International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
Volume 1, Issue 3, July 2014, PP 26-32
ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)
www.arcjournals.org

Automatic Speaker Verification System

M.N. Nachappa, A.M. Bojamma, C.N. Prasad , Nithya M

Department of Computer Science
St. Joseph's College (Autonomous)
Langford Road, Shanthinagar
Bangalore , India.
mnnachappa@gmail.com

Abstract: *Automatic Speaker Verification (ASV) systems are being used for biometric authentication even if their vulnerability to spoofing is widely acknowledged. Recent work has proposed different spoofing approaches which can be used to test vulnerabilities. A tutorial of the design and development of ASV is presented. ASV is the use of a machine to recognize a person from a spoken phrase. These systems can operate in two modes: to identify a particular person or to claim a person's claimed identity. Speech recognition and the basic components of Automatic Speaker Recognition systems are shown and design tradeoffs are discussed.*

1. INTRODUCTION

The framework of speaker recognition technology was developed in the 1960's. Since then, numerous technical groups have engaged in aggressive research and development culminating in key innovations that now make speaker recognition feasible for many applications. Speech is the principal and most inherent form of communication among humans. Because of this and the fact that speech is a primary form of personal identification (PI), people generally have no problem accepting it as a biometric. Advantages of using speech as a biometric include: it's simple to use, it feels natural to the user, it provides eyes and hands-free operation, it can easily be implemented to support remote recognition (via telephone, internet, etc.), and implementation is typically inexpensive (often requiring software only). Typical problems include: channel mismatch (e.g. different microphones for enrollment and verification), background noise, and inconsistent acoustics (e.g. lab environment for enrollment, office environment for recognition). A typical human voice is formed when acoustic waves (generated by airflow from the lungs) are carried by the trachea (wind pipe) through the vocal folds (vocal cords) out through the vocal tract. Speech features, that allow us to discriminate between speakers, are due to both physiological and behavioral aspects of the speech production system. The main physiological component of speech production is the vocal tract. As acoustic waves pass through the vocal tract, their frequency content is altered by its resonances. The main discriminating behavioral characteristics include speaking rate and dialect. From a uniqueness standpoint, some biometric "experts" feel that the human voice is not rich in discriminative features such as fingerprints and iris patterns. This makes speaker recognition a poor candidate for "identification mode" operation when there is a large database of enrollees. From a permanence standpoint, the human voice is not necessarily stable over one's lifespan. Long-term changes include aging and disease. Short-term changes include stress, colds, and allergies[2]. A system that updates the user's speaker model with each successful identification/verification might compensate for the long-term changes. The following section detailed on the principles of these concepts and its functional components: the speaker and speech recognition technologies.

2. WHY IS VOICE A GOOD FIT FOR BIOMETRIC AUTHENTICATION?

A speaker's voice is extremely difficult to forge for biometrics comparison purposes, since a myriad of qualities are measured ranging from dialect and speaking style to pitch, spectral magnitudes, and format frequencies. The vibration of a user's vocal chords and the patterns created by the physical components resulting in human speech are as distinctive as fingerprints.

Attempts to impersonate a voice or provide voice recordings to gain fraudulent authentication fail due to the distinctive details of the voiceprint used for comparison. While voice impersonations may sound like an exact match to the human ear, detailed mathematical analysis of the print tends to reveal vast differences. Likewise, voice recordings that sound like an exact match to the human ear most often reveal distortions caused in the recording process when measured for biometric authentication purposes.

To further thwart the use of pre-recorded voiceprints, authentication employs a model of voiceprint comparison known as text independent directed speech. In this model, verification is performed against a phrase that is randomly generated, instead of using a phrase known to the user ahead of time, such as an account password. The chances of a fraudulent user able to match the randomly generated phrase and provide a passable voice recording are remote.

"A Person's Voiceprint is Unique!"

Loquendo Speaker Verification makes use of the voiceprint as a biometric to verify and demonstrate an individual's identity.

3. HOW DOES IT WORK?

The voice authentication process is made up of two distinct phases:

- **Enrollment:** creates and memorizes the speaker's voiceprint. Enrollment is carried out just once, in the initial phase, and involves the recording of key words pronounced by the user;
- **Verification:** in this phase, the identity claim is accepted or rejected[7]. Verification is performed each time access to the service is required, and it compares the voice characteristics of the speaker undergoing authentication with the voiceprint previously created for that identity.

3.1. Speaker verification:

Speaker verification is defines as deciding if a speaker is whom he claims to be. This differs from speaker identification problem where it decides whether the speaker is a specific person or is among a group of persons. In speaker verification, a person makes an identity claim. Here the speaker speaks the phrase into the microphone, which in turn analyzed by the verification system that makes a decision to accept or reject user's identity claim.

Automatic Speech Verification (ASV) functioning is shown in the figure below:

Speaker Verification can be used in a wide variety of application contexts, for example: phone banking and trading, password resetting, accessing customer care services, credit card activation, transactions and payments, phone top-up.

The user who has previously enrolled in the system presents an encrypted smart card which has his identification information. He then attempts to be authenticated by speaking a phrase into the microphone. Prior to verification user must enroll in the system. During enrollment voice models are generated and stored for later verification. Speaker Verification is built on Loquendo Speech Recognition technology, and therefore it uses the high quality acoustic models from Loquendo ASR (Automatic Speech Recognition), it is able to verify both the biometric characteristics of the voice and the content of the actual utterance, significantly reducing errors.

It also uses voice recognition; it is able to verify the semantic meaning of the spoken password. If the required information is known only to the user (e.g. How much was your last online payment?), it significantly improves the overall security of the system (knowledge verification).It is available in the same languages as Loquendo ASR.

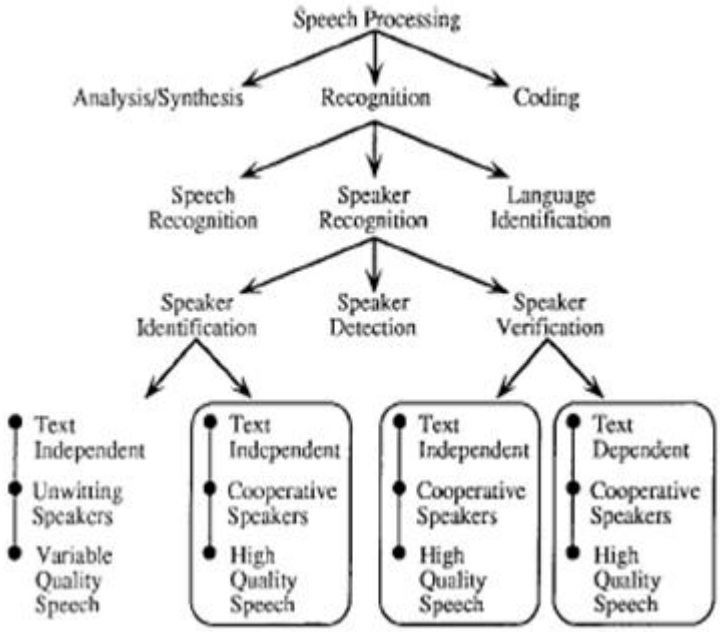


Fig1. *Speech Processing*

3.2. System Monitoring and Calibration Tools:

Loquendo Speaker Verification is equipped with tools for monitoring and calibrating the system, based on the particular security requirements of an application.

It is versatile and completely personalizes allowing system integrators to choose:

- The content of the phrase used for enrollment, and the number of repetitions required;
- The number of repetitions during the verification phase.

3.3. Efficiency and Ease of Integration:

Voiceprint size is small (tens of kilobytes), while Loquendo Speaker Verification is compatible with the major database technologies and architectures for maximum ease of integration.

3.4. Speaker Recognition

Voice recognition or speaker recognition refers to the automated method of identifying or confirming the identity of an individual based on his voice. Beware the difference between speaker recognition (recognizing who is speaking) and speech recognition (recognizing what is being said).

The voice is considered both a physiological and a behavioral biometric factor:

- The physiological component of speaker recognition is the physical shape of the subject's voice tract[9];
- The behavioral component is the physical movement of jaws, tongue and larynx.

3.5. How speaker recognition works

There exist two types of speaker recognition:

3.5.1. Text dependent (restrained)

Tthe subject has to say a fixed phrase (password) which is the same for enrollment and for verification, or the subject is prompted by the system to repeat a randomly generated phrase.

3.5.2. Text independent (unrestrained)

Recognition based on whatever words the subject sais.

Text dependent recognition has better performance for subjects that cooperate. But text independent voice recognition is more flexible that it can be used for non-cooperating individuals.

Basically identification or authentication using speaker recognition consists of four steps[1]:

- voice recording
- feature extraction
- pattern matching
- decision (accept / reject)

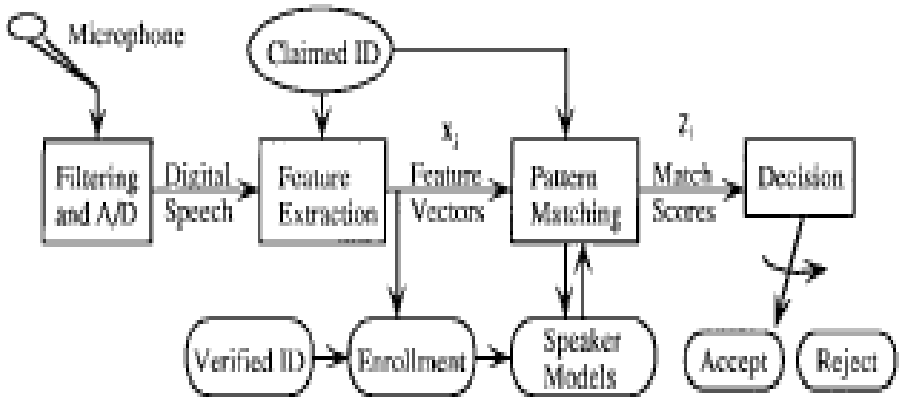
The general approach to ASV consists of five steps:

Digital speech data acquisition, Feature extraction, pattern matching, making an accept/reject decision, and enrollment to generate speaker reference models. A block diagram of this procedure is shown in fig. Feature extraction maps each interval of speech to a multidimensional feature space. (A speech interval typically spans 10–30 ms of the speech waveform and is referred to as a frame of speech.). This sequence of feature vectors is then compared to speaker models by pattern matching. This results in a match score for each vector or sequence of vectors. The match score measures the similarity of the computed input feature vectors to models of the claimed speaker or feature vector patterns for the claimed speaker. Last, a decision is made to either accept or reject the claimant according to the match score or sequence of match scores, which is a hypothesis testing problem. For speaker recognition, features that exhibit high speaker discrimination power, high inter speaker variability, and low intra speaker variability are desired[3]. Many forms of pattern matching and corresponding models are possible. Pattern-matching methods include dynamic time warping (DTW), the hidden Markov model (HMM), artificial neural networks, and vector quantization (VQ). Template models are used in DTW, statistical models are used in HMM, and codebook models are used in VQ.

3.6. Benefits of Speaker Verification:

A biometric authentication system has several benefits:

- voice authentication is versatile, simple to use and non-intrusive, which means high user acceptance
- compared to other biometric technologies, it is accurate and does not require specialized equipment - just a phone
- it is the ideal solution for overcoming security issues around remote telephone access to various types of services.



4. SUITABILITY OF SPEAKER RECOGNITION

How suitable is speaker recognition as a biometric solution? We use the following 7 criteria to evaluate the suitability of speaker recognition:

a. *Universality:* Obviously for people who are mute or having problems with their voice due to severe illness this biometric solution is not useable[2].

b. *Uniqueness :* Because of the combination of physiological and behavioral factors the voice is a unique feature of an individual, the voice has more unique features than a fingerprint.

c. *Permanence*: An issue with speaker recognition is that the voice changes with ageing, and is also influenced by factors such as sickness, tiredness, stress, etc.

d. *Collectability*: Voice recordings are easy to obtain and do not require expensive hardware. The real advantage of voice recognition is that it can be done over telephone lines or using computer microphones, with variable recording and transmission quality[3]. Pattern matching algorithms must be able to handle ambient noise and differing quality of the recordings.

e. *Acceptability*: Speaker recognition is unobtrusive, speaking is a natural process so no unusual actions are required.[5] When speaker recognition is used for surveillance applications or in general when the subject is not aware of it then the common privacy concerns of identifying unaware subjects apply.

f. *Circumvention***:** A major issue with speaker recognition is spoofing using voice recordings. The risk of spoofing with voice recordings can be mitigated if the system requests a random generated phrase to be repeated, an impostor cannot anticipate the random phrase that will be required and therefore cannot attempt a playback spoofing attack.

g. *Performance:* Robustness is very dependent on the setup, when telephone lines or computer microphones are used the algorithms will have to compensate for noise and issues with room acoustics. [6]Furthermore speaker recognition is, because the voice is a behavioral biometric, impacted by errors of the individual such as misreading and mispronunciations.

5. TYPES OF AUTOMATIC SPEECH RECOGNITION SYSTEMS

There are basically three categories of ASR systems differentiated by the degree of user training required prior to use:

- Speaker dependent
- Speaker independent and
- Speaker adaptable ASR.

Speaker dependent ASR requires speaker training or enrollment prior to use and the primary user trains the speech recognizer with samples of his or her own speech. These systems typically work well only for the person who trains it.

Speaker independent ASR does not require speaker training prior to use. The speech recognizer is pre-trained during system development with speech samples from a collection of speakers[6]. Many different speakers will be able to use this same ASR application with relatively good accuracy if their speech falls within the range of the collected sample; but ASR accuracy will generally be lower than achieved with a speaker dependent ASR system.

Speaker adaptable ASR is similar to speaker independent ASR in that no initial speaker training is required prior to use. However, unlike speaker independent ASR systems, as the speaker adaptable ASR system is being used, the recognizer gradually adapts to the speech of the user. This 'adaptation' process further refines the system's accuracy. A few types of speaker adaptable ASR systems exist differing with respect to how the adaptation is implemented. The reader is referred to the ASR review paper by Rosen and Yampolsky (2000) for further information.

ASR technologies also vary by the type of input that they can handle:

- Isolated/discrete word recognition,
- Connected word recognition, and
- Continuous speech recognition

Discrete word recognition requires a pause or period of silence to be inserted between words or utterances. Connected word recognition is an extension of discrete word recognition and requires a pause or period of silence only after a group of connected words have been spoken. For continuous speech recognition an entire phrase or complete sentences can be spoken without the need to insert pauses between words or after sentences.

6. VISUALIZATION OF THE ACOUSTIC PATTERN OF THE VOICE: LOUDNESS OF THE INPUT VS. TIME.

Depending on the application a voice recording is performed using a local, dedicated system or remotely (e.g. telephone). The acoustic patterns of speech can be visualized as loudness or frequency vs. time. Speaker recognition systems analyze the frequency as well as attributes such as dynamics, pitch, duration and loudness of the signal.

During feature extraction the voice recording is cut into windows of equal length, these cut-out samples are called frames which are often 10 to 30 ms long.

Pattern matching is the actual comparison of the extracted frames with known speaker models (or templates), these results in a matching score which quantifies the similarity in between the voice recording and a known speaker model. Pattern matching is often based on Hidden Markov Models (HMMs), a statistical model which takes into account the underlying variations and temporal changes of the acoustic pattern[7].

Alternatively Dynamic Time Warping is used, this algorithm measures the similarity in between two sequences that vary in speed or time, even if this variation is non-linear such as when the speaking speed changes during the sequence[11].

Some systems use "anti-speaker" techniques such as cohort models.

7. APPLICATION OF SPEAKER RECOGNITION:

Voice recognition is mostly used for telephone based applications, such as for telephone banking and hotel or flight bookings.

- Nuance is a US based company and a major player when it comes to speech recognition, they also developed a product for speaker recognition called Nuance Verifier.
- Voice Trust is a German company specialized in speaker recognition solutions.

8. HOW VOICE BIOMETRICS MEASURES UP?

8.1. Advantages

Voice authentication has a number of advantages. The cost of implementation is low because there is no special hardware required. A simple telephone or microphone is all that a user needs to authenticate using her voice. Other methods of biometric authentication like fingerprinting and retinal scans require special devices. Voice authentication is easy to use and easily accepted by users. It is quite natural to speak. It is not as natural to put an eye up to a reader. The concept of identifying people by voices is also quite natural. Every time someone answers a telephone call, the natural instinct is to try to identify the caller by his voice.

Perhaps most important to the future of voice biometrics is that it is the only biometric that allows users to authenticate remotely. Allowing a user to call a phone number and authenticate with her bank vocally to perform a transaction is much easier than asking the user to go to the bank in person and authenticate via fingerprint[11]. It is quick to enroll in a voice authentication system. The user is asked to speak a certain set of words or phrases, or to speak for a certain length of time. From that sample, a digital representation of the voice, called a voiceprint, is created. A good voiceprint is between 2-8 seconds of speech.

Authenticating a user is accomplished by comparing the voiceprint that was created at enrollment to a sample given when the user wants to enter the restricted area or system. Authentication is very fast; it can be completed in 0.5 seconds.

Another advantage is that the storage size of the voiceprint is small. How small it is will be vendor specific, but one vendor, Voice Security Systems, states that a user's voiceprint is less than 1K in size. This is so small that it can be stored almost anywhere: smart cards, floppy disks, databases, even on cell phones[9].

8.2. Disadvantages

Voice biometrics is not the most secure of the biometric technologies. For this reason, it is not appropriate to use them independently for authorization to systems that require high security[12].

They become more powerful when used in conjunction with another form of authorization, such as a password.

9. CONCLUSION

Biometric authentication refers to automated methods of identifying or verifying the identity of a living person in real time based on a physical characteristic or personal trait. The phrase, "living person in real time" is used to distinguish biometric authentication from forensics, which does not involve real-time identification of a living individual.

Voice biometrics is very powerful, secure and robust technology for voice verification techniques frequently used in the industry. It is cost effective, non-invasive and easy to integrate technique for authenticating people. Voice biometric can be used in any phone, smart phone or other devices easily.

The cost of deploying voice biometric authentication cost is also very low. It is quite common now to see voice recognition and authentication software being used in call centers. The customer's registration process for using their voice only requires them to say a few numbers and/or words. The next time they call in, they simply have to repeat the numbers and/or words and the software matches them up against their registration recordings. Often the voice recognition also asks the speaker to provide a password verbally in addition to their voice recognition in order to increase the likelihood that the speaker on the end of the phone is who they really are.

REFERENCES

[1] "Biometric Identification". Communications of the ACM, 43(2), p. 91-98. DOI 10.1145/328236.328110

[2] Jain, Anil K.; Ross, Arun (2008). "Introduction to Biometrics". In Jain, AK; Flynn; Ross, A. Handbook of Biometrics. Springer. pp. 1–22. ISBN 978-0-387-71040-2.

[3] "Biometrics for Secure Authentication" (PDF). Retrieved 2012-07-29.

[4] Weaver, A.C. (2006). "Biometric Authentication". Computer, 39 (2), p. 96-97. DOI 10.1109/MC.2006.47

[5] Jain, A.K.; Bolle, R.; Pankanti, S., eds. (1999). Biometrics: Personal Identification in Networked Society. Kluwer Academic Publications. ISBN 978-0-7923-8345-1.

[6] Sahoo, SoyujKumar; Mahadeva Prasanna, SR, Choubisa, Tarun (1 January 2012). "Multimodal Biometric Person Authentication : A Review". IETE Technical Review 29(1): 54. doi:10.4103/0256-4602.93139. Retrieved 23 February 2012.

[7] "Questions Raised About Iris Recognition Systems".Science Daily. 12 July 2012.

[8] Saylor, Michael (2012). The Mobile Wave: How Mobile Intelligence Will Change Everything. Perseus Books/Vanguard Press. p. 99.

[9] Bill Flook (3 October 2013). "This is the 'biometric war' Michael Saylor was talking about". Washington Business Journal.

[10] "CHARACTERISTICS OF BIOMETRIC SYSTEMS". Cernet.

[11] A. Rattani, "Adaptive Biometric System based on Template Update Procedures," PhD thesis, University of Cagliari, Italy, 2010

[12] http://zeenews.india.com/news/nation /aadhaar-scheme-does-not-violate-fundamental-rights-says-uidai_884850.html

[13] Building a Biometric National ID: Lessons for Developing Countries from India's Universal ID Program, Alan Gelb and Julia Clark, The Center for Global Development, October 2012,http://www.cgdev.org/doc /full_text/GelbClarkUID/1426583.html