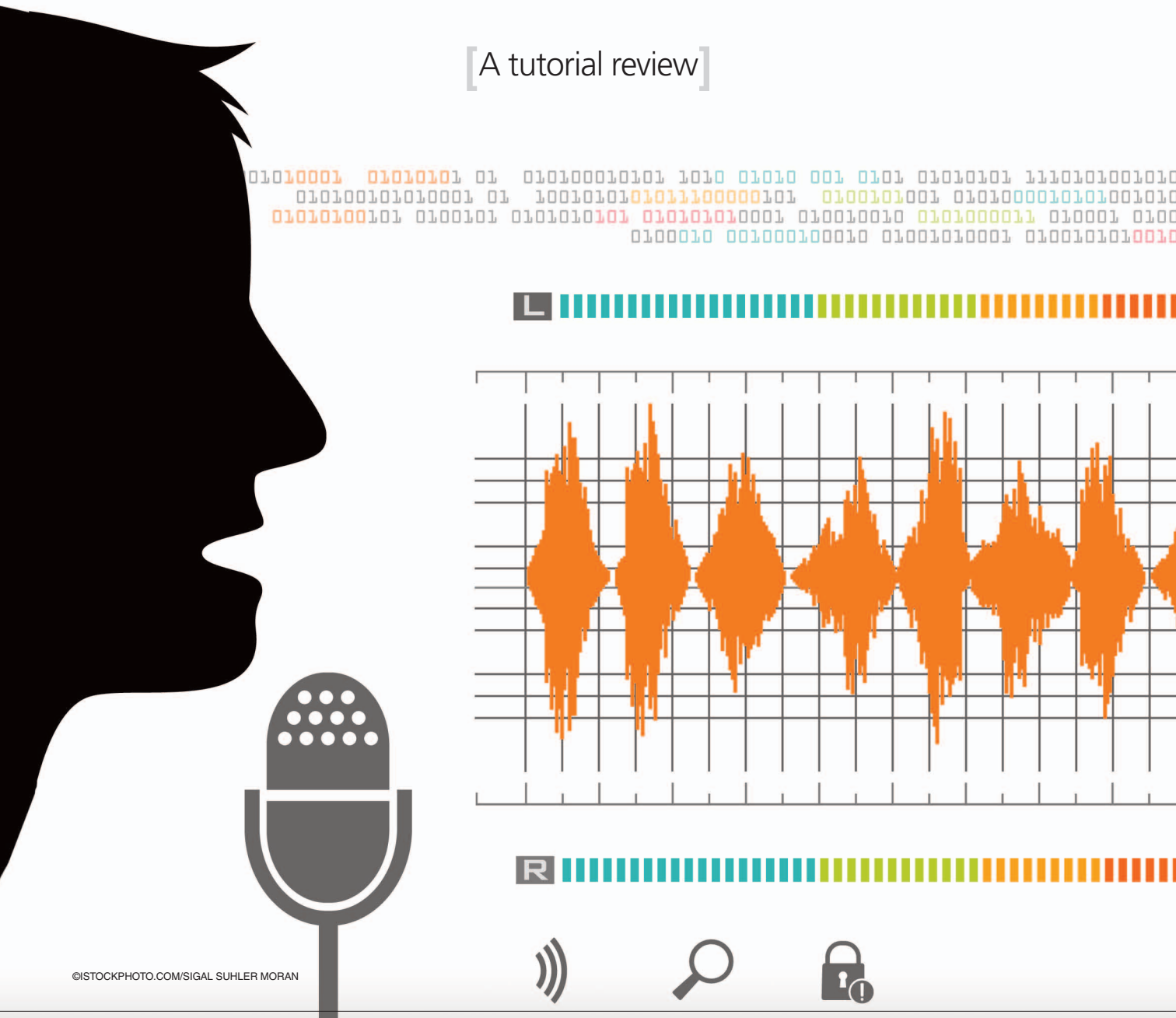
John H.L. Hansen and Taufiq Hasan

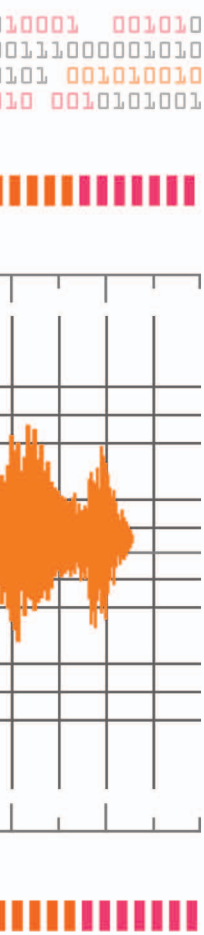
Speaker Recognition by Machines and Humans

A tutorial review



1053-5888/15©2015IEEE

Identifying a person by his or her voice is an important human trait most take for granted in natural human-to-human interaction/communication. Speaking to someone over the telephone usually begins by identifying who is speaking and, at least in cases of familiar speakers, a subjective verification by the listener that the identity is correct and the conversation can proceed. Automatic speaker-recognition systems have emerged as an important means of verifying identity in many e-commerce applications as well as in general business interactions, forensics, and law enforcement. Human experts trained in forensic speaker recognition can perform this task even better by examining a set of acoustic, prosodic, and linguistic characteristics of speech in a general approach referred to as *structured listening*. Techniques in forensic speaker recognition have been developed for many years by forensic speech scientists and linguists to help reduce any potential bias or preconceived understanding as to the validity of an unknown audio sample and a reference template from a potential suspect. Experienced researchers in signal processing and machine learning continue to develop automatic algorithms to effectively perform speaker recognition—with ever-improving performance—to the point where automatic systems start to perform on par with human listeners. In this article, we review the literature on speaker recognition by machines and humans, with an emphasis on prominent speaker-modeling techniques that have emerged in the last decade for automatic systems. We discuss different aspects of automatic systems, including voice-activity detection (VAD), features, speaker models, standard evaluation data sets, and performance metrics. Human speaker recognition is discussed in two parts—the first part involves forensic speaker-recognition methods, and the second illustrates how a naïve listener performs this task from a neuroscience perspective. We conclude this review with a comparative study of human versus machine speaker recognition and attempt to point out strengths and weaknesses of each.



INTRODUCTION

Speaker recognition and verification have gained increased visibility and significance in society as speech technology, audio content, and e-commerce continue to expand. There is an ever-increasing need to search for audio materials, and searching based on speaker identity is a growing interest. With emerging technologies such as Watson, IBM's supercomputer [1], which can compete with expert human players in the game of "Jeopardy," and Siri [2], Apple's powerful speech-recognition-based personal assistant, it is not hard to imagine a future when handheld devices will be an extension of our identity—highly intelligent, sympathetic, and fully functional personal assistants, which will not only understand the meaning of what we say but also recognize and track us by our voice or other identifiable traits.

As we increasingly realize how much sensitive information our personal handheld devices can contain, it will become critical that effective biometric authentication be an integral part of access to information and files contained on the device, with a potential range of public/private access. Because speech is the most natural means of human communication, these devices will unavoidably lean toward automatic voice-based authentication in addition to other forms of biometrics. Apple's recent iPhone models have already introduced fingerprint scanners, reflecting the industry trend. The latest Intel technology on laptops employs face recognition as the password for access. Our digital personal assistant, in theory, could also replace most forms of traditional key locks as well for our home and vehicles, again making security of such a personal device more important.

Apart from personal authentication for access control, speaker recognition is an important tool in law enforcement, national security, and forensics in general. Because of widespread availability, cell phones have become the primary means of communication for the general public, and, unfortunately, also for criminals. Unlike the domain of personal authentication for personal files/information access, these individuals usually do not want to be recognized. In such cases, many criminals may attempt to alter their voice to prevent them from being identified. This introduces additional challenges for developers of speaker-recognition technology—"Is the participant a willing individual in being assessed?" In law enforcement, any voice recorded as part of evidence may be disguised or even synthesized, to obscure recognition, adding to the difficulty of being recognized. Over a number of years, forensic speech scientists have devised different strategies to overcome these difficulties.

Interestingly, humans routinely recognize individuals by their voices with striking accuracy, especially when the degree of familiarity with the subject is high (i.e., close acquaintances or public figures). Many times, even a short nonlinguistic queue, such as a laugh, is enough for us to recognize a familiar person [3]. On the other hand, it is also common knowledge that we cannot recognize a once-heard voice very easily and sometimes have difficulty in recognizing familiar voices over the phone. With these ideas in mind, a naïve person may wonder what exactly makes speaker recognition difficult and why is it a topic of such rigorous research.

Digital Object Identifier 10.1109/MSP.2015.2462851
Date of publication: 13 October 2015

From the discussion so far, it is safe to say that speaker recognition can be accomplished in three ways.

- We can recognize familiar voices with considerable ease without any conscious training. This form of speaker recognition can be termed *naïve speaker recognition*.
- In forensic investigations, speech samples from a telephone call are often compared to recordings of potential suspects (i.e., from a phone threat, emergency 911 call, or known criminal). In these cases, trained listeners are involved in systematically comparing the speech samples to provide an informed decision concerning their similarities. We would classify this as *forensic speaker recognition*.
- Finally, we have *automatic speaker recognition,* where the complete speech analysis and decision-making process is performed using computer analysis.

In both naïve and forensic speaker recognition, humans are directly involved in the process, even though some automatic or computer-assisted means may be used to supplement knowledge extraction for the purposes of comparison in the forensic scenario. However, it should be noted that both the forensic and automatic methods are highly systematic, and the procedures from both disciplines are technical in nature.

The forensic and automatic speaker-recognition research communities have developed various methods more or less independently for several decades. Conversely, naïve recognition is a natural ability of humans—which is, at times, very accurate and effective. Recent studies on brain imaging [4], [5] have revealed many details on how we perform cognitive-based speaker recognition, which may inspire new directions for both automatic and forensic methods.

In this article, we present a tutorial review of the automatic speaker-recognition methods, especially those developed in the last decade, while providing the reader with a perspective on how humans also perform speaker recognition, especially by forensics experts and naïve listeners. The aim is to provide a discussion on the three classes of speaker recognition, highlighting the important similarities and differences among them. We emphasize how automatic techniques have evolved over time toward more current approaches. Many speech-processing techniques, such as Mel-scale filter-bank analysis and concepts in noise masking, are inspired by human auditory perception. Also, there are similarities between the methods used by forensic voice experts and automated systems—even though, in many cases, the research communities are separate. We believe that incorporating the perspective of speech perception by humans in this review, including highlights of both strengths and weaknesses in speaker recognition compared to machines, will help broaden the view of the reader and perhaps inspire new research directions in the area.

SPEAKER-RECOGNITION TASKS

First, to consider the overall research domain, it would be useful to clarify what is encompassed by the term *speaker recognition*, which consists of two alternative tasks: speaker identification and verification. In speaker identification, the task is to identify an unknown speaker from a set of known speakers. In other words, the goal is to find the speaker who sounds closest to the speech stemming from an unknown speaker within an audio sample. When all speakers within a given set are known, it is called a *closed* or *in-set scenario*. Alternatively, if the potential input test subject could also be from outside the predefined known speaker group, this becomes an open-set scenario, and, therefore, a world model or universal background model (UBM) [6] is needed. This scenario is called *open-set speaker recognition* (also *out-of-set speaker identification*).

In speaker verification, an unknown speaker claims an identity, and the task is to verify if this claim is true. This essentially comes down to comparing two speech samples/utterances and deciding if they are spoken by the same speakers. In some methods, this is done by comparing the unknown sample against two alternative models, the claimed speaker model and a world model. In the forensic scenario, the general task is to identify the unknown speaker, who is suspected of a crime, but, in many instances, verification is also necessary.

Speaker recognition can be based on an audio stream that is text dependent or text independent. This is more relevant in authentication applications—where a claimed user says something specific, such as a password or personal identification number, to gain access to some resource/information. Throughout this article, the focus will be on text-independent speaker verification, especially in the treatment of automatic systems.

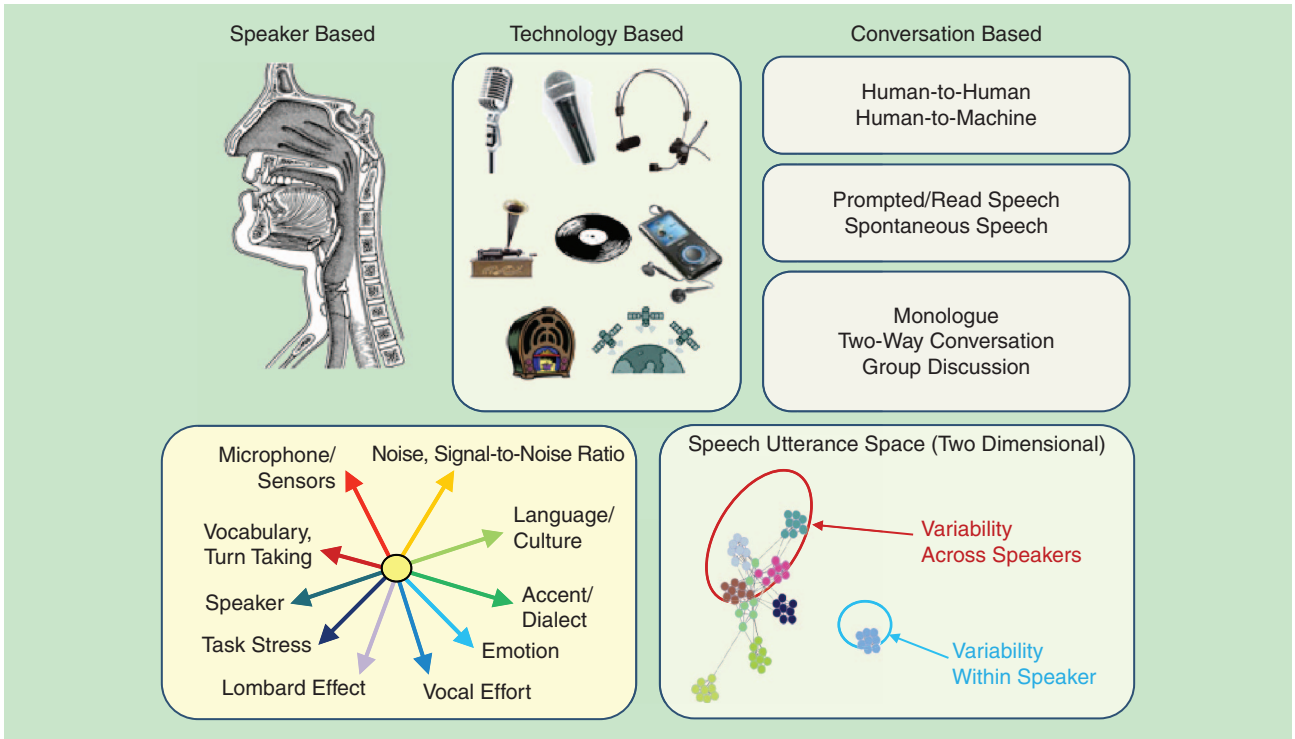
CHALLENGES IN SPEAKER RECOGNITION

Unlike other forms of biometrics (e.g., fingerprints, irises, facial features, gait, and hand geometry) [7], human speech is a performance biometric. Simply put, the identity information of the speaker is embedded (primarily) in how speech is spoken, not necessarily in what is being said (although in many voice forensic applications, it is also necessary to identify who said what within a multispeaker discussion). This makes speech signals prone to a large degree of variability. It is important to note that even the same person does not say the same words in exactly the same way every time (this is known as *style shifting* or *intraspeaker variability*) [8]. Also, various recording devices and transmission methods commonly used exacerbate the problem. For example, we may find it difficult to recognize someone's voice through a telephone or maybe when the person is not healthy (i.e., has a cold) or is performing another task or speaking with a different level of vocal effort (i.e., whispering or shouting).

SOURCES OF VARIABILITY IN SPEAKER RECOGNITION

To consider variability, Figure 1 highlights a range of factors that can contribute to mismatch for speaker recognition. These can be partitioned based on three broad classes: 1) speaker based, 2) conversation based, and 3) technology based. Also, variability for speakers can be within speakers and across speakers.

- *Speaker-based variability sources:* these reflect a range of changes in how a speaker produces speech and will affect system performance for speaker recognition. These can be thought of as *intrinsic* or *within-speaker variability* and include the following factors.



[FIG1] **Sources of variability in speaker recognition.**

- *Situational task stress*—the subject is performing some task while speaking, such as operating a vehicle (car, plane, truck, etc.), hands-free voice input (factory setting, emergency responders/fire fighters, etc.), which can include cognitive as well as physical task stress [9].
- *Vocal effort/style*—the subject alters his or her speech production from normal phonation, resulting in whispered [10], [11], soft, loud, or shouted speech; the subject alters his or her speech production mechanism to speak effectively in the presence of noise [12], known as the Lombard effect; or the subject is singing versus speaking [13].
- *Emotion*—the subject is communicating his or her emotional state while speaking (e.g., anger, sadness, happiness, etc.) [14].
- *Physiological*—the subject has some illness or is intoxicated or under the influence of medication; this can include aging as well.
- *Disguise*—the subject intentionally alters his or her voice to circumvent the system. This can be by natural means (speaking in a harsh voice to avoid detection, mimicking another person's voice, etc.) or using a voice-conversion system.

■ *Conversation-based/higher-level mode/language of speaking variability sources*: these reflect different scenarios with respect to the voice interaction with either another person or technology system, or differences with respect to the specific language or dialect spoken, and can include

- *human-to-human*: speech that includes two or more individuals interacting or one person speaking and addressing an audience
  - —language or dialect spoken
  - —if speech is read/prompted (through visual display or through headphones), spontaneous, conversational, or disguised speech
  - —monologue, two-way conversation, public speech in front of an audience or for TV or radio, group discussion
- *human-to-machine*: speech produced where the subject is directing his or her speech toward a piece of technology (e.g., cell/smart/landline telephone and computer)
  - —*prompted speech*: voice input to a computer
  - —*voice input for telephone/dialog system/computer input*: interacting with a voice-based system.

■ *Technology- or external-based variability sources*: these include how and where the audio is captured and the following issues:

- *electromechanical*—transmission channel, handset (cell, cordless, and landline) [15]–[17] microphone
- *environmental*—background noise [18] (stationary, impulsive, time-varying, etc.), room acoustics [19], reverberation [20], and distant microphone
- *data quality*—duration, sampling rate, recording quality, and audio codec/compression.

These multifaceted sources of variation pose the greatest challenge in accurately modeling and recognizing a speaker, whether automatic algorithms are used, or if human listening/assessment is performed. Given that speech will contain variability, the task of speaker verification is deciding if the variability is due to the same speaker (intra {within}-speaker) or different speakers (inter {across}-speaker).

In current automated speaker-recognition technology, various mathematical tools are used to mitigate the effects of these

variability/degradations, especially the extrinsic ones. Additive noise and transmission channel variability have received much attention recently. Intrinsic variability in speech is very difficult to quantify and account/address for in automatic assessment. Higher-level knowledge may become important in these cases. For example, even if a person's voice (spectral characteristics) may change due to his or her current health (e.g., a cold) or aging, the person's accent or style of speech remains generally the same. Forensic experts pay special attention to these details when detecting a subject's voice from potential suspects' speech recordings.

CHALLENGES IN SPEAKER RECOGNITION

Early efforts in speaker recognition involving technology focused more on the telecommunications domain, where telephone handset and communication channel variation was the primary concern. In the United States, when telephone systems were confined to handheld rotary phones in the home and public phone booths in public settings, technology- and telephony-based variability was an issue, but it was, to a large degree, significantly less important than it is today. With mobile cell phone/smartphone technology dominating the world's telecommunications market, the diversity of telephony scenarios has expanded considerably. Virtually all cell phones have a speaker option, which allows voice interaction at a distance from the microphone, and movement of the device introduces a wider range of channel variability.

Voice is also a time-varying entity. Research has shown that intersession variability, the inherent changes present within audio files captured at different times, results in changes in speaker-recognition performance. Analysis of the Multisession Audio Research Project corpus collected using the same audio equipment in the same location on a monthly basis over a 36-month period showed measurable differences in speaker-recognition performance [21], [22]. However, the changes in speaker-identification performance seem to be independent of the time difference between training and testing [21], [23]. While no aging effects were noted for the 36-month period, other research has demonstrated long-term changes in speech physiology and production due to aging [23]. More extensive research that explores the evolution of speaker structure for speaker recognition over a 20–60-year period (at least for a small subset of speakers) has shown measurable changes and suggested methods to address changes due to aging [24], [25].

These examples of variation point to the sensitivity of existing speaker-recognition technology. It is possible to employ such technology in a way that could lead to noncredible results. A recent example of how to wrongly use automatic speaker recognition was seen during the recent U.S. legal case involving George Zimmerman, who was accused of shooting Trayvon Martin during an argument [26]. In that case, a 911 emergency call captured a scream for help heard in the background. The defense team claimed that it was Zimmerman who was yelling while supposedly being attacked by Trayvon Martin, who was killed; alternatively, the prosecutors argued that it was the unarmed victim who was shouting. Parents of both parties testified that the voice heard on the 911 call belonged to their own son. Some forensic experts did attempt to use semiautomatic methods to compare the original scream and a simulated scream obtained from Zimmerman. The issue of using automatic assessment schemes for scream analysis to assess identity was controversial, as experts from the U.S. Federal Bureau of Investigation (FBI) and U.S. National Institute of Standards and Technology (NIST) testified that these methods are unreliable. A brief probe analysis of scream and speaker-recognition technology confirmed the limitations of current technology [27].

Most forensic speaker-identification scenarios, however, are not as complicated. When there is sufficient speech material available from the offender and the suspect, systematic analysis can be performed to extract speaker idiosyncratic characteristics, also known as feature parameters, from the speech data, and a comparison between the samples can be made. Also, in automatic speaker-identification systems, features designed to differentiate among speakers are first extracted and mathematically modeled to perform a meaningful comparison. Thus, in the next section, we consider what traits help identify a person from his or her speech—in other words, what are the feature parameters that we should consider in making an assessment?

SPEAKER CHARACTERIZATION: FEATURE PARAMETERS

Every speaker has some characteristic traits in his or her voice that are unique. Individual speaker characteristics may not be so easily distinguishable but are unique mainly due to speaker vocal tract physiology and learned habits of articulation. Even identical twins have differences in their voices, though studies show they have similar vocal tract shape [28] and acoustic properties [29], and it is difficult to distinguish them from a perceptual/forensics perspective [30], [31]. Researchers in voice forensics have even participated in the National Twins Day event held in Twinsburg, Ohio, [32] in an effort to capture voice and other biometrics to explore the challenges in distinguishing closely related individuals. Thus, whether recognition is performed by humans (an expert or naïve listener) or by machines, some measurable and predefined aspects of speech need to be considered to make meaningful comparisons among voices. Generally, we refer to these characterizing aspects as *feature parameters*.

One might expect that a unique voice must have unique features, but this is not always true. For example, two different speakers may have the same speaking rate (which is a valid feature parameter) but differ in average pitch. This is complicated by the variability and degradations discussed previously, which is why considering multiple feature parameters is critical.

PROPERTIES OF IDEAL FEATURES

As outlined by Nolan [33], ideally a feature parameter should

1) show high between-speaker variability and low within-speaker variability
2) be resistant to attempted disguise or mimicry
3) have a high frequency of occurrence in relevant materials
4) be robust in transmission
5) be relatively easy to extract and measure.

These properties, though mentioned in the forensic speaker-identification context, apply in general. Interestingly, Wolf [34] discussed very similar sets of properties in the context of features for

automatic speaker recognition, independently, preceding Nolan [33]. We refer to these properties as *ideal property 1–5* throughout this article. It should be reiterated that variability in features will always exist, but the important task is to determine if the origin of the variability is the same speaker or different speakers.

We now discuss various feature parameters used in forensic speaker identification, which can be and are also useful for general speech understanding. There is no fixed set of rules for what parameters should be used in forensic speaker recognition. This is largely dependent on the circumstances or availability [35]. Some forensic experts may choose parameters to compare based on the most obvious aspect of the voices under consideration. Feature parameters can be broadly classified into auditory versus acoustic, linguistic versus nonlinguistic, and short-term versus long-term features.

AUDITORY VERSUS ACOUSTIC FEATURES

Some aspects of speech are better suited for auditory analysis (i.e., through listening). Auditory features are thus defined as aspects of speech that can "be heard and objectively described" by a trained listener [36]. These can be specific ways of uttering individual speech sounds (e.g., the pronunciation of the vowel sounds in the word *hello* can be used as auditory features).

Acoustic features, on the other hand, are mathematically defined parameters derived from the speech signal using automatic algorithms. Clearly, these kinds of features are used in automatic systems, but they are also used in computer-assisted forensic speaker recognition. Fundamental frequency (F0) and formant frequency bandwidth are examples of acoustic features. Automatic systems frequently use acoustic features derived from the short-term power spectrum of speech.

Both auditory and acoustic features have their strengths and weaknesses. Two speech samples may sound very similar but have highly variant acoustic parameters [37]. Alternatively, speech samples may sound very different yet have similar acoustic features [28]. It is thus generally accepted that both auditory and acoustic features are indispensable for forensic investigations [35]. One might argue that if reverse engineering of the human auditory system [38] is fully successful, auditory features can also be extracted using automatic algorithms.

LINGUISTIC VERSUS NONLINGUISTIC FEATURES

Linguistic feature parameters can provide contrast "within the structure of a given language or across languages or dialects" [35]. They can be acoustic or auditory in nature and are classified further as phonological, morphological, and syntactic [36]. A simple example of a linguistic feature is whether the "r" sound at the end of a word, e.g., *car,* is pronounced or silent—in some dialects of English, this type of "r" sound is not pronounced (i.e., Lancashire versus Yorkshire dialects of U.K. English). This is different from an auditory analysis of how the "r" sound is pronounced, since, in this case, this speech sound will be compared across different words.

Nonlinguistic features include aspects of speech that are not related to the speech content. Typical nonlinguistic features may include: speech quality (nasalized, breathy, husky, etc.), fluency, speech pauses (frequency and type), speaking rate, average fundamental frequency, and nonspeech sounds (coughs, laughs, etc.). Again, these features can be auditory or acoustic in nature. Referring back to the Zimmerman case, the manner of screaming (i.e., loudness, pitch, and duration) could be a potential feature if it could be properly measured/parameterized.

SHORT-TERM VERSUS LONG-TERM FEATURES

Depending on the time span of the feature parameters, they can be categorized as short versus long term. Most features discussed so far are short term or segmental in nature. Popular automatic systems mostly use short-term acoustic features, especially the ones extracted from the speech spectrum. The short-term features are also effective in auditory forensic analysis, for example, direct comparison of the "r" sound and consonant–vowel transition [33].

The long-term features are usually averaged short-term parameters, (e.g., fundamental frequency, short-term spectrum). These parameters have the benefit of being insensitive to fluctuations due to individual speech sounds and provide a smoother measurement from a speech segment. The long-term features also include energy, pitch, and formant contours, which are measured/averaged over long time periods. Recent automatic systems also successfully used such features [39]–[41]. If a feature parameter is extracted from an entire speech utterance, we refer to it as an *utterance-level feature,* or *utterance feature* for short. This concept will become very useful as we proceed with the discussion to automatic systems.

FORENSIC SPEAKER RECOGNITION

While the focus in this review is on automatic machine-based speaker recognition, we also briefly consider both forensic and naïve speaker recognition. The need for forensic speaker recognition/identification arises when a criminal leaves his or her voice as evidence, be it as a telephone recording or speech heard by an earwitness. The use of technology for forensic speaker recognition has been discussed as early as 1926 [42] with speech waveforms. Later, the spectrographic representation of speech was developed at AT&T Bell Laboratories during World War II. It was popularized much later, in the 1970s, when it came to be known as the *voiceprint* [43]. As the name suggests, the voiceprint was presented as being analogous to fingerprints and with very high expectations. Later, the reliability of the voiceprint for voice identification, from its operating mechanisms to formal procedure, was thoroughly questioned and argued [44], [45], even called "an idea gone wrong" [45]. It was simply not accurate with speech being so subject to variability. Most researchers today believe it to be controversial at best. A chronological history of voiceprints can be found in [46], and an overview discussion on forensic speaker recognition can be found in [47]. Here, we present an overview with respect to current trends.

In the general domain of forensic science, the United States has recently formed the Organization of Scientific Area Committees (OSAC) (http://www.nist.gov/forensics/osac.cfm), which is overseen by NIST. The legacy structure before OSAC was Forensic Science Working Groups. The current OSAC organization was established to help formalize the process of best practices for standards as they relate to researchers, practitioners, legal and law enforcement as well as government agencies. It also allows for a

more transparent process in which experts and users of the various technologies can provide feedback and help shape best practices. Currently, OSAC is establishing a number of working documents to build consensus among the various forensic subfields. A good source of current information from OSAC is the NIST Forensic Science Publications website (http://www.nist.gov/forensics/publications.cfm).

Today, forensic speaker identification is commonly performed by expert phoneticians who generally have backgrounds in linguistics and statistics. This is a very complex procedure, and varies among practitioners. There is no standard set of procedures every practitioner agrees upon. Different aspects/features are considered when forensic experts make comparisons between utterances. The procedure is often dictated by the situation at hand—for example, if only a few seconds of screaming of the unknown speaker is available on the evidence tape, the only thing that can be done is to try to recreate a similar scream from the likely speaker (suspect) and compare, which is generally not feasible.

THE LIKELIHOOD RATIO

Regardless of the varying approaches by practitioners, forensic speaker recognition essentially entails a scientific and objective method of comparing voices (there are, apparently, people who attempt to perform this task using methods unacceptable by the general forensic community [48]). Forensic experts must testify in court concerning the similarity/dissimilarity of the speech samples in consideration in a meaningful way. However, they cannot make any categorical judgment about the voices (e.g., the two voices come from the same speaker). For this purpose, the likelihood ratio (LR) [49] measure was introduced, which forensic experts use to express the strength of their findings [50], [51]. This means that the evaluation of forensic speech samples will not yield an absolute identification or elimination of the suspect but instead provides a probabilistic confidence measure. As discussed previously, even speech samples from the same speaker will differ in realistic scenarios. The goal of the forensic voice comparison expert is thus to estimate the probability of observing the measured difference between speech samples assuming that they were spoken by 1) the same speaker and 2) different speakers [35]. The procedure for measuring the LR is given next:

- X = Speech sample recorded during a crime (evidence recording).
- Y = Speech sample obtained from suspect (exemplar).
- H_0 = The hypothesis that X and Y are spoken by the same person.
- H_1 = The hypothesis that X and Y are spoken by different persons.
- E = Observed forensic evidence (e.g., average pitch from X and Y differ by 10 Hz).

The LR formula is

$$\mathrm{LR} = \frac{p(E \mid H_0)}{p(E \mid H_1)}.$$

As an example, if the average pitch difference between two utterances is considered the feature parameter, the forensic expert first computes the probability distribution of this feature parameter for speech data collected from many same-speaker (hypothesis H_0) and different-speaker (hypothesis H_1) pairs. In the next step, given the evidence E (average pitch from X and Y differ by 10 Hz), the conditional probabilities $p(E \mid H_0)$, and $p(E \mid H_1)$, can be computed. Note that the forensic expert does not try to estimate $p(H_0 \mid E)$ (i.e., the probability that the suspect is guilty given the observed evidence). This is because this estimation is done using Bayes' theorem, which requires the prior probabilities of the hypotheses generally not provided to the expert (and are also difficult to estimate). More discussion on this can be found in [35, Ch. 4].

APPROACHES IN FORENSIC SPEAKER IDENTIFICATION

Here, we discuss general approaches taken for forensic speaker recognition. The methods described are performed by human experts, fully or partially. While full automatic approaches are also considered for forensics, we discuss automatic speaker recognition in later sections.

AUDITORY APPROACH

This approach is practiced by auditory phoneticians and involves producing a detailed transcript of the evidence tape and exemplars. Drawing on their experience, experts listen to speech samples and attempt to detect any aspects of the voices that are unusual, distinctive, or noteworthy [51]. The experience of the expert is obviously an important factor in deciding about rarity or typicality. The auditory features discussed previously are used in this approach.

The auditory approach is fully subjective, unless it is combined with other approaches. Although the LR can be used to express the outcome of the analysis, practitioners of the auditory approach generally do not use it. Instead, based on their comparison of auditory features, they present an evidentiary statement (a formal statement describing the basis of the evidence) in court.

AUDITORY-SPECTROGRAPHIC APPROACH

As discussed previously, the spectrographic approach, previously known as voiceprint analysis, is based on visual comparison of speech spectrograms. Generally, the same word or phrase is extracted from the known and questioned voices and their spectrograms are visually analyzed. Additional foil speakers' (background speakers) spectrograms are also included to facilitate in understanding similarity versus typicality. It is believed that visual comparison using spectrograms together with listening to the audio reinforces the voice identification procedure [44], [45], which is why the approach is termed *auditory-spectrographic*.

Following the controversy on voiceprints, the spectrographic method evolved in various ways. It was not evident if forensic experts could differentiate between intraspeaker (changes of speech from the same speaker) and interspeaker (changes in speech due to different speakers) variation by a general visual comparison of spectrographs. Thus, different protocols evolved that require the forensic examiner to analyze predefined aspects of the spectrographs. According to the American Board of Recorded Evidence (ABRE) protocols, the examiner is required to visually analyze and compare

aspects such as general formant shaping and positioning, pitch striations, energy distribution, word length, and coupling (nasality). It also requires auditory comparisons of pitch, stress/emphasis, speaking rate, disguise, mode, etc. [51], [52].

The auditory-spectrographic, similar to the auditory approach, is also subjective and depends heavily on the experience of the examiner. Courts in some jurisdictions do not accept testimony based on this approach. The FBI seeks advice from auditory-spectrographic experts during investigations but does not allow them to testify in court [51].

ACOUSTIC–PHONETIC APPROACH

This approach, which is commonly taken by experts trained on acoustic–phonetics, requires quantitative acoustic measurements from speech samples, and statistical analysis of the results. Acoustic features discussed previously are ones that are considered. Generally, similar phonetic units are extracted from the known and questioned speech samples, and various acoustic parameters measured from these segments are compared. The LR can be conveniently used in this approach since it is based on numerical parameters [51].

Although the acoustic–phonetic approach is a more objective approach, it does have some subjective elements. For example, an acoustic–phonetician may identify speech sounds as being affected by stress (through listening) and then perform objective analysis. However, whether the speaker was actually under stress at that moment is a subjective quantity determined by the examiner through his or her experience. It is a matter of debate if having a human element in the forensic speaker-recognition process is advantageous [51].

Forensic speaker identification will continue to be an important research area in the coming future. As evident from the discussion, the methods are evolving toward mathematical and statistical approaches, perhaps signaling that the human element in this process may actually be a source of error. The NIST has conducted studies on human-assisted speaker recognition (HASR) comparing human experts and state-of-the-art algorithms [20]. In these experiments, a set of difficult speaker pairs (i.e., same speakers that sound different in two recordings or different speakers that sound similar) were selected. The results indicated that the state-of-the-art fully automatic systems outperformed the human-assisted systems. We discuss these studies further in the "Man Versus Machine in Speaker Recognition" section.

NAÏVE SPEAKER RECOGNITION

The ability to recognize people by their voices is an acquired human trait. Research shows that we are able to recognize our mothers' voice from as early as the fetus stage [53], [54]. We analyze many different aspects of a person's voice to identify him or her, including spectral characteristics, language, prosody, and speaking style. We learn and remember these traits constantly without even putting in a conscious effort. In this section, we discuss various aspects of how a naïve listener identifies a speaker and what is currently known about the speaker-recognition process in the human brain.

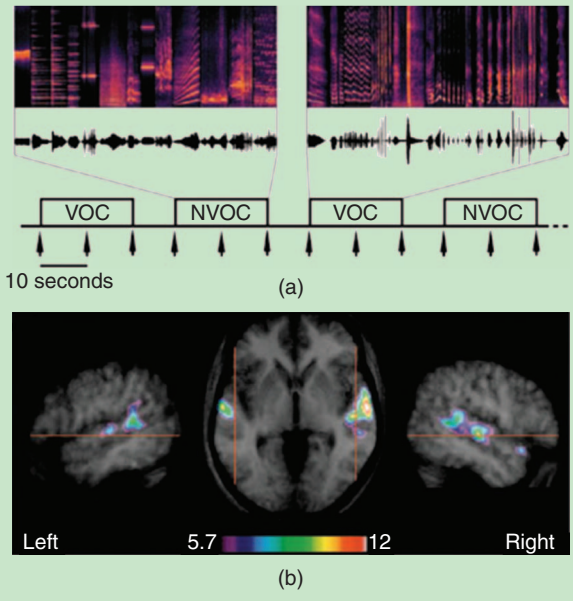
IDENTIFY SPEECH SEGMENTS

An important aspect of detecting speakers from audio samples is to first identify speech segments. Humans can efficiently distinguish between speech and nonspeech sounds from a very early age [55]. This is observed from highly voice-selective cerebral activity measured by functional magnetic resonance imaging (fMRI) in the adult human brain [4], [55], [56]. Figure 2 shows the brain regions that demonstrate higher neural activity with vocal and nonvocal stimuli. Note that in this experiment, any sound produced by a human is considered vocal (irrespective of being voiced or unvoiced), including laughs and coughs. In later sections, we discuss a very similar process required by automatic systems as a preprocessing step before performing speaker recognition.

SPEAKER RECOGNITION VERSUS DISCRIMINATION

It is obvious that we need to be familiar with a person's voice before identifying him or her. Familiarity is a subjective condition, but it is apparent that being familiar with a person depends on how much time the subject has spent in listening to that person. In other words, familiarity with a speaker depends on the amount of speech data observed by the listener. The familiar person can be a close acquaintance (e.g., a friend or relative) or someone famous (e.g., a celebrity or political leader).

Interestingly, familiar voice recognition and unfamiliar voice discrimination are known to be separate cognitive abilities [57].



[FIG2] An experiment on finding voice-selective regions of the human brain using fMRI. (a) The experimental paradigm: spectrograms (0–4 kHz) and amplitude waveforms of examples of auditory stimuli. Vocal (VOC) and nonvocal (NVOC) stimuli are presented in 20-second blocks with 10-second silence intervals. (b) Voice-sensitive activation regions in the group average: regions with significantly ($P < 0.001$) higher response to human voices than to energy-matched nonvocal stimuli are shown in color scale (t-value) on an axial slice of the group-average MRI (center) and on sagittal slices (vertical plane dividing the brain into left and right halves) of each hemisphere. (Figure adapted from [4].)

Familiar voice recognition is essentially a pattern-recognition task—humans can perform this task even if the speech signal is reversed [58]. These findings suggest that unfamiliar voice discrimination involves analysis of speech features as well as the pattern-recognition ability of the brain [57]. Forensic examiners heavily depend on the ability to discriminate since they are not usually familiar with the speakers in the speech samples involved.

The findings in [57] also imply that voice discrimination ability of the human brain is not a preprocessing step of voice recognition, since these two processes are found to be independent. For automatic systems, however, this is not usually true. The same algorithms can be used (usually with slight modification) to discriminate between speakers or identify a specific speaker. In many cases, discriminative training methods are used to learn speaker models, which can later be used to identify speakers. We discuss automatic systems further in the "Automatic Speaker Recognition" section.

FAMILIARITY WITH LANGUAGE

It is observed in [59] that humans are better at recognizing people who are familiar and speak a known language. Experiments reported in this study show that native English speakers with normal reading ability could identify voices speaking English significantly more accurately than voices speaking Chinese. Thus, the voice-recognition ability of humans depends on their familiarity with the phonology of the particular language. Humans can still recognize people speaking an unknown language, but with much lower accuracy [59].

ABSTRACT REPRESENTATIONS OF SPEECH

The human brain forms efficient abstract representations from relevant audio features that contain both phonetic and speaker identity information. These representations aid in efficient processing and high robustness due to noise and other forms of degradations. These aspects of the brain were studied in [5], where the authors have shown that it is possible to decipher both speech content and speaker identity by observing neural activity of the human listener. The brain activities were measured by fMRI and it was found that there are certain observable patterns corresponding to speech and voice stimuli elicit in the listener's auditory cortex. This is illustrated in Figure 3, where vowel (red) and speaker (blue) discriminative regions in the brain are shown.

SPEAKER RECOGNITION IN THE BRAIN: FINAL REMARKS

There is still much more to discover about the human brain and how it processes information. From what we already know, the human brain performs complex spectral and temporal audio processing [60], is sensitive to vocal stimuli [4], shows familiarity to the phonology of languages [59], and builds abstract representations of speech and speaker information that are robust to noise and other degradations [5]. Most of these abilities are highly desirable in automatic systems, especially the brain's ability to process noisy speech. It is thus natural to attempt to mimic the human brain in solving these problems. Research efforts are already underway to reverse engineer the processes performed by the human auditory pathway [38].

As discussed previously, the human brain processes familiar speakers differently than unfamiliar ones [55], [57]. This may mean that faithfully comparing human and machine performance in a speaker-recognition task can be very difficult since it is not well understood how to quantify familiarity with a person from an automatic system's perspective—what amount of data is enough for the system to be familiar with that person? Nevertheless, it will be interesting to be able to determine exactly how the human brain stores the speaker identity information of familiar speakers. These findings may lead to breakthrough algorithmic advances in the automatic speaker-recognition area.

As we conclude this section, we want to highlight the strengths and weaknesses of humans in the speaker-recognition task. Here, humans include both forensic examiners and naïve listeners.

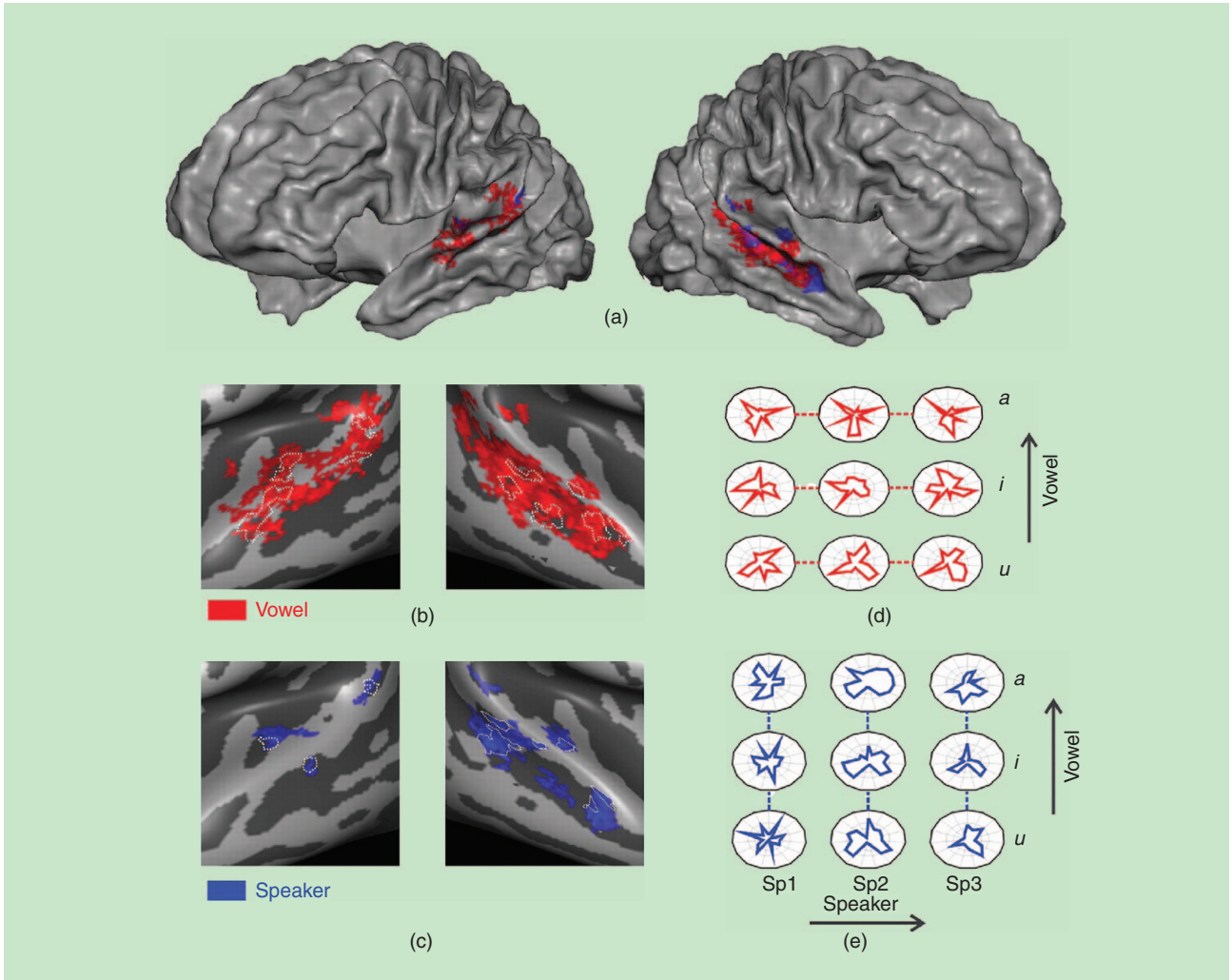
STRENGTHS OF HUMAN LISTENERS

- Humans (naïve listeners and experts alike) can identify familiar speakers with remarkable accuracy, even in challenging conditions (normal, disguised, and stressed) [61].
- Humans are good at finding the idiosyncrasies of a speaker's voice. Thus, the forensic examiner may easily identify where to look. For example, a speaker may cough in a specific manner, which a human will notice very quickly.

WEAKNESSES OF HUMAN LISTENERS

- Humans are susceptible to contextual bias [62]. For example, if the forensic examiner knows that a suspect already confessed to a crime, he is more likely to find a match between the exemplar and evidence recording.
- Humans are prone to error. The reliability of voiceprints was questioned mostly due to human errors involved in the process [46].
- Humans cannot remember a speaker's voice for a long time [63]. Memory retention ability depends on the duration of speech heard by the listener [64].
- For familiar speakers, the listener may confuse them with someone else. The subject may know that the voice is familiar but may not correctly identify exactly who the speaker is.
- Naïve listeners cannot distinguish subtle differences between voices. However, trained experts can. For example, the difference between New York and Boston accents is distinguishable by an expert but probably not by naïve listeners [35].
- Humans perform better while they are attentive. However, the attention level drops with time, and listeners tend to become fatigued after a certain time.
- The outcome of voice comparison results as LRs may not be consistent across multiple experts (or the same expert at different times).
- Human listeners (including forensic experts) may seem to identify someone from a voice recording if they are expecting to hear that person.

Concluding the discussion on speaker recognition by humans, we now move forward with the main focus of this review, which is automatic systems for speaker recognition.

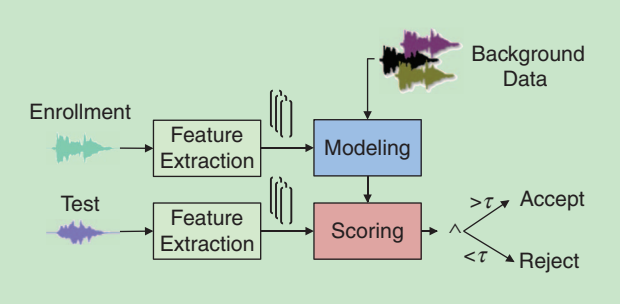


[FIG3] (a)–(c) The regions of the human brain that contribute the most in discriminating between vowels (red) and speakers (blue). (b) and (c) Enlarged representations of the auditory cortex (region of the brain sensitive to sounds). (d) and (e) Activation patterns of sounds created from the 15 most discriminative voxels (of the fMRI) for decoding (d) vowels and (e) speakers. Each axis of the polar plot forming a pattern displays the normalized activation level in a voxel. Note the similarity among the patterns of the same vowel [horizontal direction in (d)] or speaker [vertical direction in (e)]. (Figure reprinted from [5].)

AUTOMATIC SPEAKER RECOGNITION

In automatic speaker recognition, computer programs designed to operate independently with minimum human intervention identify a speaker's voice. The system user may adjust the design parameters, but to make the comparison between speech segments, all the user needs to do is provide the system with the audio recordings. In the current discussion, we focus our attention on the text-independent scenario and the speaker-verification task. Naturally, the challenges mentioned previously affect the automatic systems in the same way as they do the human listeners or forensic experts. Various speaker-verification approaches can be found in the literature that address specific challenges; see [65]–[74] for a comprehensive tutorial review on automatic speaker recognition. The research community is largely driven by standardized tasks set forth by NIST through the speaker-recognition evaluation (SRE) campaigns [75]–[78]. We discuss the NIST SRE tasks in more detail in later sections.

A simple block diagram representation of an automatic speaker-verification system is shown in Figure 4. Predefined feature parameters are first extracted from the audio recordings that are designed to capture the idiosyncratic characteristics of a



[FIG4] An overall block diagram of a basic speaker-verification system.

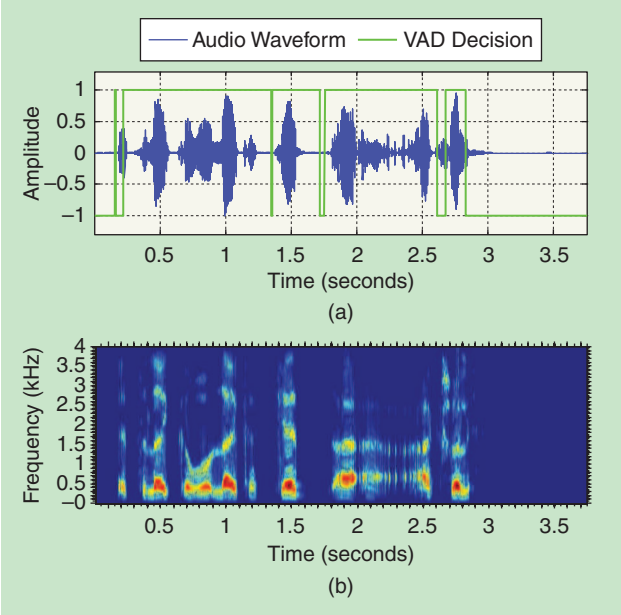
person's speech in mathematical parameters. These features obtained from an enrollment speaker are used to build/train mathematical models that summarize their speaker-dependent properties. For an unknown test segment, the same features are then extracted, and they are compared against the model of the enrollment/claimed speaker. The models are designed so that such a comparison provides a score (a scalar value) indicating whether the two utterances are from the same speaker. If this score is higher (or lower) than a predefined threshold then the system accepts (or rejects) the test speaker.

It should be noted that the block diagram in Figure 4 for speaker verification is a simplified one. As we discuss more about the standard speaker-recognition systems of today, features can be extracted from short-term segments of speech, a relatively longer duration of speech, or the entire utterance. The classification of features discussed previously also applies in this case.

In some automatic systems, the feature-extraction processes may be dependent on other speech utterances spoken by a diverse speaker population, as well as the enrollment speaker [79]. In short, the recent techniques make use of the general properties of human speech by observing many different speech recordings to make effective speaker-verification decisions. This is also intuitive, since we also learn how human speech varies across conditions over time. For example, if we only heard one language in our entire life, we would have difficulty distinguishing people speaking a different language [59].

FEATURE PARAMETERS IN AUTOMATIC SPEAKER-RECOGNITION SYSTEMS

As mentioned previously, feature parameters extracted from an entire utterance are referred to as *utterance features* in this article. This becomes more important in the automatic speaker-recognition context as many common pattern-recognition algorithms operate on fixed dimension vectors. Because of the variable length/duration property of speech, acoustic/segmental features cannot be directly used with such classifiers. However, simple methods such as averaging segmental features over time do not seem to be highly effective in this case, due to the time-varying nature and context dependency of speech [80], [81]. For example, taking speaking rate as a feature, it is obvious that two people may commonly have the same speaking rate, so this feature by itself may not be very useful. Researchers noted early on that a specific speaker's idiosyncratic features will be time varying and context/speech sound dependent [34], [66]. However, the high-level and long-term features such as dialect, accent, speaking style/rate, and prosody are also useful and can be beneficial when used together with low-level acoustic features [39], [82].



[FIG5] (a) A speech waveform with voice-activity decisions (1 versus 0 values indicate speech versus silence) and (b) a spectrogram plot of the corresponding speech waveform.

VAD

As noted previously, humans are good at distinguishing between speech and nonspeech sounds, which is also an essential part in auditory forensic speaker recognition. Clearly, in automatic systems it is also desirable that features be extracted only from speech segments of the audio waveform, which necessitates VAD [83], [84]. Detecting speech segments becomes critical when highly noisy/degraded acoustic conditions are considered. The function of VAD is illustrated in Figure 5(a), where speech presence/absence is indicated by a binary signal overlaid on the speech samples. The corresponding speech spectrogram is shown in Figure 5(b). The VAD algorithm used in this plot is presented in [83], though more advanced unsupervised solutions such as Combo-Speech Activity Detection (SAD) have recently emerged as successful in diverse audio conditions for speaker recognition [85].

SHORT-TERM FEATURES

These features refer to parameters extracted from short speech segments/frames of duration within 20–25 milliseconds. The most popular short-term acoustic features are the Mel-frequency cepstral coefficients (MFCCs) [86] and linear predictive coding (LPC)-based [87] features. For a review on different short-term acoustic features for speaker recognition, see [71] and [73]. We briefly discuss the MFCC features here. To obtain these coefficients from an audio recording, first the audio samples are divided into short overlapping segments. A typical 25-millisecond speech signal frame is shown in Figure 6(a). The signal obtained in these segments/frames is then multiplied by a window function (e.g., Hamming and Hanning), and the Fourier power spectrum is obtained. In the next step, the logarithm of the spectrum is computed and nonlinearly spaced Mel-space filter-bank analysis is performed. The logarithm operation expands the scale of the coefficients and also decomposes multiplicative components to additive [88]. The filter-bank analysis produces the spectrum energy in each channel (also known as the *filter-bank energy coefficients*), representing different frequency bands.

A typical 24-channel filter bank and its outputs are shown in Figure 6(c) and (d), respectively. As evident here, the filter bank is designed so that it is more sensitive to frequency variations in the lower end of the spectrum, similar to the human auditory system

[86]. Finally, MFCCs are obtained by performing discrete cosine transform (DCT) on the filter-bank energy parameters and retaining a number of leading coefficients. DCT has two important properties: 1) it compresses the energy of a signal to a few coefficients and 2) its coefficients are highly decorrelated. For these reasons, removing some dimensions using DCT improves modeling efficiency and reduces some nuisance components. Also, the decorrelation property of DCT helps the models that assume feature coefficients are uncorrelated. In summary, the following sequence of operations—power spectrum, logarithm, and DCT—produces the well-known cepstral representation of a signal [88]. Figure 6(e) shows the static MFCC parameters, retaining the first 12 coefficients after DCT. Generally, velocity and acceleration parameters computed across multiple frames of speech are appended to the MFCCs. These parameters (known as *deltas* and *double deltas*, respectively) represent the dynamic properties of the short-term feature coefficients.

FEATURE NORMALIZATION

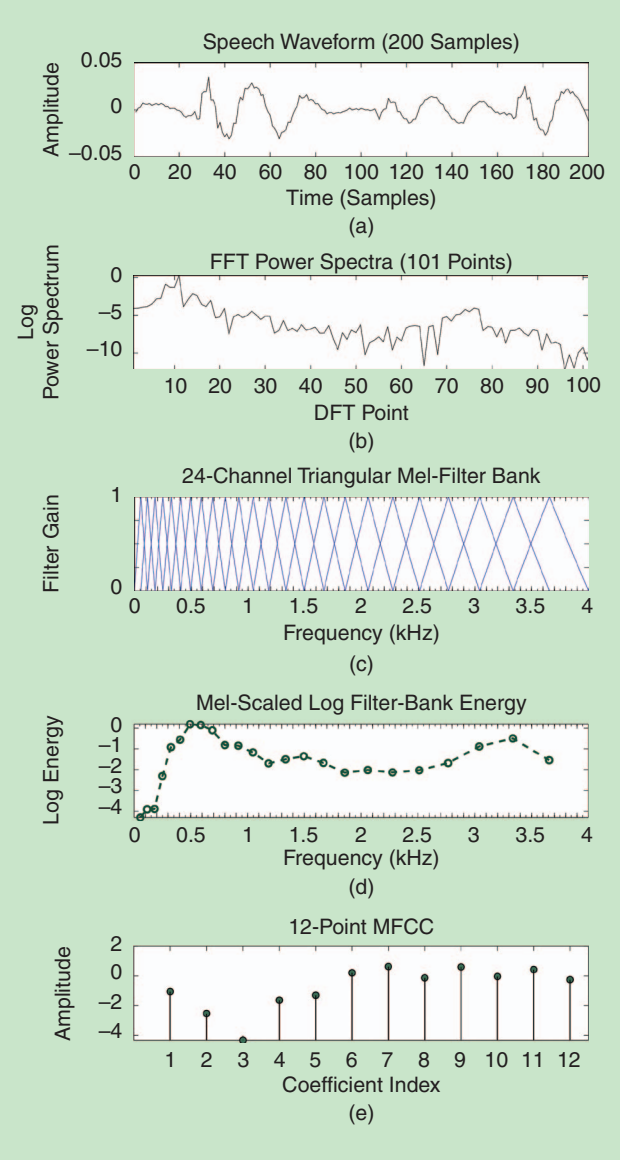
As stated previously, one of the desirable properties of acoustic features (and any feature parameter in a pattern-recognition problem) is robustness to degradation. This is one of the desirable characteristics of an ideal feature parameter [34]. In reality, it is not possible to design a feature parameter that will be absolutely unchanged in modified acoustic conditions and also provide meaningful speaker-dependent information. However, these changes can be minimized in various ways using feature-normalization techniques such as cepstral mean subtraction [89], feature warping [90], relative spectra (RASTA) processing [91], and quantile-based cepstral normalization [92]. It should be noted that normalization techniques are not designed to enhance the discriminative ability of the features (ideal property 3), rather they aim to modify the features so that they are more consistent among different speech utterances (ideal property 5). Popular normalization schemes include feature warping and cepstral mean and variance normalization.

SPEAKER MODELING

Once the audio segments are converted to feature parameters, the next task of the speaker-recognition process is modeling. In general terms, we can define modeling as a process of describing the feature properties for a given speaker. The model must also provide means of its comparison with an unknown utterance. A modeling method is robust when its characterizing process of the features is not significantly affected by unwanted distortions, even though the features are. Ideally, if features could be designed in such a way that no intraspeaker variation is present while interspeaker discrimination is maximum, the simplest methods of modeling might have sufficed. In essence, the nonideal properties of the feature extraction stage requires various compensation techniques during the modeling phase so that the effect of the nuisance variations observed in the signal are minimized during the speaker-verification process.

Most speaker-modeling techniques make various mathematical assumptions on the features (Gaussian distributed, for example). If these properties are not met by the data, we are essentially introducing imperfections during the modeling phase as well. The normalization of features can alleviate these problems to some extent, but not entirely. Consequently, mathematical models are forced to fit the features and recognition scores are derived based on these models and test data. Thus, this process introduces artifacts in the detection scores, and a family of score-normalization techniques has been proposed in the past to encounter this final-stage mismatch [17].

In summary, degradations in the acoustic signal affect features, models, and scores. Thus, improving robustness of speaker-recognition systems is important in these three domains. Recently, it has been observed that as speaker-modeling techniques are improved, score-normalization techniques become less



[FIG6] Steps in MFCC feature extraction from a speech frame: (a) 200-sample frame representing 25 milliseconds of speech sampled at a rate of 8 kHz, (b) DFT power spectrum showing first 101 points, (c) 24-channel triangular Mel-filter bank, (d) log filter-bank energy outputs from Mel-filter, and (e) 12 static MFCCs obtained by performing DCT on filter-bank energy coefficients and retaining the first 12 values.

effective [93], [94]. Similarly, we can argue that if acoustic features are improved, simple modeling techniques will be sufficient. However, from the speaker-recognition research trend in the last decade, it seems that improving feature robustness beyond a certain level (for a variety of degradations) is extremely difficult—or, in other words, data-driven modeling techniques have been more successful in improving robustness compared to new features. This is especially true if large data sets are used in training strong discriminative models. In the recent approaches for speech recognition, simple filter-bank energy features are found to be more effective than MFCCs when large neural networks are used for modeling [95]. Also, modeling techniques that aim at learning the behavior of the degradations from example speech utterances are at an advantage in improving robustness. For example, an automatic system that has observed several examples of speech recordings of different speakers in roadside noise will be better at distinguishing speakers in that environment.

In the following sections, we discuss how state-of-the-art systems have evolved during the last decade. We emphasize a few key advancements made during this time.

GAUSSIAN-MIXTURE-MODEL-BASED METHOD

A Gaussian mixture model (GMM) is a combination of Gaussian probability density functions (PDFs) generally used to model multivariate data. The GMM clusters the data in an unsupervised way (i.e., without any labeled data), but it provides a PDF of the data. Using GMMs to model a speaker's features results in a speaker-dependent PDF. Evaluating the PDF at different data points (e.g., features obtained from a test utterance) provides a probability score that can be used to compute the similarity between a speaker GMM and an unknown speaker's data. For a simple speaker-identification task, a GMM, is first obtained for each speaker. During testing, the utterance is compared against each GMM, and the most likely speaker (i.e., the highest-scoring GMM) is selected.

In text-independent speaker-recognition tasks when there is no a priori knowledge about the speech content, using GMMs to model short-term features has been found to be most effective for acoustic modeling. This is expected since the average behavior of the short-term spectral features is more speaker dependent rather than being affected by the temporal characteristics. It was first used in a speaker-recognition method in [96]. Before GMMs were introduced, the vector quantization (VQ) method [81], [97], [98] was used for speaker recognition. This technique models the speaker using a set of prototype vectors instead of PDFs. GMMs have been shown to be better speaker models compared to VQ because of their probabilistic nature for allowing greater variability. Therefore, even when the test utterance has a different acoustic condition, GMMs, being a probabilistic model, can relate to the data better than the more restrictive VQ model (see "GMM-Based Speaker Recognition: Summary").

A GMM is a mixture of Gaussian PDF parameterized by a number of mean vectors, covariance matrices, and weights of the individual mixture components. The model is represented by a weighted sum of the individual PDFs. If a random vector $\mathbf{x}_n$ can be modeled by M Gaussian components with mean vectors $\boldsymbol{\mu}_g$, covariance matrices Σ_g, where $g = 1, 2 \ldots M$ indicate the component indices, the PDF of $\mathbf{x}_n$ is given by

$$f(\mathbf{x}_n \mid \lambda) = \sum_{g=1}^{M} \pi_g \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_g, \Sigma_g), \tag{1}$$

where π_g indicates the weight of the gth mixture component. We denote the GMM model as $\lambda = \{\pi_g, \mu_g, \Sigma_g \mid g = 1 \ldots M\}$. The likelihood of a feature vector given the GMM model can be evaluated using (1). Acoustic feature vectors are generally assumed to be independent. For a sequence of feature vectors $\mathcal{X} = \{\mathbf{x}_n \mid n \in 1 \ldots T\}$, the probability of observing these features given the GMM model is computed as

$$p(\mathcal{X} \mid \lambda) = \prod_{n=1}^{T} p(\mathbf{x}_n \mid \lambda).$$

Note that the order of the features is irrelevant in calculating the likelihood, which simplifies the computation for text-dependent speaker recognition. A GMM is usually trained using the expectation-maximization (EM) algorithm [99], which iteratively increases the likelihood of the data given the model.

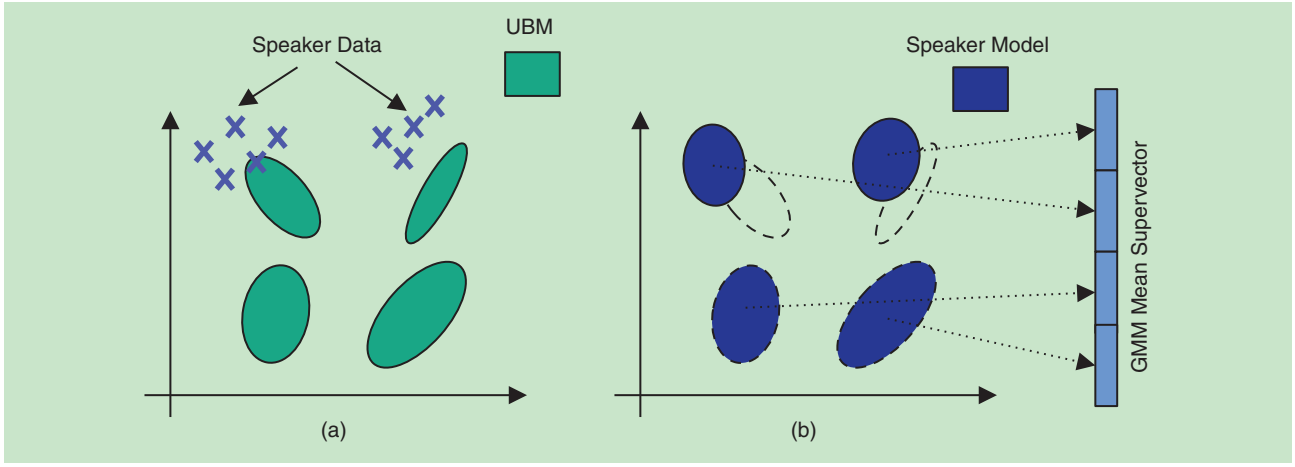
GMM-BASED SPEAKER RECOGNITION: SUMMARY

First proposed	Reynolds et al. (1995) [96]
Previous methods	Averaging of long-term features, VQ-based methods [80], [97], [98]
Proposed method	Model features using GMMs, compute similarity using feature likelihood
Why robust?	The probabilistic nature of GMM allows more variability in the data

ADAPTED GMMs: THE GMM–UBM SPEAKER-VERIFICATION SYSTEM

The GMM approach has been effective in speaker-identification tasks. For speaker verification, apart from the claimed speaker model, an alternate speaker model (representing speakers other than the target) is needed. In this way, these two models can be compared with the test data and the more likely model can be chosen, leading to an accept or reject decision. The alternate speaker model, also known as the *background* or *world model*, initiated the idea of using a UBM that represents everyone except the target speaker. It is essentially a large GMM trained to represent the speaker-independent distribution of the speech features for all speakers in general. The block diagram in Figure 4 becomes clear now since the background model is assumed to exist. Note that the UBM is assumed to be a "universal" model that serves as the alternate model for all enrolled speakers. Some methods have considered providing speaker-dependent unique background models [100], [101]. However, using a single background model has been the most effective and meaningful strategy.

The UBM was first introduced as an alternate speaker model in [102]. Later, in [6], the UBM was used as an initial model for the enrollment speaker GMMs. This concept was a significant



[FIG7] A schematic diagram of a GMM–UBM system using a four-mixture UBM. MAP adaptation procedure and supervector formation by concatenating the mean vectors are also illustrated. (a) A schematic diagram of a GMM–UBM system using a four-mixture UBM. (b)MAP adaptation procedure and supervector formation by concatenating the mean vectors are also illustrated.

advancement achieved by the so-called GMM–UBM method. In this approach, a speaker's GMM is adapted or derived from the UBM using Bayesian adaptation [103]. In contrast to performing maximum likelihood training of the GMM for an enrollment speaker, this model is obtained by updating the well-trained UBM parameters. This relation between the speaker model and the background model provides better performance than independently trained GMMs and also lays the foundation for the speaker model adaptation techniques that were developed later. We will return to these relations as we proceed. In the following subsections, we describe the formulations of this approach.

The LR Test

Given an observation O and a hypothesized speaker s, the task of speaker verification can be stated as a hypothesis test between

$$H_0 : O \text{ is from speaker } s,$$
$$H_1 : O \text{ is not from speaker } s.$$

In the GMM–UBM approach, the hypothesis H_0 and H_1 are represented by a speaker-dependent GMM λ_s and the UBM λ_0. Thus, for the set of observed feature vectors $\mathrm{X} = \{\mathbf{x}_n \,|\, n \in 1 \ldots T\}$, the LR test is performed by evaluating the following ratio:

$$\frac{p(\mathcal{X} \,|\, \lambda_s)}{p(\mathcal{X} \,|\, \lambda_0)} \begin{cases} \geq \tau \text{ accept } H_0 \\ < \tau \text{ reject } H_0 \end{cases},$$

where τ is the decision threshold. Usually, the LR test is performed in the logarithmic scale, providing the so-called log-LR

$$\Lambda(\mathcal{X}) = \log p(\mathcal{X} \,|\, \lambda_s) - \log p(\mathcal{X} \,|\, \lambda_0). \tag{2}$$

Maximum A Posteriori Adaptation of UBM

Let $\mathcal{X} = \{\mathbf{x}_n \,|\, n \in 1 \ldots T\}$ denote the set of acoustic feature vectors obtained from the enrollment speaker s. Given a UBM as in (1) and the enrollment speaker's data X, at first the probabilistic alignment of the feature vectors with respect the UBM components is calculated as

$$p(g \,|\, \mathbf{x}_n, \lambda_0) = \frac{\pi_g p(\mathbf{x}_n \,|\, g, \lambda_0)}{\sum_{g=1}^{M} \pi_g \, p(\mathbf{x}_n \,|\, g, \lambda_0)} = \gamma_n(g).$$

Next, the values of $\gamma_n(g)$ values are used to calculate the sufficient statistics for the weight, mean, and covariance parameter as

$$N_s(g) = \sum_{n=1}^{T} \gamma_n(g),$$
$$\mathrm{F}_s(g) = \sum_{n=1}^{T} \gamma_n(g) \mathbf{x}_n,$$
$$S_s(g) = \sum_{n=1}^{T} \gamma_n(g) \mathbf{x}_n \mathbf{x}_n^T.$$

These quantities are known as the zero-, first-, and second-order Baum–Welch statistics, respectively. Using these parameters, the posterior mean and covariance matrix of the features given the data vectors $\mathcal{X}$ can be found as

$$E_g[\mathbf{x}_n \,|\, \mathcal{X}] = \frac{\mathrm{F}_s(g)}{N_s(g)},$$
$$E_g[\mathbf{x}_n \, \mathbf{x}_n^T \,|\, \mathcal{X}] = \frac{\mathrm{S}_s(g)}{N_s(g)}.$$

The maximum a posteriori (MAP) adaptation update equations for weight, mean, and covariance, (3), (4), and (5), respectively, are proposed in [103] and used in [6] for speaker verification

$$\hat{\pi}_g = [\alpha_g N_s(g)/T + (1 - \alpha_g)\pi_g]\beta, \tag{3}$$
$$\hat{\mu}_g = \alpha_g E_g[\mathbf{x}_n \,|\, \mathcal{X}] + (1 - \alpha_g)\, \boldsymbol{\mu}_g, \tag{4}$$
$$\hat{\Sigma}_g = \alpha_g E_g[\mathbf{x}_n \, \mathbf{x}_n^T \,|\, \mathcal{X}] + (1 - \alpha_g)(\Sigma_g + \mu_g \mu_g^T) - \hat{\mu}_g \hat{\mu}_g^T. \tag{5}$$

The scaling factor β in (3) is computed from all the adapted mixture weights to ensure that they sum to unity. Thus, the new GMM parameters are a weighted summation of the UBM parameters and the sufficient statistics obtained from the observed data (see "GMM–UBM System: Summary"). The variable α_g is defined as

$$\alpha_g = \frac{N_s(g)}{N_s(g) + r}. \tag{6}$$

GMM–UBM SYSTEM: SUMMARY

First proposed	Reynolds et al. (2000) [6]
Previous methods	GMM models for enrollment, cohort speakers as background
Proposed method	Adapt speaker GMMs from a UBM
Why robust?	Speaker models adapted from a well-trained UBM is more reliable than directly trained GMMs for each speaker

Here, r is known as the relevance factor. This parameter controls how the adapted GMM parameter will be affected by the observed speaker data. In the original study [6], this parameter was defined differently for the model weight, mean, and covariance. However, since only adaptation of the mean vectors turned out to be the most effective, we only use one relevance factor in our discussion here. Figure 7 shows an example of MAP adaptation for a two-dimensional feature space with a four-mixture UBM case.

THE GMM SUPERVECTORS

One of the issues with speaker recognition is that the training and test speech data can be of different durations. This requires the comparison of two utterances of different lengths. Thus, one of the efforts toward effective speaker recognition has always been to obtain a fixed-dimensional representation of a single utterance [80]. This is extremely useful since many different classifiers can be used on these utterance-level features from the machine-learning literature. One effective solution to obtaining a fixed-dimensional vector from a variable-duration utterance is the formation of a GMM supervector, which is essentially a large vector obtained by concatenating the parameters of a GMM model. Generally, a GMM supervector is obtained by concatenating the GMM mean vectors of a MAP-adapted speaker model, as illustrated in Figure 7.

The term *supervector* was first used in this context for eigenvoice speaker adaptation in speech recognition applications [104]. For speaker recognition, supervectors were first introduced in [105], motivating new model adaptation strategies involving eigenvoice and MAP adaptation. Researchers realized that these large dimensional vectors are a very good platform for designing channel compensation methods. Various effective modeling techniques were proposed to operate on the supervector space. The two dominating trends observed in these efforts were based on factor analysis (FA) and support vector machines (SVMs). They will be discussed next.

GMM SUPERVECTOR SVMs

SVMs [106] are one of the most popular supervised binary classifiers in machine learning. In [107], it was observed that GMM supervectors could be effectively used for speaker recognition/verification using SVMs. The supervectors obtained from the training utterances were used as positive examples while a set of impostor utterances were used as negative examples. Channel compensation strategies were also developed in this domain, such as nuisance attribute projection (NAP) [108] and within-class covariance normalization (WCCN) [109]. Other approaches used SVM models for speaker recognition using short- and long-term features [39], [110]. However, using GMM supervectors with SVM and NAP provided the most effective solution (see "GMM-SVM System: Summary").

GMM-SVM SYSTEM: SUMMARY

First proposed	Campbell et al. (2006) [107]
Previous methods	Adapted GMM-based methods, GMM–UBM system
Proposed method	Use GMM supervector as utterance features, classify using SVMs
Why robust?	Combines the effectiveness of adapted GMM as an utterance model and the discriminating ability of the SVM

SVMs

An SVM classifier aims at optimally separating multidimensional data points obtained from two classes using a hyperplane (a high-dimensional plane). The model can then be used to predict the class of an unknown observation depending on its location with respect to the hyperplane. Given a set of training vectors and labels $(\mathbf{x}_n, y_n)$ for $n \in \{1 \ldots T\}$, where $\mathbf{x}_n \in \Re^d$ and $y_n \in \{-1, +1\}$, the goal of SVM is to learn the function $f: \Re^{\mathrm{d}} \rightarrow \Re$ so that the class label of an unknown vector $\mathbf{x}$ can be predicted as

$$I(\mathbf{x}) = \operatorname{sign}(f(\mathbf{x})).$$

For a linearly separable data set [106], a hyperplane H given by $\mathbf{w}^T\mathbf{x} + b = 0$, can be obtained that separates the two classes, so that

$$y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1, n = 1 \ldots T.$$

An optimal linear separator H provides the maximum margin between the classes, i.e., the distance between H and the projections of the training data from the two different classes are maximum. The maximum margin is found to be $2/\|\mathbf{w}\|$ and data points $\mathbf{x}_n$ for which $y_n(\mathbf{w}^T\mathbf{x}_n + b) = 1$, (i.e., points that lie on the margins, are known as *support vectors*). In a simple two-dimensional case, the operation of SVM is illustrated in Figure 8. When training data are not linearly separable, the features can be mapped into a higher-dimensional space using Kernel functions where the classes become linearly separable. For more details on SVM training and kernels, refer to [106] and [111]. Compensation strategies that are developed for SVM-based speaker recognition (e.g., NAP and WCCN) are discussed in later sections.

FA OF THE GMM SUPERVECTORS

FA aims at describing the variability in high-dimensional observable data vectors using a lower number of unobservable/hidden

variables. For speaker recognition, the idea of explaining the speaker- and channel-dependent variability using FA in the GMM supervector space was first discussed in [112]. Many variants of FA methods were employed since then, which finally led to the current state-of-the-art i-vector approach [79]. In this section, we discuss these methods briefly to illustrate how the techniques have evolved.

[TABLE 1] A SUMMARY OF THE LINEAR STATISTICAL MODELS IN SPEAKER RECOGNITION.

MODEL	FORMULATION	REMARKS
CLASSICAL MAP	$\mathbf{m}_s = \mathbf{m}_0 + \mathbf{D}\mathbf{z}_s$	$\mathbf{D}$ IS DIAGONAL, $\mathbf{z}_s \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$
EIGENVOICE	$\mathbf{m}_s = \mathbf{m}_0 + \mathbf{V}\mathbf{y}_s$	$\mathbf{V}$ IS LOW RANK, $\mathbf{y}_s \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$
EIGENCHANNEL	$\mathbf{m}_{s,h} = \mathbf{m}_0 + \mathbf{D}\mathbf{z}_s + \mathbf{U}\mathbf{x}_h$	$\mathbf{U}$ IS LOW RANK, $(\mathbf{z}_s, \mathbf{x}_h) \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$
JFA	$\mathbf{m}_{s,h} = \mathbf{m}_0 + \mathbf{U}\mathbf{x}_h + \mathbf{V}\mathbf{y}_s + \mathbf{D}\mathbf{z}_{s,h}$	$\mathbf{U}, \mathbf{V}$ ARE LOW RANK, $(\mathbf{x}_h, \mathbf{y}_s, \mathbf{z}_{s,h}) \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$
i-VECTOR	$\mathbf{m}_{s,h} = \mathbf{m}_0 + \mathbf{T}\mathbf{w}_{s,h}$	$\mathbf{T}$ IS LOW RANK, $\mathbf{w}_{s,h} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$

Linear Distortion Model

In the discussions to follow, a speaker-dependent GMM supervector $\mathbf{m}_s$ is generally assumed to be a linear combination of four components. These components are as follows:

1) speaker-/channel-/environment-independent component $(\mathbf{m}_0)$
2) speaker-dependent component $(\mathbf{m}_{\mathrm{spk}})$
3) channel-/environment-dependent component $(\mathbf{m}_{\mathrm{chn}})$
4) residual $(\mathbf{m}_{\mathrm{res}})$.

Component 1 is usually obtained from the UBM and is a constant. Components 2–4 are random vectors and are responsible for variability in the supervectors due to different phenomena. Using this model, a GMM supervector obtained from speaker s and session h is written as

$$\mathbf{m}_{s,h} = \mathbf{m}_0 + \mathbf{m}_{\mathrm{spk}} + \mathbf{m}_{\mathrm{chn}} + \mathbf{m}_{\mathrm{res}}. \tag{7}$$

For acoustic features of dimension d and a UBM with M mixture components, these GMM supervectors are of dimension $(Md \times 1)$. As an example, the speaker- and channel-independent supervector $\mathbf{m}_0$ is the concatenation of the UBM mean vectors. We denote the subvectors of $\mathbf{m}_0$ for the gth mixture as $\mathbf{m}_{0[g]}$, which equals μ_g. In the following sections, we discuss how well-known linear Gaussian models, including FA, can be used to develop methods based on this generic decomposition of the GMM supervectors. A summary of the various linear statistical models in speaker recognition is included in Table 1, which highlights both formulation and specifics on matrix/model traits.

Classical MAP Adaptation

We revisit the MAP adaptation technique discussed previously in the GMM–UBM system. If we examine the adaptation equation (4), which is used to update the mean vectors, it is clear that this is a linear combination of two components: one is speaker dependent and the other is independent. In a more generalized way, MAP adaptation can be represented as an operation on the GMM mean supervector as:

$$\mathbf{m}_s = \mathbf{m}_0 + \mathbf{D}\mathbf{z}_s, \tag{8}$$

where $\mathbf{D}$ is $(Md \times Md)$ a diagonal matrix and $\mathbf{z}_s$ is a $Md \times 1$ standard normal random vector. We dropped the subscript due to session h for simplicity. According to the linear distortion model of (7), $\mathbf{m}_{\mathrm{spk}} = \mathbf{D}\mathbf{z}_s$. As discussed in [113], in the special case when we set

$$\mathbf{D}^2 = (1/r)\,\Sigma,$$

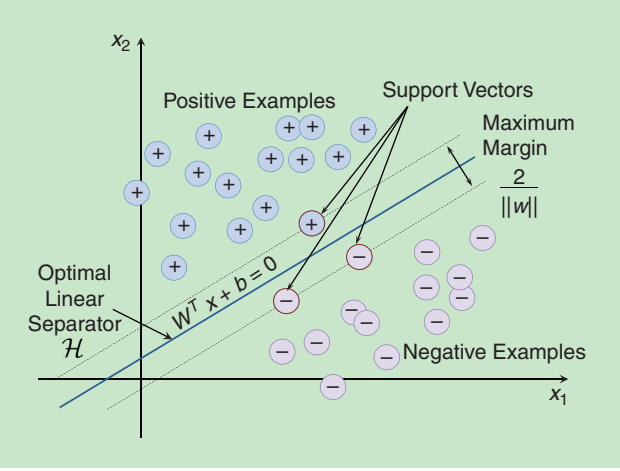
the MAP adaptation equations given in (4) [6] arises from (8), where r is the relevance factor in (6).

Eigenvoice Adaptation

Perhaps the first FA-related model used in speaker recognition was the eigenvoice method [105]. The eigenvoice method was initially proposed for speaker adaptation in speech recognition [114]. In essence, this method restricts the speaker model parameters to lie in a lower dimensional subspace, which is defined by the columns of the eigenvoice matrix. In this model, a speaker-dependent GMM mean supervector $\mathbf{m}_s$ is expressed as

$$\mathbf{m}_s = \mathbf{m}_0 + \mathbf{V}\mathbf{y}_s, \tag{9}$$

where $\mathbf{m}_0$ is the speaker-independent supervector obtained from the UBM, the columns of the matrix $\mathbf{V}$ spans the speaker subspace, and $\mathbf{y}_s$ are the standard normal hidden variables known as speaker factors. Here, we dropped the subscript h for simplicity. In accordance with the linear distortion model in (7), the speaker-dependent component is $\mathbf{m}_{\mathrm{spk}} = \mathbf{V}\mathbf{y}_s$. Note that this model does not have a residual noise term as in probabilistic PCA (PPCA) [115] or FA. This means that the eigenvoice model is essentially equivalent to PCA. The model covariance is $\mathbf{V}\mathbf{V}^T$. Since supervectors are usually of a large dimension, a full rank sample covariance matrix, i.e., the supercovariance matrix, is difficult to estimate with limited amount of data. Thus, EM algorithms [116], [117] are used to estimate the eigenvoices. The speaker factors need to be estimated for an enrollment speaker. Computation



[FIG8] A conceptual illustration of an SVM classifier: Positive (+) and negative (–) examples are correspondingly labeled, with the optimal linear separator and support vectors shown.

of the likelihood is carried out as provided in [16, eq. (19)], using the adapted supervector.

This model implies that the adaptation of the GMM supervector parameters is restricted by the eigenvoice matrix. The advantage with this model is that when a small amount of data is available for adaptation, the adapted model is more robust as it is restricted to live in the speaker-dependent subspace, being less affected by nuisance directions. However, the eigenvoice model does not model the channel or intraspeaker variability.

Eigenchannel Adaptation

Similar to adapting the UBM toward a speaker model, a speaker model can also be adapted to a channel model [105]. This can be useful when an unseen channel distortion is observed during testing, and the enrollment speaker model can be adapted to that channel. Similar to the eigenvoice model, the channel variability can also be assumed to lie in a subspace spanned by the principal eigenvectors of the channel covariance matrix. According to our distortion model (7), for a specific channel h, the term $\mathbf{m}_{\text{chn}} = \mathbf{U}\mathbf{x}_h$, where $\mathbf{U}$ is a low-rank matrix that spans the channel subspace, and $\mathbf{x}_h \in \mathrm{N}(0, \mathbf{I})$ are the channel factors. When eigenchannel adaptation is combined with classical MAP, we obtain the model for speaker- and session-dependent GMM supervector

$$\mathbf{m}_{s,h} = \mathbf{m}_0 + \mathbf{D}\mathbf{z}_s + \mathbf{U}\mathbf{x}_h. \tag{10}$$

More details on training the hyperparameters $\mathbf{D}$ and $\mathbf{U}$ can be found in [113]. Likelihood computation can be carried out in a similar way as the eigenvoice method.

Joint FA

The joint FA (JFA) model is formulated by combining both eigenvoice and eigenchannel together, which is accomplished by MAP adaptation for a single model (see "JFA: Summary"). This model assumes that both speaker and channel variability lie in lower dimensional subspaces of the GMM supervector space. These subspaces are spanned by the matrices V and U, as before. The model assumes, for a randomly chosen utterance obtained from speaker s and session h, that its GMM mean supervector can be represented by

$$\mathbf{m}_{s,h} = \mathbf{m}_0 + \mathbf{U}\mathbf{x}_h + \mathbf{V}\mathbf{y}_s + \mathbf{D}\mathbf{z}_{s,h}. \tag{11}$$

JFA: SUMMARY

First proposed	Kenny et al. (2004) [118]
Previous methods	MAP adapted GMM, GMM-SVM approach
Proposed method	Model speaker and channel variability in GMM supervectors
Why robust?	Exploits the behavior of speakers' features in variety of channel conditions learned using FA

Thus, this is the only model so far that considers all four components of the linear distortion model we discussed previously. Indeed, JFA was shown to outperform the other contemporary methods. More details on implementation of JFA can be found in [16] and [118].

The i-Vector Approach

As discussed previously, SVM classifiers on GMM supervectors have been a very successful approach for robust speaker recognition. FA based methods (especially the JFA technique) were also among state-of-the-art systems. In an attempt to combine the strengths of these two approaches, Dehak et al. [79], [119], [120] attempted to use JFA as a feature extractor for SVMs. In their initial attempt [119], the speaker factors estimated using JFA were used as features for SVM classifiers. Observing the fact that the channel factors also contain speaker-dependent information, the speaker and channel factors were combined into a single space termed the *total variability space* [79], [120]. In this FA model, a speaker- and session-dependent GMM supervector is represented by

$$\mathbf{m}_{s,h} = \mathbf{m}_0 + \mathbf{T}\mathbf{w}_{s,h}. \tag{12}$$

The hidden variables $\mathbf{w}_{s,h} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ in this case are called *total factors*. Similar to all of the FA methods above, the hidden variables are not observable but can be estimated by their posterior expectation. The estimates of the total factors, which can be used as features to the next stage of classifiers, came to be known as the *i-vectors*. The term *i-vector* is a short form of "identity vector," regarding the speaker-identification application, and also of "intermediate vectors," referring to its intermediate dimension between those of a supervector and an acoustic feature vector [79] (see "The i-Vector System: Summary").

THE i-VECTOR SYSTEM: SUMMARY

First proposed	Dehak et al. (2009) [79]
Previous methods	JFA and GMM–SVM-based approaches
Proposed method	Reduce supervector dimension using FA before classification
Why robust?	i-vectors effectively summarize utterances and allows using compensation methods that were not practical in large dimensional supervectors

Unlike JFA or other FA methods, the i-vector approach does not make a distinction between speaker and channel. It is simply a dimensionality reduction method of the GMM supervector. In essence, (12) is very similar to a PCA model on the GMM supervectors. The $\mathbf{T}$ matrix is trained using the same algorithms as for the eigenvoice model, except that each utterance is assumed to be obtained from a different speaker.

Mismatch Compensation In i-Vector Domain

The i-vector approach itself does not perform any compensation; on the contrary, it only provides a meaningful lower-dimensional ($400 \cong 800$) representation of a GMM supervector. Thus, it has

most of the advantages of the supervectors, but because of its lower dimension, many conventional compensation strategies can be applied to speaker recognition, which were previously not practical with the large-dimensional supervectors.

LINEAR DISCRIMINANT ANALYSIS

Linear discriminant analysis (LDA) is a commonly employed technique in statistical pattern recognition that aims at finding linear combinations of feature coefficients to facilitate discrimination of multiple classes. It finds orthogonal directions in the feature space that are more effective in discriminating the classes. Projecting the original features in these directions improve classification accuracy. Let D indicate the set of all development utterances, $\mathbf{w}_{s,i}$ indicates an utterance feature (e.g., supervector or i-vector) obtained from the ith utterance of speaker s, n_s denotes the total number of utterances belonging to speaker s, and S is the total number of speakers in D. The between-and within-class covariance matrices are given by

$$\mathbf{S}_b = \frac{1}{S}\sum_{s=1}^{S}(\bar{\mathbf{w}}_s - \bar{\mathbf{w}})(\bar{\mathbf{w}}_s - \bar{\mathbf{w}})^T \quad \text{and} \tag{13}$$

$$\mathbf{S}_w = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}(\mathbf{w}_{s,i} - \bar{\mathbf{w}}_s)(\mathbf{w}_{s,i} - \bar{\mathbf{w}}_s)^T, \tag{14}$$

where the speaker-dependent and speaker-independent mean vectors are given by

$$\bar{\mathbf{w}}_s = \frac{1}{n_s}\sum_{i=1}^{n_s}\mathbf{w}_{s,i} \quad \text{and}$$

$$\bar{\mathbf{w}} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}\mathbf{w}_{s,i},$$

respectively. The LDA optimization thus aims at maximizing the between class variance while minimizing the within-class variance (due to channel variability). The projections obtained from this optimization are found by the solution of the following generalized eigenvalue problem:

$$\mathbf{S}_b\mathbf{v} = \Lambda\mathbf{S}_w\mathbf{v}. \tag{15}$$

Here, Λ is the diagonal matrix containing the eigenvalues. If the matrix $\mathbf{S}_w$ is invertible, this solution can be found by finding the eigenvalues of the matrix $\mathbf{S}_w^{-1}\mathbf{S}_b$. Generally, the first $k < R$ eigenvalues are used to prepare a matrix $\mathbf{A}_{\text{LDA}}$ of dimension $R \times k$ given by

$$\mathbf{A}_{\text{LDA}} = [\mathbf{v}_1 \ldots \mathbf{v}_k],$$

where $\mathbf{v}_1 \ldots \mathbf{v}_k$ denote the first k eigenvectors obtained by solving (15). The LDA transformation of the utterance feature $\mathbf{w}$ is thus obtained by

$$\Phi_{\text{LDA}}(\mathbf{w}) = \mathbf{A}_{\text{LDA}}^{\text{T}}\,\mathbf{w}.$$

NAP

The NAP algorithm was originally proposed in [108]. In this approach, the feature space is transformed using an orthogonal projection in the channel's complementary space, which depends only on the speaker (assuming that other variability in the data is insignificant). The projection is calculated using the within-class covariance matrix. Define a $d \times d$ projection matrix [108] of co-rank $k < d$

$$\mathbf{P} = \mathbf{I} - \mathbf{u}_{[k]}\mathbf{u}_{[k]}^T,$$

where $\mathbf{u}_{[k]}$ is a rectangular matrix of low rank whose columns are the k principal eigenvectors of the within-class covariance matrix $\mathbf{S}_w$ given in (14). NAP is performed on $\mathbf{w}$ as

$$\Phi_{\text{NAP}}(\mathbf{w}) = \mathbf{P}\mathbf{w}.$$

WCCN

This normalization was originally proposed for improving robustness in the SVM-based speaker-recognition framework [109] using a one-versus-all decision approach. The WCCN projection aims at minimizing the false-alarm and miss-error rates during SVM training.

The implementation of the strategy begins with using a data set D similar to the one that was described in the previous section. The within-class covariance matrix $\mathbf{S}_w$ is calculated using (14), and the WCCN projection is performed as

$$\Phi_{\text{WCCN}}(\mathbf{w}) = A_{\text{WCCN}}^T\,\mathbf{w},$$

where $\mathbf{A}_{\text{WCCN}}$ is computed through the Cholesky factorization of $\mathbf{S}_w^{-1}$ such that

$$\mathbf{S}_w^{-1} = \mathbf{A}_{\text{WCCN}}\,\mathbf{A}_{\text{WCCN}}^{\text{T}}.$$

In contrast to LDA and NAP, the WCCN projection conserves the directions of the feature space.

SPEAKER VERIFICATION USING i-VECTORS

After i-vectors were introduced, in essence, many previously available pattern-recognition methods were effectively applied in this domain. We discuss some of the popular methods of classification using i-vectors.

SVM Classifier

As discussed previously, the i-vector representation was discovered in an attempt to utilize JFA as a feature extractor for SVMs. Thus, initially i-vectors were used with SVMs with different kernel functions [79]. The idea is the same as SVM with GMM supervectors, except that the i-vectors are now used as utterance-dependent features. Because of the lower dimension of the i-vectors compared to supervectors, the application of LDA and WCCN projections together became more effective and were well suited.

Cosine Distance Scoring

In [79], the cosine similarity measure-based scoring was proposed for speaker verification. In this measure, the match score between a target and test i-vector $\mathbf{w}_{\text{target}}$ and $\mathbf{w}_{\text{test}}$ is computed as their normalized dot product

$$\text{CDS}(\mathbf{w}_{\text{target}}, \mathbf{w}_{\text{test}}) = \frac{\mathbf{w}_{\text{target}} \cdot \mathbf{w}_{\text{test}}}{\|\mathbf{w}_{\text{target}}\|\,\|\mathbf{w}_{\text{test}}\|}.$$

Probabilistic Linear Discriminant Analysis

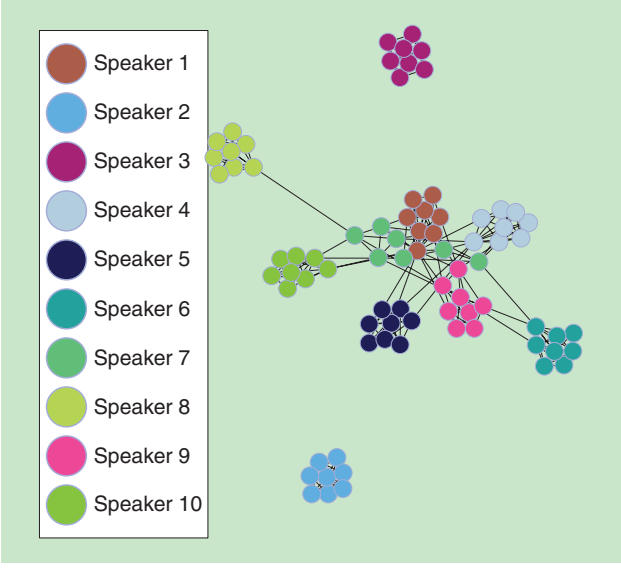
Probalistic LDA (PLDA) was first used for session variability compensation for facial recognition [121]. This essentially follows the same modeling assumptions as JFA, i.e., a pattern vector contains class-dependent and session-dependent variabilities, both lying in lower-dimensional subspaces. An i-vector extracted from utterance u is decomposed as

$$\mathbf{w}_{s,h} = \mathbf{w}_0 + \Phi\beta_s + \Gamma\alpha_h + \varepsilon_{s,h}. \tag{16}$$

Here, $\mathbf{w}_0 \in \mathbf{R}^R$ is the speaker-independent mean i-vector, Φ is the $R \times N_{ev}$ low-rank matrix representing the speaker-dependent basis functions/eigenvoices, Γ is the $R \times N_{ec}$ low-rank matrix spanning the channel subspace, $\boldsymbol{\beta}_s \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is an $N_{ev} \times 1$ hidden variable (i.e., speaker factors), $\boldsymbol{\alpha}_s \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is an $N_{ec} \times 1$ hidden variable (i.e., channel factors), and $\boldsymbol{\varepsilon}_{h,s} \in \mathrm{R}^R$ is a random vector representing the residual noise.

PLDA was first introduced in speaker verification in [94] using a heavy-tailed distribution assumption on i-vectors instead of a Gaussian assumption. Later, it was shown that when i-vectors are length normalized (i.e., they are divided by their corresponding vector length) [122], a Gaussian PLDA model performs equivalent to its heavy-tailed version. Since the latter is computationally more expensive, Gaussian PLDA models are more commonly used. Also, the use of a full-covariance noise model for $\boldsymbol{\varepsilon}_{h,s}$ is feasible in this formulation that allows one to drop the eigenchannel term ($\Gamma\boldsymbol{\alpha}_h$) from (16) without loss of performance. In this case, the PLDA model would be as follows:

$$\mathbf{w}_{s,h} = \mathbf{w}_0 + \Phi\boldsymbol{\beta}_s\boldsymbol{\varepsilon}_{s,h}.$$



[FIG9] A graphical representation of 79 utterances spoken by ten individuals collected from the NIST SRE 2004 corpus. The i-vector representation is used for each segment; the plot is generated using GUESS, an open-source graph exploration software [123] that can visualize higher-dimensional data using distance measures between samples.

We note that, though developed independently, the JFA model is very similar to PLDA. Looking at (11) and (16) and comparing the terms makes this clear. The obvious difference between these models is that JFA models the GMM supervectors, while PLDA models i-vectors. Since i-vectors are essentially dimensionality reduced versions of supervectors (incurring loss of information), JFA, in principle, should be better in modeling the within- and between-speaker variations. However, in reality, the amount of labeled training data is limited, and due to the large number of parameters in JFA, it cannot be trained as effectively as a PLDA model on lower dimensional i-vectors (using the same amount of labeled data). Besides, the total variability model (i-vector extractor) can be trained on unlabeled data sets, which are available in larger amounts.

Although the model equations are identical, there are significant differences in the training process of the two models. Since JFA was designed for GMM supervectors, the formulations involved processing the acoustic speech frames and their statistics in different mixtures of the UBM. Unlike i-vectors, the GMM supervectors are not extracted first before JFA training—instead, JFA operates directly on the acoustic features and can provide similarity scores between two utterances from their corresponding feature streams. This dependence on acoustic features (and the various order statistics) makes the scoring process more computationally expensive for JFA. For PLDA, the input features are i-vectors that are extracted beforehand, and, during the scoring process, only two i-vectors from the corresponding utterances are required—not the acoustic features. This makes PLDA much simpler in implementation.

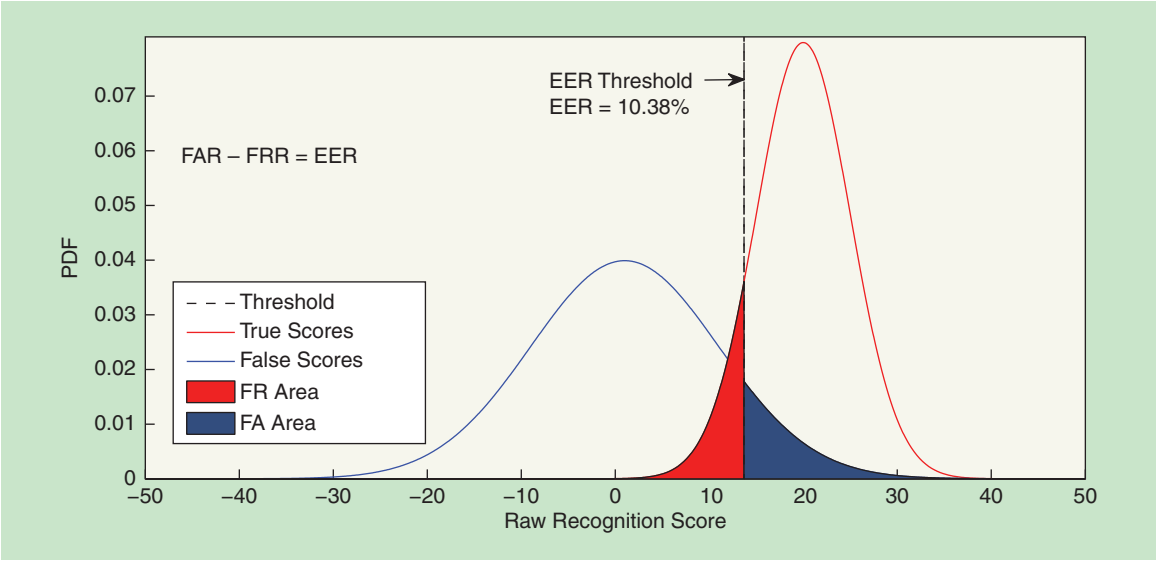
It can be argued that, with a sufficiently large labeled data set, JFA can outperform an i-vector-PLDA system. However, we are not aware of such results reported at this time.

PERFORMANCE EVALUATION IN STANDARDIZED DATA SETS

Evaluating the performance of a speaker-verification task using a standardized data set is a very important element of the research cycle. Over the years, new data sets and performance metrics have been introduced to match realistic scenarios. These, in turn, motivated researchers to discover new strategies to address the challenges, compare results among peers, and exchange ideas.

THE NIST SRE CHALLENGE

NIST has been organizing an SRE campaign for the past several years aiming at providing standard data sets, verification tasks, and performance metrics for the speaker ID community (Figure 9). Every year's evaluation introduces new challenges for the research community. These challenges include newly introduced recording conditions (e.g., microphone, handset, and room acoustics), short test utterance duration, varying vocal effort, artificial and real-life additive noise, restrictions or allowances in data-utilization strategy, new performance metrics to be optimized, etc. It is clear that the performance metric defined for a speaker-recognition task depend on the data set and train-test pairs of speech (also known as *trials*) used for the evaluation. A sufficient number of such trials needs to be provided for a statistically significant evaluation measure [78]. The performance measures can be based on hard verification decisions or soft scores, they may require log-LR as scores, and depend on the



[FIG10] An illustration of target and nontarget score distributions and the decision threshold. Areas under the curves with blue and red colors represent FAR and FRR errors, respectively.

prior probability of encountering a target speaker. For a given data set and task, systems evaluated using a specific error/cost criteria can be compared. Before discussing the common performance measures, we introduce the type of errors encountered in speaker verification.

TYPES OF ERRORS

There are mainly two types of errors in speaker verification (or any other biometric authentication) when a hard decision is made by the automatic system. From the speaker authentication point of view, we define them as

- *false accept (FA)*: granting access to an impostor speaker
- *false reject (FR)*: denying access to a legitimate speaker.

From the speaker-detection point of view (a target speaker is sought), these are called *false-alarm* and *miss errors*, respectively. According to these definitions, two error rates are defined as

$$\text{False-Acceptance Rate (FAR)} = \frac{\text{Number of FA errors}}{\text{Number of impostor attempts}}$$

$$\text{False-Rejection Rate (FRR)} = \frac{\text{Number of FR errors}}{\text{Number of legitimate attempts}}.$$

Speaker-verification systems generally output a match score between the training speaker and the test utterance. This is true for most two-class recognition/binary detection problem. This score is a scalar variable that represents the similarity between the enrolled speaker and the test speaker, with higher values indicating the speakers are more similar. To make a decision, the system needs to use a threshold (τ) as illustrated in Figure 10. If the threshold is too low, there will be a lot of FA errors, whereas if the threshold is too high, there will be too many FR/miss errors.

EQUAL ERROR RATE

The equal error rate (EER) is defined as the FAR and FRR values when they become equal. That is, by changing the threshold, we find a point where the FAR and FRR become equal. This is shown in Figure 10. The EER is a very popular performance measure for speaker-verification systems. Only the soft scores from the automatic system are required to compute the EER. No actual hard decisions are made. It should be noted that operating a speaker-verification system on the threshold corresponding to the EER might not be desirable for practical purposes. For high-security applications, one should set the threshold higher, lowering the FA errors at the cost of miss errors. However, for high convenience, the threshold may be set lower. Let us discuss some examples. In authenticating users for bank accounts, security is of utmost importance. It is thus better to deny access to the legitimate user (and ask other forms of verification) as opposed to granting access to an impostor. On the contrary, for an automated customer service, denying a legitimate speaker will cause inconvenience and frustration to the user. In this case, accepting an illegitimate speaker is not as critical as in high-security applications.

DETECTION COST FUNCTION

This is, in fact, a family of performance measures introduced by NIST over the years. As mentioned before, the EER does not differentiate between the two errors, which sometimes is not a realistic performance measure. The detection cost function (DCF), thus, introduces numerical costs/penalties for the two types of errors (FA and miss). The a priori probability of encountering a target speaker is also provided. The DCF is computed over the full range of decision threshold values as

$$\text{DCF}(\tau) = C_{\text{MISS}} P(\tau) P_{\text{Target}} + C_{\text{FA}} P_{\text{FA}}(\tau)(1 - P_{\text{Target}}).$$

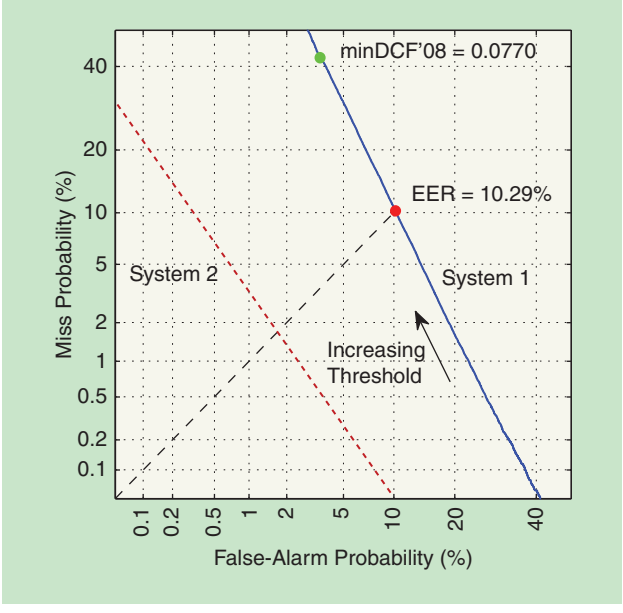
Here,

C_{Miss} = Cost of a miss/FR error
C_{FA} = Cost of an FA error
P_{Target} = Prior probability of target speaker.
$P_{\text{Miss}}(\tau)$ = Probability of (Miss | Target, Threshold = τ)
$P_{\text{FA}}(\tau)$ = Probability of (FA | Nontarget, Threshold = τ).

Usually, the DCF is normalized by dividing it by a constant [77]. The probability values here can be computed using the distribution of true and impostor scores and computing the areas under the curve as shown in Figure 10. The first three quantities above (C_{Miss}, C_{FA}, and P_{Target}) are predefined. Generally, the goal of the system designer is to find the optimal threshold value that minimizes the DCF.

In NIST SRE 2008, these DCF parameters were set as $C_{Miss} = 10$, $C_{FA} = 1$, and $P_{Target} = 0.01$. The values of the costs indicate that the system is penalized ten times more for making a miss error rather than an FA error. As a real-world example, when detecting a known criminal's voice from evidence recordings, it may be better to have false positives (e.g., to suspect and investigate an innocent speaker) than to miss the target speaker (e.g., to be unable to detect the criminal at all). If we ignore P_{Target} for the moment, setting a lower threshold (τ) would be beneficial since, in this case, the system will tolerate more FAs but will not miss too many legitimate speakers [$P_{Miss}(\tau)$ will be lower], yielding a lower DCF value for that threshold. Now, the value of the prior ($P_{Target} = 0.01$) indicates that a target speaker will be encountered by the system once in every 100 speaker-verification attempts. If this condition is considered independently, it is better to have a higher threshold since most of the attempts will be from impostors ($P_{Nontarget} = 0.99$). However, when all three parameters are considered together, finding the optimal threshold requires sweeping through all the DCF values.

By processing the DCF, two performance measures are derived: 1) the minimum DCF (MinDCF) and 2) the actual DCF (ActDCF). The MinDCF is the minimum value of DCF that can be obtained by changing the threshold, τ. The MinDCF parameter can be computed only when the soft scores are provided by the systems. When the system provides hard decisions, the actual DCF is used where the probability values involved (in the DCF equation) are simply computed by counting the errors. Both of these performance measures have been extensively used in the NIST evaluations. The most recent evaluation in 2012 introduced a DCF that is a dependent on two different operating points (two sets of error costs and target priors) instead of one.



[FIG11] DET curves of two speaker-verification systems (System 1 and System 2). In System 1, the points on the curve corresponding to the threshold that yields the EER and minimum DCF (as in NIST SRE 2008), and the direction of an increasing threshold are shown. Being closer to the origin, System 2 shows a better performance.

It is important to note here that the MinDCF (or ActDCF) parameter is not an error rate in the general sense. Thus, its interpretation is not straightforward. Obviously, the lower MinDCF, the better the system performance. However, the exact value of the MinDCF can only be used to compare other systems evaluated using the same trials and performance measure. Generally, when the system EER improves, the DCF parameters also improve. An elaborate discussion on the relationship between EER and DCF can be found in [124].

DETECTION ERROR TRADEOFF CURVE

When speaker-verification performance needs to be evaluated in a range of operating points, the detection error tradeoff (DET) curve is generally employed. The DET curve is a plot of the errors FAR versus FRR/miss. An example DET curve is shown in Figure 11. As the system performance improves, the curve moves toward the origin. As illustrated in Figure 11, the DET curve corresponding to System 2 is closer to the origin and thus represents a better system. The EER and minDCT points are shown on the DET curve of System 1.

During the preparation of the DET curve, the cumulative density functions (CDFs) of the true and impostor scores are transformed to normal deviates. This means that the true/impostor score CDF value for a given threshold is transformed by a standard normal inverse CDF (ICDF) and the resulting values are used to make the plot. This transform yields a linear DET curve when the two distributions are normal and have equal variances. Thus, even though the labels indicate the axis as error probabilities, they are actually plotted according to the corresponding normal deviate values.

RECENT ADVANCEMENTS IN AUTOMATIC SPEAKER RECOGNITION

In recent years, considerable research progress has been made in spoofing and countermeasures [125], [126], back-end classifiers [127], [128], compensation for short utterances [129]–[131], score calibration and fusion [132], [133], deep neural network (DNN) [134]–[136], and alternate acoustic modeling [137] techniques. In this section, we briefly discuss some of these topics and their possible implications in the speaker-recognition research.

NIST i-VECTOR MACHINE-LEARNING CHALLENGE AND BACK-END PROCESSING

The most recent NIST-sponsored evaluation, the i-Vector Machine-Learning Challenge, focused on back-end classifiers. In this paradigm, instead of audio data, i-vectors from speech utterances were provided to the participants [138]. In this way, the entry barrier to the evaluation was reduced as many machine-learning-focused research groups were able to participate without expertise in audio/speech processing. Significant performance

improvements were observed from top-performing systems compared to the baseline system provided by NIST [138]. Since only i-vectors were provided by NIST, the algorithmic improvements are all due to modeling and back-end processing of i-vectors. In addition, the i-vectors provided by NIST did not have any speaker labels, which also generated new ideas on utilizing unlabeled data in speaker recognition [139].

DURATION VARIABILITY COMPENSATION

Duration variability is one of the problems that has received considerable attention in recent years. Since the advent of GMM supervectors and i-vectors, variable-duration utterances could be mapped to a fixed-dimensional pattern. This has been a significant advancement since various machine-learning tools were being applied to these vectors, especially i-vectors due to their smaller dimensions. However, it is clear that an i-vector extracted from a short utterance will not be as representative of a speaker compared to the one extracted from a longer utterance. Duration mismatch between train and test is thus a major problem. One way to mitigate this problem is by including short utterances in the PLDA training [130], [140]. Alternatively, this can be addressed in the score domain [130]. In [141], Kenny et al. propose that i-vectors extracted from short utterances are less reliable and incorporates this variability by including a noise term into the PLDA model. In [142], a DNN-based method was proposed for speaker recognition in short utterances where the content of the test utterance was searched in the enrollment data to be compared.

DNN-BASED METHODS

In the last few years, DNNs have been tremendously successful at many speech-processing tasks, most prominently in speech recognition [143], [144]. Naturally, DNNs have also been used in speaker recognition. Works by Kenny et al. [136] have shown improvements in extracting Baum–Welch statistics for speaker recognition using DNNs. DNNs have also been incorporated for multisession speaker recognition [134] as well as phonetically aware DNNs for noise-robust speaker recognition [135]. DNNs have also been used to extract front-end features, also known as *bottle-neck features* [145]. Since, there are an extensive set of literature on deep learning [143], [146] and its application in speaker recognition is relatively new, we have not included a discussion on DNNs in this tutorial.

[TABLE 2] THE SPEAKER-RECOGNITION PROCESS: MAN VERSUS MACHINE.

ASPECT	HUMANS	MACHINES
TRAINING	SPEAKER RECOGNITION IS AN ACQUIRED HUMAN TRAIT AND REQUIRES TRAINING.	REQUIRES SUFFICIENT DATA TO TRAIN THE RECOGNIZERS.
VAD	DIFFERENT PARTS OF THE HUMAN BRAIN ARE ACTIVATED WHEN SPEECH AND NONSPEECH STIMULI ARE PRESENTED.	SPEECH SIGNAL PROPERTIES AND STATISTICAL MODELS ARE USED TO DETECT PRESENCE OR ABSENCE OF SPEECH.
AUDIO PROCESSING	THE HUMAN BRAIN PERFORMS BOTH SPECTRAL AND TEMPORAL PROCESSING. IT IS NOT KNOWN EXACTLY HOW THE AUDIO SIGNAL DEVELOPS THE SPEAKER- OR PHONEME-DEPENDENT ABSTRACT REPRESENTATIONS/MODELS.	ACOUSTIC FEATURE PARAMETERS DEPENDING ON SPECTRAL AND TEMPORAL PROPERTIES OF THE AUDIO SIGNAL ARE UTILIZED FOR RECOGNITION.
HIGH-LEVEL FEATURES	WE CONSIDER LEXICON, INTONATION, PROSODY, AGE, GENDER, DIALECT, SPEAKING RATE, AND MANY OTHER PARALINGUISTIC ASPECTS OF SPEECH TO REMEMBER A PERSON'S VOICE.	RECENT ALGORITHMS HAVE INCORPORATED PROSODY, PRONUNCIATION, DIALECT, AND OTHER HIGH-LEVEL FEATURES FOR SPEAKER IDENTIFICATION.
COMPACT REPRESENTATION	THE HUMAN BRAIN FORMS SPEAKER-DEPENDENT, EFFICIENT ABSTRACT REPRESENTATIONS. THESE ARE INVARIANT TO CHANGES OF THE ACOUSTIC INPUT, PROVIDING ROBUSTNESS TO NOISE AND SIGNAL DISTORTION.	HIGH-LEVEL FEATURES ARE EXTRACTED THAT SUMMARIZE THE VOICE CHARACTERISTICS OF A SUBJECT. THESE ARE EXTRACTED IN A WAY TO MINIMIZE SESSION VARIABILITY DUE TO NOISE OR DISTORTION.
LANGUAGE DEPENDENCE	SPEAKER RECOGNITION BY HUMANS IS BETTER IF THEY KNOW THE LANGUAGE BEING SPOKEN.	AUTOMATIC SYSTEM'S PERFORMANCE IS DEGRADED IF THERE IS A MISMATCH IN TRAINING AND TEST LANGUAGE.
FAMILIAR VERSUS UNFAMILIAR SPEAKERS	HUMANS ARE EXTREMELY GOOD AT IDENTIFYING FAMILIAR VOICES, BUT NOT SO FOR UNFAMILIAR ONES.	MACHINES PROVIDE CONSISTENT PERFORMANCE WHEN ADEQUATE AMOUNT OF DATA IS PROVIDED. FAMILIARITY CAN BE RELATED TO THE AMOUNT OF TRAINING DATA.
IDENTIFICATION VERSUS DISCRIMINATION	THE HUMAN BRAIN PROCESSES THESE TWO TASKS DIFFERENTLY.	IN MOST CASES, THE SAME ALGORITHM (WITH SLIGHT MODIFICATION) CAN BE USED TO IDENTIFY AND DISCRIMINATE BETWEEN SPEAKERS.
MEMORY RETENTION	HUMANS' ABILITY TO REMEMBER A PERSON'S VOICE DEGRADES WITH TIME.	A COMPUTER ALGORITHM CAN STORE THE MODELS OF A PERSON INDEFINITELY IF PROVIDED SUPPORT.
FATIGUE	HUMAN LISTENERS CANNOT PERFORM AT THE SAME LEVEL FOR A LONG DURATION.	COMPUTERS DO NOT HAVE ISSUES WITH FATIGUE. LONG RUNTIMES MAY CAUSE OVERHEATING IF NECESSARY PRECAUTIONS ARE NOT TAKEN.
IDENTIFY IDIOSYNCRASIES	HUMANS ARE VERY GOOD AT IDENTIFYING CHARACTERISTIC TRAITS OF A VOICE.	THE MACHINE ALGORITHMS HAVE TO BE SPECIFICALLY TOLD WHAT TO LOOK FOR AND COMPARE.
MISMATCHED CONDITIONS	HUMANS RELY MORE ON PARALINGUISTIC ASPECTS OF SPEECH IN SEVERE MISMATCHED CONDITIONS.	AUTOMATIC SYSTEMS ARE TRAINED ON VARIOUS ACOUSTIC CONDITIONS, AND USUALLY ARE MORE ROBUST.
SUSCEPTIBILITY TO BIAS	HUMAN JUDGMENT CAN BE BIASED BY CONTEXTUAL INFORMATION.	AUTOMATIC ALGORITHMS CAN BE BIASED TOWARD THE TRAINING DATA.

MAN VERSUS MACHINE IN SPEAKER RECOGNITION

In this section, we attempt to compare the speaker-recognition task as performed by humans and the state-of-the-art algorithms. First we must realize that it is very difficult to do such comparisons in a statistically meaningful manner. This is because getting humans to evaluate a large number of utterances reliably is quite challenging. However, attempts have been made to make such comparisons in the past [147]–[149]. In the majority of these cases, especially in the recent ones, the speaker-recognition performance of humans was found to be inferior to that of automatic systems.

In [150], the authors compared the speaker-recognition performance of human listeners to a typical algorithm (automatic system is not mentioned in the paper) using a subset of NIST SRE 1998 data. Opinions of multiple human listeners were combined to form the final speaker-recognition decision. Results showed that humans are as good as the best system and outperformed standard algorithms especially when there is a mismatch in the telephone channel (a different number was used to make the phone call).

Recently, NIST presented speaker-recognition tasks for evaluating systems that combined human and machines [20]. The task, known as the HASR, was designed in a way such that the most difficult test samples are selected for the evaluation (channel mismatch, noise, same/different speakers that sound highly dissimilar/similar, etc.). However, the total number of trials in these experiments was very low compared to evaluations designed for automatic systems. One of the motivations of this study was to evaluate if automatic systems have become good enough, in other words, is it beneficial to keep humans involved in the process? The HASR study was repeated during the 2012 NIST SRE where both noisy and channel degraded speech data were encountered.

Interestingly, machines consistently performed better than human-assisted approaches in the given NIST HASR tasks [151]–[155]. In [156], it was even claimed that by combining multiple naïve listeners' decisions, the HASR 2010 task can be performed as well as forensic experts, which somewhat undermines the role of a forensic expert. In [157], it was shown that human and machine decisions were complementary, meaning that in some cases the humans correctly identified a speaker where the automatic system failed, and vice versa. However, the HASR tasks were exceptionally difficult for human listeners because of the severe channel mismatch, unfamiliarity with the speakers, noise, and other factors. A larger and more balanced set of trials should be used for a proper evaluation of human performance. Following the HASR paradigm, further research focused on how humans can aid the decision of an automatic system, especially in the context of forensic speaker recognition [157]. An i-vector system [79] with a PLDA classifier was used in this particular study.

The performance of humans and machines was compared in a forensic context in [149], where 45 trials were used (nine target and 36 nontarget). The human system consisted of a panel of listeners whereas a GMM–UBM-based system [6] was used for the automatic system. Here again, the automatic system outperformed the human panel of listeners. However, the results should be interpreted with caution since the number of trials here was low.

In [158], human speaker-recognition performance was compared with automatic algorithms in presence of voice mimicry. A GMM–UBM system and an i-vector-based system were used in the study. The speech database consisted of five Finnish public figures and their voices were mimicked by a professional voice actor. The results show that humans are more likely to make errors when impersonation is done. On average, the automatic algorithm performed better than the human listeners.

Although most experiments so far show human performance to be inferior to automatic algorithms, these cannot be considered as definitive proof that machines are always better than humans. In many circumstances, humans will perform better, especially when paralinguistic information becomes important. As discussed previously, humans perform exceptionally well in recognizing familiar speakers. To the best of our knowledge, a comparison of familiar speaker recognition versus automatic algorithm (with sufficient training data) has not been performed yet. Thus, for familiar speakers, humans may perform much better than state-of-the-art algorithms—and this should motivate researchers to discover how the human brain stores familiar speakers' identity information. In HASR, the goal was to have humans assist the automatic system. On the other hand, automatic systems inspired by the forensic experts' methodology have already been investigated [159], where speaker nativeness, dialect, and other demographic information were considered. A generic comparison between how humans and machines perform speaker recognition is provided in Table 2.

CONCLUSIONS

A substantial amount of work still needs to be done to fully understand how the human brain makes decisions about speech content and speaker identity. However, from what we know, it can be said that automatic speaker-recognition systems should focus more on high-level features for improved performance. Humans are effective at effortlessly identifying unique traits of speakers they know very well, whereas automatic systems can only learn a specific trait if a measurable feature parameter can be properly defined. Automatic systems are better at searching over vast collections of audio and, perhaps, at being able to more effectively set aside those audio samples which are less likely to be speaker matches; whereas humans are better at comparing a smaller subset and overcoming microphone or channel mismatch more easily. It may be worthwhile to investigate what it really means to "know" a speaker from the perspective of an automatic system. The search for alternative compact representations of speakers and audio segments emphasizing the identity relevant parameters while suppressing the nuisance components will always be an ongoing challenge for system developers.

AUTHORS

John H.L. Hansen (John.Hansen@utdallas.edu) received his Ph.D. and M.S. degrees in electrical engineering from the Georgia Institute of Technology and his B.S.E.E. degree from Rutgers University, New Jersey. He is with the University of Texas at Dallas, where he is associate dean for research and professor of electrical engineering. He holds the Distinguished Chair in Telecommunications and oversees

the Center for Robust Speech Systems. He is an International Science Congress Association (ISCA) fellow; a past chair of the IEEE Speech and Language Technical Committee; a coorganizer and technical chair of the IEEE International Conference on Acoustics, Speech, and Signal Processing 2010; a coorganizer of IEEE Spoken Language Technology 2014; and the general chair/organizer of ISCA Interspeech 2002. He has supervised more than 70 Ph.D./M.S. students and coauthored more than 570 papers in the field. His research interests include digital speech processing, speech and speaker analysis, robust speech/speaker/language recognition, speech enhancement for hearing loss, and robust hands-free human-interaction in car environments. He is a Fellow of the IEEE.

Taufiq Hasan (taufiq.hasan@utdallas.edu) received his B.S. and M.S. degrees in electrical and electronic engineering (EEE) from Bangladesh University of Engineering and Technology, Dhaka, Bangladesh, in 2006 and 2008, respectively. He earned his Ph.D. degree from the Department of Electrical Engineering, Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UT Dallas), Richardson, in 2013. From 2006 to 2008, he was a lecturer in the Department of EEE, United International University, Dhaka. He served as the lead student from the Center for Robust Speech Systems at UT Dallas during the 2012 National Institute of Standards and Technology Speaker Recognition Evaluation submissions. Currently, he works as a research scientist at Robert Bosch Research and Technology Center in Palo Alto, California. His research interests include speaker recognition in mismatched conditions, speech recognition, enhancement and summarization, affective computing, and multimodal signal processing.

REFERENCES

[1] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, et al., "Building Watson: An overview of the DeepQA project," *AI Mag*., vol. 31, no. 3, pp. 59–79, 2010.

[2] J. Aron, "How innovative is Apple's new voice assistant, Siri?" *New Sci*., vol. 212, no. 2836, pp. 24, 29 Oct. 2011.

[3] A. Eriksson, "Tutorial on forensic speech science," in *Proc. European Conf. Speech Communication and Technology*, Lisbon, Portugal, 2005, pp. 4–8.

[4] P. Belin, R. J. Zatorre, P. Lafaille, P. Ahad, and B. Pike, "Voice-selective areas in human auditory cortex," *Nature*, vol. 403, pp. 309–312, Jan. 2000.

[5] E. Formisano, F. De Martino, M. Bonte, and R. Goebel, "'Who' is saying 'what'? Brain-based decoding of human voice and speech," *Science*, vol. 322, pp. 970–973, Nov. 2008.

[6] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process*., vol. 10, no. 1–3, pp. 19–41, Jan. 2000.

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

[8] P. Eckert and J. R. Rickford, *Style and Sociolinguistic Variation*. Cambridge, U.K.: Cambridge Univ. Press, 2001.

[9] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Commun*., vol. 20, no. 1–2, pp. 151–173, Nov. 1996.

[10] X. Fan and J. H. L. Hansen, "Speaker identification within whispered speech audio streams," *IEEE Trans. Audio, Speech, Lang. Process*., vol. 19, no. 4, pp. 1408–1421, May 2011.

[11] C. Zhang and J. H. L. Hansen, "Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing," *IEEE Trans. Audio, Speech, Lang. Process*., vol. 19, no. 4, pp. 883–894, May 2011.

[12] 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. Audio, Speech, Lang. Process*., vol. 17, no. 2, pp. 366–378, Feb. 2009.

[13] M. Mehrabani and J. H. L. Hansen, "Singing speaker clustering based on subspace learning in the GMM mean supervector space," *Speech Commun*., vol. 55, no. 5, pp. 653–666, June 2013.

[14] J. H. Hansen, C. Swail, A. J. South, R. K. Moore, H. Steeneken, E. J. Cupples, T. Anderson, C. R. Vloeberghs, et el., "The impact of speech under 'stress' on military speech technology," NATO Project Rep., no. 104, 2000.

[15] 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," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'95)*, pp. 329–332.

[16] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process*., vol. 15, no. 4, pp. 1435–1447, 2007.

[17] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Process*., vol. 10, no. 1–3, pp. 42–54, Jan. 2000.

[18] 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. Speech Audio Process*., vol. 2, no. 2, pp. 245–257, 1994.

[19] Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process*., vol. 15, no. 7, pp. 2023–2032, 2007.

[20] C. Greenberg, A. Martin, L. Brandschain, J. Campbell, C. Cieri, G. Doddington, and J. Godfrey, "Human assisted speaker recognition in NIST SRE10," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010, pp. 180–185.

[21] 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," in *Proc. Interspeech*, Brighton, U.K., pp. 1811–1814, 2009.

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

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

[24] F. Kelly, A. Drygajlo, and N. Harte, "Speaker verification with long-term ageing data," in *Proc. Int. Conf. Biometrics,* 2012, pp. 478–483.

[25] F. Kelly, A. Drygajlo, and N. Harte, "Speaker verification in score-ageing-quality classification space," *Comput. Speech Lang*., vol. 27, no. 5, pp. 1068–1084, Aug. 2013.

[26] Wikipedia Contributors. George Zimmerman. Entry on Wikipedia (2014, Nov. 29). [Online]. Available: http://en.wikipedia.org/wiki/George_Zimmerman

[27] J. H. L. Hansen and N. Shokouhi. (2013, Nov. 29). Speaker identification: Screaming, stress and non-neutral speech, is there speaker content?. [Online]. *IEEE SLTC Newsletter*. Available: http://www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2013-11/SpeakerIdentification/

[28] F. Nolan and T. Oh, "Identical twins, different voices," *Int. J. Speech Lang. Law*, vol. 3, no. 1, pp. 39–49, 1996.

[29] W. D. Van Gysel, J. Vercammen, and F. Debruyne, "Voice similarity in identical twins," *Acta Otorhinolaryngol. Belg*., vol. 55, no. 1, pp. 49–55, 2001.

[30] K. M. Van Lierde, B. Vinck, S. De Ley, G. Clement, and P. Van Cauwenberge, "Genetics of vocal quality characteristics in monozygotic twins: a multiparameter approach," *J. Voice*, vol. 19, no. 4, pp. 511–518, Dec. 2005.

[31] D. Loakes, "A forensic phonetic investigation into the speech patterns of identical and non-identical twins," *Int. J. Speech Lang. Law*, vol. 15, no. 1, pp. 97–100, 2008.

[32] *Twins Day. Twinsburg, Ohio* (2015). [Online]. Available: http://www.twinsdays.org

[33] F. Nolan, *The Phonetic Bases of Speaker Recognition*. Cambridge, U.K.: Cambridge Univ. Press, 1983.

[34] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *J. Acoust. Soc. Amer*, vol. 51, no. 68, p. 2044, 1972.

[35] P. Rose, *Forensic Speaker Identification*. Boca Raton, FL: CRC Press, 2004.

[36] P. Rose, "The technical comparison of forensic voice samples," in *Expert Evidence*, 1 ed. Sydney, Australia: Thompson Lawbook Co., 2003, Chap. 99, pp. 1051–6102.

[37] F. Nolan, "The limitations of auditory-phonetic speaker identification," in *Texte Zur Theorie Und Praxis Forensischer Linguistik*, H. Kniffka, Ed. Berlin, Germany: De Gruyter, 1990, pp. 457–479.

[38] L. Watts, "Reverse-engineering the human auditory pathway," in *Advances in Computational Intelligence*. New York: Springer, 2012, pp. 47–59.

[39] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition," *Speech Commun*., vol. 46, no. 3–4, pp. 455–472, July 2005.

[40] A. G. Adami, R. Mihaescu, D. A. Reynolds, and J. J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'03)*, pp. 788–791.

[41] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Trans. Audio, Speech Lang. Process*., vol. 15, no. 7, pp. 2095–2103, 2007.

[42] J. H. Wigmore, "A new mode of identifying criminals," *17 Amer Inst. Crim. L. Criminology 165*, vol. 17, no. 2, pp. 165–166, Aug. 1926.

[43] L. G. Kersta, "Voiceprint identification," *Police L. Q.*, vol. 3, no. 5, 1973–1974.

[44] B. E. Koenig, "Spectrographic voice identification: A forensic survey," *J. Acoust. Soc. Amer*, vol. 79, no. 6, pp. 2088–2090, 1986.

[45] H. F. Hollien, *Forensic Voice Identification*. New York: Academic Press, 2002.

[46] L. Yount, *Forensic Science: From Fibers to Fingerprints*. New York: Chelsea House, 2007.

[47] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J. Bonastre, and D. Matrouf, "Forensic speaker recognition," *IEEE Signal Process. Mag.*, vol. 26, no. 2, pp. 95–103, 2009.

[48] G. S. Morrison, "Distinguishing between science and pseudoscience in forensic acoustics," in *Proc. Meetings on Acoustics*, 2013, pp. 060001.

[49] J. Neyman and E. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philos. Trans. R. Soc.,* vol. 231, pp. 289–337, Jan. 1933.

[50] G. S. Morrison, "Forensic voice comparison and the paradigm shift," *Sci. Justice*, vol. 49, no. 4, pp. 298–308, 2009.

[51] G. S. Morrison, "Forensic voice comparison," in *Expert Evidence 99*, 1 ed. London: Thompson Reuters, 2010, Chap. 99, p. 1051.

[52] R. H. Bolt, *On the Theory and Practice of Voice Identification*. Washington, DC: National Academy of Sciences, 1979.

[53] B. S. Kisilevsky, S. M. Hains, K. Lee, X. Xie, H. Huang, H. H. Ye, K. Zhang, and Z. Wang, "Effects of experience on fetal voice recognition," *Psychol. Sci.*, vol. 14, no. 3, pp. 220–224, May 2003.

[54] M. Latinus and P. Belin, "Human voice perception," *Curr. Biol.*, vol. 21, no. 4, pp. R143–R145, Feb. 2011.

[55] P. Belin, P. E. Bestelmeyer, M. Latinus, and R. Watson, "Understanding voice perception," *Br. J. Psychol.*, vol. 102, no. 4, pp. 711–725, Nov. 2011.

[56] J. Cacioppo, *Foundations in Social Neuroscience*. Cambridge, MA: MIT Press, 2002.

[57] D. Van Lancker and J. Kreiman, "Voice discrimination and recognition are separate abilities," *Neuropsychologia*, vol. 25, no. 5, pp. 829–834, 1987.

[58] D. Van Lancker, J. Kreiman, and K. Emmorey, "Familiar voice recognition: Patterns and parameters. Part I: Recognition of backward voices," *J. Phonetics*, vol. 13, no. 1, pp. 19–38, 1985.

[59] T. K. Perrachione, S. N. Del Tufo, and J. D. Gabrieli, "Human voice recognition depends on language ability," *Science*, vol. 333, no. 6042, pp. 595–595, July 2011.

[60] R. J. Zatorre and P. Belin, "Spectral and temporal processing in human auditory cortex," *Cereb. Cortex*, vol. 11, pp. 946–953, Oct. 2001.

[61] H. Hollien, W. Majewski, and P. Hollien, "Perceptual identification of voices under normal, stress, and disguised speaking conditions," *J. Acoust. Soc. Amer*, vol. 56, no. S53, 1974.

[62] S. M. Kassin, I. E. Dror, and J. Kukucka, "The forensic confirmation bias: Problems, perspectives, and proposed solutions," *J. Appl. Res. Memory Cognit.*, vol. 2, no. 1, pp. 42–52, Mar. 2013.

[63] G. Papcun, J. Kreiman, and A. Davis, "Long-term memory for unfamiliar voices," *J. Acoust. Soc. Amer*, vol. 85, no. 2, pp. 913, Feb. 1989.

[64] S. Cook and J. Wilding, "Earwitness testimony: Never mind the variety, hear the length," *Appl. Cognit. Psychol.*, vol. 11, no. 2, pp. 95–111, Apr. 1997.

[65] A. E. Rosenberg, "Automatic speaker verification: A review," *Proc. IEEE*, vol. 64, no. 4, pp. 475–487, 1976.

[66] B. S. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE*, vol. 64, no. 4, pp. 460–475, 1976.

[67] G. R. Doddington, "Speaker recognition—Identifying people by their voices," *Proc. IEEE*, vol. 73, no. 11, pp. 1651–1664, 1985.

[68] J. M. Naik, "Speaker verification: A tutorial," *IEEE Commun. Mag.*, vol. 28, no. 1, pp. 42–48, 1990.

[69] S. Furui, "Speaker-dependent-feature extraction, recognition and processing techniques," *Speech Commun.*, vol. 10, no. 5–6, pp. 505–520, Dec. 1991.

[70] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Process. Mag.*, vol. 11, no. 4, pp. 18–32, 1994.

[71] R. J. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: A feature-based approach," *IEEE Signal Process. Mag.*, vol. 13, no. 5, p. 58, 1996.

[72] S. Furui, "Recent advances in speaker recognition," *Pattern Recog. Lett.*, vol. 18, no. 9, pp. 859–872, Sept. 1997.

[73] J. P. Campbell Jr., "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[74] F. Bimbot, J. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, et al., "A tutorial on text-independent speaker verification," *EURASIP J. Appl. Signal Process.*, vol. 2004, pp. 430–451, Apr. 2004.

[75] A. F. Martin and M. A. Przybocki, "The NIST speaker recognition evaluations: 1996-2001," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Crete, Greece, pp. 1–5, 2001.

[76] C. S. Greenberg and A. F. Martin, "NIST speaker recognition evaluations 1996-2008," in *Proc. SPIE Defense, Security, and Sensing*, 2009, pp. 732411–732411-12.

[77] A. F. Martin and C. S. Greenberg, "The NIST 2010 speaker recognition evaluation." in *Proc. Interspeech*, 2010, pp. 2726–2729.

[78] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation–overview, methodology, systems, results, perspective," *Speech Commun.*, vol. 31, no. 2–3, pp. 225–254, June 2000.

[79] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, 2011.

[80] J. Markel, B. Oshika, and A. Gray, Jr., "Long-term feature averaging for speaker recognition," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 25, no. 4, pp. 330–337, 1977.

[81] K. Li and E. Wrench Jr., "An approach to text-independent speaker recognition with short utterances," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'83)*, 1983, pp. 555–558.

[82] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, et el., "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'03*), pp. 784–787.

[83] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[84] F. Beritelli and A. Spadaccini, "The role of voice activity detection in forensic speaker verification," in *Proc. Digital Signal Processing*, 2011, pp. 1–6.

[85] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 197–200, 2013.

[86] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[87] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer*, vol. 87, no. 4, pp. 1738, Apr. 1990.

[88] A. V. Oppenheim and R. W. Schafer, "From frequency to quefrency: A history of the cepstrum," *IEEE Signal Process. Mag.*, vol. 21, no. 5, pp. 95–106, 2004.

[89] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[90] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Crete, Greece, pp. 1–5, 2001.

[91] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[92] H. Boril and J. H. L. Hansen, "Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 18, no. 6, pp. 1379–1393, 2010.

[93] P. Matejka, O. Glembek, F. Castaldo, M. J. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, "Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'11*), pp. 4828–4831.

[94] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010.

[95] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech Lang. Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[96] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[97] F. Soong, A. Rosenberg, L. Rabiner, and B. Juang, "A vector quantization approach to speaker recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'85*), pp. 387–390.

[98] D. Burton, "Text-dependent speaker verification using vector quantization source coding," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 35, no. 2, pp. 133-143, 1987.

[99] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc. Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.

[100] A. E. Rosenberg and S. Parthasarathy, "Speaker background models for connected digit password speaker verification," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'96*), pp. 81–84.

[101] A. E. Rosenberg, J. DeLong, C. Lee, B. Juang, and F. K. Soong, "The use of cohort normalized scores for speaker verification." in *Proc. Int. Conf. Spoken Language Processing*, 1992, pp. 599–602.

[102] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proc. Eurospeech*, pp. 963–966, 1997.

[103] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[104] R. Kuhn, P. Nguyen, J. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "eigenvoices for speaker adaptation." in *Proc. Int. Conf. Spoken Language Processing*, 1998, pp. 1774–1777.

[105] P. Kenny, M. Mihoubi, and P. Dumouchel, "New MAP estimators for speaker recognition," in *Proc. Interspeech*, Geneva, Switzerland, pp. 2964–2967, 2003.

[106] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learning*, vol. 20, no. 3, pp. 273–297, Sept. 1995.

[107] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308–311, 2006.

[108] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'05)*, pp. 629–632.

[109] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition." in *Proc. Interspeech*, Pittsburgh, PA, pp. 1471–1474, 2006.

[110] W. M. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'02)*, pp. 161–164.

[111] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[112] P. Kenny and P. Dumouchel, "Disentangling speaker and channel effects in speaker verification," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'04)*, pp. 37–40.

[113] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[114] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 6, pp. 695–707, 2000.

[115] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Comput.*, vol. 11, no. 2, pp. 443–482, Feb. 1999.

[116] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.

[117] S. Roweis, "EM algorithms for PCA and SPCA," *Adv. Neural Inform. Process. Syst.*, vol. 10, no. 1, pp. 626–632, 1998.

[118] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," Tech. Rep. CRIM-06/08-13, CRIM, Montreal, Quebec, Canada, 2005.

[119] N. Dehak, P. Kenny, R. Dehak, O. Glembek, P. Dumouchel, L. Burget, V. Hubeika, and F. Castaldo, "Support vector machines and joint factor analysis for speaker verification," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'09)*, pp. 4237–4240.

[120] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification." in *Proc. Interspeech*, 2009, pp. 1559–1562.

[121] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE Int. Conf. Computer Vision*, 2007, pp. 1–8.

[122] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, 2011, pp. 249–252.

[123] E. Adar, "GUESS: A language and interface for graph exploration," in *Proc. ACM's Special Interest Group on Computer-Human Interaction*, 2006, pp. 791–800.

[124] N. Brummer, "Measuring, refining and calibrating speaker and language information extracted from speech," Ph.D. diss. Stellenbosch, Univ. Stellenbosch, 2010.

[125] Z. Wu, C. E. Siong, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition." in *Proc. Interspeech*, Portland, OR, pp. 1700–1703, 2012.

[126] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: a survey," *Speech Commun.*, vol. 66, pp. 130–153, Feb. 2015.

[127] G. Liu and J. H. L. Hansen, "An investigation into back-end advancements for speaker recognition in multi-session and noisy enrollment scenarios," *IEEE/ACM Trans. Audio, Speech Lang. Processing*, vol. 22, no. 12, pp. 1978–1992, 2014.

[128] S. Novoselov, T. Pekhovsky, K. Simonchik, and A. Shulipa, "RBM-PLDA subsystem for the NIST i-Vector Challenge," in *Proc. Interspeech*, Singapore, Sept. 14–18, 2014, pp. 378–382.

[129] A. K. Sarkar, D. Matrouf, P. Bousquet, and J. Bonastre, "Study of the effect of I-vector modeling on short and mismatch utterance duration for speaker verification," in *Proc. Interspeech*, Portland, OR, pp. 2662–2665, 2012.

[130] T. Hasan, R. Saeidi, J. H. Hansen, and D. A. van Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'13)*, pp. 7663–7667.

[131] R. Travadi, M. Van Segbroeck, and S. Narayanan, "Modified-prior i-vector estimation for language identification of short duration utterances," in *Proc. Interspeech*, pp. 3037–3041, 2014.

[132] G. S. Morrison, "Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio," *Aust. J. Foren. Sci.*, vol. 45, no. 2, pp. 173–197, Dec. 2013.

[133] V. Hautamäki, K. Lee, D. A. van Leeuwen, R. Saeidi, A. Larcher, T. Kinnunen, T. Hasan, S. O. Sadjadi, et al., "Automatic regularization of cross-entropy cost for speaker recognition fusion." in *Proc. Interspeech*, 2013, pp. 1609–1613.

[134] O. Ghahabi and J. Hernando, "i-Vector modeling with deep belief networks for multi-session speaker recognition," in *Proc. Odyssey:The Speaker and Language Recognition Workshop*, Joensuu, Finland, June 16–19, 2014, pp. 305–310.

[135] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'14)*, pp. 1695–1699.

[136] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep neural networks for extracting Baum-Welch statistics for speaker recognition," in *Odyssey: The Speaker and Language Recognition Workshop*, Joensuu, Finland.

[137] T. Hasan and J. H. L. Hansen, "Maximum likelihood acoustic factor analysis models for robust speaker verification in noise," *IEEE/ACM Trans. Audio, Speech Lang. Processing*, vol. 22, no. 2, pp. 381–391, 2014.

[138] C. S. Greenberg, D. Bansé, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, T. Kinnunen, A. F. Martin, A. McCree, et al., "The NIST 2014 speaker recognition i-vector machine learning challenge," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, pp. 224–230, 2014.

[139] G. Liu, C. Yu, A. Misra, N. Shokouhi, and J. H. Hansen, "Investigating state-of-the-art speaker verification in the case of unlabeled development data," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Joensuu, Finland, pp. 118–122, 2014.

[140] T. Hasan, S. O. Sadjadi, G. Liu, N. Shokouhi, H. Boril, and J. H. Hansen, "CRSS systems for 2012 NIST speaker recognition evaluation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'13)*, pp. 6783–6787.

[141] P. Kenny, T. Stafylakis, P. Ouellet, M. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'13)*, pp. 7649–7653.

[142] N. Scheffer and Y. Lei, "Content matching for short duration speaker recognition," in *Proc. Interspeech*, Joensuu, Finland, 2014, pp. 1317–1321.

[143] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[144] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech Lang. Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[145] T. Yamada, L. Wang, and A. Kai, "Improvement of distant-talking speaker identification using bottleneck features of DNN," in *Proc. Interspeech*, Lyon, France, 2013, pp. 3661–3664.

[146] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[147] A. Schmidt-Nielsen and T. H. Crystal, "Human vs. machine speaker identification with telephone speech." in *Proc. ICSLP*, 1998.

[148] D. O'Shaughnessy, *Speech Communication: Human and Machine*. India: Universities Press, 1987.

[149] J. Lindh and G. S. Morrison, "Humans versus machine: Forensic voice comparison on a small database of swedish voice recordings," in *Proc. Int. Congress of Phonetic Sciences (ICPhS)*, 2011, p. 4.

[150] A. Schmidt-Nielsen and T. H. Crystal, "Speaker verification by human listeners: Experiments comparing human and machine performance using the NIST 1998 speaker evaluation data," *Digital Signal Process.*, vol. 10, no. 1–3, pp. 249–266, Jan. 2000.

[151] V. Hautamäki, T. Kinnunen, M. Nosratighods, K. Lee, B. Ma, and H. Li, "Approaching human listener accuracy with modern speaker verification," in *Proc. Interspeech* 2010, Makuhari, Chiba, Japan, Sept. 26–30, 2010, pp. 1473–1476.

[152] R. Schwartz, J. P. Campbell, W. Shen, D. E. Sturim, W. M. Campbell, F. S. Richardson, R. B. Dunn, and R. Granville, "USSS-MITLL 2010 human assisted speaker recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'11)*, pp. 5904–5907.

[153] D. Ramos, J. Franco-Pedroso, and J. Gonzalez-Rodriguez, "Calibration and weight of the evidence by human listeners. The ATVS-UAM submission to NIST human-aided speaker recognition 2010," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'11)*, pp. 5908–5911.

[154] J. Kahn, N. Audibert, S. Rossato, and J. Bonastre, "Speaker verification by inexperienced and experienced listeners vs. speaker verification system," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'11)*, pp. 5912–5915.

[155] C. S. Greenberg, A. F. Martin, G. R. Doddington, and J. J. Godfrey, "Including human expertise in speaker recognition systems: Report on a pilot evaluation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'11)*, pp. 5896–5899.

[156] W. Shen, J. Campbell, D. Straub, and R. Schwartz, "Assessing the speaker recognition performance of naïve listeners using mechanical turk," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'11)*, pp. 5916–5919.

[157] R. G. Hautamäki, V. Hautamäki, P. Rajan, and T. Kinnunen, "Merging human and automatic system decisions to improve speaker recognition performance," in *Proc. Interspeech*, pp. 2519–2523, 2013.

[158] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, and A. Laukkanen, "Comparison of human listeners and speaker verification systems using voice mimicry data," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Joensuu, Finland, 2014, pp. 137–144.

[159] K. J. Han, M. K. Omar, J. Pelecanos, C. Pendus, S. Yaman, and W. Zhu, "Forensically inspired approaches to automatic speaker recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'11)*, pp. 5160–5163.

[160] NIST OSAC—The Organization of Scientific Area Committees. (2015). [Online]. Available: http://www.nist.gov/forensics/osac.cfm

[161] NIST OSAC. (2015). NIST forensic science publications. [Online]. Available: http://www.nist.gov/forensics/publications.cfm

[SP]