech was recorded. In the case of recorded telephone calls, the two sides of the conversation are often recorded on different channels and so can be separately easily. Failure to separate the speech of the speaker of interest from all other speakers could degrade the performance SR systems. In cases where the speech is in a language unfamiliar to the examiner, the examiner may need to consult someone fluent in that language to identify which utterances belong to the speaker of interest or else recuse him or herself from the case.

5.2 Duration, Long

In situations where the quantity of speech available is greater than required, the examiner may select only some portions of speech to be used for comparison. In such cases, examiners may select speech to minimize mismatches between questioned, known, and relevant population data samples as described in §6. For example, if the speaker of interest briefly whispers but otherwise speaks in a normal mode, and the content of the whispered speech is not itself critical to the examination, the examiner may choose to remove the whispered utterances because system performance is degraded in such cases. Efforts should be made to avoid cutting speech in the midst of an utterance when feasible. Many automated SR systems employ voice activity detection; thus, natural pauses should be included as a part of the prepared sample. Examiners should guard against potential cognitive bias in subjective decisions including the decision of which utterances to include in an examination.

5.3 Duration, Short

Very short durations can be a limiting factor for an automated system. Duration of the samples compared is one of several factors affecting system performance. Some commercial systems will not process samples below a certain duration. There are no mitigation measures for this situation. Speech may not be replicated or concatenated to itself to increase the apparent number or duration of audio samples.

5.4 Noise, Reverberation and Non-Speech

Background noise, room acoustics, signal and speech distortions, and similar problems can interfere with the analysis, as can non-speech sounds such as laughter, coughs, throat clearing, lip sounds, whistles, screams and breathing. The examiner should determine the extent to which these may affect the analysis. At the extremes, brief, occasional distortions could simply be removed from the sample or suppressed prior to analysis while pervasive noise that impedes intelligibility or the examiner's ability to separate speakers could block further analysis. While stationary background noises can be suppressed to a great extent, reverberation resulting from room size, shape and materials as well as location of the microphone is an example of a nonlinear distortion which is not as easy to suppress. The capabilities and limitations of the SR system being used should also be taken into account when considering noise and non-speech. Although non-speech vocalizations could indeed be characteristic of a speaker, the absence of models for these in the relevant population data may prevent the examiner from determining if they are in fact characteristic.

5.5 Impoverished Lexicon

Speech samples in which the lexicon is impoverished, for example where the speaker answers questions in exclusively single words, provide very little of the linguistic and acoustic data needed to analyze a speaker's speech. Examiners should seek, where possible, speech samples that have greater lexical variety. Where this is not possible, examiners may find a degradation in system performance.

5.6 Format

Audio format conversion, for our purposes including changes to encoding scheme, sample rate, sample size and number of channels (e.g., stereo, monaural), may be necessary to produce an audio sample compatible with the SR system in use. Such conversion may also have been carried out as part of an earlier diarization process. However, audio quality should be maintained to the degree possible given the constraints of the system. That is, the format conversion must not reduce the quality of the samples or introduce artifacts unless required to meet the operating conditions of the system used.

5.7 Other Data Processing

SR systems may perform preprocessing such as voice activity detection, tone removal or speech enhancement, which may improve system performance. Some systems offer the option to activate, deactivate or configure parameters associated with such automated preprocessing. Speech preprocessing, especially using non-automated methods, may increase intelligibility for humans but reduce SR system performance. Research is ongoing concerning the impact of pre-processing upon system performance so except as required to meet SR system requirements, pre-processing should only be applied where the scientific literature has demonstrated improvement.

Regardless of the amount or type of pre-processing performed on speech samples, all selection, removal and/or enhancement procedures should be documented in a manner consistent with appropriate evidence handling procedures. In addition, the integrity of the original digital data should be protected so that alterations are transparent and correctible.

6. Data Properties and Relevant Population Data Selection

Speech is a highly variable phenomenon. Even the same speaker uttering the same phrase twice in rapid succession will produce small variations in speech. That variation grows with changes in the speaker's interlocutors, physical and emotional state, and age. Given that variability, the task of SR is deciding what variability is due to within-speaker versus across-speaker differences. The multifaceted sources of variation pose some of the greatest challenges to accurately modeling and recognizing a speaker regardless of the approach used.

The performance of SR systems depends strongly on the properties of the recordings to be compared. It is imperative that the system in use be tested and validated using relevant population data with similar properties to those of the case. Thus, it is necessary to understand and evaluate the properties of the case recordings to determine whether an examination can or should be conducted using the samples and to select the relevant population data according to the requirements of the system in use. If certain properties of the samples have not been adequately tested, or if they will degrade system performance to an unacceptable level, the examination should be terminated.

Properties of the speech sample that can influence SR system performance may depend upon the speaker (intrinsic), the situation in which the speech is uttered (contextual) or the method by

which and physical environment in which the speech samples were obtained (extrinsic). Each of these are detailed below.

6.1 Mismatch of Conditions

As mentioned above, SR systems are sensitive to the intrinsic, contextual, and extrinsic conditions of audio samples. Some factors such as noise and reverberation may negatively impact system performance by their mere presence. Despite advancements in SR technology, background or channel noise have near infinite variability of type, level, and combination, such that there will always be instances when they negatively impact an audio sample to render it unintelligible to humans and/or unsuitable for automated analysis.

Mismatches of other conditions can degrade the system's ability to discriminate speakers effectively. Mismatched conditions occur when the audio samples of an investigation differ sufficiently in ways that potentially degrade SR system performance. For example, the questioned speaker recording might have occurred in a public setting characterized by street noise while the known speaker recording was made in a quiet, controlled environment. Mismatch of conditions can occur between the questioned and known samples of the case data, between the case data and relevant population data, within the relevant population data, between the case data and the data used during SR system model training and within the data used to train system models. Furthermore, the conditions under which a recording, generally the questioned recording, was made may not be known, thus complicating the search for comparable data. In any of these cases, mismatched conditions can have varying levels of impact on the discriminative power and calibration of a forensic SR system and thus the reliability of the outcome.

Unlike other forms of biometrics, such as fingerprint, iris, face, and hand geometry [9], human speech is a performance biometric, combining both physiological phenomena and behavioral traits. Simply put, the identity information of the speaker is embedded (primarily) in how speech is spoken, though the speaker's lexical choice can also be incorporated into the analysis and the degree of overlap in the content of speech between questioned and known recordings certainly impact on SR effectiveness [10] [11]. The behavioral component makes speech signals prone to greater variability such that even the same person would not say the same words in the same way every time (this is known in different bodies of literature as "style-shifting" or "intra-speaker variability") [12] . Differences in recording devices and transmission methods only exacerbate a problem already inherent in SR [13] [14].

Given the behavioral component of speech, some characteristics of an audio sample are prone to variability over the duration of the recording. As an example of this within-session variability [12], a person may speak in a neutral tone at the beginning of a recording but with anger at another moment. In such cases, it may not be possible to locate relevant population data with the same transition in conditions, and therefore, a suggested protocol would be to analyze the neutral and angered parts independently or select only the neutral part for comparison.

The extent to which mismatched conditions impact SR must be determined empirically through method validation, either with or without mismatch compensation methods. The degree of impact may be dependent on the SR system employed but is very often a function of the number of mismatched factors and severity of the mismatches. Forensic applications typically encounter mismatch in one or more factors, and while it is generally not possible to remove mismatch from audio with processing (speech enhancement to remove additive noise being a possible exception), other mitigation approaches such as selection of a subset of the questioned speaker data and careful selection of the relevant population data may be available to reduce the impact of mismatch. The examiner is again reminded that selecting a subset of case data or relevant population data is susceptible to unconscious bias.

The following section details characteristics that may be relevant to forensic SR and whether the degree of impact is known and, if so, with what confidence.

6.2 Impact of Variability on Speaker Recognition

Any assessment of the impact of condition variability on forensic speaker recognition should include both the degree of impact and the frequency with which it is encountered in casework. A 2017 OSAC survey of forensic speaker recognition experts, including practitioners, researchers, and legal experts, identified twenty-seven factors that impact forensic casework and asked respondents to comment, via 10-point Likert scales, on the importance of those factors, how well understood they were in the research literature and whether guidance on dealing with the conditions existed and had been tested. The survey asked respondents to consider frequency in casework in their assessment of importance. At least 49 experts were invited to complete the survey and permitted to answer whatever portion they believed they could. 16 experts responded. Table 1 shows the ten conditions that were judged most important. The Survey reports as high and low scores those that are >0.5 standard deviations above or below the mean as well as the five factors with most and least agreement among scores again standard deviations.

emotional/cognitive stress	physical effort/stress
vocal speaking effort	voice variability due to distance from microphone
emotion	reverberation
acoustic background noise	cross-language comparisons
speaking style	language/dialect variation

Table 1: Importance of Conditions affecting Forensic Speaker Recognition per OSAC 2017 Survey

The scientific literature on SR system performance contributes to our knowledge of the impact of any condition or mismatch. Impact is the effect that the presence of a condition, or a mismatch in that condition, has on the accuracy and reliability of a given SR system. Categorizing the impact of a particular condition, even as simply as low, moderate, or high, is only as reliable as the data that supports it. For a variety of reasons, the conditions considered important or frequent by forensic practitioners are not always those most well understood in the scientific literature. Our scientific understanding of many conditions rated as important by practitioners may be weak or non-existent. Indeed, the factor deemed most

important in the OSAC survey of practitioners, emotional stress, is one for which there is little empirical evidence of impact which may be due to the relative dearth of research on emotional stress in SR.

To assess the empirical support for any estimate of impact, the following are considered:

- number of speakers and number of sessions per speaker
- number of different corpora used, diversity of corpus conditions (languages, environment, channels, durations, etc.)
- naturalness of speech recorded (versus prompted or acted speech)
- diversity and independence of systems evaluated
- confidence of results

For each condition in the following sections, the goal is to provide information about the relative importance of the phenomenon, and to assess the degree of impact on SR, where feasible. It is important to note that large gaps remain in our understanding of 1) how conditions impact system performance, 2) how mismatches in conditions affect results and especially 3) how combinations of mismatches interact (i.e., most studies consider single condition mismatches, and results may not carry over when multiple mismatch issues are present simultaneously). For those conditions for which evidence concerning impact is weak, the following sections refrain from providing guidance based on intuition, and instead provide a possible way forward by defining the steps to achieving sufficient understanding, for example, by pointing to data sources from which one can obtain benchmark findings.

Below, impact is estimated as low, moderate, or high where low represents a limited and probably acceptable negative impact on speaker recognition technology, moderate means caution is needed, high means the variability in the factor will severely degrade the performance of most SR systems. To be precise, in the sections below, impact should always be interpreted as negative impact. As a visual aid, that scale is associated with traffic lights as follows:

low 

moderate 

high 

low to moderate 

moderate to high 

Some of the factors that affect speech are intrinsic to the speaker such as age, gender, language, and health. Vocabulary and other aspects of the way speech is produced, for example the average number of words uttered per minute, can be associated with the speaker and are sometimes labeled “speaking style”. Other aspects of “style” may be sensitive to the context in which speech is produced (§6.4). Still other factors, called extrinsic, relation to the

location in which the recording was made, the equipment used and other factors external to the speaker.

6.3 Speaker-Intrinsic Variability

This section describes a range of speaker intrinsic variations in how speech is produced which impact SR system performance to different degrees. Many factors are not present in unison, but rather combined with other, related factors.
Vocal Effort/Style refers to alterations in speech production from normal phonation, such as whispered speech [15] [16], soft, loud or shouted speech, singing [17], or Lombard speech which occurs when speakers alter their production to speak effectively in the presence of noise [18] or other effects such as room reverberation. Speech can be placed on a scale of vocal effort: whispered - soft - neutral - loud - shouted. Although vocal effort is among the most important factors according to the 2017 OSAC survey, little guidance is available. Nearly all speech for SR model training has been collected with normal (neutral) vocal effort.

Soft - Neutral - Loud Speech

There is some change in SR performance for soft and loud utterances. Generally, the impact may be low to moderate depending on the extent of the mismatch including the actual duration of speech produced with vocal effort different from neutral. In 1999-2000, the NATO RSG.10 project considered the impact of soft and loud speech on SR performance in a limited probe study [19], and later with both neutral and matched trained models [20]. Loud speech had moderately degraded performance, but the actual study data was limited in size.

IMPACT: generally low impact if soft or loud speech is not sustained across the entire speech sample, perhaps moderate for the greatest deviations from neutral.

Whispered Speech

Whispered speech has been shown to have a strong negative impact on speaker recognition performance for most speakers [15] [16] when models are trained on neutral speech.

IMPACT: high

Shouted Speech

Here we distinguish shouted speech, where the speaker is attempting to utter words during vocalization, from screaming where there is no attempt to produce words. Despite the dearth of studies on shouted speech, it is expected to have a strong negative impact on performance for an SR system trained on neutral speech by analogy to what has been found for screaming [5] [7] and loud speech [20].

IMPACT: high

Lombard Effect

Speech produced in the presence of noise and other signal disturbances such as room reverberation results in the Lombard Effect in which the speaker alters his or her speech production in a subconscious response to improve intelligibility [18] [21]. Lombard speech has a strong negative impact on speech system performance [18] [22] [23] [24], (EER increases from 7.0 to 26.9% in some studies). While the Lombard Effect has been recognized since 1911 [25], only recently has it been shown that there are a range of "flavors" of Lombard effect speech based on different types and levels of noise exposure [18], resulting in changes in acoustic–phonetic characteristics based on excitation structure, duration, energy histogram, and spectral tilt. It is possible to mitigate this impact by adapting neutral speaker models with Lombard speech data of limited duration [18]. A range of compensation and normalization strategies have also been explored to help reduce Lombard Effect production differences [23] [22] [24] [26]. As with other variations in vocal effort, very little data exhibiting the Lombard Effect is available for training SR models.

IMPACT: high, even more so in severe noise levels/noise types where Lombard Effect speech will shift further from neutral speech conditions. 

Singing vs. Speaking

The relatively small amount of research on of singing versus speaking the same text suggests there is a high impact. Singing changes the time-frequency structure of a speaker's voice. A recent series of studies explored singing versus speaking the same audio content in four languages [27] [28] [29] [30]. SR performance suffered significantly when speaker models were trained on spoken/read speech and tested with the same speech in singing. It is possible to reduce this impact by using alternate clustering methods (i.e., PLDA) [31]. Another study explored how acoustic (e.g., MFCCs) and facial (e.g., lip and eye content/structure) features could be employed to assess singing quality [32]. Forensic use cases might include those in which obstructions prevent facial recognition but provide enough features to augment speaker recognition. It is suggested that employing an objective means of assessing singing quality could represent a secondary meta-data source that could be leveraged with more traditional speaker recognition solutions to improve speaker recognition with singing audio content. Speaker recognition solutions today, trained on spoken text content, however, degrade in performance is tested with the same singing content, and presumably even more when singing text content is open [33] [34]. The degree of impact of singing is mitigated by its infrequency in forensic scenarios.

IMPACT: Impact is high on the rare occasions when it occurs. 

Emotion

Emotional state (e.g., anger, happiness, sadness, joy, despair, etc.) can have a significant, sometimes audible, impact on speech output. However, emotion is not a one-dimensional factor; different emotions have different effects on speech and different speakers have different ways of conveying their emotional state when speaking. Notwithstanding its complexity, research on emotion is in great demand as casework often contains emotionally

charged speech and emotion is prone to within-session variation (i.e., emotional state changes within a single recording).

Listeners are generally better able to perceive some emotions (i.e., happiness, sadness, anger, fear) than others which can be viewed as more abstract and difficult to define (i.e., disdain, disgust, melancholy, anxiety, etc.). Emotion (anger) and stress were documented as having an impact on SR systems [19]. However, few studies have effectively addressed this for SR. These and its audible effect on speech suggests that emotion could have a strong negative impact on SR performance. Where possible, a mitigation strategy may be, prior to comparison, to separate speech samples from the case data according to the emotion exhibited, matching the subsets to relevant population data that exhibit similar emotion. The reader is again cautioned that distinguishing emotional from neutral speech is a subjective decision susceptible to unconscious bias as mentioned in §3. At the time of writing, most publicly available corpora contain predominantly neutral emotional state. The dearth of data containing emotional speech appropriate for speaker recognition system development is a critical issue we draw to the attention of funding body program officers.

IMPACT: Respondents to the 2017 OSAC survey agreed that emotion was among the most important factors in their casework though there is little understanding of emotion in the scientific literature or guidance on how to cope with emotion in casework. Unassessed. Further research is needed.

Language and Dialect

There may be a mismatch between the language or dialect of the questioned and known recordings or between the language or dialect of the system training data and that of the questioned or known recording or both. In addition, the language or dialect may change within a session (known as **code-switching**). One study [35] that employs language mismatching relative to small male relevant population data sets for Turkish and Dutch and another [11] showed that the phonetic content of samples being compared may impact an SR system's ability to discriminate speakers. Two recent studies have used larger data sets. One used data from a NIST Speaker Recognition Evaluation (SRE) campaign with 265 speakers to develop compensation methods to address and reported an EERi increase from 2.41% for English matched data to 5.51% for language mismatched data falling to 3.99% with compensation [36]. A second study also considered language mismatch using NIST SRE and CRSS Bi-LING corpora again confirmed moderate to high negative impact for language mismatch, which was suggested to be more pronounced as language pairs were further apart based on a language family tree analysis [37]. Although limited, additional studies exist assessing SR for speakers with two distinct languages. Much less is known or quantified concerning code-switching when one language is spoken primarily with brief switches into another language. However, a mitigating strategy in such cases may be to remove speech containing brief or infrequent switches into another language or dialect of the same language.

IMPACT: There was disagreement among 2017 OSAC Survey respondents concerning the importance of this factor though its score was above the mean. The evidence from the scientific literature is not yet adequate to offer guidance. Further research is needed.

Gender

The gender of the speaker will typically not affect SR unless the gender of the questioned recording is not obvious, or the investigation does not account for gender. There may be a mismatch when the system is trained on speech from both genders in an effort to be gender-independent or when the case data has one gender and relevant population data has another.

IMPACT: A level of low to moderate based on an SR validation report [38], which focused on Dutch telephone data; further studies are required to increase confidence in estimates of the factor's impact.

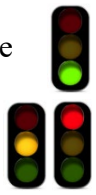
Time Difference/Aging

Time Difference and Aging mismatches both refer to the time difference between when the questioned recording and the known recording were made. If the time difference is short (e.g., hours or days), this is referred to as "session mismatch" but can also be called "time difference". "Aging mismatch" typically involves speech physiology changes which are on a much longer time scale (i.e., years vs. days). The time difference present in the case data should be represented in the relevant population data. As the time difference increases, so too does the potential for change in a speaker's voice due to the effects of vocal aging.

Vocal aging can be defined as a combination of physiological, cognitive, and environmental factors that result in vocal change. While progressive vocal change occurs throughout adulthood, the effects of vocal aging are most discernible during childhood, adolescence, and old age, where there can be rapid physiological and cognitive changes. The acoustic effects of vocal aging have been well-investigated [39] [40].

Research into the effect of vocal aging on speaker recognition has indicated the scores obtained for same-speaker comparisons decrease as the time difference between the samples under comparison increases, and that this decrease is proportional to the size of the time difference [14] [41]. Consequently, at a time difference of 1-3 years, aging has a measurable, negative impact on speaker recognition performance [14] and at larger time differences, from 4 years up to several decades, the impact of aging becomes progressively more severe [42] [14] [43] [44]. Research has also shown that the age difference between different-speaker comparisons can affect speaker recognition performance [45]. Another study, using the MARP corpus, found an approximately equal degradation in SR performance for any time differences greater than 1 month and less than 15-24 months suggesting that these differences may reveal the effect of session mismatch rather than aging [13].

IMPACT: A time difference of weeks to months is typical in casework; however, time differences of several years are not uncommon. Low impact can be expected for comparisons with a 1–3-year time difference, progressively increasing to moderate to high for greater time differences.



This is based on a small number of studies on limited datasets: MARP [14], [46], [13], TCDSA [42] [14] [43], and FAME! [44]. Further research is required to increase confidence

in this assessment of the impact of vocal aging, and to continue the development of approaches [42], [14] that compensate for its detrimental impact.

Physiological

When the subject has a respiratory illness or other condition that alters breathing, or is intoxicated, this changes the speech of a subject relative to their normal speech. Description of physiological factors sometime include the impact of aging, including the aging of vocal folds and vocal articulators as discussed above. They can also include altered physical traits associated with vocal muscle control due to medicine/drugs or breathing condition. One recent study considered a longitudinal study of astronauts' speech during NASA Apollo-11 mission and showed significant changes in speaker models across various phases of the mission [47]. In general, however, the impact of physiological factors on SR is not well understood due to the lack of data though the factor is not uncommon in casework.

IMPACT: 2017 OSAC Survey respondents disagreed as to the importance of this factor though they gave a low score to our understanding of it from the research literature. Unassessed due to lack of data. Further research is needed.

Situational Task Stress

Situation Task Stress refers to when a person is performing some task while speaking which causes a cognitive load, physical task stress, emotional stress, or a combination of these. Examples of increased cognitive load include operating a vehicle or playing a video game. Physical task stress may arise when jogging, exercising or doing something that exerts physical load on the body. Emotional stress may be experienced by emergency responders, fire fighters, jet fighter pilots, individuals engaged in a crime, etc. [26] This factor is moderately prone to within-recording variation. All types of stress may appear at times in casework. Numerous studies have shown the impact of physical, emotional, and task-based stress on SR performance. Studies have shown that physical task stress directly alters speech production, affecting speech quality [48] and phonemic structure that in turn affects speaker recognition features [49]. Physical task stress also has a measurable negative impact on speaker recognition performance [50]. Stress in general, including situational and emotional, also impacts speech technology including SR [51] [20].

IMPACT: Responses to the 2017 OSAC Survey strongly agreed on the high importance of what the survey called 'emotional/cognitive stress' and gave consistently low scores to both our understanding of the factor in the research and the availability of guidance. Regarding the survey condition "physical effort/stress", responses gave slightly lower but still high scores on importance and below average score on research understanding and the availability of guidance. Within the scientific literature the degree of impact has varied from low to high, depending on the situational stress context.

Disguise

Speakers may intentionally alter their voice by natural means (mimicking another's voice) or by using a voice conversion system in an attempt to avoid recognition. Disguise is prone to within-session variability if the attempt employs natural means. There have been a number of studies in the area of Anti-Spoofing, where a computer alters/disguises the identity of a speaker or simply playback [52], [53], [54]. In spite of these and other studies, further research is needed in this domain, and there is some uncertainty as to the frequency of computer altered disguise in actual forensic cases.

IMPACT: There was disagreement about the importance of this factor in the 2017 OSAC survey though respondents gave low scores to our research understanding and agreed on very low scores for the availability of guidance. This factor has not been adequately assessed in the scientific literature. Further research is needed.

6.4 Speaker-Context Variability

Forensic SR applications often compare questioned speech samples, where the situations in which the speech was uttered are uncontrolled or even unknown, with known speech samples recorded in a controlled situation, e.g., reading the transcript of the questioned sample. This mismatch of the *context* within which the speakers are producing speech presents a potential challenge for SR systems.

The context in which an interaction takes place can alter the way a speaker produces speech. Context has been treated differently in taxonomies of the factors that affect SR systems because while the context is external to the speaker, the speaker's reaction to the context is speaker specific (i.e., contextual factors can be viewed as either intrinsic or extrinsic factors, neither or a combination of both). Context has a number of dimensions including the number of interlocutors and the degree to which the speech content is prepared or spontaneous and natural. For example, speakers may read text aloud or respond to questions or other prompts or produce a spontaneous monolog interacting with a human or a computer, smart device or piece of speech technology. Speakers may interact with one or more others via a telephone or other communication devices with or without a video component. Speakers might address a small group or give a public speech to an audience or engage in a formal group meeting or an unplanned conversation. The interlocutors may or may not speak the same linguistic variety (language, dialect) as the speaker. Speakers might attempt to disguise their voices or speak in a different dialect. Speakers may be familiar or unfamiliar with the interlocutors. The interlocutors may be socially subordinate or superior to the speaker or may be authority figures. Speakers may have a positive, neutral or antagonistic relation with their interlocutors. All represent contextual factors that can affect speaking style [55] [56]. The impact of context mismatches on SR system performance is not yet well understood.

For example, one study [57] analyzed the number of syllables in the lexicon of all unique words ('types') uttered in the Switchboard corpus [58] of two-person telephone conversations and then counted the frequency of those words as uttered in the same corpus ('tokens'). The study found that although only 22% of the word types in the lexicon were monosyllabic, they comprised 81% of the word tokens uttered. Conversely, words of 4, 5, or 6 syllables were rarely uttered in the corpus (1.16% of all tokens) though they comprise

13.52% of all word types. Timing and syllable structure analysis were used to surmise that the 911 Olympic Park Bombing caller was most likely reading from a prepared text vs. speaking spontaneously [59].

IMPACT: A single study based on a small number of speakers with very clean data showed no significant impact (i.e., low) of mismatch between read/interview and conversational speech [60]. However, it is well known from sociolinguistic research that word choice and pronunciation change based on context. At the moment, however, the actual impact on SR system performance has not been addressed sufficiently. This mismatch needs further investigation using different data sets and greater variability in speech style. The case where speech is human-to-human conversational vs. human-to-machine or read/prompted, suggest some low to moderate impact, but again more extensive research is needed. Changes in speech style are not very prone to within-session variability provided the context does not also change.

6.5 Speaker-Extrinsic Variability

Extrinsic properties are external to the speaker and generally related to the location and recording conditions in which the audio sample is collected including the type and placement of the microphone, how it was connected to the recording device, whether any audio compression was used, and the presence of noise and reverberation at the recording location. This section covers a range of differences in how audio is captured and whether they impact SR system performance.

Electromechanical

Electromechanical factors include transmission channel, handset (cell, cordless, landline) [61] [62] [63] and microphone. Specific examples of transmission channel mismatches may include different communication paths, such as the public service telephone network (PSTN) versus a digital connection used by a mobile phone or a voice over internet protocol (VoIP) communication used during a Skype call [64]. Microphone mismatches may occur when acquiring the data from different devices such as a landline phone, mobile phone, speaker phone, headset microphone, PC microphone, etc., as these devices will each have their own impact on the signal. While much progress has been made in limiting the impact of device and channel mismatches, the problem has not yet been solved. In general, device and channel mismatches will have a moderate impact, though it can be severe in extreme cases.

IMPACT: Much of the early corpus development supporting speaker recognition technologies focused on microphone and channel differences. The 2017 OSAC Survey gave high scores to our understanding of several factors that we consider together in this section. The survey called these “device-induced voice variability”, “the impact of microphone types”, “the impact of recording dynamic range”, “the impact of audio format”. The Survey also indicated there was guidance available on “the impact of microphone types”, “analog transmission and reception effects”, “the impact of recording dynamic range”, and “digital communications”. Regarding importance, “analog transmission/reception effects” received low scores. Impact is generally moderate.

Environmental Noise

There are numerous potential sources of environmental noise [65] that can include random (white) noise, other people talking (babble), music, and other types of noise with specific frequency content (colored noise) that may come from machinery or other mechanical sources. The recording environment such as whether the recording was made indoors, outdoors or in a moving vehicle can have an impact as can the size, shape [66] and amount of reverberation [67] of a room even in the indoor condition. Some of these effects can be mitigated by close talking or directional microphones but this is not always the case. As with channel and device mismatches, noise robustness is a topic that has been the focus of much research in the speaker recognition community. While the problem has not yet been solved, noise is generally considered to have a moderate impact on accuracy depending on the signal-to-noise ratio measured within the audio samples.

IMPACT: OSAC 2017 Survey respondents agreed that "acoustic background noise" was of above average importance in case work but also recognized a strong understanding of the phenomena in the research literature. They gave very low importance scores to "non-stationary (intermittent) noise". While there was disagreement among scores on our understanding of the condition in the research literature, scores were below average. Audio samples with high signal-to-noise ratios will exhibit low impact due to noise [68] [69], while very low SNR audio samples can experience high negative impact [68] [69] [70].

Data quality

Withing SR, data quality typically refers to the duration [71], sampling rate and audio compression employed during recording. Duration was introduced in §5.2 and §5.3.

Standard audio sampling rates used to sufficiently capture speaker information in the speech content tend to be at least 8 kHz (for telephone) or higher. This should be considered as the minimum sampling rate used for performing speaker recognition comparisons as lower sampling rates may exclude frequency bands with higher formant frequencies that have been found relevant for speaker recognition [72]. Up-sampling audio from a lower sample rate to 8 kHz is not sufficient for meeting the minimum requirement as it does not replace the information in the frequency bands that were absent in the lower-sampled audio. In cases where a digital audio sample is being created from an analog tape or a live source, the sampling rate should be at least 8 kHz but should ideally have the highest bandwidth possible, or at least the highest accepted by the SR system, to provide it as much information as possible.

IMPACT: OSAC 2017 Survey respondents gave high scores for the availability and maturity of guidance on duration. Impact is low with sampling rates 8kHz and above

Audio compression can also impact SR performance and is related to transmission channel (electromechanical) differences [64] [73]. For example, given one audio sample acquired over a telephone network that is sampled at 8 kHz with G.711 coding (i.e., u-law with 64 kilobits per second) and another audio sample acquired over a VoIP channel with G.729

coding (i.e., CS=ACELP with 8 KB per second), the difference in quality will degrade performance [74] [64].

IMPACT: ranging from low to moderate to high [64].

The duration of speech content is directly correlated with the amount of information provided by the speaker [71]. In speech samples of short duration (i.e., < 20 seconds), the coverage of individual sounds and combinations is limited while longer speech samples (>60 seconds) tend to provide reasonable phonetic coverage from the speaker. Greater overlap in phonetic coverage enable better comparison of the questioned and known speech samples. Based on the OSAC survey, variability in duration has a major negative impact on SR reliability. The impact of duration variability can be mitigated by reducing the duration of system training data and relevant population data to match the duration of the case data if the duration of the case data is adequate to support the investigation. However, if the decision as to which samples are removed is made directly by a human, rather than by some random algorithmic process, then it is susceptible to unconscious cognitive bias and great care should be taken. There is no mitigation for having samples of too short duration; samples cannot be improved by concatenating them to themselves as doing so does not add to the phonetic coverage.

IMPACT: moderate to high (Duration) [71].

6.6 Comparable Test Data, Relevant Populations

In current SR technology, various mathematical tools are used to mitigate the effects of the variability described above, particularly extrinsic sources of variability such as noise and transmission channel. Intrinsic variability is more difficult to quantify and thus address in automatic assessment.

For a forensic speaker recognition examination to produce interpretable and relevant results, the system must be trained, tested and validated using relevant population data with similar characteristics to those found in the case data. If the set of available relevant population data is not sufficiently similar to the speech samples from the case, the results of the forensic comparison will be misleading and unreliable. Therefore, examinations should be terminated when no representative relevant population data is available.

No dataset or collection of datasets is likely to include all intrinsic and extrinsic properties of the known and questioned samples. However, the examiner should search available data sets to find data matching as many of the relevant characteristics as possible. If this search yields a sufficient quantity of audio samples that match the most important intrinsic and extrinsic properties of the known and questioned samples, then the evaluation can proceed. If there is an insufficient quantity of audio samples matching these intrinsic and extrinsic properties, then a structured data collection may be required to provide a sample of the relevant population, or the examination should be terminated. Specific details regarding what constitutes sufficiently representative relevant population data in terms of extrinsic and intrinsic features and their impact on SR is a subject of ongoing research.

It may be possible to simulate certain extrinsic properties within the relevant population data set such as codec, duration, reverberation, background noise, etc. However, before using such simulated data in casework, it is necessary to empirically validate that the simulation method(s) yield comparable results to data collected in the simulated conditions. Some techniques for artificial simulation of case conditions may yield unrealistic or misleading results. More research is needed in the effectiveness of various simulation techniques.

Limitations that the examiner should take into account during relevant population data selection include those listed above as having moderate to high impact on SR performance as well as those that have not yet been assessed. In general, a larger set of relevant population data is better. When used to train systems components, the size of the relevant population data correlates with the system accuracy. When the relevant population data is used to calculate accuracy baselines for case conditions, then a larger relevant population data set can reduce statistical variation and provide more meaningful results. However, it has been shown that above a certain number of speakers, adding extra does not further contribute to SR performance. The number of speakers where this happens may differ depending on the case circumstances and the specific way the relevant population data is used in the SR system. Including speech samples from the relevant population data that are mismatched to the questioned and known samples on conditions listed above as having moderate to high impact can negatively affect system performance even in a large set of relevant population data.

Higher-level knowledge has become increasingly important. For example, a person's voice (spectral characteristics) may change on a day-to-day basis due to their health (e.g., a cold) or over time, due to aging, stylistic variation or age grading. Forensic experts pay special attention to these details when comparing speech samples.

7. Further Recommendations and/or Path Forward

The current process of selecting relevant population data is dependent on individual case conditions and available corpora and can therefore be highly variable. As a result, there is no guarantee of consistent implementation across practitioners. However, while the variability in case conditions is unlikely to change rapidly, new corpora are often developed. SRE corpora created by NIST from data collected by the Linguistic Data Consortium (LDC) and the Voices Obscured in Complex Environmental Settings (VOiCES) dataset from IQT Labs and SRI, International are examples. As a companion to the present document, OSAC-SR is compiling a comprehensive list of publicly available corpora relevant to SR. In addition, the proliferation of synthetic speech may eventually offer significant advantages in terms of quickly and easily creating large numbers of voice samples that can be tailored to meet situational needs.

SR systems and capabilities will continue to evolve and improve. While case data will probably always need some degree of pre-processing, newer SR systems will help the forensic audio practitioner by natively accepting a wider variety of multimedia file formats, sample rates and sample sizes. There is also active research in terms of calibrating the system at run time to fit the characteristics of the data being tested.

Finally, practitioners are strongly encouraged to make use of the OSAC-SR document titled "Validation of forensic speaker recognition for the purpose of informing legal admissibility decisions". This process provides the examiner a means of objectively characterizing system performance.

Works Cited

[1] Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council, "Strengthening Forensic Science in the United States: A Path Forward," August 2009. [Online]. Available: https://www.ojp.gov/pdffiles1/nij/grants/228091.pdf. [Accessed 6 June 2021].

[2] National Commission on Forensic Science, "Ensuring That Forensic Analysis Is Based Upon Task-Relevant Information," [Online]. Available: https://www.justice.gov/archives/ncfs/page/file/641676/download. [Accessed 8 June 2022].

[3] Wikipedia Contributors, "George Zimmerman Entry on Wikipedia," [Online]. Available: http://en.wikipedia.org/wiki/George_Zimmerman. [Accessed 29 November 2018].

[4] *State of Florida v. George Zimmerman, case no. 12-CF-1083-A, Order excluding the testimony of audio expert witnesses from trial , dated 6/22/2013 (Fla. Cir. Ct., 18th J. Dist., Seminole County).,* 2013.

[5] J. H. L. Hansen and N. Shokouhi, "Speaker identification: Screaming, stress and non-neutral speech, is there speaker content?," *IEEE SLTC Newsletter November 2013.*

[6] M. K. Nandwana, A. Ziaei and J. H. L. Hansen, "Robust Unsupervised Detection of Human Screams in Noisy Acoustic Environments," *Proc. IEEE ICASSP-2015,* pp. 161-164, 2015.

[7] J. H. L. Hansen, M. Nandwana and N. Shokouhi, "Analysis of human scream and its impact on text-independent speaker verification," *Journal of the Acoustical Society of America,* vol. 141, no. 4, p. 2957–2967, 2017.

[8] M. Nandwana, H. Boril and J. H. L. Hansen, "A New Front-End for Classification of Non-Speech Sounds: A Study on Human Whistle," in *Proceedings ISCA INTERSPEECH*, Dresden, Germany, 2015.

[9] [Online]. Available: www.biometrics.gov.

[10] A. Lawson and M. Huggins, "Triphone-Based Confidence System for Speaker Identification," in *Interspeech*, Jeju, Korea, 2004.

[11] M. Ajili, J.-F. Bonastre, B. K. Waad, R. Solange and K. Juliette, "Phonetic content impact on forensic voice comparison," *Proc. IEEE Spoken Language Technology Workshop (SLT),* pp. 210-217, 2016.

[12] A. Lawson, A. Stauffer, B. Smolenski and e. al., "Long Term Examination of Intra-Session and Inter-Session Speaker Variability," in *Interspeech*, Brighton, UK., 2009.

[13] K. W. Godin and J. H. L. Hansen, "Session variability contrasts in the MARP corpus," *Proc. Interspeech,* pp. 298-301, 2010.

[14] F. Kelly and J. H. L. Hansen, "Score-Aging Calibration for Speaker Verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 24, no. 12, pp. 2414-2424, 2016.

[15] X. Fan and J. H. L. Hansen, "Speaker identification within whispered speech audio streams," *IEEE Trans. on Audio, Speech, and Lang. Process.,* vol. 19, pp. 1408-1421, 2011.

[16] C. Zhang and J. H. L. Hansen, "Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing," *IEEE Trans. on Audio, Speech, and Lang. Process.,* vol. 19, pp. 883-894, 2011.

[17] M. Mehrabani and J. H. L. Hansen, "Singing speaker clustering based on subspace learning in the GMM mean supervector space," *Speech Commun.,* vol. 55, pp. 653-666, 2013.

[18] J. H. L. Hansen and V. Varadarajan, "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," *IEEE Trans. on Audio, Speech, and Lang. Process.,* vol. 17, no. 2, pp. 366-378, 2009.

[19] J. H. L. Hansen, C. Swail, A. J. South, R. K. Moore, H. Steeneken, E. J. Cupples, T. Anderson, C. R. Vloeberghs, I. Trancoso and P. Verlinde, "The impact of speech under 'stress' on military speech technology," *NATO Project Report,* 2000.

[20] H. J. M. Steeneken and J. H. L. Hansen, "Speech Under Stress Conditions: Overview of the Effect of Speech Production on Speech System Performance," in *IEEE Inter. Conf. on Acoustics, Speech, Signal Processing, vol. 4*, Phoenix, Arizona, 1999.

[21] J. H. L. Hansen and T. Hasan, "Speaker Recognition by Machines and Humans: A Tutorial Review," *IEEE Signal Processing Magazine,* pp. 74-99, November 2015.

[22] J. H. L. Hansen and D. Cairns, "ICARUS: A Source Generator Based Real-time System for Speech Recognition in Noise, Stress, and Lombard Effect," *Speech Communication,* vol. 16, no. 4, pp. 391-422, 1995.

[23] H. Boril and J. H. L. Hansen, "Unsupervised Equalization of Lombard Effect for Speech Recognition in Noisy Adverse Environments," *IEEE Trans. Audio, Speech, and Language Processing,* vol. 18, no. 6, pp. 1379-1393, 2010.

[24] J. H. L. Hansen, "Morphological Constrained Enhancement with Adaptive Cepstral Compensation (MCE-ACC) for Speech Recognition in Noise and Lombard Effect," *IEEE Transactions on Speech & Audio Processing, SPECIAL ISSUE: Robust Speech Recognition,* vol. 2.

[25] E. Lombard, "Le signe de l'elevation de la voix," *Annales Des Maladies de l'Oreille, Du Larynx, Du Nez Et Du Pharynx,* vol. 37, pp. 101-119, 1911.

[26] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Communication,* vol. 20, pp. 151-173, 1996.

[27] M. Mehrabani and J. H. L. Hansen, "Language Identification: Analysis of Singing Speakers," in *IEEE ICASSP-2011*, Prague, Czech Republic, May 22-27, 2011.

[28] M. Mehrabani and J. H. L. Hansen, "Speaker Clustering for a Mixture of Singing and Reading," in *ISCA Interspeech-2012*, Portland, OR, Sept. 9-13, 2012.

[29] M. Mehrabani and J. H. L. Hansen, "Dimensionality Analysis of Singing Speech Based on Locality Preserving Projections," in *ISCA INTERSPEECH*, Lyon, France, August 25-29, 2013.

[30] M. Mehrabani and J. H. L. Hansen, "Singing Speaker Clustering Based on Subspace Learning in the GMM Mean Supervector Space," *Speech Communication,* vol. 55, pp. 653-666, 2013.

[31] F. Bahmaninezhad and J. H. L. Hansen, "i-Vector/PLDA speaker recognition using support vectors with discriminant analysis.," in *Proceedings ICASSP*, 2017.

[32] M. Bokshi, F. Tao, C. Busso and J. H. L. Hansen, "Assessment and Classification of Singing Quality Based on Audio-Visual Features," in *IEEE Visual Communications and Image Processing (VCIP)*, Saint Petersburg, Florida, December 10-13, 2017.

[33] J. Hansen, M. Bokshi and S. Khorram, "Speech Variability: A cross-language study on acoustic variations of speaking versus untrained singing," *Journal of the Acoustical Society of America,* vol. 148, no. 2, pp. 829-844, August 2020.

[34] M. Mehrabani and J. Hansen, "Singing Speaker Clustering Based on Subspace Learning in the GMM Mean Supervector Space," *Speech Communication,* vol. 55, pp. 653-666, February 2013.

[35] D. Van der Vloed, M. Jessen and S. Gfroerer, "Experiments with Two Forensic Automatic Speaker Comparison Systems Using Reference Populations that (Mis) Match the Test Language," in *AES*, Arlington, VA, 2017.

[36] A. Misra and J. H. L. Hansen, "Compensating for Language Mismatch in Speaker Verification," *Speech Communication,* vol. 96, pp. 58-66, 2018.

[37] A. Misra and J. H. L. Hansen, "Spoken Language Mismatch in Speaker Verification: An Investigation with NIST-SRE and CRSS Bi-Ling Corpora," in *IEEE SLT-2014: Spoken Language Technology Workshop*, Lake Tahoe, 2014.

[38] D. Van der Vloed, "Validation Report," 2018.

[39] L. A. Ramig and R. L. Ringel, "Effects of physiological aging on selected acoustic characteristics of voice," *J. of Speech, Lang., and Hearing Research,* vol. 26, pp. 22-30, 1983.

[40] E. T. Stathopoulos, J. E. Huber and J. E. Sussman, "Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4-93 years of age," *J. Speech, Language, Hearing Research,* vol. 54, p. 1011–1021, 2011.

[41] G. S. Morrison and F. Kelly, "A statistical procedure to adjust for time-interval mismatch in forensic voice comparison," *Speech Communication,* vol. 112, pp. 15-21, 2019.

[42] F. Kelly, A. Drygajlo and N. Harte, "Speaker verification in score-ageing-quality classification space," *Comput. Speech Lang.,* vol. 27, pp. 1068-1084, 2013.

[43] R. Rhodes, "Aging effects on voice features used in forensic speaker comparison," *International Journal Of Speech Language and The Law,* vol. 24, no. 2, pp. 177-199, 2017.

[44] E. Yilmaz, D. Jelske, H. van de Velde, F. Kampstra, J. Algra, H. van den Heuvel and D. A. van Leeuwen, "Longitudinal Speaker Clustering and Verification Corpus with Code-Switching Frisian-Dutch Speech," in *Interspeech*, 2017.

[45] G. R. Doddington, "The effect of target/non-target age difference on speaker recognition performance," in *Odyssey*, Singapore, 2012.

[46] A. D. Lawson, A. Stauffer, E. J. Cupples, S. J. Wenndt, W. Bray and J. J. Grieco, "The multi-session audio research project (MARP) corpus: Goals, design and initial findings," *Proc. Interspeech, Brighton, UK, 2009. .*

[47] C. Yu and J. H. L. Hansen, "A study of voice production characteristics of astronaut speech during Apollo-11 for long-term speaker modeling in space," *," Journal of the Acoustical Society of America,* vol. 141, no. 3, pp. 1605-1614, 2017.

[48] K. W. Godin and J. H. L. Hansen, "Physical Task Stress and Speaker Variability in Voice Quality," *EURASIP Journal of Speech, Audio, and Music Processing,* vol. 29, pp. 1-13, 2015.

[49] K. W. Godin and J. H. L. Hansen, "The effects of Physical Task Stress on Phone Classes of American English," *Journal of the Acoustical Society of America,* vol. 130, no. 6, pp. 3992-3998, 2011.

[50] C. Zhang, G. Liu, C. Yu and J. H. L. Hansen, "i-Vector Based Physical Task Stress Detection with Different Fusion Strategies," in *ISCA INTERSPEECH-2015*, Dresden, Germany, 2015.

[51] J. H. L. Hansen, A. Sangwan and Kim, "Chapter 5: Speech Processing for Robust Speaker Recognition: Advancements for Speech Under Stress, Lombard Effect, and Emotion," in *Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism (eds. H. Patil, A. Neustein)*, Springer, 2012, pp. 103-123.

[52] A. Misra, S. Ranjan, C. Zhang and J. H. L. Hansen, "Anti-spoofing System: An Investigation of measures to Detect Synthetic And Human Speech," in *ISCA INTERSPEECH-2015*, Dresden, Germany, 2015.

[53] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu and S. Marcel, "Joint Speaker Verification and Antispoofing in the i -Vector Space," *IEEE Trans. Information Forensics and Security,* pp. 821-832, 2015.

[54] Z. Wu, E. S. Chng and H. Li, "Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition," in *ISCA INTERSPEECH-2012*, Portland, OR, 2012.

[55] P. Eckert and J. R. Rickford, Style and Sociolinguistic Variation, Cambridge University Press, 2001.

[56] J. H. L. Hansen and H. Boril, "On the issues of intra-speaker variability and realism in speech, speaker, and language recognition tasks," *Speech Communication,* vol. 101, pp. 94-108, 2018.

[57] S. Greenberg, "On the Origins of Speech Intelligibility in the Real World," in *ISCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, France, 1997.

[58] J. J. Godfrey, E. C. Holliman and J. McDaniel, "SWITCHBOARD: telephone speech corpus for research and development," in *Proceedings ICASSP - IEEE International Conference on Acoustics, Speech, and Signal Processing*, San Francisco, CA, 1992.

[59] B. Pellom and J. H. L. Hansen, "Voice Analysis in Adverse Conditions: The Centennial Olympic Park Bombing 911 Call," in *IEEE Midwest Symposium on Circuits & Systems*, Sacramento, CA, 1997.

[60] M. K. Nandwana, L. Ferrer, M. McLaren, D. Castan and A. Lawson, "Analysis of Critical Metadata Factors for the Calibration of Speaker Recognition Systems," in *Proc. Interspeech*, Graz, Austria, 2019.

[61] D. A. Reynolds, M. A. Zissman, T. F. Quatieri, G. C. O'Leary and B. A. Carlson, "The effects of telephone transmission degradations on speaker recognition performance," *Proc. IEEE ICASSP,* pp. 329-332, 1995.

[62] P. Kenny, G. Boulianne, P. Ouellet and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. on Audio, Speech, and Lang. Process.,* vol. 15, pp. 1435-1447, 2007.

[63] R. Auckenthaler, M. Carey and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing,* vol. 10, pp. 42-54, 2000.

[64] M. McLaren, V. Abrash, M. Graciarena, Y. Lei and J. Pesan, "Improving Robustness to Compressed Speech in Speaker Recognition," in *Interspeech*, Lyon, France, 2013.

[65] R. C. Rose, E. M. Hofstetter and D. A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Trans. on Speech and Audio Process.,* vol. 2, pp. 245-257, 1994.

[66] Q. Jin, T. Schultz and A. Waibel, "Far-field speaker recognition," *IEEE Trans. on Audio, Speech, and Lang. Process.,* vol. 15, pp. 2023-2032, 2007.

[67] C. Greenberg, A. Martin, L. Brandschain, J. Campbell, C. Cieri, G. Doddington and J. Godfrey, "Human assisted speaker recognition in NIST SRE10," *Proc. ISCA Odyssey, Brno, Czech Republic,* pp. 180-185, 2010.

[68] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot and N. Scheffer, "Promoting robustness for speaker modeling in the community: the PRISM evaluation set," in *Proc. of NIST SRE 2011 Workshop*, 2011.

[69] M. McLaren, Y. Lei and L. Ferrer, "Advances in deep neural network approaches to speaker recognition," *IEEE international conference on acoustics, speech and signal processing (ICASSP),* pp. 4814-4818, 2015.

[70] M. McLaren, Y. Lei, N. Scheffer and L. Ferrer, "Application of convolutional neural networks to speaker recognition in noisy conditions," in *Proc. Interspeech*, 2014.

[71] M. I. Mandasari, R. Saeidi, M. McLaren and D. A. van Leeuwen, "Quality measure functions for calibration of speaker recognition systems in various duration conditions," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 21, no. 11, pp. 2425-2438, 2013.

[72] M. Sambur, "Selection of Acoustic Features for Speaker Identification," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 23, no. 2, 1975.

[73] A. Stauffer and A. Lawson, "Speaker Recognition on Lossy Compressed Speech using the SPEEX Codec," in *Interspeech*, Brighton, UK., 2009.

[74] J. Silovsky, P. Cerva and J. Zdansky, "Assessment of Speaker Recognition on Lossy Codecs Used for Transmission of Speech," in *53rd International Symposium ELMAR*, Zadar, Croatia, 2011.

End Notes

[i] EER or Equal Error Rate is the point on a systems operating curve where the number of false positives a system commits is equal to the number of false negatives it commits on data where speakers are known. EER is commonly used in the evaluation of Speaker Recognition systems. It should be noted that for some applications users of a system may prefer many fewer false positives even at the expense of many more false negative or the reverse. Where technology is evaluated for a specific application, it is possible to evaluate multiple systems using a specific operating point, for example the number of false negatives for a given false positive rate. However, in evaluations for which no specific application is identified, ERR is more commonly used.