Voice recognition

How speech can open doors

by Ravi Das

Voice recognition is a biometric technology that finds its origin in World War II. At that time, research showed that there are variations in the intensity of various sounds in a person's voice and at different frequency levels. In this article, Ravi Das discusses how voice recognition works, the factors which affect it, its (dis)advantages and various market applications. The article concludes with a case study.

Ravi Das *is the President/CEO of Apollo Biometrics, Inc., a security consultancy firm with offices in Chicago & New York City. Ravi holds a Master of Science Degree in Agribusiness Economics from Southern Illinois University and a Master of Business Administration (specialising in Management Information Systems) from Bowling Green State University.*

In the 1940s, the variations in sound intensity propelled the idea of using the voice to confirm or verify the identity of a particular individual. The research and development into voice recognition which started during World War II continued well into the 1960's. The voice spectrographs which were used at the time utilised statistical modelling as a means of biometric template creation, rather than using the traditional biometric approaches, such as using the finger, hand, face, or the eye. In fact, the first known voice recognition systems was called the Forensic Automatic Speaker Recognition, or FASR for short.

In today's biometric world, voice recognition can be considered to be both a behavioural-based and a physical-based biometric. The acoustic properties of a particular person's voice are a direct function of the shape of the individual's mouth, as well as the length and the quality of the vocal cords (the physical component). But at the same time the behavioural data of an individual's voice which include such variables as the pitch, volume, and rhythm of the voice are present in the template as well.

Voice recognition: how it works

Voice production is a facet of life which we take for granted every day, although the actual process is complicated. The production of sound originates at the vocal cords, in between which is a gap. When we speak, the muscles that control the vocal cords contract. As a result, the gap narrows, and as we exhale, our breathe passes through the gap, which creates sound.

The unique patterns of an individual's voice are produced by the vocal tract, which consists of the laryngeal pharynx, oral pharynx, oral cavity, nasal pharynx, and the nasal cavity. These patterns can be used by voice recognition systems. Even though people may sound alike amongst one another to the human ear, everybody, to some degree, has a different or unique annunciation in their speech.

The first step in voice recognition is for an individual to produce an actual voice sample. To ensure a good quality voice sample for the voice recognition system to capture, the individual usually recites some sort of text, which can either be a verbal phrase, a series of numbers, or a text passage, which they usually have to repeat a number of times.

The most common devices used to capture an individual's voice samples are computer microphones, smartphones, and landline-based telephones. As a result, a key advantage of voice recognition is that it can leverage existing telephony technology, with minimal disruption to an entity's business processes. In terms of noise disruption, computer microphones and cell phones create the most, and landline-based telephones create the least.

It is very important that the medium which is used for the purposes of verification and identification is the same as the one used to enrol the end-user into the voice recognition system. For example, if a smartphone was used to create the enrolment template, then the same smartphone should be used in subsequent verification/identification transactions in the voice recognition system. Likewise, if a landline phone was used initially, it should be used again to confirm the identity of the individual.

After they are collected, the voice samples are converted from an analogue format to a digital format for processing. These raw voice data types are used as input for a spectrograph, which is a visualisation of the acoustic properties of the individual's voice. The next steps are the unique feature extractions from the voice samples, and the creation of the enrolment and verification templates.

The extraction algorithms look for the unique patterns in the individual's voice samples. In order to create the templates, a statistical 'model' of the voice is created. There are two statistical modelling techniques which

are primarily utilized when formulating the voice recognition biometric templates:

- Hidden Markov Model (HMM);
  This is a statistical model which is used with text dependent voice recognition systems (when the end-user is given the specific verbiage to recite). This type of model displays such variables as the changes and fluctuations of the voice over a certain period of time, which is a direct function of the pitch, duration, dynamics, and the quality of the person's speaking voice.
- Gaussian Mixture Model (GMM).
  This is a state mapping model, in which various types and kinds of vector states are created which represent the unique sound characteristics of the particular individual. Unlike the HMM, the GMM is devoted exclusively for use by text independent voice recognition systems (when the end-user is not given the specific verbiage to recite).

Factors affecting voice recognition

There are numerous variables which can affect the quality of voice samples, such as mispronounced verbal phrases, different telephony media used for enrolment and verification (using a landline telephone for the enrolment process, but a mobile phone for verification), as well as the emotional and physical conditions of the individual. Other factors include poor room sound acoustics and the age of the individual, since the vocal tract changes as we get older.

Advantages and disadvantages of voice recognition

The advantages and the disadvantages of voice recognition can be judged against seven major criteria:

- Universality;
  Voice recognition is not language dependent, which is probably its biggest strength. Theoretically, as long as an individual can speak, they can be easily enrolled into a voice recognition system.
- Uniqueness;
  Unlike some of the other biometrics, such as the iris and the retina, the voice does not possess as many unique features. This results in a lack of rich information and data.
- Permanence;
  The voice can change for many reasons, such as age, fatigue, any disease affecting the vocal cords, medication, but also someone's emotional state.
- Collectability;
  This is probably the biggest disadvantage of voice recognition. Any type or kind of variability in the medium which is used to collect the raw voice sample can greatly affect or skew the voice recognition biometric templates. Therefore, it is of upmost importance to ensure that the same type and kind of collection medium is used for both the enrolment

and verification stages. There should be no interchangeability involved whatsoever.

- Performance;
  Because of the variableness involved when collecting the raw voice samples, it is difficult to gauge how well a voice recognition system can actually perform. Also, the enrolment and verification template sizes can be very large, ranging from 1.5 kB to 3 kB.
- Acceptability;
  This is one of the strongest advantages of voice recognition. The technology is very non-intrusive, and it can be deployed in a manner which is very covert to the end-user.
- Resistance to circumvention.
  Compared to the other biometric technologies, voice recognition, to a certain degree, can be easily spoofed by mimicking someone's voice acoustics. This is in large part due to the lack of unique features in the voice itself.

Market applications

The market applications of voice recognition are much more limited than those of the other biometric technologies of fingerprint, iris, and vein pattern recognition. One of the biggest reasons for this is that there are not many biometric vendors actually developing voice recognition solutions. But, overall, voice recognition is starting to gain some serious traction, as businesses and governments worldwide have started to become aware of its ease of deployment and the other advantages it possesses.

Probably one of the biggest market applications for voice recognition is that of financial trading. Many of the brokerage institutions now offer voice recognition to their customers as a means of quick verification. Rather than wasting a customer's time by having them enter their social security number, a customer can now be identified very quickly by the use of voice

recognition. What would normally take minutes with the traditional means of security is now literally reduced to just mere seconds. As a result, financial transactions for the customer can occur at a very quick pace.

In terms of other emerging applications, voice recognition is starting to be used on smartphones as a means of verification, instead of using a PIN code. Voice recognition is also being used in correctional facilities to monitor the telephone privileges of the inmates, the railroad system, border protection and control, as well as certain types and kinds of physical and logical access entries.

Case study

Pronexus, Inc. is a Canadian company which specialises in creating Interactive Voice Response Platform (IVR) solutions. The company develops both Automatic Speech Recognition (ASR), and Text to Speech (TTS) software applications. IVR is a technology which allows for the interaction of computers by humans via their voice. Due to an increased demand for IVR-based solutions by their clients, the possibility of voice recognition was explored.

After conducting an exhaustive search, the company finally decided upon the use of VBVoice™, developed by LumenVox, LLC. Its primary advantage is that software developers can easily create and deploy IVR-based applications, based upon the exacting needs of clients.

VBVoice™ provides a plug and play environment, and also interacts with Microsoft Visual Studio. As a result, software code for dynamic web pages can be created with voice recognition embedded into it, as an extra means of security. With the use of the Media Resource Control Protocol (MRCP), clients can now quickly and easily deploy their customised IVR applications. Therefore, time is not wasted in installing local licenses onto each and every computer or server.

Examples of IVR-based solutions which have been created using VBVoice™ include:

- Scaller;
  The identity of a student is verified using their voice, in order for them to further access university resources and tools in their educational process.
- Utilities On Call;
  Through the use of voice recognition, routine customer requests are now fully automated, thus freeing up resources to handle more complex customer issues and projects. This solution also allows for the utility companies to implement automated bill collection, coupled with instant payment options, such as mobile/virtual payments and e-checking.
- CallAssure.
  With this IVR-based platform, automated outreach medical services are provided to patients. With the use of voice recognition, once the identity of the patient is confirmed, the following medical services are offered:
  - chronic disease management;
  - post-discharge follow-up appointment setting and reminders;
  - medication dosing;
  - smoking cessation.