

---

# Approaches to Speaker Recognition: A Primer

Anil Alexander & J Keith McElveen

Clarifying Technologies Ltd., Oxford, United Kingdom

## Overview

This article seeks to give the reader a basic overview of the application of speaker recognition to forensic tasks. We discuss how an individual's speech is related to his or her identity, where speaker recognition can be applied and what the common approaches to speaker recognition are. We discuss the considerations in performing speaker recognition in the typically uncontrolled recording conditions common to law enforcement and forensic tasks. These conditions present challenging problems to both human and computer-based approaches. Automatic speaker recognition techniques (i.e., the computer-based recognition of speakers) are discussed.

Speaker recognition methods (human and computer-based) can successfully be applied to investigative tasks. The advantages of using automatic, computer-based speaker recognition, such as the objective analysis of large amounts of audio data and reducing the dependence on human expertise in the language under consideration, are also presented.

## How is speech related to identity?

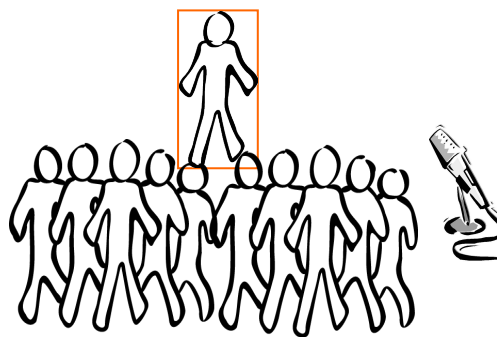
Human speech is a complex signal and is influenced by several physiological, psychological and environmental factors. The acoustic signal we produce when we speak is determined by, among other things, the physiology of the vocal tract and articulators, age, gender, native

language and dialect acquisition and regional traits. These factors are constantly changing and influencing the way we communicate. They help shape the distinctive 'identity' of the speech of different individuals.

## How is voice different from other human biometric characteristics?

Voice differs from other biometric characteristics of human beings, like fingerprints or DNA, as it changes over time; depends on the health and emotional state of the speaker; and can be altered, at will, by the speaker (disguise or impersonation). The variability in the human voice makes it a less powerful biometric than the so called unique characteristics like fingerprints or DNA. Voice is also not considered to be an 'intrusive' biometric, as it does not involve direct contact with the individual (as in the case of DNA, retinal scans and fingerprints).

## When and how is speaker recognition useful?



Speaker recognition can be employed in situations involving courts and police

---

investigative agencies, as well as private organizations. Applications of this technology include:

- Courts, to settle a challenge about the alleged speaker at the source of a questioned recording.
- Wire-tapping or body-wire techniques used by the police to collect information from or about suspected persons (especially in the absence of simultaneous video recordings).
- Erroneous ear-witness identification by naive listeners.
- Threats and warnings received by the police and investigative agencies.
- Counter-terrorism investigations with audio and video threats and propaganda recordings
- Analysis of surveillance audio recordings for criminal intelligence information
- Security portal or checkpoint verification
- Private organizations and companies for internal inquiry in cases of harassment allegations, corruption, etc.

### **What is speaker recognition and what are the different approaches to carry it out?**

‘Speaker recognition’ is an umbrella term to include all the many different tasks of differentiating between people based on their voices. Speaker recognition is traditionally divided into *speaker verification* and *speaker identification*. Speaker verification involves a single speaker whose voice is compared with that of the test utterance and the identity is ‘verified’. In speaker identification, however, the voice in the

test utterance is compared with a large group of speakers, and the individual from whom it is most probable to have come from, is ‘identified’. A useful way of differentiating between the two is that speaker verification answers the question ‘Is this John?’ while speaker identification answers the question ‘Who is this?’

Speaker recognition outside of a controlled environment is a difficult task, typically requiring trained and careful analysis by both humans and by machines. The approaches commonly used in speaker recognition include *aural-perceptual*, *auditory instrumental* (both of which are performed by humans), and *automatic* (performed by machine) approaches.

**Aural-perceptual** (also known as auditory analysis) methods basically rely on careful listening to recordings by trained phoneticians, where the perceived differences in the speech samples are used to estimate the extent of similarity between voices. Some of the features considered in aural-perceptual analysis include dialect and ‘sociolect’, speech defects and voice quality. In addition to these, the phonetician may listen for differences in the rate of speech and intonation, pauses, articulation and diction. Also, higher level characteristics, such as idiomatic and linguistic characteristics, as well as the prosody of the speech, are indicative of a speaker’s identity. Almost all of the features that the forensic phonetician measures are at a level above spectral characteristics. With this approach, a subjective probability of the similarity of the two voices can be established. The aural-perceptual approach has its limitations, and, in

---

traditional phonetic analysis, it is used mainly to extract features of interest, which are then analysed using the auditory-instrumental approach.

The **auditory-instrumental** approach involves measuring various acoustic parameters such as average fundamental frequency (F0), vowel formants, pitch contour, spectral energy, etc. and then performing statistical analyses of the data. The means and variances of these parameters are compared. The 'voiceprint' technique, which is often discussed in the mainstream media and can be considered one of the many auditory-instrumental approaches.

The use of spectrograms, also known as 'voiceprints', for speaker recognition has come under considerable criticism from the scientific community in recent years. In 1976, at the request of the FBI, the National Academy of Sciences investigated and concluded that there were serious technical uncertainties concerning the use of spectrograms and that they should be used only with utmost caution in forensic applications. Federal government agencies do not base testimony on voiceprints, but do use it as an investigative aid. Most states do not accept their use in court [Schwartz, 2006].

**Forensic automatic speaker recognition** is an established term used when automatic or computer-based speaker recognition methods are adapted to forensic applications. In automatic speaker recognition, the statistical models of acoustic parameters of the speaker's voice and the acoustic parameters of questioned recordings are compared. The quantified degree of similarity between speaker-dependent features extracted from the questioned

recording (or trace) and speaker-dependent features extracted from the recorded speech of a known suspect, represented by his/her model, is calculated in order to evaluate the evidence.

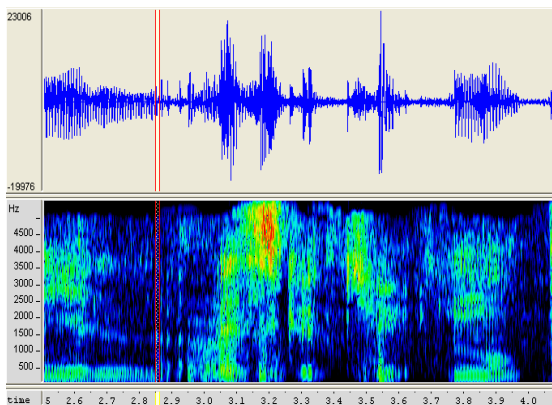
The advantage of using computerized, automatic systems for speaker recognition is that large amounts of audio recordings can be processed with simple, objective measures of the similarity between subject and questioned recordings. Nowadays, state-of-the-art automatic speaker recognition systems perform very well in discriminating between voices of speakers under controlled recording conditions. Although automatic speaker recognition has been demonstrated to be highly accurate in controlled conditions, a degradation of this performance can be observed in adverse conditions.

### **What kind of features should be chosen for speaker recognition?**

The human speech-production mechanism is driven by the acoustic excitation of the vocal tract by airflow from the lungs, passing through the vocal cords. The sound produced can be classified into voiced and unvoiced, which differ in the kind of excitation of the vocal tract required to produce them. The response of the vocal tract to this excitation contributes to the distinctiveness of a speaker's voice. It is important to note that the vocal tract and speech production mechanism are only some of the aspects that contribute to individuality of speech. As discussed earlier, dialectal, sociolectal, idiomatic and linguistic features provide important clues to the identity of the speaker. Phonetic, linguistic and semi-automatic methods concentrate on these aspects of

speech while engineering or computer-based automatic methods traditionally focus on the vocal tract and speech production mechanism.

The likely shape of the vocal tract can be approximately estimated from the analysis of the spectral shape of the voice signal. In automatic speaker recognition, coefficients representing the sounds, taking into consideration the vocal tract shape and excitation, are parameterized and used as features. Small sections or windows (of the order of a few hundredths of a second) of the sound are used to analyze the speech and to extract features. One such section is illustrated below:

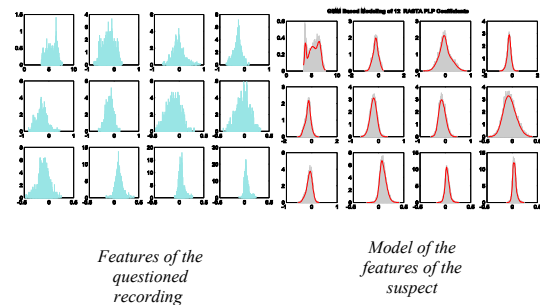


Ideally, the features chosen for speaker recognition must satisfy the following criteria [Wolf, 1972]:

- *Have lower within-speaker (within-source) variability and relatively higher between speakers (between-sources) variability.*
- *Be stable over time.*
- *Be difficult to disguise or mimic.*
- *Be robust to transmission and noise.*
- *Be relatively easy to extract and measure, and should occur frequently in the speech samples.*

Using these features, models are created

for each speaker and these models represent their speech. These representations of speech are complex statistical and mathematical models. Test recordings are then compared with these models to obtain a ‘match score’ which indicates how close the two recordings are. Modeling techniques include, ‘vector quantization’, Gaussian mixture modeling (GMM), hidden Markov models (HMM), Support Vector Machines (SVM), neural networks, etc. The National Institute of Standards [NIST], USA, conducts regular evaluations of speaker recognition technologies, and in recent years, GMM has been the leading and most widely used algorithm which models the probabilities of observing these features.



Automatic speaker recognition has been demonstrated to be highly accurate in controlled conditions. This performance is better than human recognition (with normal listeners unfamiliar with the test speakers) in matched test and comparison recording conditions [Alexander et al, 2005]. In addition, large volumes of audio recordings can be analyzed in a short time using automatic recognition compared to human analysis.

**Are some speakers more difficult to identify than others?**

Yes, some speakers are easy to classify and the others more difficult. *Doddington et al (1998)*, proposed an

---

interesting classification of speakers into different categories on the basis of the difficulty of recognizing them using an automatic system. Each category of speakers has been assigned to a different animal.



- **Sheep:** They represent the majority of the population and are the default speaker type. Most automatic systems for verification and recognition perform reasonably well for them with a low number of false acceptances and false rejections.
- **Goats:** The goats represent a section of the speakers who are particularly difficult to recognize. The goats account for a disproportionate share of false rejections.
- **Lambs:** The speakers whose voices are most easily imitated by another speaker are called lambs, and the presence of lambs tends to increase the false acceptance rate and can represent a system weakness.
- **Wolves:** The wolves are the speakers whose voices are complementary to the lambs, in the sense that the characteristic features of their voices are exceptionally similar to the

features of other speakers. Their speech is often likely to be recognized as that of some other speaker. Wolves would account for a disproportionately large share of the false acceptances. Like the lambs, these speakers also represent a potential system weakness.

### **Why is automatic speaker recognition an attractive option for forensic tasks?**

- Automatic speaker recognition has been shown to perform with a high accuracy in controlled recording conditions.
- Although there is a decrease in accuracy in uncontrolled recording conditions, with compensation for these conditions, automatic speaker recognition still provides useful information that can help with case investigations.
- Forensic cases often include large amounts of audio data which are difficult to evaluate within the time constraints of an investigation. Automatic systems are useful in these situations.
- Automatic recognition can complement the traditional aural-perceptual and semi-automatic speaker recognition techniques used in forensic speaker recognition. For instance, the automatic system can be used for a quick quality control check of the results from traditional analysis methods.
- It is less dependent on language, dialect and its nuances, and is useful when there is a need to analyze speech in languages

---

where sufficient expertise is unavailable.

## **Conclusion**

- Speaker recognition is a challenging problem for both human and automatic systems in the particularly adverse recording conditions that are encountered in law-enforcement and forensic investigations.
- Speaker recognition technology (human and computer-based) can be applied to investigative tasks, but should be used for evidential purposes with utmost care.
- Automatic, computer-based speaker recognition gives an objective way of quickly processing large amounts of audio data and reducing the dependence on the often difficult to find human expertise in the language under consideration.

## **Contact Information:**



[www.clarifyingtech.com](http://www.clarifyingtech.com)  
info@clarifyingtech.com