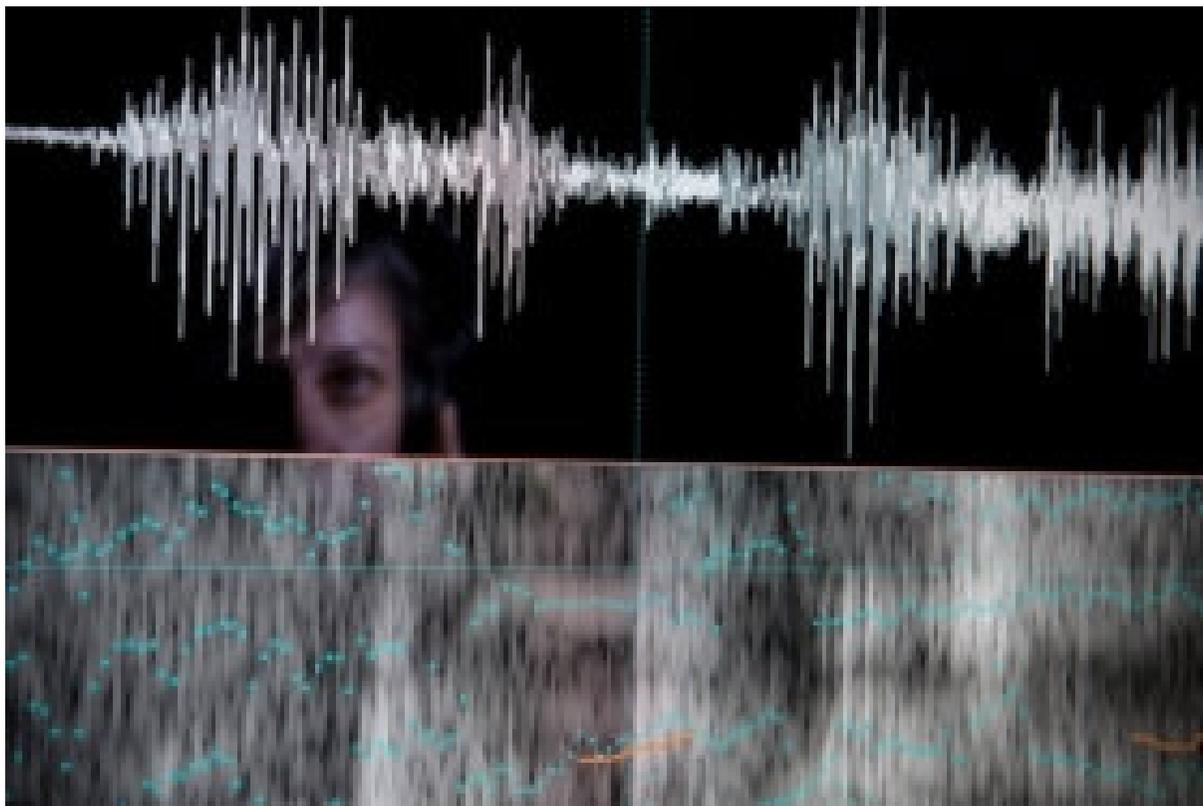


POLICY & ETHICS

Voice Analysis Should Be Used with Caution in Court

Although voice recognition is often presented as evidence in legal cases, its scientific basis can be shaky

By Michele Catanzaro, Elisabetta Tola, Philipp Hummel, Astrid Viciano on January 25, 2017

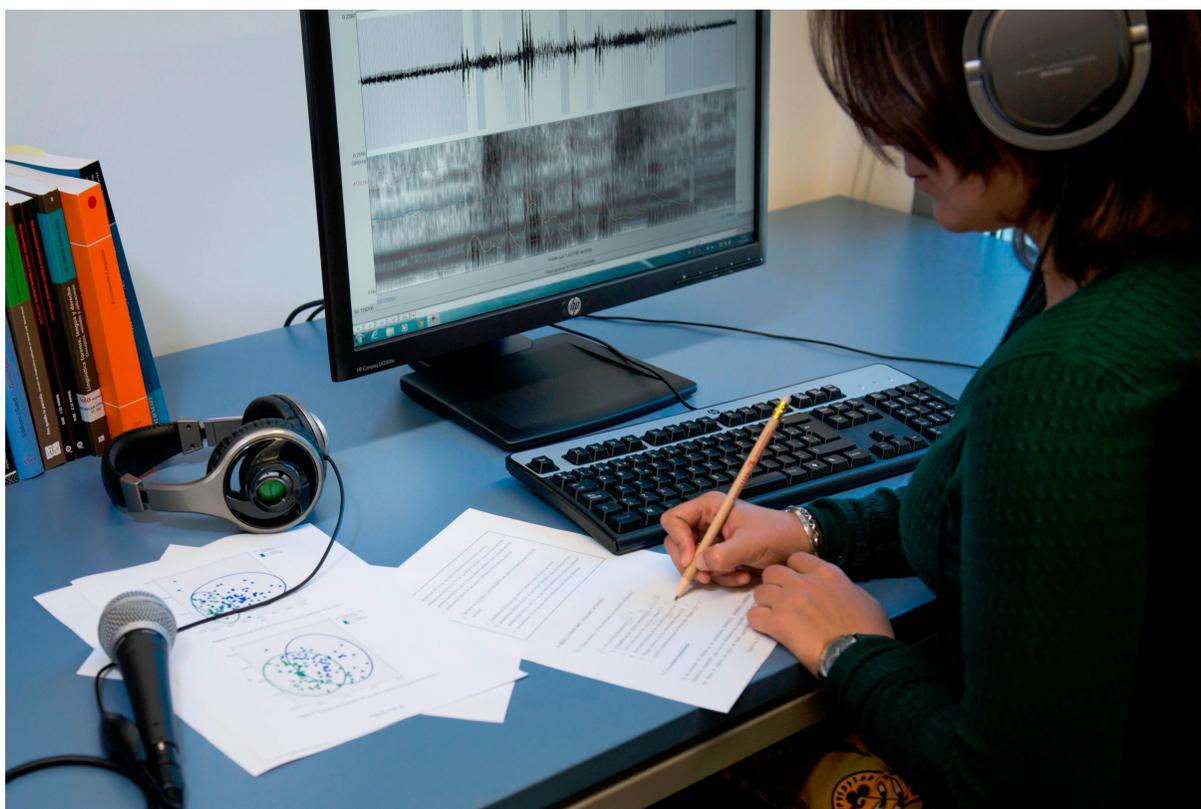


Credit: Gianluca Battista

Voice recognition has started to feature prominently in intelligence investigations. Examples abound: When ISIS released the video of journalist James Foley being beheaded, experts from all over the world tried to identify the masked terrorist known as Jihadi John by analyzing the sound of his voice. Documents disclosed by Edward Snowden revealed that the U.S. National Security Agency has analyzed and extracted the content of millions of phone conversations. Call centers at banks are using voice biometrics to authenticate users and to identify potential fraud.

But is the science behind voice identification sound? Several articles in the scientific literature have warned about the quality of one of its main applications: forensic phonetic expertise in courts. We have compiled two dozens judicial cases from around the world in which forensic phonetics were controversial. Recent figures published by INTERPOL indicate that half of forensic experts still use audio techniques that have been openly discredited.

For years, movies and television series like CSI paint an unrealistic picture of the “science of voices.” In the 1994 movie *Clear and Present Danger* an expert listens to a brief recorded utterance and declares that the speaker is “Cuban, aged 35 to 45, educated in the [...] eastern United States.” The recording is then fed to a supercomputer that matches the voice to that of a suspect, concluding that the probability of correct identification “is 90.1 percent.” This sequence sums up a good number of misimpressions about forensic phonetics, which have led to errors in real-life justice. Indeed, that movie scene exemplifies the so-called “CSI effect”—the “phenomenon in which judges hold unrealistic expectations of the capabilities of forensic science,” says Juana Gil Fernandez, a forensic speech scientist at the Consejo Superior de Investigaciones Cientificas (Superior Council of Scientific Investigations) in Madrid, Spain.



A voice analyst at work in a speech forensics laboratory in Spain. Credit: Gianluca Battista

In 1997 the French Acoustical Society issued a public request to end the use of forensic voice science in the courtroom. The request was a response to the case of Jerome Prieto, a man who spent 10 months in prison because of a controversial police investigation that erroneously identified Prieto's voice in a phone call claiming credit for a car bombing. There are plenty of troubling examples of dubious forensics and downright judicial errors, which have been documented by Hearing Voices, a science journalism project on forensic science carried out by the authors of this article in 2015 and 2016.

It's impossible to know how many voice investigations are conducted each year because no country keeps a register, but Italian and British experts estimate that in their respective countries there must be hundreds per year. The process usually involves at least one of the following tasks: transcribing a recorded voice, comparing an intercepted voice to that of a suspect, putting the suspect's voice in a lineup of different voices, profiling a speaker based on dialect or language spoken, interpreting noises or verifying the authenticity of a recording.

The recorded fragments subject to analysis can be phone conversations, voice mail, ransom demands, hoax calls and calls to emergency or police numbers. One of the main hurdles voice analysts have to face is the poor quality of recorded fragments. "The telephone signal does not carry enough information to allow for fine-grained distinctions of speech sounds. You would need a band twice as broad to tell certain consonants apart, such as *f* and *s* or *m* and *n*," said Andrea Paoloni, a scientist at the Ugo Bordononi Foundation and the foremost forensic phoneticist in Italy until his death in November 2015. To make things worse, recorded messages are often noisy, short and can be years or even decades old. In some cases, simulating the context of a phone call can be particularly challenging. Imagine recreating a call placed in a crowded movie theater, using an old cell phone or one made by an obscure foreign brand.

In a 1994 article in the Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, the expert Hermann Künzel estimated that 20 percent of the fragments analyzed by the German federal police contained

only 20 seconds of usable voice. Nevertheless, many forensic experts are willing to work on sound excerpts that are of extremely low quality. In the famous case of George Zimmerman, the neighborhood watch coordinator who in 2012 shot the young African American Trayvon Martin in Sanford, Fla., one expert stated that he could extract a voice profile and even interpret the screams that could be heard in the background of an emergency call.

Unfortunately, these errors are not isolated exceptions. A survey published in June 2016 in the journal *Forensic Science International* by INTERPOL, the international organization that represents the police forces of 190 countries, showed that half of the respondents (21 out of 44)—belonging to police forces from all over the world—employ techniques that have long been known to have shaky scientific grounds. One example is the simplest and oldest voice recognition method: unaided listening, leading to subjective judgement by a person with a “trained ear” or even to the opinion of victims and witnesses.

In 1992 Guy Paul Morin, a Canadian, was sentenced to life imprisonment for the rape and murder of a nine-year-old girl. In addition to other evidence, the victim’s mother said she had recognized Morin’s voice. Three years later, a DNA test exonerated Morin as the murderer. This kind of mistake is not surprising. In a study published in “Forensic Linguistics” in 2000, a group of volunteers who knew one another listened to anonymous recordings of the voices of various members of the group. The rate of recognition was far from perfect, with a volunteer failed to recognize even his own voice.

This does not imply, however, that automated methods are always more accurate than the human ear. Actually, the first instrumental technique used in forensic phonetics has been denied any scientific basis for a number of years, even though some of its variations are still in use, according to the INTERPOL report. We are referring to *voiceprinting*, or spectrogram matching, in which a human observer compares the spectrograms of a word pronounced by the suspect with the same word pronounced by an intercepted speaker. A spectrogram is a graphic representation of the frequencies of the voice spectrum, as they change in time while a word or sound is produced.

Voiceprinting gained notoriety with the 1962 publication of a paper by Lawrence G. Kersta, a scientist at Bell Labs, in the journal *Nature*. But in 1979, a report by the National Science Foundation declared that voiceprints had no scientific basis: the authors wrote that spectrograms are not very good at differentiating speakers and they are too variable. “Spectrogram matching is a hoax, pure and simple. Comparing images is just as subjective as comparing sounds,” said Paoloni. Nevertheless, the technique still maintains a lot of credibility. In 2001, after DNA testing, David Shawn Pope of the U.S. was acquitted of aggravated sexual assault after spending 15 years in prison. The conviction was partly based on voiceprint analysis.

SOUNDS INTERPRETED DIFFERENTLY

The scientific community has explicitly discredited some voice analysis techniques, but is still far from reaching a consensus on the most effective method for identifying voices. There are two schools of thought, says Juana Gil Fernandez. “Linguists support the use of semi-automatic techniques that combine computerized analysis and human interpretation, while engineers attribute more importance to automatic systems.”

Semi-automatic techniques are still the most widely used. These methods are called “acoustic-phonetic” because they combine measurements obtained by listening (acoustic) with the output of automated sound analysis (phonetics). Experts who rely on acoustic-phonetic methods usually start by listening to the recording and transcribing it into phonetic transcription. Then they identify a number of features of the voice signal. The high level features are linguistic: for example, a speaker’s choice of words (lexicon), sentence structure (syntax), the use of filler words such as “um” or “like,” and speech difficulties such as stuttering. The sum of these characteristics is the idiolect—a person’s specific, individual way of speaking. Other high level qualities are the so-called suprasegmental features: voice quality, intonation, number of syllables per second and so on.

Lower-level characteristics, or segmental features, mostly reflect voice physiology, and are better measured with specific software. One basic feature is the fundamental

frequency. If the voice signal is divided into segments a few milliseconds long, each segment will contain a vibration with an almost perfectly periodic waveform. The frequency of this vibration is the fundamental frequency, which corresponds to the vibration frequency of the vocal folds, and contributes to what we perceive as the timbre or tone of a specific voice. The average fundamental frequency of an adult male is about 100 hertz, and that of an adult female is about 200 hertz. It can be hard to use this feature to pin down a speaker. On the one hand, it varies very little between different speakers talking in the same context. On the other hand, the fundamental frequency of the same speaker changes dramatically when he or she is angry, or shouting to be heard over a bad phone line.

Other segmental features commonly measured are vowel formants. When we produce a vowel, the vocal tract (throat and oral cavity) behaves like a system of moving pipes with specific resonances. The frequencies of these resonances (called the formants) can be plotted in a graph that represents a specific “vowel space” for each speaker, and the graph can be compared to that of other speakers.

In spite of its popularity, the acoustic-phonetic method raises some issues. Because it is semi-automatic, it leaves margin to subjective judgement, and sometimes experts working on the same material using a similar technique can reach discordant conclusions. In addition, there are very few data on the range and distribution in the general population of phonetic features other than the fundamental frequency. For these reasons, the most rigorous experts say that we can never be sure of the identity of a speaker based on voice alone. At most, we can say that two voices are compatible.

AUTOMATED SYSTEMS CAN PRODUCE FALSE POSITIVES

In the 1990s a new system that minimized human judgment started to gain popularity: automatic speaker recognition. In ASR the recordings are processed by software that extracts features from the signal, categorizes them and matches them to the features in a voice databank. Most algorithms work by dividing the signal into brief time windows and extracting the corresponding spectra of frequencies. The spectra then undergo mathematical transformations that extract parameters, called

cepstral coefficients, related to the geometric shape of the vocal tract. Cepstral coefficients provide a model of the speaker's vocal tract shape. "What we do is very different from what linguists do," says Antonio Moreno, vice president of Agnitio, the Spanish company that produces Batvox, the most widely used ASR system, according to INTERPOL. "Our system is much more precise, is measurable and can be reproduced: two different operators will get the same result from the system."

Linguists disagree. "The positive side of ARS is that it needs less human input.... The negative side is that cepstral coefficients reflect the geometry of the human vocal tract, but we are not too different from one another, so the system tends to make false hits," says Peter French from the University of York, president of the International Association for Forensic Phonetics and Acoustics (IAFPA) and director of J.P. French Associates, the main forensic phonetics company in the U.K. "I believe that automatic systems should be combined with human intervention," French says.

Other experts are more extreme in their criticism: "At the moment ASR does not have a theoretical basis strong enough to justify its use in real-life cases," states Sylvia Moosmuller, an acoustic scientist at the Austrian Academy of Sciences. One of the main reasons for skepticism is the fact that most ASR algorithms are trained and tested on a voice database from the U.S. National Institute of Standards and Technology (NIST). The database is an international standard, but it includes only studio recordings of voices that fail to approximate the complexity of real life, with speakers using different languages, communication styles, technological channels and so on.

"In fact, what the program is modeling is not a voice, but a session, made up of voice, communication channel and other variables," Moreno says. At the beginning, voice verification analysts tried to replicate the context in which a voice had been recorded. But about 10 years ago they changed approach, and instead resorted to algorithms that reduced the impact of recording conditions, called compensation techniques. "In the NIST database, the same speaker is recorded through many different channels, and many different speakers are recorded through the same channel", Moreno explains. "Compensation techniques are tested on this dataset, and allow us to disentangle the speaker's characteristics from that of the session." In other words, a

program trained with this method should be able to identify the same speaker in two different phone calls, one placed by landline, for example, and the other by cell phone.

Moreno believes that automatic speaker identification “is more than ready to produce valid results and improve the reliability of forensic evaluations.” However, he admits that ASR “is one of the many techniques available to experts, and the techniques complement each other: the more advanced labs have interdisciplinary groups.”

The main problem with ASR may lie not in the software itself but in the person using it. “It takes a voice scientist. You cannot just place any operator in front of a computer.... These programs are like airplanes: you can buy a plane in one day, but you cannot learn how to fly in three weeks,” says Didier Meuwly, of the Netherlands Forensic Institute. Yet companies sell as much as possible, and they end up selling software to customers who are not experts in forensic voice matching, says Geoffrey Stewart Morrison, a professor of linguistics at the University of Alberta, Canada. Agnitio offers a three-year course, but so far only 20 to 25 percent of the hundreds of Batvox users have completed it. The Batvox tool can cost up to 100,000 Euros.

MODERN STATISTICAL ANALYSES NEEDED

.....

Irrespective of the analysis method, forensic phonetics suffers from an even deeper scientific problem. Overall, the discipline has not gone through the paradigm shift in the statistical approach to data that more advanced techniques, such as forensic DNA testing, have already adopted: the shift to Bayesian statistics.

One example of this approach is presented by Morrison, the flag bearer of Bayesian statistics in forensic phonetics and a co-author of the INTERPOL study. “Imagine we found a size 9 shoe print at a crime scene, and we have a suspect who wears size 9 shoes. In another case we find a size 15 shoe print, and the suspect wears a size 15. In the second case, the evidence against the suspect is stronger, because a size 15 is less common than a size 9,” says Morrison. In other words, it’s not enough to measure the similarity between two shoe prints (or two voices, or two DNA samples). Analysts also have to take into account how typical those footprints (or voices, or DNA) are.

For voice, the problem can be framed as follows: If a suspect and a criminal are the same person, how likely is the similarity between the two voices? And if they are not the same person, how likely is the similarity? The ratio of these two probabilities is called the likelihood ratio, or strength of evidence. The higher the strength of evidence (for example, for voices that are very similar and very atypical), the stronger the evidence.

A higher or lower likelihood ratio can increase or diminish the likelihood of culpability, but the probability is also dependent on other cues and evidence, forensic and not. As is typical of Bayesian statistics, the probability is not calculated once and for all, but is constantly adjusted as new evidence is discovered.

In the guidelines for forensic science published in June 2015, the European Network of Forensic Science Institutes recommends the use of a Bayesian framework, and especially of the likelihood ratio. However, according to the INTERPOL report, only 18 of the 44 experts surveyed had made the switch.

One serious obstacle interferes with the application of Bayesian statistics: It is difficult to estimate how typical a voice is, because there are no statistical norms on the distribution of voice features. “If you have a database of two million finger prints you can be quite confident of the reliability of your estimates, but voice databases are much smaller,” said Paoloni. For example, the DyViS databank used in the U.K. includes 100 male speakers, most of them educated at Cambridge. Moreno is certain that some police databanks, which are not public, contain thousands of voices, and that some organizations have databases reaching hundreds of thousands of speakers.

“In the era of big data, the most reasonable thing to do would be to set up a corpus with a large amount of data,” modeled on the platforms that provide online services, said Paoloni. Given that there is nothing similar, Morrison’s recipe is to collect recordings of speakers within populations relevant for each case, based on demographic features (gender, language, dialect and so on) and speaking style (tired, excited, sleepy) and more. The problem, however, “is that many laboratories say that they don’t have any kind of database,” according to Daniel Ramos, a scientist at the

Universidad Autónoma of Madrid who also collaborates with a Spanish police force, the Guardia Civil.

Our investigation into the state of the art of forensic phonetics has shown some limitations of the science of voice identification, and suggests that the results of its application should be considered with extreme caution. “In my opinion, nobody should be condemned because of a voice,” concluded Paoloni. “*In dubio pro reo*--when in doubt, judge in favor of the accused. With voice, the likelihood of error is too high for a judge to ever be able to state that someone is guilty ’beyond any reasonable doubt.’”

This article originally appeared in *Le Scienze*, and is translated and adapted with permission. It was developed with the support of Journalismfund.eu.

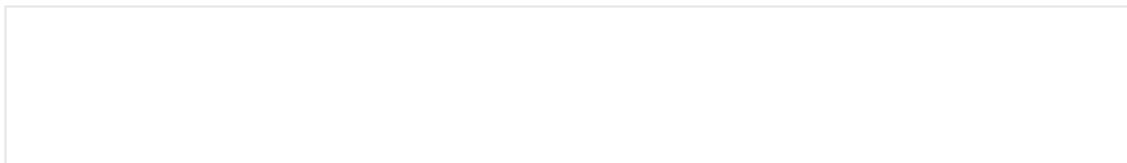
FURTHER READING

INTERPOL Survey of the Use of Speaker Identification by Law Enforcement Agencies. Morrison G. S., Sahito F. H., Jardine G., Djokic D., Clavet S., Berghs S., Goemans Dorny C., in *Forensic Science International*, Vol. 263, pp. 92-100, June 2016. <http://dx.doi.org/10.1016/j.forsciint.2016.03.044>.

Forensic Speaker Recognition. Meuwly D., in Wiley Encyclopedia of Forensic Science, 2009.

Interpreting Evidence: Evaluating Forensic Science in the Courtroom. Robertson B., Vignaux G.A., John Wiley and Sons, 1995.

The website of Hearing Voices, with cases, techniques and legislation:
<http://formicablu.github.io/hearingvoices/en>.



ADVERTISEMENT | REPORT AD

ABOUT THE AUTHOR(S)

Michele Catanzaro

Michele Catanzaro is a physicist and freelance journalist. He is based in Barcelona and writes for "Nature," El Periodico and other media.

Elisabetta Tola

Elisabetta Tola is a microbiologist, journalist and science communicator. She is the director of the communication agency FormicaBlu in Bologna, Italy.

Philipp Hummel

Philipp Hummel is a physicist and freelance journalist. He lives in Berlin.

Astrid Viciano

Astrid Viciano is a physician and writes for the weekend science supplement of the Suddeutsche Zeitung in Munich, Germany.

Scientific American is part of Springer Nature, which owns or has commercial relations with thousands of scientific publications (many of them can be found at www.springernature.com/us). Scientific American maintains a strict policy of editorial independence in reporting developments in science to our readers.

© 2017 SCIENTIFIC AMERICAN, A DIVISION OF NATURE AMERICA, INC.

ALL RIGHTS RESERVED.