
# Speaker Recognition

## Introduction

Speaker, or voice, recognition is a biometric modality that uses an individual's voice for recognition purposes. (It is a different technology than "speech recognition", which recognizes words as they are articulated, which is not a biometric.) The speaker recognition process relies on features influenced by both the physical structure of an individual's vocal tract and the behavioral characteristics of the individual.

A popular choice for remote authentication due to the availability of devices for collecting speech samples (e.g., telephone network and computer microphones) and its ease of integration, speaker recognition is different from some other biometric methods in that speech samples are captured dynamically or over a period of time, such as a few seconds. Analysis occurs on a model in which changes over time are monitored, which is similar to other behavioral biometrics such as dynamic signature, gait, and keystroke recognition.

## History

Speaker verification has co-evolved with the technologies of speech recognition and speech synthesis because of the similar characteristics and challenges associated with each. In 1960, Gunnar Fant, a Swedish professor, published a model describing the physiological components of acoustic speech production, based on the analysis of x-rays of individuals making specified phonic sounds.[1] In 1970, Dr. Joseph Perkell used motion x-rays and included the tongue and jaw[1] to expand upon the Fant model. Original speaker recognition systems used the average output of several analog filters to perform matching, often with the aid of humans "in the loop".[2,3,4,5,6] In 1976, Texas Instruments built a prototype system that was tested by the U.S. Air Force and The MITRE Corporation.[1,7] In the mid 1980s, the National Institute of Standards and Technology (NIST) developed the NIST Speech Group to study and promote the use of speech processing techniques. Since 1996, under funding from the National Security Agency, the NIST Speech Group has hosted yearly evaluations, the NIST Speaker Recognition Evaluation Workshop, to foster the continued advancement of the speaker recognition community.[8]


## Approach

The physiological component of voice recognition is related to the physical shape of an individual's vocal tract, which consists of an airway and the soft tissue cavities from which vocal sounds originate.[1] To produce speech, these components work in combination with the physical movement of the jaw, tongue, and larynx and resonances in the nasal passages. The acoustic patterns of speech come from the physical characteristics of the airways. Motion of the mouth and pronunciations are the behavioral components of this biometric.

There are two forms of speaker recognition: text dependent (constrained mode) and text independent (unconstrained mode). In a system using "text dependent" speech, the individual presents either a fixed (password) or prompted ("Please say the numbers '33-54-63'") phrase that is programmed into the system and can improve performance especially with cooperative users. A "text independent" system has no advance knowledge of the presenter's phrasing and is much more flexible in situations where the individual submitting the sample may be unaware of the collection or unwilling to cooperate, which presents a more difficult challenge.[9]

Speech samples are waveforms with time on the horizontal axis and loudness on the vertical access. The speaker recognition system analyzes the frequency content of the speech and compares characteristics such as the quality, duration, intensity dynamics, and pitch of the signal.[1]

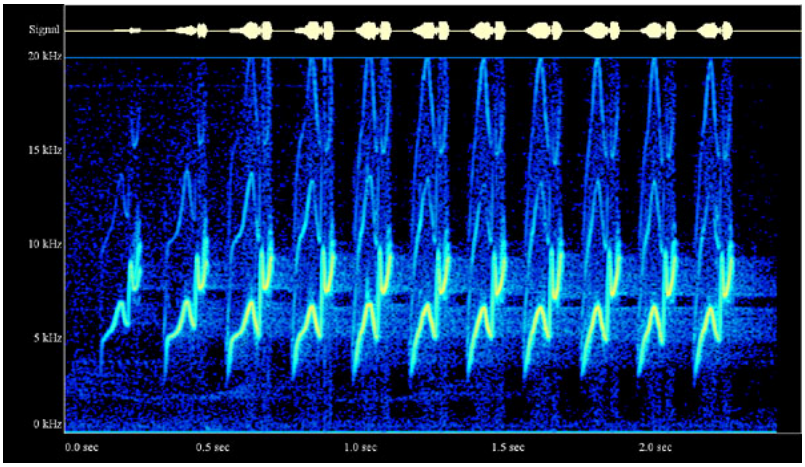


Figure 1: Voice Sample: The voice input signal (top of image) shows the input loudness with respect to the time domain. The lower image (blue) depicts the spectral information of the voice signal. This information is plotted by displaying the time versus the frequency variations.[10]




In “text dependent” systems, during the collection or enrollment phase, the individual says a short word or phrase (utterance), typically captured using a microphone that can be as simple as a telephone. The voice sample is converted from an analog format to a digital format, the features of the individual’s voice are extracted, and then a model is created. Most “text dependent” speaker verification systems use the concept of Hidden Markov Models (HMMs), random based models that provide a statistical representation of the sounds produced by the individual. The HMM represents the underlying variations and temporal changes over time found in the speech states using the quality/duration/intensity dynamics/pitch characteristics mentioned above.[9] Another method is the Gaussian Mixture Model, a state-mapping model closely related to HMM, that is often used for unconstrained “text independent” applications. Like HMM, this method uses the voice to create a number of vector “states” representing the various sound forms, which are characteristic of the physiology and behavior of the individual.[1] These methods all compare the similarities and differences between the input voice and the stored voice “states” to produce a recognition decision.

After enrollment, during the recognition phase, the same quality/duration/loudness/pitch features are extracted from the submitted sample and compared to the model of the claimed or hypothesized identity and to models from other speakers. The other-speaker (or “anti-speaker”) models contain the “states” of a variety of individuals, not including that of the claimed or hypothesized identity.[9] The input voice sample and enrolled models are compared to produce a “likelihood ratio,” indicating the likelihood that the input sample came from the claimed or hypothesized speaker. If the voice input belongs to the identity claimed or hypothesized, the score will reflect the sample to be more similar to the claimed or hypothesized identity’s model than to the “anti-speaker” model.[9]

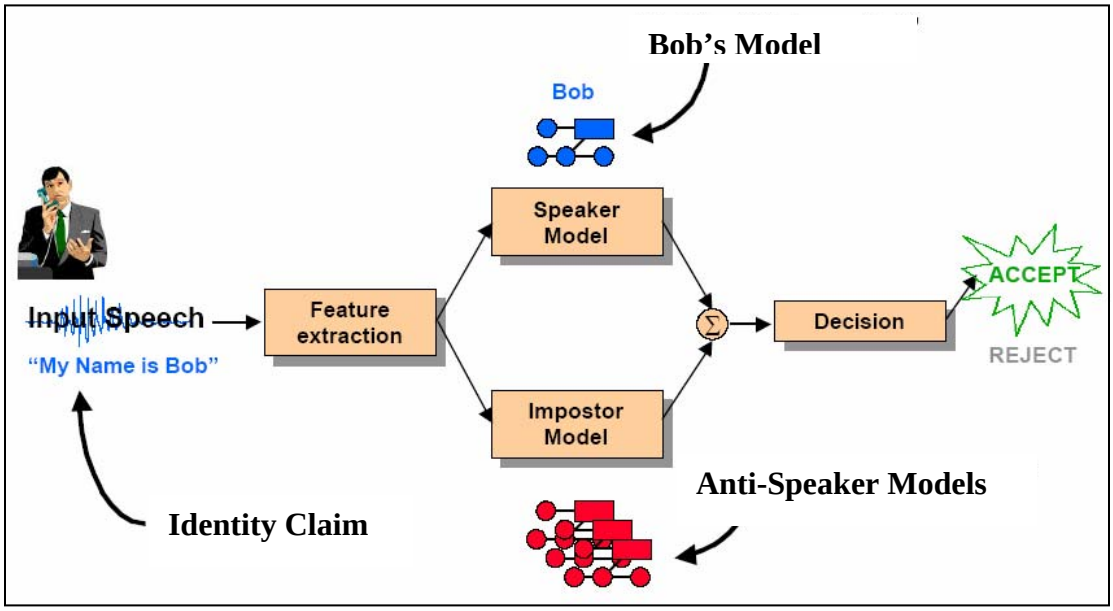


Figure 2: Speaker Verification.[11]

The seemingly easy implementation of speaker recognition systems contributes to their process's major weakness — susceptibility to transmission channel and microphone variability and noise. Systems can face problems when end users have enrolled on a clean landline phone and attempt verification using a noisy cellular phone. The inability to control the factors affecting the input system can significantly decrease performance. Speaker verification systems, except those using prompted phrases, are also susceptible to spoofing attacks through the use of recorded voice. Anti-spoofing measures that require the utterance of a specified and random word or phrase are being implemented to combat this weakness. For example, a system may request a randomly generated phrase, such as "33-54-63," to prevent an attack from a pre-recorded voice sample. The user cannot anticipate the random sample that will be required and therefore cannot successfully attempt a "playback" spoofing attack on the system.

Current research in the area of "text independent" speaker recognition is mainly focused on moving beyond the low-level spectral analysis previously discussed.[9] Although the spectral level of information is still the driving force behind the recognitions, fusing higher level characteristics with the low level spectral information is becoming a popular laboratory technique.[9] (Examples of higher level characteristics include: prosodic




characteristics such as rhythm, speed, modulation and intonation, based on personality type and parental influence; and semantics, idiolects, pronunciations and idiosyncrasies, related to birthplace, socio-economic status, and education level.) Higher level characteristics can be combined with the underlying low-level spectral information to improve the performance of "text independent" speaker recognition systems.

## United States Government Evaluations

Since 1996, the National Institute of Standards and Technology (NIST) has been conducting an ongoing series of yearly evaluations called the NIST Speaker Recognition Evaluations (http://www.nist.gov/speech/tests/spk/index.htm), which serve as test beds to compare and collaborate on research efforts across the community. The purpose of the evaluations is to determine the current state of the art, to cultivate technology growth, and to identify the most dominant and promising algorithmic approach to the problems facing speaker recognition.[8]

## Standards Overview

Standards play an important role in the development and sustainability of technology, and work in the international and national standards arena will facilitate the improvement of biometrics. The major standards work in the area of speaker recognition involves the Speaker Verification Application Program Interface (SVAPI), which is used by technology developers and allows for compatibility and interoperability between various vendors and networks.

Standards, such as INCITS 398-2005 Common Biometric Exchange Formats Framework (CBEFF), deal specifically with the data elements used to describe the biometric data in a common way, but may not yet apply to speaker recognition techniques.

## Summary

Thanks to the commitment of researchers and the support of NSA and NIST, speaker recognition will continue to evolve as communication and computing technology advance. Their determination will help to further develop the technology into a




reliable and consistent means of identification for use in remote recognition.

## Document References

[1] John D. Woodward, Jr., Nicholas M. Orlans, and Peter T. Higgins, Biometrics (New York: McGraw Hill Osborne, 2003).

[2] Potter, Kopp, and Green, Visible Speech (1947).

[3] S. Pruzansky, "Pattern-matching procedure for automatic talker recognition," JASA (26) 1963: 403-406.

[4] K. P. Li, et al, "Experimental studies in SV using an adaptive system," JASA (40) 1966: 966-978.

[5] P. D. Bricker and S. Pruzansky, "Effects of stimulus content and duration on talker identification," JASA (40) 1966: 1441-1449.

[6] K. Stevens, et al, "Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material," JASA (44) 1968: 1596-1607.

[7] W. Haberman and A. Fejfar, "Automatic ID of Personnel through Speaker and Signature Verification – System Description and Testing," 1976 Carnahan Conference on Crime Countermeasures, May 1976, University of Kentucky.

[8] "NIST Speaker Recognition Evaluations" 25 April 2005, NIST Speech Group 23 June 2005 <http://www.nist.gov/speech/tests/spk/index.htm>.

[9] Douglas A. Reynolds, "Automated Speaker Recognition: Current Trends and Future Direction," Biometrics Colloquium 17 June 2005.

[10] "Audio Spectrum Analysis," Spectrogram Version 11: A Product of Visualization Software LLC by Richard Horne <http://www.visualizationsoftware.com/gram.html>.

[11] Douglas A. Reynolds (M.I.T. Lincoln Laboratory) and Larry P. Heck (Nuance Communications), "Automatic Speaker Recognition: Recent Progress, Current Applications and Future Trends" 19 February 2000 Presented at the AAAS 2000 Meeting: Humans, Computers and Speech Symposium 19 February 2000 <http://www.ll.mit.edu/IST/pubs/aaas00-dar-pres.pdf>.