# 1

# Speaker Recognition

Homayoon Beigi
*Recognition Technologies, Inc.*
*U.S.A.*

## 1. Introduction

Speaker Recognition is a multi-disciplinary technology which uses the vocal characteristics of speakers to deduce information about their identities. It is a branch of biometrics that may be used for *identification*, *verification*, and *classification* of individual speakers, with the capability of *tracking*, *detection*, and *segmentation* by extension.

A speaker recognition system first tries to model the vocal tract characteristics of a person. This may be a mathematical model of the physiological system producing the human speech or simply a statistical model with similar output characteristics as the human vocal tract. Once a model is established and has been associated with an individual, new instances of speech may be assessed to determine the likelihood of them having been generated by the model of interest in contrast with other observed models. This is the underlying methodology for all speaker recognition applications. The earliest known papers on speaker recognition were published in the 1950s (Pollack et al., 1954; Shearme & Holmes, 1959).

Initial speaker recognition techniques relied on a human expert examining representations of the speech of an individual and making a decision on the person's identity by comparing the characteristics in this representation with others. The most popular representation was the *formant* representation. In the recent decades, fully automated speaker recognition systems have been developed and are in use (Beigi, 2011).

There have been a number of tutorials, surveys, and review papers published in the recent years (Bimbot et al., 2004; Campbell, 1997; Furui, 2005). In a somewhat different approach, we have tried to present the material, more in the form of a comprehensive summary of the field with an ample number of references for the avid reader to follow. A coverage of most of the aspects is presented, not just in the form of a list of different algorithms and techniques used for handling part of the problem, as it has been done before.

As for the importance of speaker recognition, it is noteworthy that *speaker identity* is the only biometric which may be easily tested (identified or verified) remotely through the existing infrastructure, namely the telephone network. This makes speaker recognition quite valuable and unrivaled in many real-world applications. It needs not be mentioned that with the growing number of cellular (mobile) telephones and their ever-growing complexity, speaker recognition will become more popular in the future.

There are countless number of applications for the different branches of speaker recognition. If audio is involved, one or more of the speaker recognition branches may be used. However, in terms of deployment, speaker recognition is in its early stages of infancy. This is partly due to unfamiliarity of the general public with the subject and its existence, partly because of the limited development in the field. These include, but are certainly not limited to, *financial*,

*forensic and legal* (Nolan, 1983; Tosi, 1979), *access control and security*, *audio/video indexing and diarization*, *surveillance*, *teleconferencing*, and *proctorless distance learning* Beigi (2009).
Speaker recognition encompasses many different areas of science. It requires the knowledge of *phonetics*, *linguistics* and *phonology*. *Signal processing* which by itself is a vast subject is also an important component. *Information theory* is at its basis and *optimization theory* is used in solving problems related to the training and matching algorithms which appear in *support vector machines (SVMs)*, *hidden Markov models (HMMs)*, and *neural networks (NNs)*. Then there is *statistical learning theory* which is used in the form of *maximum likelihood estimation*, *likelihood linear regression*, *maximum a-posteriori probability*, and other techniques. In addition, *Parameter estimation* and *learning* techniques are used in *HMM, SVM, NN*, and other underlying methods, at the core of the subject. *Artificial intelligence* techniques appear in the form of *sub-optimal searches* and *decision trees*. Also *applied math*, in general, is used in the form of *complex variables theory*, *integral transforms*, *probability theory*, *statistics*, and many other mathematical domains such as *wavelet analysis*, etc.
The vast domain of the field does not allow for a thorough coverage of the subject in a venue such as this chapter. All that can be done here is to scratch the surface and to speak about the inter-relations among these topics to create a complete speaker recognition system. The avid reader is recommended to refer to (Beigi, 2011) for a comprehensive treatment of the subject, including the details of the underlying theory.
To start, let us briefly review different biometrics in contrast with speaker recognition. Then, it is important to clarify the terminology and to describe the problems of interest by reviewing the different manifestations and modalities of this biometric. Afterwards, some of the challenges faced in achieving a practical system are listed. Once the problems are clearly posed and the challenges are understood, a quick review of the production and the processing of speech by humans is presented. Then, the state of the art in addressing the problems at hand is briefly surveyed in a section on theory. Finally, concluding remarks are made about the current state of research on the subject and its future trend.

## 2. Comparison with other biometrics

There have been a number of biometrics used in the past few decades for the recognition of individuals. Some of these markers have been discussed in other chapters of this book. A comparison of voice with some other popular biometrics will clarify the scope of its practical usage. Some of the most popular biometrics are *Deoxyribonucleic Acid (DNA)*, *image-based* and *acoustic ear recognition*, *face recognition*, *fingerprint* and *palm recognition*, *hand* and *finger geometry*, *iris* and *retinal recognition*, *thermography*, *vein recognition*, *gait*, *handwriting*, and *keystroke recognition*.
Fingerprints, as popular as they are, have the problem of not being able to identify people with damaged fingers. These are, for example, construction workers, people who work with their hands, or maybe people without limbs, such as those who have either lost their hands or their fingers in an accident or those who congenitally lack fingers or limbs. According to the National Institute of Standards and Technology (NIST), this is about 2% of the population! Also, latex prints of finger patterns may be used to spoof some sensors.
People, with damaged irides, such as some who are blind, either congenitally or due to an illness like glaucoma, may not be recognized through iris recognition. It is very hard to tell the size of this population, but they certainly exist. Additionally, one would need a high quality image of the iris to perform recognition. Acquiring these images is quite problematic. Although there are long distance iris imaging cameras, their field of vision may easily be

blocked by uncooperative users through the turning of the head, blinking, rolling of the eyes, wearing of hats, glasses, etc. The image may also not be acceptable due to lighting and focus conditions. Also, irides tend to change due to changes in lighting conditions as the pupils dilate or contract. It is also possible to spoof some iris recognition systems, either by wearing contact lenses or by simply using an image of the target individual's irides.

Of course, there is also a percentage of the population who are unable to speak, therefore they will not be able to use speaker recognition systems. The latest figures for the population of deaf and mute people in the United States reflected by the US Census Bureau set this percentage at 0.4% for deaf and mute individuals (USC, 2005). Spoofing, using recordings is also a concern in practical speaker recognition systems.

In terms of public acceptance, fingerprint recognition has long been associated with criminology. Due to these legacy associations, many individuals are wary of producing a fingerprint for fear of its malicious usage or simply due to the criminal connotation it carries. As an example, a few years ago, the United States government required capturing the image and fingerprint of all tourists entering the nation's airports. This action offended many tourists to the point that some countries such as Brazil placed a reciprocal system in place only for U.S. citizens entering their country. Many people entering the U.S. felt like they were being treated as criminals, only based on the act of fingerprinting. Of course, since many other countries have been adopting the fingerprint capture requirement, it is being tolerated by travelers much better, around the world.

Because facial, iris, retinal images, and fingerprints have a sole purpose of being used in recognition, they are somewhat harder to capture. In general, the public is more wary of providing such information which may be archived and misused. On the other hand, speech has been established for communication and people are far less likely to be concerned about parting with their speech. Even in the technological arena, the use of speech for telephone communication makes it much more socially acceptable.

Speaker recognition can also utilize the widely available infrastructure that has been around for so long, namely the telephone network. Speech may be used for doing remote recognition of the individual using the existing telephone network and without the need for any extra hardware or other apparatus. Also, speaker recognition, in the form of tracking and detection may be used to do much more than simple identification and verification of individuals, such as a full diarization of large media databases. Another attractive point is that cellular telephone and PDA-type data security needs no extra hardware, since cellular telephones already have speech capture devices, namely microphones. Most PDAs also contain built-in microphones. On the other hand, for fingerprint and image recognition, a fingerprint scanner and a camera would have to be present.

Multimodal biometrics entail systems which combine any two or more of these or other biometrics. These combinations increase the accuracy of the identification or verification of the individual based on the fact that the information is obtained through different, mostly independent sources. Most practical implementations of biometric system will need to utilize some kind of multimodal approach; since any one technique may be bypassed by the eager impostor. It would be much more difficult to fool several independent biometric systems simultaneously. Many of the above biometrics may be successfully combined with speaker recognition to produce viable multimodal systems with much higher accuracies. (Viswanathan et al., 2000) shows an example of such a multimodal approach using speaker and image recognition.

## 3. Terminology and manifestations

In addressing the act of *speaker recognition* many different terms have been coined, some of which have caused great confusion. *Speech recognition* research has been around for a long time and, naturally, there is some confusion in the public between *speech* and *speaker* recognition. One term that has added to this confusion is *voice recognition*.

The term *voice recognition* has been used in some circles to double for *speaker recognition*. Although it is conceptually a correct name for the subject, it is recommended that the use of this term is avoided. *Voice recognition*, in the past, has been mistakenly applied to *speech recognition* and these terms have become synonymous for a long time. In a speech recognition application, it is not the voice of the individual which is being recognized, but the contents of his/her speech. Alas, the term has been around and has had the wrong association for too long.

Other than the aforementioned, a myriad of different terminologies have been used to refer to this subject. They include, *voice biometrics*, *speech biometrics*, *biometric speaker identification*, *talker identification*, *talker clustering*, *voice identification*, *voiceprint identification*, and so on. With the exception of the term *speech biometrics* which also introduces the addition of a speech knowledge-base to speaker recognition, the rest do not present any additional information.

### 3.1 Speaker enrollment

The first step required in most manifestations of speaker recognition is to enroll the users of interest. This is usually done by building a mathematical model of a sample speech from the user and storing it in association with an identifier. This model is usually designed to capture statistical information about the nature of the audio sample and is mostly irreversible – namely, the enrollment sample may not be reconstructed from the model.

### 3.2 Speaker identification

There are two different types of speaker identification, *closed-set* and *open-set*. Closed-set identification is the simpler of the two problems. In close-set identification, the audio of the test speaker is compared against all the available speaker models and the speaker ID of the model with the closest match is returned. In practice, usually, the top best matching candidates are returned in a ranked list, with corresponding confidence or likelihood scores. In closed-set identification, the ID of one of the speakers in the database will always be closest to the audio of the test speaker; there is no rejection scheme.

One may imagine a case where the test speaker is a 5-year old child where all the speakers in the database are adult males. In closed-set Identification, still, the child will match against one of the adult male speakers in the database. Therefore, closed-set identification is not very practical. Of course, like anything else, closed-set identification also has its own applications. An example would be a software program which would identify the audio of a speaker so that the interaction environment may be customized for that individual. In this case, there is no great loss by making a mistake. In fact, some match needs to be returned just to be able to pick a customization profile. If the speaker does not exist in the database, then there is generally no difference in what profile is used, unless profiles hold personal information, in which case rejection will become necessary.

Open-set identification may be seen as a combination of closed-set identification and speaker verification. For example, a closed-set identification may be conducted and the resulting ID may be used to run a speaker verification session. If the test speaker matches the target speaker based on the ID, returned from the closed-set identification, then the ID is accepted

and passed back as the true ID of the test speaker. On the other hand, if the verification fails, the speaker may be rejected all-together with no valid identification result. An open-set identification problem is therefore at least as complex as a speaker verification task (the limiting case being when there is only one speaker in the database) and most of the time it is more complex. In fact, another way of looking at verification is as a special case of open-set identification in which there is only one speaker in the list. Also, the complexity generally increases linearly with the number of speakers enrolled in the database since theoretically, the test speaker should be compared against all speaker models in the database – in practice this may be avoided by tolerating some accuracy degradation (Beigi et al., 1999).

### 3.3 Speaker verification (authentication)

In a generic speaker verification application, the person being verified (known as the test speaker), identifies himself/herself, usually by non-speech methods (e.g., a username, an identification number, et cetera). The provided ID is used to retrieve the enrolled model for that person which has been stored according to the enrollment process, described earlier, in a database. This enrolled model is called the *target speaker model* or the *reference model*. The speech signal of the test speaker is compared against the target speaker model to verify the test speaker.

Of course, comparison against the target speaker's model is not enough. There is always a need for contrast when making a comparison. Therefore, one or more competing models should also be evaluated to come to a verification decision. The competing model may be a so-called (universal) background model or one or more cohort models. The final decision is made by assessing whether the speech sample given at the time of verification is closer to the target model or to the competing model(s). If it is closer to the target model, then the user is verified and otherwise rejected.

The speaker verification problem is known as a one-to-one comparison since it does not necessarily need to match against every single person in the database. Therefore, the complexity of the matching does not increase as the number of enrolled subjects increases. Of course in reality, there is more than one comparison for speaker verification, as stated – comparison against the target model and the competing model(s).

#### 3.3.1 Speaker verification modalities

There are two major ways in which speaker verification may be conducted. These two are called the *modalities* of speaker verification and they are *text-dependent* and *text-independent*. There are also variations of these two modalities such as *text-prompted*, *language-independent text-independent* and *language-dependent text-independent*.

In a purely *text-dependent* modality, the speaker is required to utter a predetermined text at enrollment and the same text again at the time of verification. Text-dependence does not really make sense in an identification scenario. It is only valid for verification. In practice, using such text-dependent modality will be open to *spoofing* attacks; namely, the audio may be intercepted and recorded to be used by an impostor at the time of the verification. Practical applications that use the text-dependent modality, do so in the text-prompted flavor. This means that the enrollment may be done for several different textual contents and at the time of verification, one of those texts is requested to be uttered by the test speaker. The chosen text is the prompt and the modality is called *text-prompted*.

A more flexible modality is the *text-independent* modality in which case the texts of the speech at the time of enrollment and verification are completely random. The difficulty with this

method is that because the texts are presumably different, longer enrollment and test samples are needed. The long samples increase the probability of better coverage of the idiosyncrasies of the person's vocal characteristics.
The general tendency is to believe that in the text-dependent and text-prompted cases, since the enrollment and verification texts are identical, they can be designed to be much shorter. One must be careful, since the shorter segments will only examine part of the dynamics of the vocal tract. Therefore, the text for text-prompted and text-dependent engines must still be designed to cover enough variation to allow for a meaningful comparison.
The problem of spoofing is still present with text-independent speaker verification. In fact, any recording of the person's voice should now get an impostor through. For this reason, text-independent systems would generally be used with another source of information in a multi-factor authentication scenario.
In most cases, *text-independent* speaker verification algorithms are also *language-independent*, since they are concerned with the vocal tract characteristics of the individual, mostly governed by the shape of the speaker's vocal tract. However, because of the coverage issue discussed earlier, some researchers have developed text-independent systems which have some internal models associated with phonemes in the language of their scope. These techniques produce a text-independent, but somewhat language-dependent speaker verification system. The language limitations reduce the space and, hence, may reduce the error rates.

### 3.4 Speaker and event classification

The goal of classification is a bit more vague. It is the general label for any technique that pools similar audio signals into individual bins. Some examples of the many classification scenarios are gender classification, age classification, and event classification. Gender classification, as is apparent from its name, tries to separate male speakers and female speakers. More advanced versions also distinguish children and place them into a separate bin; classifying male and female is not so simple in children since their vocal characteristics are quite similar before the onset of puberty. Classification may use slightly different sets of features from those used in verification and identification, depending on the problem at hand. Also, either there may be no enrollment or enrollment may be done differently. Some examples of special enrollment procedures are, *pooling enrollment data* from like classes together, using *extra features in supplemental codebooks* related to specific natural or logical specifics of the classes of interest, etc.(Beigi, 2011)
Although these methods are called speaker classification, sometimes, the technique are used for doing event classification such as classifying speech, music, blasts, gun shots, screams, whistles, horns, etc. The feature selection and processing methods for classification are mostly dependent on the scope and could be different from mainstream speaker recognition.

### 3.5 Speaker segmentation, diarization, detection and tracking

Automatic segmentation of an audio stream into parts containing the speech of distinct speakers, music, noise, and different background conditions has many applications. This type of segmentation is elementary to the practical considerations of speaker recognition as well as speech and other audio-related recognition systems. Different specialized recognizers may be used for recognition of distinct categories of audio in a stream.
An example is the ever-growing tele-conferencing application. In a tele-conference, usually, a host makes an appointment for a conference call and notifies attendees to call a telephone number and to join the conference using a special access code. There is an increasing

interest from the involved parties to obtain transcripts (minutes) of these conversations. In order to fully transcribe the conversations, it is necessary to know the speaker of each statement. If an enrolled model exists for each speaker, then prior to identifying the active speaker (*speaker detection*), the audio of that speaker should be segmented and separated from adjoining speakers. When speaker segmentation is combined with speaker identification and the resulting index information is extracted, the process is called *speaker diarization*. In case one is only interested in a specific speaker and where that speaker has spoken within the conversation (the timestamps), the process is called *speaker tracking*.

### 3.6 Knowledge-based speaker recognition (speech biometrics)

A knowledge-based speaker recognition system is usually a combination of a speaker recognition system and a speech recognizer and sometimes a natural language understanding engine or more. It is somewhat related to the *text-prompted* modality with the difference that there is another abstraction layer in the design. This layer uses knowledge from the speaker to test for liveness or act as an additional authentication factor. As an example, at the enrollment time, specific information such as a Personal Identification Number (PIN) or other private data may be stored about the speakers. At the verification time, randomized questions may be used to capture the test speaker's audio and the content of interest. The content is parsed by doing a transcription of the audio and using a natural language understanding (Manning, 1999) system to parse for the information of interest. This will increase the factors in the authentication and is usually a good idea for reducing the chance of successful impostor attacks – see Figure 1.

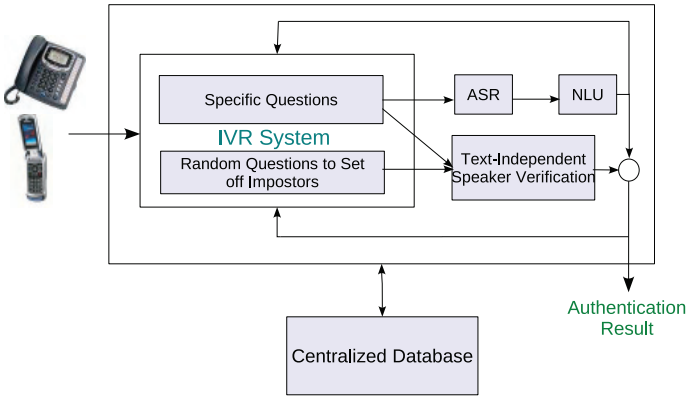



Fig. 1. A practical speaker recognition system utilizing speech recognition and natural language understanding

## 4. Challenges of speaker recognition

Aside from its positive outlook such as the established infrastructure and simplicity of adoption, speaker recognition, too, is filled with difficult challenges for the research community. Channel mismatch is the most serious difficulty faced in this technology. As an example, assume using a specific microphone over a channel such as a cellular communication channel with all the associated band-limitations and noise conditions in one session of using

a speaker recognition system. For instance, this session can be the enrollment session for instance.

Therefore, all that the system would learn about the identity of the individual is tainted by the channel characteristics through which the audio had to pass. On the hand, at the time of performing the identification or verification, a completely different channel could be used. For example, this time, the person being identified or verified may call from his/her home number or an office phone. These may either be digital phones going through voice T1 services or may be analog telephony devices going through analog switches and being transferred to digital telephone company switches, on the way.

They would have specific characteristics in terms of dynamics, cut-off frequencies, color, timber, etc. These channel characteristics are basically modulated with the characteristics of the person's vocal tract. Channel mismatch is the source of most errors in speaker recognition.

Another problem is signal variability. This is by no means specific to speaker recognition. It is a problem that haunts almost all biometrics. In general, an abundance of data is needed to be able to cover all the variations within an individual's voice. But even then, a person in two different sessions, would possibly have more variation within his/her own voice than if the signal is compared to that of someone else's voice, who possesses similar vocal traits.

The existence of wide intra-class variations compared with inter-class variations makes it difficult to be able to identify a person accurately. *Inter-class* variations denote the difference between two different individuals while *intra-class* variations represent the variation within the same person's voice in two different sessions.

The signal variation problem, as stated earlier, is common to most biometrics. Some of these variations may be due to aging and time-lapse effects. Time-lapse could be characterized in many different ways (Beigi, 2009). One is the aging of the individual. As we grow older, our vocal characteristics change. That is a part of aging in itself. But there are also subtle changes that are not that much related to aging and may be habitual or may also be dependent on the environment, creating variations from one session to another. These short-term variations could happen within a matter of days, weeks, or sometimes months. Of course, larger variations happen with aging, which take effect in the course of many years.

Another group of problems is associated with background conditions such as ambient noise and different types of acoustics. Examples would be audio generated in a room with echos or in a street while walking and talking on a mobile (cellular) phone, possibly with fire trucks, sirens, automobile engines, sledge hammers, and similar noise sources being heard in the background. These conditions affect the recognition rate considerably. These types of problems are quite specific to speaker recognition. Of course, similar problems may show up in different forms in other biometrics.

For example, analogous conditions in image recognition would show up in the form of noise in the lighting conditions. In fingerprint recognition they appear in the way the fingerprint is captured and related noisy conditions associated with the sensors. However, for biometrics such as fingerprint recognition, the technology may more readily dictate the type of sensors which are used. Therefore, in an official implementation, a vendor or an agency may require the use of the same sensor all around. If one considers the variations across sensors, different results may be obtained even in fingerprint recognition, although they would probably not be as pronounced as the variations in microphone conditions.

The original purpose of using speech has been to be able to convey a message. Therefore, we are used to deploying different microphones and channels for this purpose. One person, in general uses many different speech apparatuses such as a home phone, cellphone, office

phone, and even a microphone headset attached to a computer. We still expect to be able to perform reasonable speaker recognition using this varied set of sensors and channels. Although, as mentioned earlier, this becomes an advantage in terms of ease of adoptability of speaker recognition in existing arenas, it also makes the speaker recognition problem much more challenging.
Another problem is the presence of vocal variations due to illness. Catching a cold causes changes to our voice and its characteristics which could create difficulties in performing accurate speaker recognition. Bulk of the work in speaker recognition research is to be able to alleviate these problems, although not every problem is easily handled with the current technology.

## 5. Human speech generation and processing

A human child develops an inherent ability to identify the voice of his/her parents before even learning to understand the content of their speech. In humans, speaker recognition is performed in the right (less dominant) hemisphere of the brain, in conjunction with the functions for processing pitch, tempo, and other musical discourse. This is in contrast with most of the language functions (production and perception) in the brain which are processed by the *Broca* and *Wernicke* areas in the left (dominant) hemisphere of the *cerebral cortex* (Beigi, 2011).
Speech generation starts with the speech content being developed in the brain and processed through the nervous system. It includes the intended message which is created in the brain. The abstraction of this message is encoded into a code that will then produce the language (language coding step). The brain will then induce neuro-muscular activity to start the vocal tract in vocalizing the message. This message is transmitted over a channel starting with the air surrounding the mouth and continuing with electronic devices and networks such as a telephone system to transmit the coded message.
The resulting signal is therefore transmitted to the air surrounding the ear, where vibrations travel through different sections of the outer and the middle ear. The *cochlear* vibrations excite the *cilia* in the inner ear, generating neural signals which travel through the *Thalamus* to the brain. These signals are then decoded by different parts of the brain and are decoded into linguistic concepts which are understood by the receiving individual.
The intended message is embedded in the abstraction which is deduced by the brain from the signal being presented to it. This is a very complex system where the intended message generally contains a very low bit-rate content. However, the way this content undergoes transformation into a language code, neuro-muscular excitation, and finally audio, increase the bit-rate of the signal substantially, generating great redundancy.
Therefore a low information content is encoded to travel through a high-capacity channel. This small amount of information may easily be tainted by noise throughout this process.
Figure 2 depicts a control system representation of speech production proposed by (Beigi, 2011). Earlier, we considered the transformation of a message being formed in the brain into a high-capacity audio signal. In reality, the creation of the audio signal from the fundamental message formed in the brain may be better represented using a control system paradigm.
Let us consider the *Laplace transform* of the original message being generated in the brain as $U(s)$. We may lump together the different portions of the nervous system at work in generating the control signals which move the vocal tract to generate speech, into a controller block, $G_c(s)$. This block is made up of $G_b(s)$ which makes up those parts of the nervous system

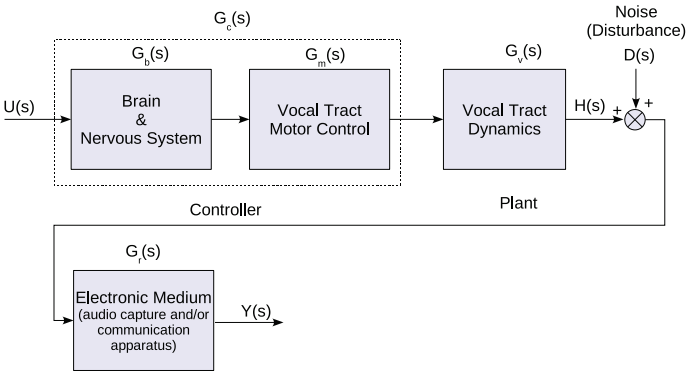



Fig. 2. Control system representation from (Beigi, 2011)

in the brain associated with generating the motor control signals and $G_m(s)$ which is the part of the nervous system associated with delivering the signal to the muscles in the vocal tract. The output of $G_c(s)$ is delivered to the vocal tract which is seen here as the plant. It is called $G_v(s)$ and it includes the moving parts of the vocal tract which are responsible for creating speech. The output, $H(s)$, is the *Laplace transform* of the speech wave, exciting the transmission medium, namely air. At this point we may model all the noise components and disturbances which may be present in the air surrounding the generated speech. The resulting signal is then transformed by passing through some type of electronic medium through audio capture and communication. The resulting signal, $Y(s)$ is the signal which is used to recognize the speaker.

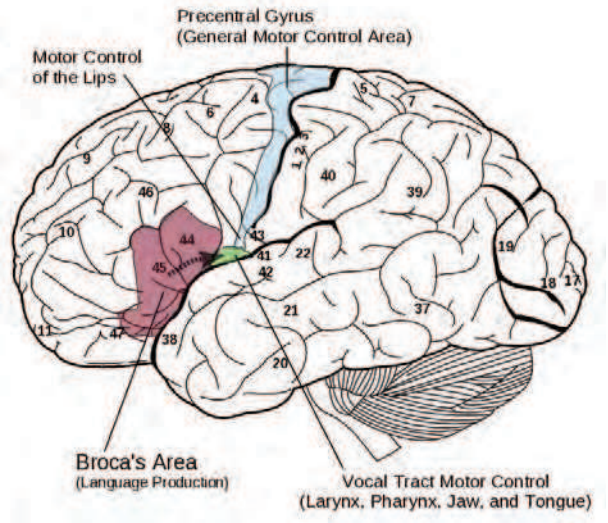



Fig. 3. Speech production in the Cerebral Cortex – from (Beigi, 2011)

Figure 3, borrowed from (Beigi, 2011), shows the superimposition of the interesting parts of the brain associated with producing speech. *Broca's area* which is part of the frontal lobe is

associated with producing the *language code* necessary for *speech production*. It may be seen as a part of the $G_b(s)$ in the control system representation of speech production. The *Precentral Gyrus* shown in the blue color is a long strip in the frontal lobe which is responsible for our *motor control*. The lower part of this area which is adjacent to Broca's area is further split into two parts. The lower part of the blue section is responsible for lip movement. The green part is associated with control of our *larynx*, *Pharynx*, *jaw*, and *tongue*. Together, these parts make up part of the $G_m(s)$ which is the second box in the controller. Note the proximity of these control regions to Broca's area, which is the coding section. Due to the slow transmission of chemical signals, the brain has evolved to allow for messages to travel quickly from $G_b(s)$ to $G_m(s)$ by utilizing proximity.
Note that *Broca's area* is also connected to the *language perception* section known as *Wernicke's area*. This will allow the feedback and refinement of the outgoing message. The vocal tract produces a carrier signal, based on its inherent dynamics, which is modified by the signal being generated by the $G_c(s)$. This is the actual plant which was called $G_v(s)$ in the control paradigm (Figure 2).
In *text-independent speaker recognition*, we are only concerned with learning the characteristics of the carrier signal in $G_v(s)$. *Speech recognition*, on the other hand, is concerned with decoding the intended message produced by Broca's area. This is why the signal processing is quite similar between the two disciplines, but in essence each discipline is concerned with a different part of the signal. The total time-signal is therefore a convolution of these two signals. The separation of these convolved signals is quite challenging and the results are therefore tainted in both disciplines causing a major part of the recognition error. Other sources are due to many complex disturbances along the way.
Figure 4 shows the major portion of the vocal tract which begins with the trachea and ends at the mouth and at the nose. It has a very plastic shape in which many of the cavities can change their shapes to be able to adjust the plant dynamics of Figure 2.

## 6. Theory and current approaches

The plasticity of the shape of the vocal tract makes the speech signal a non-stationary signal. This means that any segment of it, when compared to an adjacent segment in the time domain, has substantially different characteristics, indicating that the dynamics of the system producing these sections varies with time.
As mentioned in the *Introduction*, the first step is to store the vocal characteristics of the speakers in the form of speaker models in a database, for future reference. To build these models, certain features should be defined such that they would best represent the vocal characteristics of the speaker of interest. The most prevalent features used in the field happen to be identical to those used for speech recognition, namely, Mel Frequency Cepstral Coefficients (MFCCs) – see (Beigi, 2011).

### 6.1 Sampling

A Discrete representation of the signal is used for Automatic Speaker Recognition. Therefore we need to utilize the sampling theorem to help us determine the appropriate sampling frequency to be used for converting the continuous speech signal into its discrete signal representation.
One must therefore ensure that the sampling rate is picked in accordance with the guidelines set by the *Whittaker-Kotelnikoff-Shannon (WKS)* sampling theorem (Beigi, 2011). The WKS sampling theorem requires that the sampling frequency be at least two times the *Nyquist*

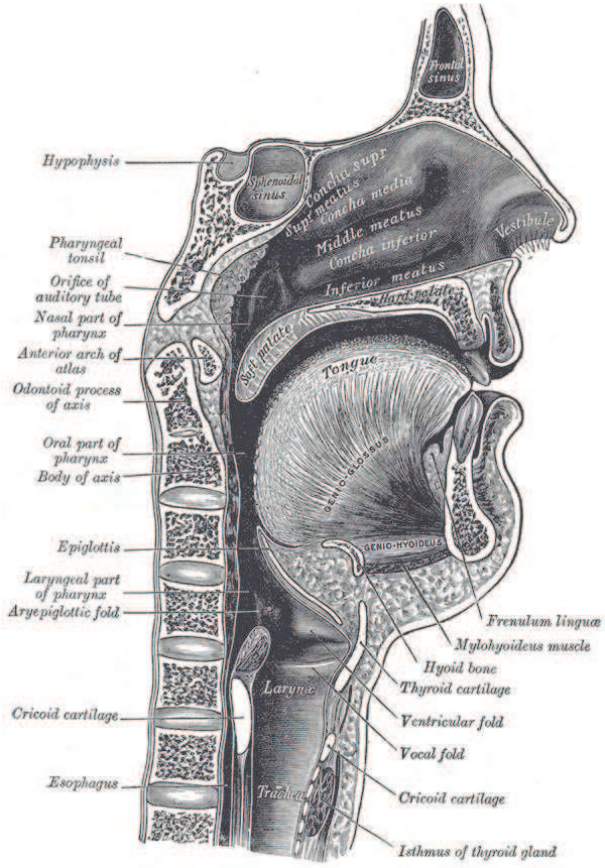



Fig. 4. Sagittal section of Nose, Mouth, Pharynx, and Larynx; *Source: Gray's Anatomy (Gray, 1918)*

*Critical frequency*. The *Nyquist critical frequency* is really the highest frequency content of the analog signal. For simplicity, normally an *ideal sampler* is used, which acts like the multiplication of an impulse train with the analog signal, where the impulses happen at the chosen sampling frequency.
In this representation, each sample has a zero width and lasts for an instant. The sampling theorem may be stated in words by requiring that the sampling frequency be greater than or equal to the *Nyquist rate*. The *Nyquist rate*, is defined as two times the *Nyquist critical frequency*.

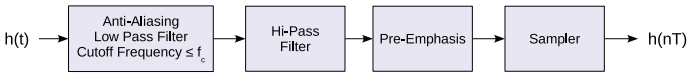



Fig. 5. Block diagram of a typical sampling process

Figure 5 shows a typical sampling process which starts with an analog signal and produces a series of discrete samples at a fixed frequency, representing the speech signal. The discrete samples are usually stored using a Codec (Coder/Decoder) format such as *linear PCM*, *μ-Law*,

*a-Law*, etc. Standardization is quite important for interoperability with different recognition engines (Beigi & Markowitz, 2010).
There are different forms for representing the speech signal. The simplest one is the *speech waveform* which is basically the plot of the sampled points versus time. In general, the amplitude is normalized to dwell between −1 and 1. In its quantized form, the data is stored in the range associated with the quantization representation. For example, for a *16-bit signed linear PCM*, it would go from −32768 to 32767.

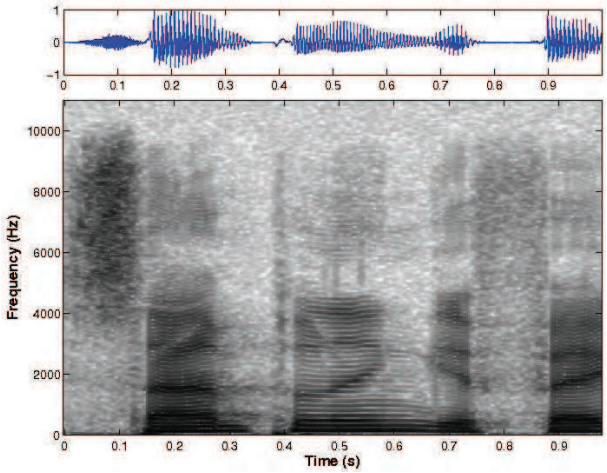



Fig. 6. Narrowband spectrogram of a speech signal

Another representation is, so-called, the *spectrogram* of the signal. Figure 6 shows the *narrowband spectrogram* of a signal. A sliding widow of 23 ms was used for to generate this figure. The spectrogram shows the frequency content of the speech signal as a function of time. It is really a three-dimensional representation where the z-axis is depicted by the darkness of the points on the figure. The darker the pixel, the stronger the signal strength in that frequency for the time slice of choice. An artifact of the narrowband spectrogram is the existence of the horizontal curved lines across time. A *speech waveform* representation has also been plotted on top of the spectrogram of Figure 6 to show the relation between different points in the waveform with their corresponding frequency content in the spectrogram.
The system of Figure 5 should be designed so that it reduces *aliasing*, *truncation*, *band-limitation*, and *jitter* by choosing the right parameters, such as the sampling rate and volume normalization. Figure 7 shows how most of the fricative information is lost going from a 22 kHz sampling rate to 8 kHz. Normal telephone sampling rates are at best 8 kHz. Mostly everyone is familiar with having to qualify fricatives on the telephone by using statements such as “S” as in “Sam” and “F” as in “Frank”.

### 6.2 Feature extraction

*Cepstral coefficients* have fallen out of studies in exploring the arrival of echos in nature (Bogert et al., 1963). They are related to the spectrum of the log of spectrum of a speech signal. The frequency domain of the signal in computing the *MFCCs* is warped to the M̂elody (Mel) scale. It is based on the premise that human perception of pitch is linear up to 1000 Hz and then becomes nonlinear for higher frequencies (somewhat logarithmic). There are models of the

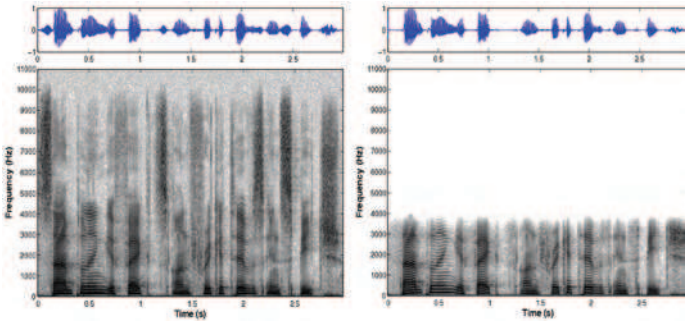


Fig. 7. Utterance: "Sampling Effects on Fricatives in Speech", sampled at 22kHz (left) and 8kHz (right)

human perception based on other warped scales such as the *Bark scale*. There are several ways of computing Cepstral Coefficients. They may be computed using the *Direct Method*, also known as *Moving Average (MA)* which utilizes the *Fast Fourier Transform (FFT)* for the first pass and the *Discrete Cosine Transform (DCT)* for the second pass to ensure real coefficients. This method usually entails the following steps:

1. Framing – Selecting a sliding section of the signal with a fixed width in time which is then moved with some overlap. The sliding window is generally about 30*ms* with an overlap of about 20*ms* (10*ms* shift).
2. Windowing – A window such as a Hamming, Hann, Welch, etc. is used to smooth the signal for the computation of the *Discrete Fourier Transform (DFT)*.
3. FFT – The *Fast Fourier Transform (FFT)* is generally used for approximating the DFT of the windowed signal.
4. Frequency Warping – The FFT results are warped in the frequency domain in accordance with the Melody (Mel) or Bark scale.
5. MFCC – The Mel Frequency Cepstral Coefficients (MFCC) are computed.
6. Mel Cepstral Dynamics – Delta and Delta-Delta Cepstra are computed based on adjacent MFCC values.

Some use the *Linear Predictive*, also known as *AutoRegressive (AR)* features by themselves: *Linear Predictive Coefficients (LPC)*, *Partial Correlation (PARCOR)* – also known as *reflection coefficients*, or *log area ratios*. However, mostly the LPCs are converted to cepstral coefficients using autocorrelation techniques (Beigi, 2011). These are called Linear Predictive Cepstral Coefficients (LPCCs). There are also the Perceptual Linear Predictive (PLP) (Hermansky, 1990) features, shown in Figure 9. PLP works by warping the frequency and spectral magnitudes of the speech signal based on auditory perception tests. The domain is changed from magnitudes and frequencies to loudness and pitch (Beigi, 2011).
There have been an array of other features used such as *wavelet filterbanks* (Burrus et al., 1997), for example in the form of *Mel-Frequency Discrete Wavelet Coefficients* and *Wavelet Octave Coefficients of Residues (WOCOR)*. There are also *Instantaneous Amplitudes and Frequencies* which are in the form of *Amplitude Modulation (AM)* and *Frequency Modulation (FM)*. These features come in different flavors such as *Empirical Mode Decomposition (EMD)*, *FEPSTRUM*, *Mel Cepstrum Modulation Spectrum (MCMS)*, and so on (Beigi, 2011).

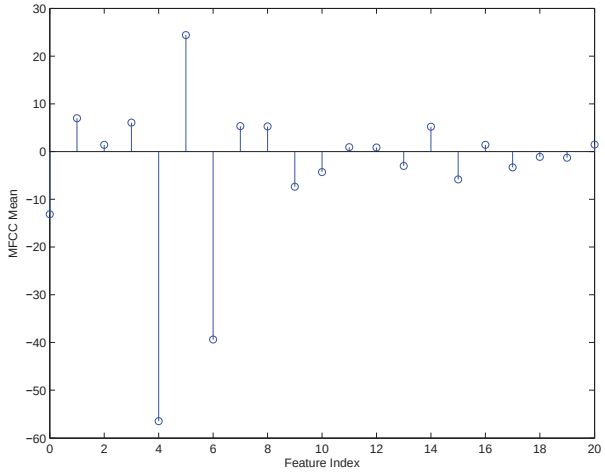



Fig. 8. A sample MFCC vector – from (Beigi, 2011)

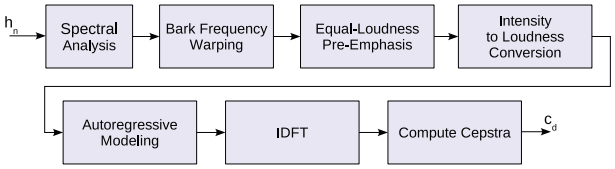



Fig. 9. A typical Perceptual Linear Predictive (PLP) system

It is important to note that most audio segments include a good deal of silence. Addition of features extracted from silent areas in the speech will increase the similarity of models, since silence does not carry any information about the speaker's vocal characteristics. Therefore, *Silence Detection (SD)* or *Voice Activity Detection (VAD)* (Beigi, 2011) is quite important for better results. Only segments with vocal signals should be considered for recognition. Other preprocessing such as *Audio Volume Estimation* and normalization and *Echo Cancellation* may also be necessary for obtaining desirable result (Beigi, 2011).

### 6.3 Speaker models

Once the features of interest are chosen, models are built based on these features to represent the speakers' vocal characteristics. At this point, depending on whether the system is text-dependent (including text-prompted) or text-independent, different methods may be used. Models are usually based on HMMs, GMMs, SVMs, and NNs.

#### 6.3.1 Gaussian Mixture Models (GMM)

In general, there are many different modeling scenarios for speaker recognition. Most of these techniques are similar to those used for speech recognition modeling. For example, a *multi-state ergodic Hidden Markov Models* is usually used for text-dependent speaker recognition since there is textual context. As a special case of Hidden Markov Models, *Gaussian Mixture Models (GMM)* are used for doing text-independent speaker recognition. This is probably the most popular technique which is used in this field. GMMs are basically single-state degenerate HMMs.

The models are tied to the type of learning that is done. A popular technique is the use of a Gaussian Mixture Model (GMM) (Duda & Hart, 1973) to represent the speaker. This is mostly relevant to the text-independent case which encompasses speaker identification and text-independent verification. Even text-dependent techniques can use GMMs, but, they usually use a GMM to initialize Hidden Markov Models (HMMs) (Poritz, 1988) built to have an inherent model of the content of the speech as well. Many speaker diarization (segmentation and ID) systems use GMMs. To build a Gaussian Mixture Model of a speaker's speech, one should make a few assumptions and decisions. The first assumption is the number of Gaussians to use. This is dependent on the amount of data that is available and the dimensionality of the feature vectors.
Standard clustering techniques are usually used for the initial determination of the Gaussians. Once the number of Gaussians is determined, some large pool of features is used to train these Gaussians (learn the parameters). This step is called training. The models generated by training are called by many different names such as *background models*, *universal background models (UBM)*, *speaker independent models*, *Base models*, etc.
In a GMM, the models are parameters for collections of multi-variate normal density functions which describe the distribution of the Mel-Cepstral features (Beigi, 2011) for speakers' enrollment data. This distribution is represented by Equation 1.

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \quad (1)$$
$$where \begin{cases} \mathbf{x},\ \boldsymbol{\mu} \in \mathcal{R}^d \\ \mathbf{\Sigma} : \mathcal{R}^d \mapsto \mathcal{R}^d \end{cases}$$

In Equation 1, $\boldsymbol{\mu}$ is the mean vector where,

$$\boldsymbol{\mu} \triangleq \mathcal{E}\{\mathbf{x}\} \triangleq \int_{-\infty}^{\infty} \mathbf{x}\, p(\mathbf{x}) d\mathbf{x} \quad (2)$$

The so-called "Sample Mean" approximation for Equation 2 is,

$$\boldsymbol{\mu} \approx \frac{1}{N} \sum_{i=0}^{N-1} \mathbf{x}_i \quad (3)$$

where $N$ is the number of samples and $\mathbf{x}_i$ are the Mel-Cepstral feature vectors (Beigi, 2011).
The Variance-Covariance matrix of a multi-dimensional random variable is defined as,

$$\mathbf{\Sigma} \triangleq \mathcal{E}\left\{(\mathbf{x}-\mathcal{E}\{\mathbf{x}\})(\mathbf{x}-\mathcal{E}\{\mathbf{x}\})^T\right\} \quad (4)$$
$$= \mathcal{E}\left\{\mathbf{x}\mathbf{x}^T\right\} - \boldsymbol{\mu}\boldsymbol{\mu}^T \quad (5)$$

This matrix is called the *Variance-Covariance* since the diagonal elements are the variances of the individual dimensions of the multi-dimensional vector, $\mathbf{x}$. The off-diagonal elements are the covariances across the different dimensions. Some have called this matrix the *Variance* matrix. Mostly in the field of Pattern Recognition it has been referred to, simply, as the *Covariance* matrix which is the name we will adopt here.
The Unbiased estimate of $\mathbf{\Sigma}$, $\tilde{\mathbf{\Sigma}}$ is given by the following expression,

$$\tilde{\mathbf{\Sigma}} = \frac{1}{N-1}\sum_{i=0}^{N-1}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \tag{6}$$

$$= \frac{1}{N-1}\left[\mathbf{S}_{xx} - N(\boldsymbol{\mu}\boldsymbol{\mu}^T)\right] \tag{7}$$

where the sample mean $\boldsymbol{\mu}$ is given by Equation 3 and the second order sum matrix, $\mathbf{S}_{xx}$ is given by,

$$\mathbf{S}_{xx} = \sum_{i=0}^{N-1}\mathbf{x}_i\mathbf{x}_i^T \tag{8}$$

After the training is done, generally, the basis for a speaker independent model is built and stored in the form of the above statistics. At this stage, depending on whether a *universal background model* (UBM) (Reynolds et al., 2000) or *cohort models* are desired, different processing is done. For a UBM, a pool of speakers is used to optimize the parameters of the Gaussians as well as the mixture coefficients, using standard techniques such as *maximum likelihood estimation (MLE)*, *Maximum a-Posteriori (MAP)* adaptation and *Maximum Likelihood Linear Regression (MLLR)*. There may be one or more *Background models*. For example, some create a single background model called the UBM, others may build one for each gender, by using separate male and female databases for the training. *Cohort models*(Beigi et al., 1999) are built in a similar fashion. A cohort is a set of speakers that have similar vocal characteristics to the target speaker. This information may be used as a basis to either train a Hidden Markov Model including textual context, or to do an expectation maximization in order to come up with the statistics for the underlying model.
At this point, the system is ready for performing the enrollment. The enrollment may be done by taking a sample audio of the target speaker and adapting it to be optimal for fitting this sample. This ensures that the likelihoods returned by matching the same sample with the modified model would be maximal.

#### 6.3.2 Support vector machines

*Support vector machines (SVMs)* have been recently used quite often in research papers regarding speaker recognition. Although they show very promising results, most of the implementations suffer from huge optimization problems with large dimensionality which have to be solved at the training stage. Results are not substantially different from GMM techniques and in general it may not be warranted to use such costly optimization calculations.
The claim-to-fame of *support vector machines* (*SVMs*) is that they determine the boundaries of classes, based on the training data, and they have the capability of *maximizing the margin* of class separability in the feature space. (Boser et al., 1992) states that the number of parameters used in a support vector machine is automatically computed (see *Vapnik-Chervonenkis* (*VC*) *dimension* (Burges, 1998; Vapnik, 1998)) to present a solution in terms of a linear combination of a subset of observed (training) vectors, which are located closest to the decision boundary. These vectors are called *support vectors* and the model is known as a *support vector machine*.
Vapnik (Vapnik, 1979) pioneered the statistical learning theory of *SVMs*, which is based on minimizing the classification error of both the training data and some unknown (held-out) data. Of course, the core of support vector machines and other *kernel* techniques stems from

much earlier work on setting up and solving *integral equations*. Hilbert (Hilbert, 1912) was one of the main developers of the formulation of *integral equations* and *kernel transformations*.
One of the major problems with SVMs is their intensive need for memory and computation power at the training stage. Training of SVMs for speaker recognition also suffers from these limitations. To address this issue, new techniques have been developed to split the problem into smaller subproblems which would then be solved in parallel as a network of problems. One such technique is known as *cascade SVM* (Tveit & Engum, 2003) for which certain improvements have also been proposed in the literature (Zhang et al., 2005).
Some of the shortcomings of *SVMs* have been addressed by combining them with other learning techniques such as *fuzzy logic* and *decision trees*. Also, to speed up the training process, several techniques based on the decomposition of the problem and selective use of the training data have been proposed.
In application to speaker recognition, experimental results have shown that *SVM* implementations of speaker recognition are slightly inferior to *GMM* approaches. However, it has also been noted that systems which combine *GMM* and *SVM* approaches often enjoy a higher accuracy, suggesting that part of the information revealed by the two approaches may be complementary (Solomonoff et al., 2004). For a detailed coverage, see (Beigi, 2011).
In general SVMs are two-class classifiers. That's why they are suitable for the speaker verification problem which is a two-class problem of comparing the voice of an individual to his/her model versus a background population model. N-class classification problems such as speaker identification have to be reduced to N two-class classification problems where the $i^{th}$ two-class problem compares the $i^{th}$ class with the rest of the classes combined (Vapnik, 1998). This can become quite computationally intensive for large-scale speaker identification problems. Another problem is that the Kernel function being used by SVMs is almost magically chosen.

#### 6.3.3 Neural networks

Another modeling paradigm is the neural network perspective. There are quite a number of different neural networks and related architectures such as feed forward networks, TDNNs, probabilistic random access memory or pRAM models, Hierarchical Mixtures of Experts or HMEs, etc. It would take an enormous amount of time to go through all these and other possibilities. See (Beigi, 2011) for details.

#### 6.3.4 Model adaptation (enrollment)

For a new person being enrolled in the system, the base speaker-independent models are modified to match the *a-posteriori statistics* of the enrolled person or target speaker's sample enrollment speech. This is done by any technique such as *maximum a-posteriori probability estimation (MAP)*, for example, using *expectation maximization (EM)*, or *maximum likelihood linear regression* for text-independent systems or simply by modifying the counts of the transitions on a *hidden Markov model (HMM)* for text-dependent systems.

## 7. Speaker recognition

At the identification and verification stage, a new sample is obtained for the test speaker. In the identification process, the sample is used to compute the likelihood of this sample being generated by the different models in the database. The identity of the model that returns the highest likelihood is returned as the identity of the test speaker. In identification, the

results are usually ranked by these likelihoods. To ensure a good dynamic range and better discrimination capability, log of the likelihood is computed.

At the verification stage, the process becomes very similar to the identification process described earlier, with the exception that instead of computing the log likelihood for all the models in the database, the sample is only compared to the model of the target speaker and the background or cohort models. If the target speaker model provides a better log likelihood, the test speaker is verified and otherwise rejected. The comparison is done using the *Log Likelihood Ratio (LLR)* test.

An extension of speaker recognition is diarization which includes segmentation followed by speaker identification and sometimes verification. The segmentation finds abrupt changes in the audio stream. *Bayesian Information Criterion (BIC)* (Chen & Gopalakrishnan, 1998) and Generalized Likelihood Ratio (GLR) techniques and their combination (Ajmera & McCowan, 2004) as well as other techniques (Beigi & Maes, 1998) have been used for the initial segmentation of the audio. Once the initial segmentation is done, a limited speaker identification procedure allows for tagging of the different parts with different labels. Figure 10 shows such a results for a two-speaker segmentation.

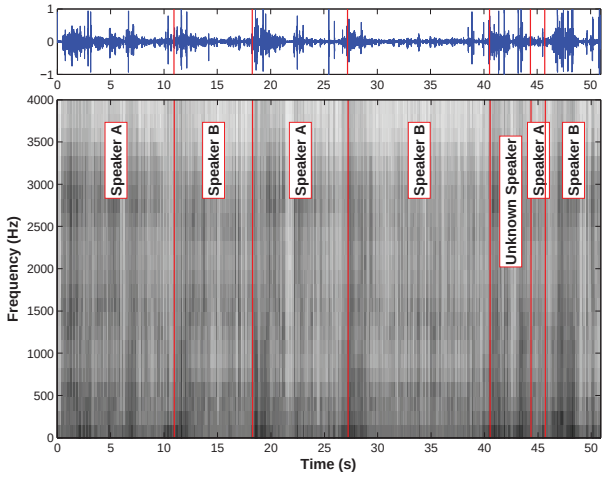



Fig. 10. Segmentation and labeling of two speakers in a conversation using turn detection followed by identification

### 7.1 Representation of results

Speaker identification results are usually presented in terms of the error rate. They may also be presented as the error rate based on the true result being present in the top $N$ matches. This case is usually more prevalent in the cases where identification is used to prune a large set of speakers to only a handful of possible matches so that another expert system (human or machine) would finalize the decision process.

In the case of speaker verification, the method of presenting the results is somewhat more controversial. In the early days in the field, a *Receiver Operating Characteristic (ROC)* curve was used (Beigi, 2011). For the past decade, the *Detection Error Trade-Off (DET)* curve (Martin et al., 1997; Martin & Przybocki, 2000) has been more prevalent, with a measurement of the cost of producing the results, called the *Detection Cost Function (DCF)* (Martin & Przybocki, 2000).

Figures 11 and 12 show sample DET curves for two sets of data underscoring the difference in performances. Recognition results are usually quite data-dependent. The next section will speak about some open problems which degrade results.

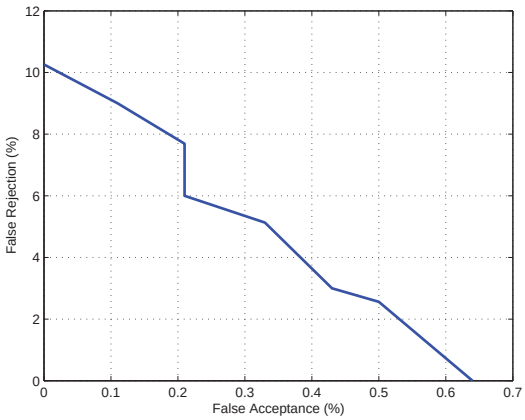



Fig. 11. DET Curve for quality data

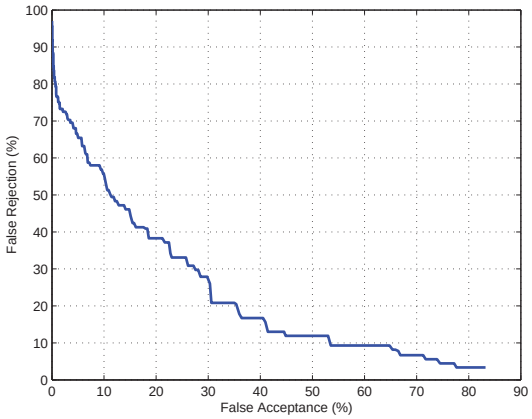



Fig. 12. DET Curve for highly mismatched and noisy data

There is a controversial operating point on the DET curve which is usually marked as the point of comparison between different results. This point is called the *Equal Error Rate (EER)* and signifies the operating point where the false rejection rate and the false acceptance rate are equal. This point does not carry any real preferential information about the "correct" or "desired" operating point. It is mostly a point of convenience which is easy to denote on the curve.

## 8. State of the art

In designing a practical speaker recognition system, one should try to affect the interaction between the speaker and the engine to be able to capture as many vowels as possible. Vowels are periodic signals which carry much more information about the resonance subtleties of the vocal tract. In the text-dependent and text-prompted cases, this may be done by actively designing prompts that include more vowels. For text-independent cases, the simplest way is to require more audio in hopes that many vowels would be present. Also, when speech recognition and natural language understanding modules are included (Figure 1), the conversation may be designed to allow for higher vowel production by the speaker.
As mentioned earlier, the greatest challenge in speaker recognition is *channel-mismatch*. Considering the general communication system given by Figure 13, it is apparent that the channel and noise characteristics at the time of communication are modulated with the original signal. Removing these channel effects is the most important problem in information theory. This is of course a problem where the goal is to recognize the message being sent. It is, however, a much bigger problem when the quest is the estimation of the model that generated the message – as it is with the speaker recognition problem. In that case, the channel characteristics have mixed in with the model characteristics and their separation is nearly impossible. Once the same source is transmitted over an entirely different channel with its own noise characteristics, the problem of learning the source model becomes even harder.

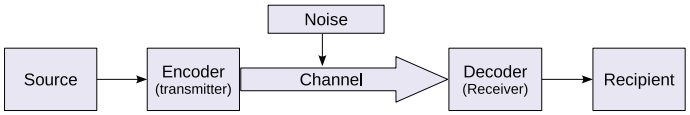



Fig. 13. One-way communication

Many techniques are used for resolving this problem, but it is still the most important source of errors in speaker recognition. It is the reason why most systems that have been trained on a predetermined set of channels, such as landline telephone, could fail miserably when cellular (mobile) telephones are used. The techniques that are being used in the industry are listed here, but there are more techniques being introduced every day:

- **Spectral Filtering and Cepstral Liftering**
  - Cepstral Mean Subtraction (CMS) or
    Cepstral Mean Normalization (CMN) (Benesty et al., 2008)
  - Cepstral Mean and Variance Normalization (CMVN) (Benesty et al., 2008)
  - Histogram Equalization (HEQ) (de la Torre et al., 2005) and
    Cepstral Histogram Normalization (CHN) (Benesty et al., 2008)
  - AutoRegressive Moving Average (ARMA) (Benesty et al., 2008)
  - RelAtive SpecTrAl (RASTA) Filtering (Hermansky, 1991; van Vuuren, 1996)
  - J-RASTA (Hardt & Fellbaum, 1997)
  - Kalman Filtering (Kim, 2002)
- **Other Techniques**
  - Vocal Tract Length Normalization (VTLN) – first introduced for
    speech recognition: (Chau et al., 2001) and later for speaker recognition (Grashey & Geibler, 2006)
  - Feature Warping (Pelecanos & Sridharan, 2001)

- Feature Mapping (Reynolds, 2003)
- Speaker Model Synthesis (SMS) (R. et al., 2000)
- Speaker Model Normalization (Beigi, 2011)
- H-Norm (Handset Normalization) (Dunn et al., 2000)
- Z-Norm and T-Norm (Auckenthaler et al., 2000)
- Joint Factor Analysis (JFA) (Kenny, 2005)
- Nuisance Attribute Projection (NAP) (Solomonoff et al., 2004)
- Total Variability (i-vector) (Dehak et al., 2009)

Recently, depending on whether GMMs are used or SVMs, the two techniques of *joint factor analysis (JFA)* and *nuisance attribute projection (NAP)* have been used respectively, in most research reports.

Joint factor analysis (JFA) (Kenny, 2005) is based on factor analysis (FA) (Jolliffe, 2002). *FA* is a linear transformation which makes the assumption of having an *explicit model* which differentiates it from principal component analysis (PCA) and linear discriminant analysis (LDA). In fact in some perspective, it may be seen as a more general version of *PCA*. *FA* assumes that the underlying random variable is composed of two different components.

The first component is a random variable, called the *common factors*, which has a lower dimensionality compared to the combined random state, $X$, and the observation, $Y$. It is called the vector of *common factors* since the same vector, $\Theta : \boldsymbol{\theta} : \mathcal{R}^1 \mapsto \mathcal{R}^M, M <= D$, is a component of all the samples of $\mathbf{y}_n$.

The second component is the, so called, vector of *specific factors*, or sometimes called the *error* or the *residual* vector. It is denoted by $E : (\mathbf{e})_{D1}$. Therefore, this linear *FA* model for a specific random variable, $\tilde{Y} : \tilde{\mathbf{y}} : \mathcal{R}^q \mapsto \mathcal{R}^D$, related to the observed random variable $Y$ may be written as follows,

$$\tilde{\mathbf{y}}_n = \mathbf{V}\boldsymbol{\theta}_n + \mathbf{e}_n \tag{9}$$

where $\mathbf{V} : \mathcal{R}^M \mapsto \mathcal{R}^D$ is known as the *factor loading* matrix and its elements, $(\mathbf{V})_{dm}$, are known as the *factor loadings*. Samples of random variable $\Theta : (\boldsymbol{\theta}_n)_{M1}, n \in \{1, 2, \cdots, N\}$ are known as the vectors of *common factors*, since due to the linear combination nature of the factor loading matrix, each element, $(\boldsymbol{\theta})_m$, has a hand in shaping the value of (generally) all $(\tilde{\mathbf{y}}_n)_d, d \in \{1, 2, \cdots, D\}$. Samples of random variable $E : \mathbf{e}_n, n \in \{1, 2, \cdots, N\}$ are known as vectors of *specific factors*, since each element, $(\mathbf{e}_n)_d$ is specifically related to a corresponding, $(\tilde{\mathbf{y}}_n)_d$.

JFA uses the concept of FA to split the space of the model parameters into speaker model parameters and channel parameters. It makes the assumption that the channel parameters are normally distributed, have a smaller dimensionality, and are common to all training samples. The model parameters, on the other hand, are common for each speaker. This separation allows for learning the channel characteristics in the form of separate model parameters, hence producing pure and somewhat channel-independent speaker models.

*Nuisance attribute projection* (*NAP*) (Solomonoff et al., 2004) is a method of modifying the original *kernel*, being used for the *support vector machine* (*SVM*) formulation, to one with the capability of telling specific channel information apart. The premise behind this approach is that by doing so, in both training and recognition stages, the system will not have the ability to distinguish channel specific information. This channel specific information is what is dubbed nuisance by (Solomonoff et al., 2004). *NAP* is a projection technique which assumes that most of the information related to the channel is stored in specific low-dimensional subspaces

of the higher dimensional space to which the original features are mapped. Furthermore, these regions are assumed to be somewhat distinct from the regions which carry speaker information.
Some even more recent developments have been made in speaker modeling. The *identity vector* or *i-vector* is a new representation of a speaker in a space of speakers called the *total variability space*. This model came from an observation by (Dehak et al., 2009) that the *channel space* in JFA still contained some information which may be used to distinguish speakers. This triggered the following representation of the GMM supervector of means ($\boldsymbol{\mu}$) which contains both speaker- and channel-dependent information.

$$\boldsymbol{\mu} = \boldsymbol{\mu}_I + \mathbf{T}\boldsymbol{\omega} \tag{10}$$

In Equation 10, $\boldsymbol{\mu}$ is assumed to be *normally distributed* with $\mathcal{E}\{\boldsymbol{\mu}\} = \boldsymbol{\mu}_I$, where $\boldsymbol{\mu}_I$ is the GMM supervector computed over the speaker- and channel-independent model which may be chosen to be the *universal background model*. The covariance for $\boldsymbol{\mu}$ is assumed to be $Cov(\boldsymbol{\mu}) = \mathbf{T}\mathbf{T}^T$, where $\mathbf{T}$ is a low-rank matrix, and $\boldsymbol{\omega}$ is the *i-vector* which is a standard normally distributed vector ($p(\boldsymbol{\omega}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$). The *i-vector* represents the coordinates of the speaker in the, so-called, *total variability space*.

## 9. Future of the research

There are many challenges that have not been fully addressed in different branches of speaker recognition. For example, the large-scale speaker identification problem is one that is quite hard to handle. In most cases when researchers speak of large-scale in the identification arena, they speak of a few thousands of enrolled speakers. As the number of speakers increases to millions or even billions, the problem becomes quite challenging. As the number of speakers increases, doing an exhaustive match through the whole population becomes almost computationally implausible. Hierarchical techniques (Beigi et al., 1999) would have to be utilized to handle such cases. In addition, the speaker space is really a continuum. This means that if one considers a space where speakers who are closer in their vocal characteristics would be placed near each other in that space, then as the number of enrolled speakers increases, there will always be a new person that would fill in the space between any two neighboring speakers. Since there are intra-speaker variabilities (differences between different samples taken from the same speaker), the intra-speaker variability will be at some point more than inter-speaker variabilities, causing confusion and eventually identification errors. Since there are presently no large databases (in the order of millions and higher), there is no indication of the results, both in terms of the speed of processing and accuracy.
Another challenge is the fact that over time, the voice of speakers may change due to many different reasons such as illness, stress, aging, etc. One way to handle this problem is to have models which constantly adapt to changes (Beigi, 2009).
Yet another problem plagues speaker verification. Neither background models nor cohort models are error-free. Background models generally smooth out many models and unless the speaker is considerably different from the norm, they may score better than the speaker's own model. This is especially true if one considers the fact that nature is usually Gaussian and that there is a high chance that the speaker's characteristics are close to the smooth background model. If one were to only test the target sample on the target model, this would not be a problem. But since a test sample which is different from the target sample (used for creating the model) is used, the intra-speaker variability might be larger than the inter-speaker variability between the test speech and the smooth background model.

There are, of course, many other open problems. Some of these problems have to do with acceptable noise levels until break-down occurs. Using a cellular telephone with its inherently bandlimited characteristics in a very noisy venue such as a subway (metro) station is one such challenge.

Given the number of different operating conditions in invoking speaker recognition, it is quite difficult for technology vendors to provide objective performance results. Results are usually quite data-dependent and different data sets may pronounce particular merits and downfalls of each provider's algorithms and implementation. A good speaker verification system may easily achieve an 0% EER for clean data with good inter-speaker variability in contrast with intra-speaker variability. It is quite normal for the same "good" system to show very high equal error rates under severe conditions such as high noise levels, bandwidth limitation, and small relative inter-speaker variability compared to intra-speaker variability. However, under most controlled conditions, equal error rates below 5% are readily achieved. Similar variability in performance exists in other branches of speaker recognition, such as identification, etc.

## 10. References


Ajmera, J. & McCowan, I.and Bourlard, H. (2004). Robust speaker change detection, *IEEE Signal Processing Letters* 11(8): 649–651.

Auckenthaler, R., Carey, M. & Lloyd-Thomas, H. (2000). Score normalization for text-independent speaker verification systems, *Digital Signal Processing* 10(1–3): 42–54.

Beigi, H. (2009). Effects of time lapse on speaker recognition results, *16th Internation Conference on Digital Signal Processing*, pp. 1–6.

Beigi, H. (2011). *Fundamentals of Speaker Recognition*, Springer, New York. ISBN: 978-0-387-77591-3.

Beigi, H. & Markowitz, J. (2010). Standard audio format encapsulation (safe), *Telecommunication Systems* pp. 1–8. 10.1007/s11235-010-9315-1.
URL: *http://dx.doi.org/10.1007/s11235-010-9315-1*

Beigi, H. S., Maes, S. H., Chaudhari, U. V. & Sorensen, J. S. (1999). A hierarchical approach to large-scale speaker recognition, *EuroSpeech 1999*, Vol. 5, pp. 2203–2206.

Beigi, H. S. & Maes, S. S. (1998). Speaker, channel and environment change detection, *Proceedings of the World Congress on Automation (WAC1998)*.

Benesty, J., Sondhi, M. M. & Huang, Y. (2008). *Handbook of Speech Processing*, Springer, New york. ISBN: 978-3-540-49125-5.

Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska-Delacrétaz, D. & Reynolds, D. (2004). A tutorial on text-independent speaker verification, *EURASIP Journal on Applied Signal Processing* 2004(4): 430–451.

Bogert, B. P., Healy, M. J. R. & Tukey, J. W. (1963). The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking, *in* M. Rosenblatt (ed.), *Time Series Analysis*, pp. 209–243. Ch. 15.

Boser, B. E., Guyon, I. M. & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers, *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery* 2: 121–167.

Burrus, C. S., Gopinath, R. A. & Guo, H. (1997). *Introduction to Wavelets and Wavelet Transforms: A Primer*, Prentice Hall, New york. ISBN: 0-134-89600-9.

Campbell, J.P., J. a. (1997). Speaker recognition: a tutorial, *Proceedings of the IEEE* 85(9): 1437–1462.

Chau, C. K., Lai, C. S. & Shi, B. E. (2001). Feature vs. model based vocal tract length normalization for a speech recognition-based interactive toy, *Active Media Technology*, Lecture Notes in Computer Science, Springer, Berlin/Heidelberg, pp. 134–143. ISBN: 978-3-540-43035-3.

Chen, S. S. & Gopalakrishnan, P. S. (1998). Speaker, environemnt and channel change detection and clustering via the bayesian inromation criterion, *IBM Techical Report, T.J. Watson Research Center*.

Dehak, N., Dehak, R., Kenny, P., Brummer, N., Ouellet, P & Dumouchel, P. (2009). Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification, *Interspeech*, pp. 1559–1562.

de la Torre, A., Peinado, A. M., Segura, J. C., Perez-Cordoba, J. L., Benitez, M. C. & Rubio, A. J. (2005). Histogram equalization of speech representation for robust speech recognition, *IEEE Transaction of Speech and Audio Processing* 13(3): 355–366.

Duda, R. O. & Hart, P. E. (1973). *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York. ISBN: 0-471-22361-1.

Dunn, R. B., Reynolds, D. A. & Quatieri, T. F. (2000). Approaches to speaker detection and tracking in conversational speech, *Digital Signal Processing* 10: 92–112.

Furui, S. (2005). 50 years of progress in speech and speaker recognition, *Proc. SPECOM*, pp. 1–9.

Grashey, S. & Geibler, C. (2006). Using a vocal tract length related parameter for speaker recognition, *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, pp. 1–5.

Gray, H. (1918). *Anatomy of the Human Body*, 20th edn, LEA and FEBIGER, Philadelphia. Online version, New York (2000).
URL: *http://www.Bartleby.com*

Hardt, D. & Fellbaum, K. (1997). Spectral subtraction and rasta-filtering in text-dependent hmm-based speaker verification, *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, Vol. 2, pp. 867–870.

Hermansky, H. (1990). Perceptual linear predictive (plp) analysis of speech, 87(4): 1738–1752.

Hermansky, H. (1991). Compensation for the effect of the communication channel in the auditory-like analysis of speech (rasta-plp), *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH-91)*, pp. 1367–1370.

Hilbert, D. (1912). *Grundzüge Einer Allgemeinen Theorie der Linearen Integralgleichungen (Outlines of a General Theory of Linear Integral Equations)*, Fortschritte der Mathematischen Wissenschaften, heft 3 (Progress in Mathematical Sciences, issue 3), B.G. Teubner, Leipzig and Berlin. In German. Originally published in 1904.

Jolliffe, I. (2002). *Principal Component Analysis*, 2nd edn, Springer, New york.

Kenny, P. (2005). Joint factor analysis of speaker and session varaiability: Theory and algorithms, *Technical report*, CRIM.
URL: *http://www.crim.ca/perso/patrick.kenny/FAtheory.pdf*

Kim, N. S. (2002). Feature domain compensation of nonstationary noise for robust speech recognition, *Speech Communication* 37(3–4): 59–73.

Manning, C. D. (1999). *Foundations of Statistical Natural Language Processing*, The MIT Press, Boston. ISBN: 0-26-213360-1.

Martin, A., Doddington, G., Kamm, T., Ordowski, M. & Przybocki, M. (1997). The det curve in assessment of detection task performance, *Eurospeech 1997*, pp. 1–8.

Martin, A. & Przybocki, M. (2000). The nist 1999 speaker recognition evaluation – an overview, *Digital Signal Processing* 10: 1–18.

Nolan, F. (1983). *The phonetic bases of speaker recognition*, Cambridge University Press, New York. ISBN: 0-521-24486-2.

Pelecanos, J. & Sridharan, S. (2001). Feature warping for robust speaker verification, *A Speaker Odyssey - The Speaker Recognition Workshop*, pp. 213–218.

Pollack, I., Pickett, J. M. & Sumby, W. (1954). On the identification of speakers by voice, *Journal of the Acoustical Society of America* 26: 403–406.

Poritz, A. B. (1988). Hidden markov models: a guided tour, *International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1988)*, Vol. 1, pp. 7–13.

R., T., B., S. & L., H. (2000). A model-based transformational approach to robust speaker recognition, *International Conference on Spoken Language Processing*, Vol. 2, pp. 495–498.

Reynolds, D. A. (2003). Channel robust speaker verification via feature mapping, *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, Vol. 2, pp. II–53–6.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted gaussian miscture models, *Digital Signal Processing* 10: 19–41.

Shearme, J. N. & Holmes, J. N. (1959). An experiment concerning the recognition of voices, *Language and Speech* 2: 123–131.

Solomonoff, A., Campbell, W. & Quillen, C. (2004). Channel compensation for svm speaker recognition, *The Speaker and Language Recognition Workshop Odyssey 2004*, Vol. 1, pp. 57–62.

Tosi, O. I. (1979). *Voice Identification: Theory and Legal Applications*, University Park Press, Baltimore. ISBN: 978-0-839-11294-5.

Tveit, A. & Engum, H. (2003). Parallelization of the incremental proximal support vector machine classifier using a heap-based tree topology, *Workshop on Parallel Distributed Computing for Machine Learning*.

USC (2005). Disability census results for 2005, World Wide Web.
URL: *http://www.census.gov*

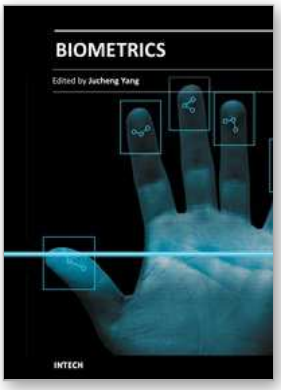
van Vuuren, S. (1996). Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch, *International Conference on Spoken Language Processing (ICSLP)*, pp. 784–787.

Vapnik, V. N. (1979). *Estimation of Dependences Based on Empirical Data*, russian edn, Nauka, Moscow. English Translation: Springer-Verlag, New York, 1982.

Vapnik, V. N. (1998). *Statistical learning theory*, John Wiley, New York. ISBN: 0-471-03003-1.

Viswanathan, M., Beigi, H. S. & Maali, F. (2000). Information access using speech, speaker and face recognition, *IEEE International Conference on Multimedia and Expo (ICME2000)*.

Zhang, J.-P., Li, Z.-W. & Yang, J. (2005). A parallel svm training algorithm on large-scale classification problems, *International Conference on Machine Learning and Cybernetics*, Vol. 3, pp. 1637–1641.




**Biometrics**

Edited by Dr. Jucheng Yang

ISBN 978-953-307-618-8
Hard cover, 266 pages
**Publisher** InTech
**Published online** 20, June, 2011
**Published in print edition** June, 2011

Biometrics uses methods for unique recognition of humans based upon one or more intrinsic physical or behavioral traits. In computer science, particularly, biometrics is used as a form of identity access management and access control. It is also used to identify individuals in groups that are under surveillance. The book consists of 13 chapters, each focusing on a certain aspect of the problem. The book chapters are divided into three sections: physical biometrics, behavioral biometrics and medical biometrics. The key objective of the book is to provide comprehensive reference and text on human authentication and people identity verification from both physiological, behavioural and other points of view. It aims to publish new insights into current innovations in computer systems and technology for biometrics development and its applications. The book was reviewed by the editor Dr. Jucheng Yang, and many of the guest editors, such as Dr. Girija Chetty, Dr. Norman Poh, Dr. Loris Nanni, Dr. Jianjiang Feng, Dr. Dongsun Park, Dr. Sook Yoon and so on, who also made a significant contribution to the book.

**How to reference**

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Homayoon Beigi (2011). Speaker Recognition, Biometrics, Dr. Jucheng Yang (Ed.), ISBN: 978-953-307-618-8, InTech, Available from: http://www.intechopen.com/books/biometrics/speaker-recognition

INTECH
open science | open minds

**InTech Europe**
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

**InTech China**
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821