Chapter 1
Introduction

What I know is not what others knew; what they shall know is not what I know! They wrote what they knew and I write what I know; in hopes that they will write what they shall know!
Homayoon Beigi – July 31, 2009

1.1 Definition and History

Speaker recognition, sometimes referred to as *speaker biometrics*, includes *identification*, *verification* (*authentication*), *classification*, and by extension, *segmentation*, *tracking* and *detection* of speakers. It is a generic term used for any procedure which involves knowledge of the identity of a person based on his/her voice.

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 we steer away from using this term. *Voice recognition* [37, 46, 48], 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, there have been a myriad of different terminologies used to refer to this subject. These include, *voice biometrics* [74], *speech biometrics* [8, 43], *biometric speaker identification* [16, 35], *talker identification* [1, 11], *talker clustering* [25], *voice identification* [70], *voiceprint identification* [36], and so on. With the exception of the term *speech biometrics* which also introduces the addition of a speech knowledge-base to speaker recognition task, the rest do not present any additional information.

H. Beigi, *Fundamentals of Speaker Recognition*,
DOI 10.1007/978-0-387-77592-0_1, © Springer Science+Business Media, LLC 2011

Part of the problem is that there has been no standard reference for the subject. In fact, this is the first text book addressing *automatic speaker recognition*. Of course, there have been other texts on the subject such as Nolan's, *The Phonetic bases of Speaker Recognition* [49] and Tosi's, *Voice Identification: Theory and Legal Applications* [70]. These books are quite valuable, but have had a completely different viewpoint. They have deeply delved into the phonetic and psychological aspects of speaker recognition and have discussed it in its forensic and legal applications in so far as human experts can tell speakers apart. Yet, no complete treatment of the *automatic speaker recognition* class of problems has been produced in textbook form until now. It should be mentioned that although there has been no textbook, the author estimates that there are in excess of 3500 research papers, to date, on the subject. The earliest known papers on speaker recognition were published in the 1950s. [54, 63] In the course of writing this book (about 3 years), more than 2400 publications were reviewed, some in more detail than others.

To avoid any further confusion, the author proposes standard usage of the most popular and concise terms for the subject in addressing this discipline. These terms are *speaker recognition* for the whole class of problems and *speaker identification*, *speaker verification*, *speaker classification*, *speaker segmentation*, *speaker tracking*, and *speaker detection* for the specific branches of the discipline. Of course there are other combinations of speaker recognition ideas with other knowledge sources such as *speaker diarization* and *speech biometrics*.

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 [45, 24] or simply a statistical model with similar output characteristics as the human vocal tract. [8] 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.

In 2006, in the movie, *Mission Impossible III*, Tom Cruise claims the identity of Philip Seymour Hoffman by putting on a mask of his face as it is customary in all *Mission Impossible* programs. However, this time, he forces the person being impersonated to read an excerpt (similar to the enrollment in speaker recognition) and uploads the audio to a remote notebook computer which builds a model of the person's voice. The model parameters are in-turn transmitted to a device on Tom Cruise's neck, located over his trachea. This device adaptively modifies his vocal characteristics to mimic the voice of Mr. Hoffman. In this scenario, the objective is to *spoof* the most familiar speaker recognition engine, namely the human perception. Of course, this idea is not new to the movie industry. In the 1971 James Bond film, "Diamonds are Forever," too, Blofeld who is Sean Connery's nemesis uses a cassette tape which includes the resonance information (formants) for the voice of Mr. Whyte to modify his vocal characteristics to those of the space program admin-

istrator.

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.

1.2 Speaker Recognition Branches

The speaker recognition discipline has many branches which are either directly or indirectly related. In general, it manifests itself in 6 different ways. The author categorizes these branches into two different groups, *Simple* and *Compound*. *Simple* speaker recognition branches are those which are self-contained. On the other hand, *Compound* branches are those which utilize one or more of the simple manifestations possibly with added techniques. The *Simple* branches of speaker recognition are *speaker verification*, *speaker identification*, and *speaker classification*. By the above definition, the *Compound* branches of speaker recognition are *speaker segmentation*, *speaker detection*, and *speaker tracking*. Currently, speaker verification (speaker authentication) is the most popular branch due to its importance in security and access control and the fact that it is an easier problem to handle than the first runner up, *speaker identification*. The reason for the difficulty in handling speaker identification will be made apparent later in Sections 1.2.2 and 17.3.

1.2.1 Speaker Verification (Speaker 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). By non-speech, we are strictly talking about content-based methods; such information may still be delivered using the speech medium, but the speech *carrier signal* is not directly used for identification, in the, so called, non-speech methods. The provided ID is used to retrieve the model for that person from a database. This model is called the *target speaker model*.[1] Then, 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

[1] In some circles this is referred to as the *reference model*, but *target speaker model* is used here.

model is not enough.

There is always a need for contrast when making a comparison. Take the evolution of the monotheistic religions in human history. The first known monotheistic religion, by some accounts, Zoroastrianism which developed in Iran derived its concepts from the older religions of Indo-European. In its initial developments, darkness was attributed to Ahriman and light was attributed to Ahura-Mazda (God). Initially, these two forces were almost equally powerful. Even later, when the role of Ahura-Mazda became much more important, hence the creation of a monotheistic religion, darkness was still deemed necessary to be able to contrast light and goodness. This ideology found itself in following monotheistic religions such as Judaism, Christianity and Islam. The devil was always a philosophical necessity to give followers of these religions an appreciation for the good forces. This stems from the need for contrasting poles in order to assess the quantitative closeness of something to one pole. Namely, the definition of something being good, needs to include how bad it is not.

In analogy, imagine trying to assess the brightness of an object. It would be hard to come up with a measure without having an opposite sense, which would be darkness. We can have a model for brightness and we can compare to it, but we will not be able to make any quantitative judgment of the amount of light without having a model for darkness (or zero light). The same is true for speaker verification (or any other verification system). To be able to get a quantitative assessment of the likeness of the test speaker to the target speaker we would have to know how the test speaker is unlike other speakers. This is partly due to the fuzzy nature of speech. It is impossible for two instances of speech to be identical due to many reasons including the content of speech, the nature of speech (low information content being transmitted by a high capacity signal), and many other reasons.

To properly assess the closeness of the test speaker to a target speaker, there are several approaches. The Two major approaches, in the literature, deal with the said contrast by introducing one or more competing models. The first method uses a *Background Model* or a *Universal Background Model*.[57] This is usually a model based on data from a large population. The idea behind it is that, if the test speaker is closer to the average population than the target speaker, then he/she is most likely not the target speaker.

The second method uses a, so called, *cohort model*.[8] The members of the cohort of the target speaker are speakers who sound similar to the target speaker. The philosophy behind this approach is that if the test speaker happens to be closer to the target speaker compared to the cohort, then most likely the test speaker is the same as the target speaker. In this method, there is no need to involve the rest of the population. The comparison is done between the target speaker and his/her cohort.

As we have seen, the speaker verification process involves a small number of comparisons (generally two); so as the population grows, the amount of computation needed for the recognition stays constant. This is in part responsible for its popularity among vendors – namely, it is an easier problem to solve. Of course, this should not be interpreted as the problem being generally easy. Again, as it was stated, there is a relative degree to everything. It is easier than speaker identification, but it certainly has its own share of problems which make it quite challenging.

1.2.2 Speaker Identification (Closed-Set and Open-Set)

There are two different types of speaker identification, *closed-set* and *open-set*. Closed-set identification is the simpler of the two problems. In closed-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.[2] Note that 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 and where all the speakers in the database are adult males. 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 or diversion to a different profile 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 it is 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 gen-

[2] In practice, usually, the top best matching candidates are returned in a ranked list with corresponding confidence or likelihood scores.

erally increases linearly with the number of speakers enrolled in the database, since, theoretically, the test speaker should be compared against all the speaker models in the database.[3]

1.2.3 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.

As it will be made more clear in section 17.5, classification may use slightly different sets of features from those used in verification and identification. For instance, vowels and fricatives have much more information regarding the gender of the speaker since they carry a lot more information about the fundamental frequency of the vocal tract and its higher harmonics. These harmonic variations stem from the variations in the vocal tract lengths [17] and shapes among adult males and females, and children. For example, the fundamental frequencies of the vocal tracts of males, females, and children lie around 130 Hz, 220 Hz, and 265 Hz respectively.[53] Pitch has therefore been, quite popularly, used to determine the gender of speakers.

To classify people into different age groups, also, specialized features have been studied. Some such features are *jitter and shimmer* which are defined based on pitch variations.[47] Spectral envelopes have also been used for performing such classification.[32] Of course, the classic features used in verification and identification are still used with good results.[52, 65]

Similar to the above examples of gender and age classification, it is more of an art to come up with the proper features when looking for specific features which would be able to classify audio events such as blasts, gun shots, music, screams, whistles, horns, etc. For this reason, there is no cookbook method which can be used to classify such events, giving classification the vagueness of which we spoke at the beginning of this section.

---

[3] In practice, this may be avoided by tolerating some accuracy degradation.[7]

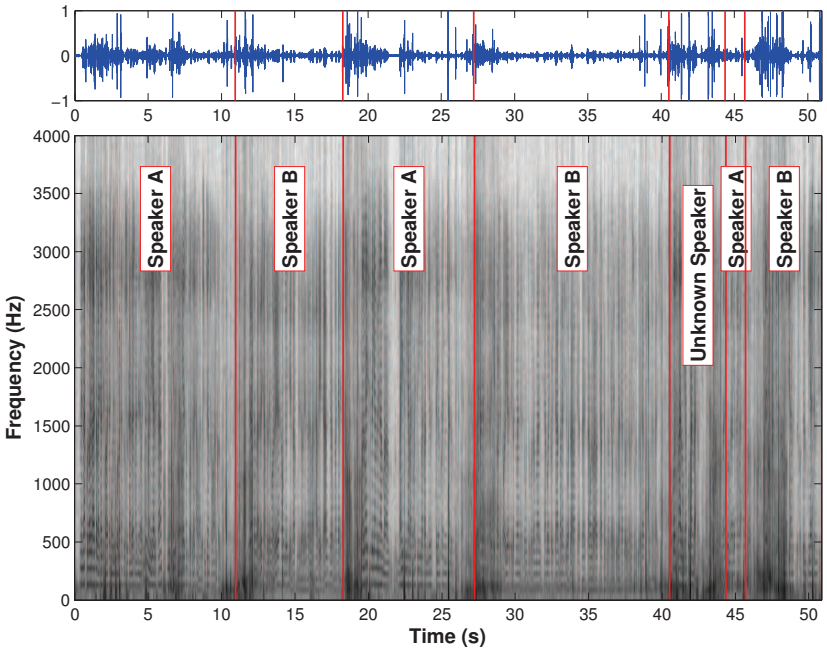

Fig. 1.1: Open-Set Segmentation Results for a Conference Call
Courtesy of Recognition Technologies, Inc.

1.2.4 Speaker Segmentation

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 of such tasks is audio transcription, like the *ARPA HUB4* evaluation task consisting of automatic transcription of radio broadcast news shows from the *Market Place* program.[3, 4]

A typical radio broadcast news contains speech and non-speech signals from a large variety of sources like clean speech, band-limited speech such as telephony sources, music segments, speech over music, speech over ambient noise, speech over speech, etc. The segmentation challenge is to be able to separate the speech produced by different speakers from each other. It is also desirable to separate, music and other non-speech segments.

It is worth noting that most speech recognizers will break down if they are presented with music instead of speech. Therefore, it is important to separate the music from recognizable speech. In addition, one may wish to remove all the music and only store the speech in an archiving scenario to save space.

Another example is the ever-growing tele-conferencing application. An array of conference calling systems have been established which allow telephone conversations among multiple speakers. 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, the audio of that speaker should be segmented and separated from adjoining speakers.

Here, we consider speaker segmentation a type of speaker recognition since the process of segmenting audio is quite similar to other speaker recognition techniques. Normally, statistical models of the local characteristics of two adjoining segments of audio are created. Based on the difference between the underlying model parameters and features that are appropriate for modeling speaker characteristics, an assessment of the similarity of the two segments is made. If the segments are deemed sufficiently dissimilar, a segmentation mark is realized at this point of transition.

Once the basic segmentation points are identified, it is useful to classify the data into segments associated with known speakers. Even if the speaker identities are unknown, by knowing the number of speakers who have participated in the conversation, one may classify the speakers with some common label for identical speakers. As it will be seen in Section 17.4, knowledge of the number of speakers in the conversation is quite helpful and hard to estimate. The underlying difficulty in the estimation of the number of speakers is linked to the "contrast" argument which was made earlier, in Section 1.2.1.

We qualify the operation of tagging the speakers as *speaker classification*, since it performs a classification of the speech associated with an unknown number of unknown speakers – see Sections 1.2.3 and 17.5. If the speaker identities are known, then the sub-problem becomes an identification problem – see Sections 1.2.2 and 17.3. If all the speakers in a conversation have been enrolled in the system, then it may not be necessary to know the total number of speakers in the conversation in contrast with the case where the speakers are unknown and would have to be tagged by the system as speaker A, speaker B, etc.

Note that the segmentation problem has two stages: a basic stage which entails the elementary segmentation of the audio into small pieces with uniform production properties and a more advanced stage which labels these segments and sometimes merges similar adjoining segments. Some references in the literature only consider

the initial stage as segmentation, but since as it shall be seen later, there will be feedback between these two stages, they are usually inseparable. Here, it is preferred to include the whole process under the auspices of one category, called *speaker segmentation*.

1.2.5 Speaker Detection

Speaker detection is the act of detecting one or more specific speaker in a stream of audio. Therefore, the underlying theory encompasses segmentation as well as identification and/or verification of speakers. Enrollment data is usually necessary. The choice of speaker identification, verification, or to use them both is mostly dependent on the problem formulation. For example, if there are many speakers in a stream of audio and it is known that there will always be a speaker from the database speaking at any time, with no possibility of extraneous data such as music, then closed-set identification may be applied to the results of the speaker segmentation to identify the speaker for each segment.

A more complex problem would be one in which there are speakers outside the known set of speakers or there may be music or other types of audio in the stream. In this case, if the list of speakers to detect is not large, then a verification session may be conducted on each segment for every one of the members of the list. On the other hand if the list is large, then an identification may be conducted and the result of the identification may be used as the claimed ID of a subsequent verification. If the identified speaker is verified and if it is a member of the list of speakers to detect, then the result is returned. One can imagine many different possible scenarios in which a combination of speaker identification and verification may be used in conjunction with the results of segmentation.

1.2.6 Speaker Tracking

Speaker Tracking is somewhat similar to speaker detection with the subtle difference that one or more of the speakers are tracked across the stream. In this case, one may envision conditions where no enrollment data is available, but not only is the audio segmented into single speaker segments, but the segments are also tagged with labels signifying the individual speakers in the stream. If enrollment data is available, then the speaker labels for the segments may be adjusted to reflect the true speakers of those segments from the enrollment database. However, in the general sense, it may not be necessary to have specific labels portraying real speakers. In most cases, as long as the different speakers in the stream are identified with general labels such

as alphanumeric tags, then the goal is achieved. The most important application of tracking is the tagging of speakers in a conversation such as a telephone conference call.

1.3 Speaker Recognition Modalities

Theoretically speaker recognition may be implemented using different modalities which are tied to the use of linguistics, context, and other means. However, in the practical sense these modalities are only relevant for speaker verification. In this branch, based on the requirements of the application, different sources of information may be mixed in with the acoustic information present from the vocal tract. Of course, identification and other branches may also be able to use some extra information for improving performance, but the following modalities are most relevant to speaker verification. Although, they are appended with the phrase, "speaker recognition," to show generality.

1.3.1 Text-Dependent Speaker Recognition

In the 1992 film, *Sneakers*, Robert Redford plays the role of an expert who is hired to find vulnerabilities in security systems. When the plot plays out, he tries to access a secure laboratory which is protected with many means including a *text-dependent* speaker verification system. To be able to access the lab, the following was to be spoken: "Hi. My name is Werner Brandes. My voice is my passport. Verify me." He pulls it off by having expected this system and preparing for it.

Since sophisticated digital recording was not common-place at that time, he sends out a woman with the lab owner, Werner Brandes. She is wearing a recording device and tries to ask Mr. Brandes different questions in the course of the evening for which he would have to say the words in the expected prompt in a scattered fashion. The audio tape is then spliced to create the expected prompt. This could be achieved much more easily these days with the existence of small and high quality digital recording devices and digital editing techniques.

This, so called, *liveness challenge*, is the very reason why the *text-dependent* modality is flawed and should not be used in any serious application. So, why do people still work on this type of recognizer? The answer is, because of its relative high accuracy. As we shall see in the next few sections, the liveness issue is still a problem with other speaker recognition modalities, but it is remedied in simple fashion in those modalities. Also, Section 1.3.3 shows that it is possible to view

the *text-prompted* modality as an extension of *text-dependent* recognition. The *text-prompted* modality may be used to remedy the liveness problem of *text-dependent* systems.

The text-dependent modality only applies to the speaker verification branch. Most other branches cannot be used with specific phrases since they happen in a more passive manner where a recognizer listens to an utterance and makes a decision.

1.3.2 Text-Independent Speaker Recognition

Text-independent speaker recognition is the most versatile of the modalities. It is also the only viable modality which may be used in all branches of speaker recognition. There are different degrees of text-independence. Some recognizers are completely text- and language-independent. There are engines which are somewhat language-dependent. Some are completely language-dependent, but text-independent to a degree. Chapter 17 will discuss these different possibilities in more detail.

A purely text-independent and language-independent system only relies on the vocal tract characteristics of the speaker and makes no assumption about the context of the speech. One of the most important problems plaguing text-independent systems is the possibility of a poor coverage of the part of speech. Take an enrollment utterance for example. Under the auspices of a text-independent process, generally, there is no constraint on the enrollment text. Also, a common goal of all recognizers is to try to minimize the length of enrollment and test segments.

As enrollment and test data lengths are reduced, the possibility of a common coverage of the phonetic space is reduced. Therefore, parts of the test utterance may never have been seen at the enrollment time. So, it is plausible that the phones in the enrollment utterance of a non-target speaker and the test utterance of the target speaker have more in common, acoustically, than the enrollment utterance of the target speaker and the test segment for that speaker. This commonality may contribute toward a mis-recognition. To account for this problem, most text-independent speaker recognition engines require longer enrollment and test utterances to be able to achieve acceptable accuracies.

Text-independent speaker recognition also suffers from the liveness assessment problem described earlier (see Section 1.3.1). In this case, one does not even need to have specific words spoken by the individual to be able to spoof the system. Any type of high quality recording of the individual's voice will be enough. Sec-

tions 1.3.3 and 1.3.4 show different solutions to the liveness problem.

1.3.3 Text-Prompted Speaker Recognition

Text-prompted speaker recognition, as it may be apparent from its name, prompts the speaker to say a specific phrase at the time of testing. It was mainly developed to combat spoofing from impostors. If the speaker is not anticipating the text of the prompt, he/she will not be able to prepare for fooling the system. Take the example of section 1.3.1. If the system, being utilized, had used a text-prompted engine, the system would not have been easily fooled.

There are generally two main approaches to the design of a *text-prompted* system. The first method would modify a *text-dependent* system to generate somewhat random phrases for its prompts. These systems will randomly generate a phrase and then build the *text-dependent* language model for that prompt. The response will therefore have to match the vocal characteristics of the target speaker as well as the context of the prompted phrase. This process will be discussed in more detail in Chapter 17.

One of the main advantages of doing *text-dependent* recognition is the sufficience of shorter enrollment texts. However, to be able to perform *text-prompted* recognition through the text-dependent approach, more enrollment data has to be collected to cover the most common phones and phone sequences. Section 17.2.1 sheds more light onto this process.

The second approach would also generate a random prompt, but will use the combination of a *text-independent* speaker recognition engine and a speech recognizer to perform recognition. To recognize an individual, the vocal characteristics of the individual would have to be matched. Also, the recognized text coming from the speech recognizer has to match the expectation of the prompted phrase.

Note that *text-prompted* recognition only makes sense for speaker verification. Most other branches of speaker recognition cannot use specific prompts. This is also the case for *text-dependent* recognition. It is true that one can dream up special cases where other branches of speaker recognition may be used with dictated prompts, but it will not be true in general.

1.3.4 Knowledge-Based Speaker Recognition

A knowledge-based speaker recognition system is usually a combination of a *text-independent* or *text-prompted* speaker recognition system with a speech recognizer and sometimes a natural language understanding engine or more. It is somewhat related to the basic *text-prompted* modality with the difference that there is another abstraction layer in the design. This layer uses knowledge, supplied by the speaker, to test for liveness.

Consider the very familiar security check used by many financial institutions in which they ask for responses to questions that only the caller should know. This has advanced through the years from the old "Mother's Maiden Name," "Last Four Digits of a Social Security Number," and so on to the more versatile systems which require the client to come up with questions, the answers to which are only apparent to him/her. There have also been newer incarnations, especially on Internet-Based systems, where at the time of enrollment, the system asks the user to pick an image. Later, at the verification time, the user is asked to divulge what he/she had chosen at the time of enrollment. Since the choice would have to be limited to a few images, the systems would allow impostors in, with the probability of at best $\frac{1}{Numberofchoices}$, even if the impostor would pick an image randomly.

Therefore, such systems are usually used to strengthen other authentication processes. An example of such *fusion* is to use the image-based authentication in place of the submit button in another type of authentication. However, for personal use, I have found it hard to remember what image I had used at the time of enrollment, especially with systems which are seldom used.

Similar ideas may be used in conjunction with speaker recognition so that the vocal tract characteristics are used in addition to the knowledge from the individual. This is sometimes seen as a *fusion* of two different information sources, but its main objective is to handle the liveness issue described earlier.

In this scenario, the speaker has to provide his/her voice as well as some knowledge-base to the system so that he/she may be challenged with proper questions assessing his/her liveness. This scenario is true for a speaker verification system; however, knowledge-based systems may also be relevant for other branches of speaker recognition, but in a slightly different way. An example is the use of the content of speech as well as the vocal characteristics to come up with a match in a speaker indexing problem. Imagine hours upon hours of audio from a newswire or a similar source. The objective may be to find a specific speaker (speaker detection), but with additional constraints.

For instance, the user may be searching for a certain speaker (speaker detection) when that speaker is talking about a specific topic. This is an example of knowledge-based speaker recognition, where there is no need for the enrollment within the con-

text, since it follows the search topic as recognized by a speech recognition engine, with possible natural language understanding (NLU) capabilities. Also, the speaker voice model may be an excerpt in the same stream, chosen by the user.[72]

1.4 Applications

There is truly no limit to the applications 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. Also, there are some deterrents that feed the skeptics, the most important of which are channel mismatch and quality issues. These topics have been discussed in detail in Sections 22.4 and 22.12 respectively.

Some of the major applications of speaker recognition have been discussed in the following few sections. This list is by no means complete and it is not in any specific order. These examples have been chosen in an effort to try and cover some of the most popular applications. Also, some attention has been paid to covering examples for the different branches of speaker recognition.

1.4.1 Financial Applications

It is hard to lump everything having to do with financial institutions in one category. There are so many different applications that basically span all the branches of speaker recognition. However, here, we will try to cover some of the more popular applications which have direct relevance to the experiences of all of us as users of financial services.

Most of us may have been in the position of contacting a financial institution for questions regarding our accounts. These may be credit-card accounts or simply standard bank accounts. Since financial data is sensitive and should only be accessed by the owners of the accounts, there are usually a number of procedures which are used by financial companies to establish the identity of the individual (on the telephone or in person).

At the present, most institutions provide fully automated account information, accessible through the telephone. They usually require your account number and a pin number to establish your identity. Then full access is granted to the account which could be detrimental if the wrong person gains access. Pin numbers have also

been limited to 4 digits by most financial institutions to be compatible with an international standard. Many of these institutions also disallow the use of 0 or 1 at the beginning and the end of the pin number, considerably reducing the number of permutations. Add to this the fact that most people use easy-to-remember numbers such as birthdays or important dates in their immediate family and you have a recipe for a simple breach of security.

Also, when speaking to a customer support representative, no pin number is necessary. The customer is asked for the account number in addition to some very simple questions, the answers to which are quite easily obtainable. Examples of these questions are "Mother's Maiden Name" which is public knowledge for most people, "Favorite Color" which happens to be blue for over 40% of the population, "Last Four Digits of Social Security Number" which is also something quite accessible to the persistent impostor, etc. Some variations may exist to make these questions somewhat harder to answer. However, to retain the loyalty of their clients, these institutions cannot make the questions too hard to answer.

An important security breach which is hardly considered these days is the possibility of sniffing DTMF sequences by tapping into the telephone line of an individual while pin-based authentications are performed. This is quite simple and does not require much skill. The tapping may be done close to the source (close to the user) or in more serious cases close to the institution performing the authentication. Once the DTMF information is recorded, it may readily be catalogued and used by impostors with dire consequences. Another potential security breach is the readily available personal information being sold by special sites on the Internet for typical $19.95 prices. These are some of the reasons for the enormous number of identity thefts being reported.

Speaker verification is a great match for this type of access problem. It may be used in an automatic fashion. It requires an enrollment which may be performed once a rigorous authentication is conducted by an agent. It can also be used in conjunction or in lieu of existing security protocols. In some cases such as any customer-agent interaction, verification may be done in a passive manner, namely, the verification engine can listen in on the conversation between the agent and the customer. Since the agent is an employee of the institution, his/her voice may be pre-enrolled into the system so that it would be possible to separate the audio of the customer using speaker segmentation followed by verification. In most cases, separate-channel recording will even alleviate the separation problem. Once the customer's audio is isolated, it may be verified as the conversation between the agent and the customer is in progress. The agent may obtain feedback on the possibility of fraud on his/her computer screen which allows for further scrutiny of the caller without the caller even being aware.

Another application is in passive fraud control. Instead of running speaker verification on the valid users, one may hold a list of known fraudsters' voices. In most

cases, a limited number of professional fraudsters call frequently to try and fool the security system. An identification engine can try to alert the security agents of these attacks by listening in on all established communication channels. This can also run in parallel with the verification process described earlier. Imagine a fraudster who has successfully been able to assume a valid client's identity. This fraudster may have also gone through the enrollment process and enrolled himself/herself as the true client. Having this parallel fraud monitor can alert the agents in the institution that there is a likelihood of fraud.

These and other applications could prove to be priceless for financial institutions where fraud could be quite costly. Also, the same argument applies to many other similar institutions with the same security and access requirements, such as health institutions which are required by law to keep the health status and personal information of their customers extremely confidential. There are U.S. government mandates trying to limit the requests by agents for personal information such as social security information, in an effort to reduce the chance of misappropriation of such information.

1.4.2 Forensic and Legal Applications

Speech has a unique standing due to its non-intrusive nature. It may be collected without the speaker's knowledge or may even be processed as a biometric after it has been collected for other purposes. This makes it a prime candidate for forensic and legal applications which deal with passive recognition of the speakers or non-cooperative users. Passive recognition involves tasks in which the application does not generally dictate the flow and type of data being processed. For example, even if speaker verification is used for some specific needs in Forensics, it cannot generally be text-prompted or text-dependent. It will have to be mainly utilized on whatever audio is available and most of the time it is done without the knowledge of the speaker, in which case even the knowledge-based modality does not apply in the strict sense. It is true that information about textual contents of the audio may also be used, in addition to the vocal tract information, but it will be done in the most independent sense.

There are other biometrics such as fingerprint and DNA recognition that also allow some degree of passivity in terms of data collection. That is why they have been successfully used in forensics and legal applications. However, they are not as convenient as speech which may be collected, intercepted, and transmitted much more effectively with the existing infrastructure.

Forensic applications rely on a few different modalities of speaker recognition. Based on the above discussion, speaker identification [10, 19, 21, 27, 51] is the main

modality of interest in these applications. For example, it may be used to identify an individual against a list of suspects. In addition, *speaker segmentation* and *classification* can be quite useful. Segmentation is needed to separate the audio of the target individual in a stream of audio consisting of several sources of speech and other types of audio. Classification can help in categorizing the segmented audio. Also, it may be used to look for anomalies such as abrupt events (gun shots, blasts, screaming, et cetera).

1.4.3 Access Control (Security) Applications

Access control is another place where speech may be utilized as a very effective biometric. Entering secure locations is only a small part of the scope of speaker biometrics. In that domain they compete head-to-head with most other biometrics and possess pros and cons like any other – see Section 1.5. Where speaker recognition truly excels with respect to other biometrics is in remote access control in which the user is not physically at the location where access should take place. Take, for example, the case of accessing sensitive data through a telephony network or a conventional computer network. With any biometric the biometric sensor has to be located where the user is. This, for speech, is a microphone which is readily available in many devices we use. Most other biometric sensors have had no other use in the past and therefore will have to be exclusively utilized for the biometric at hand. In addition, the telephony network is so well distributed that it would be hard to imagine an application that does not have access, at least, to a telephone.

Most access control cases would utilize the speaker verification branch of this biometric. Sections 1.3 and 17.2 describe speaker verification and its modalities in detail.

1.4.4 Audio and Video Indexing (Diarization) Applications

Indexing is a major application of speaker recognition which involves many of its branches, in addition to other technologies such as speech recognition. It has also, quite successfully, been fused with other biometrics such as face recognition.[73] It requires segmentation of the audio stream, detection and/or tracking of speakers, sometimes with the added burden of identifying these speakers. Generally speaking, detection or tracking may have localized scope over a limited set of speakers, such as the distinct speakers which may be identified in a stream of audio. After performing such localized tasks, sometimes a global identification may be required. In some cases, though, the speaker list, used for detection and tracking, is not limited to a

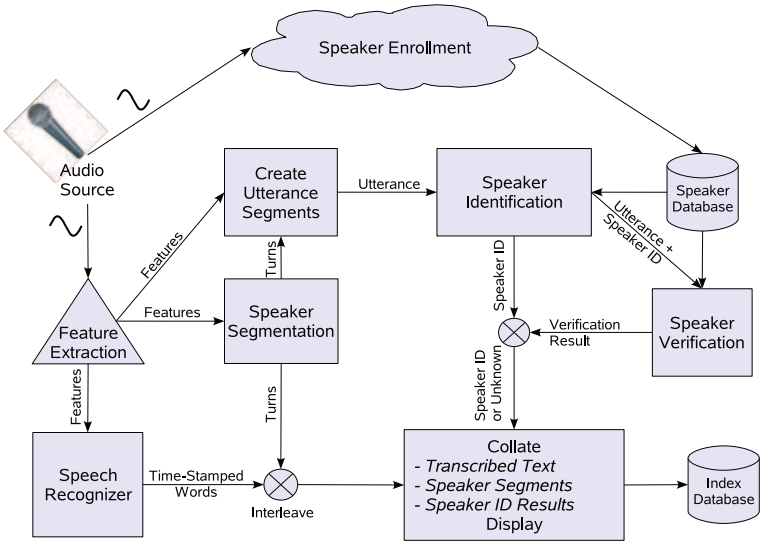

Fig. 1.2: Diagram of a Full Speaker Diarization System including Transcription for the Generation of an Indexing Database to be Used for Text+ID searches

small set. Figure 1.2 shows the diagram for a full *diarization* system, including the transcription of the audio. The results are used to build an indexing database which allows for doing searches based on the text, the speakers or the combination of these two sources.

1.4.5 Surveillance Applications

Surveillance applications (lawful intercept) are really very similar to forensic applications discussed in Section 1.4.2. All surveillance applications, by definition, have to be conducted in a passive manner as discussed in the Forensics section. Unfortunately, they can sometimes be misappropriated by some governments and private organizations due to their relative ease of implementation. There have been many controversies on this type of intercept especially in the last few years. However, if done lawfully, they could be implemented with great efficiency.

An obvious case is one where a system would be searching on telephone networks for certain perpetrators which have been identified by the legal process and need to be found at large. Of course, speaker segmentation would be essential in any such application. Also, identification would have to be used to achieve the final goal. Essentially, the subtle difference between forensic and surveillance applica-

tions is that the former deals with identification while the latter requires the compound branch of speaker recognition, *speaker tracking*.

Another method in which speaker recognition can help in the field of surveillance is at a capacity very similar to that of speaker indexing (Section 1.4.4). Imagine having to transcribe the speech of a target speaker. On the other hand, transcribing the audio of other individuals, not allowed by the legal intercept rules, may not be acceptable. Speaker recognition can concentrate the transcription effort (which is also usually quite costly) to the speech of the target individual.

1.4.6 Teleconferencing Applications

Teleconferencing can also benefit from different branches of speaker recognition. It used to be the case that teleconferencing was limited to large corporations which had the infrastructure to perform such meetings. With the increasing number of free and paid sites, many more teleconferencing sessions are taking place daily. Many of these sites are looking for ways to improve their services in hope of being more competitive in this growing market. Many are looking for speaker diarization capabilities to add as a service. The application discussed in Section 1.4.4 may be used for providing this value-added service. In addition, companies may implement their own diarization systems to make sure that the minutes of the meetings are made available to the individuals on the conference call as well as the companies' archives for future reference.

1.4.7 Proctorless Oral Testing

Figure 1.3 shows a distant learning application of speaker recognition. This is used for performing proctorless oral language proficiency testing. These tests take place on a telephone network. The candidate is usually in a different location from the tester. There is also a set of second tier raters who offer supplementary opinions about the rating of the candidate. In one such application, the candidate is matched by the testing office to a tester for the specific language of interest. Most of the time the tester and the candidate are not even in the same country.

The date of the test is scheduled at the time of matching the tester and the candidate. In addition, the candidate is asked to speak into the Interactive Voice Response (IVR) system which is enabled by speaker recognition technology to be enrolled in the speaker recognition system. The speaker recognition system will then enroll the candidate and save the resulting speaker model for future recognition sessions. Once

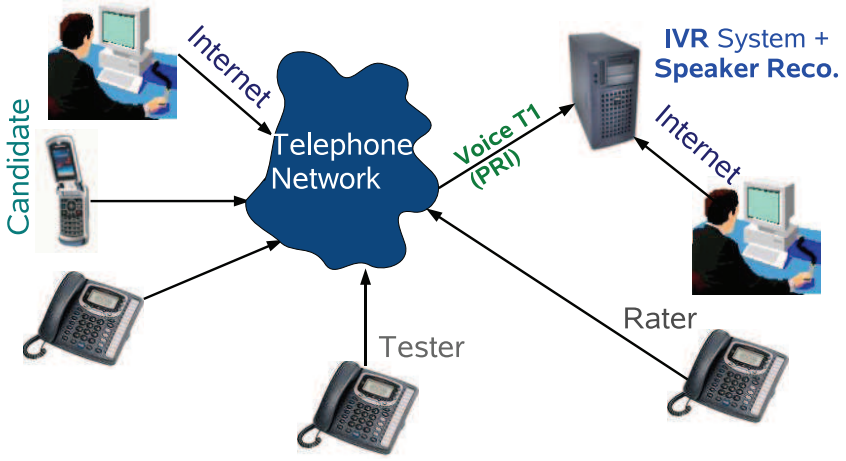

Fig. 1.3: Proctorless Oral Language Proficiency Testing
Courtesy of Recognition Technologies, Inc.

it is time for the candidate to call in for performing the oral exam, he/she calls the IVR system and enters a test code which acts as the key into the database holding the candidate's test details. The candidate is first asked to say something so that a verification process may be conducted on his/her voice. The ID of the candidate is known from the test code entered earlier, so verification may be performed on the audio.

If the candidate is verified, he/she is connected to the tester and the oral examination takes place. In the process of taking the oral examination, the speaker recognition system which is listening in on the conversation between the candidate and the tester keeps doing further verifications. Since the tester is known, the audio of the candidate may be segmented and isolated by the recognition engine from the conversation, to be verified. This eliminates the need for a proctor to be present with the candidate at the time of the examination which reduces the cost of the test. The conversation is also recorded and made available for the second tier rater(s) to be evaluated further.

This application makes use of speaker verification and segmentation. It also does speaker tracking to make sure the candidate has not handed the phone to another person in the middle of the test.

1.4.8 Other Applications

The applications of speaker recognition are not limited to those described here. There are countless other applications which are either known or will be made apparent as more advancement is made in this subject and more people are made aware of its existence. This is one of the many goals of this textbook, namely to promote awareness of the speaker recognition discipline so that it may be applied in new fields.

1.5 Comparison to Other Biometrics

In 1997, Bruce Feirstein touched upon an array of biometrics including *speaker verification* in the screenplay of the James Bond movie *Tomorrow Never Dies*, acted by Pierce Brosnan as James Bond (007). It is certain that speaker recognition is not alone in the biometric arena. Also, it is conceded that there is no single way to solve any problem. In fact there is a place for every kind of biometric to be used and as we will discuss later, it is most of the time beneficial to combine a few techniques to achieve better performance. All of us use a combination of biometric measures in our daily lives to make decisions about the identity of those around us. It is important to be well informed about the strengths and weaknesses of all that is available to us to be able to make a decision about the combination of systems we would need to utilize for any specific application. Although this book concentrates on speaker recognition, in this section, we attempt to review the most popular biometrics and try and compare them with speaker recognition whenever possible.

Many have attempted to categorize the different types of biometrics into two different categories, *Behavioral* and *Physiological*. The problem with this kind of categorization is that just like any other clustering, things are not always so clear cut. Many biometrics have elements from both categories. In fact, the specific treatment of some biometrics may place them in any of the two categories or the combination of the two.

Take, for example, the subject matter of this book. *Speaker recognition* could be construed as a behavioral biometric if the recognition system concentrates on the transitions of the audio and the manner of speaking. In contrast, it may be considered as a physiological biometric based on the characteristics of the vocal tract. In fact, most *text-independent* treatments of the subject view it as a physiological biometric. *Text-dependent* systems, in addition to the physiological information, also use many behavioral tips to make their assessments.

Other, so called, behavioral biometrics also include considerable amounts of physiological information. Behavior is something that may be consciously adapted,

however, many of the characteristics of our voice, signature (handwriting), gait (style of walking), keystrokes, and so on may only be changed in a limited fashion and in accordance with our physiology. In general, there are certain biometrics which are purely physiological. *DNA*, *fingerprint*, *palm*, *iris*, *retina*, *thermogram*, and *vein* are some such biometrics.

Of course, there will always be more inventive biometric techniques some of which will not be so practical, such as the *lip identification* system, as featured by *Eric Horsted*, the writer of "A Taste of Freedom[4]," an episode of the animated television series, *Futurama*. Although it may be possible to do that, having to kiss a glass scanner may not be so sanitary! In some cases, although the sensors for a specific biometric may be harmless, public perception will dictate its success. An example is *retina* imaging which happens to be very accurate and predates *iris* imaging, but since the sensors shine a laser onto the retina, the public has been quite resistant toward accepting this biometric.

With the exception of a few, most biometrics are not usable by every member of the population since they rely on body parts or features which may be lacking, defective or disabled in part of the population. For speaker recognition to be useful, the person has to be able to speak. In most cases, hearing as a feedback measure is also necessary. It is hard to find any separate census information regarding the hearing impaired and mute persons. There have been numbers available since 1850 for the deaf and mute as a whole. Based on the US Census Bureau results, the percentage of the population who was both deaf and mute was 0.04% in 1850 [29], 0.07% in 1880 [29] and 0.4% in 2005 [71]. It is hard to compare these numbers since the number in 2005 has two categories of hearing impaired individuals. The number quoted here is those over the age of 15 who were able to hear any conversation at all. There is another number also reported for those who can hear conversations partially. We are not considering those individuals. However, most of the growth in the percentage may possibly be attributed to more proper census practices. The 2005 number is presumed to be more accurate.

1.5.1 Deoxyribonucleic Acid (DNA)

The idea behind *DNA recognition* is to start with a target sequence of the 4 nucleotides which make up the coding of a *DNA strand*, namely, *A*, *C*, *G*, and *T*. At the time of recognition, one or more samples of a DNA or its fragments are first replicated using *polymerase chain reaction* (*PCR*). Table 1.1 shows the 4 *DNA nucleotides* and the corresponding industrial *triphosphates* used in the replication

[4] Episode 59, fourth season. The series was created by Matt Groening and developed by him and David X. Cohen. this episode originally aired on December 22, 2002. [76]

process.[23] This will replicate the original sample, increasing the number by a few orders of magnitude. Then, an *hybridization process* is used to compare the replicated sequence with the sample sequence.

Abbreviation	Short Name	Chemical Formula	Triphosphate	Triphosphate Formula
A	Adenine	$C_5H_5N_5$	dATP[a]	$C_{10}H_{13}N_5O_{12}P_3Na_3$
C	Cytosine	$C_4H_5N_3O$	dCTP[b]	$C_9H_{13}N_3O_{13}P_3Na_3$
G	Guanine	$C_5H_5N_5O$	dGTP[c]	$C_{10}H_{13}N_5O_{13}P_3Na_3$
T	Thymine	$C_5H_6N_2O_2$	dTTP[d]	$C_{10}H_{14}N_2O_{14}P_3Na_3$

[a] dATP: Deoxyadenosine Triphosphate

[b] dCTP: Deoxycytidine Triphosphate

[c] Deoxyguanosine Triphosphate

[d] Thymidine Triphosphate

Table 1.1: DNA Nucleotides

Currently, the *DNA recognition* procedure is at its infancy. It could potentially be very accurate once the system matures. At the present, chips are being developed to aid in the hybridization process of the recognition. However, with current technology, only *single strands of DNA* (*ssDNA*) of *pathogenic bacteria* can be recognized using electronic technology.[28, 66] Still the *PCR* and *hybridization processes* are done separately and manual intervention is necessary. Work is being done to create a single chip capable of doing the whole recognition process.[66]

Fortunately, *DNA* seems to be one of those biometrics which is available for every human being. It makes *DNA* a powerful biometric, but there are certainly some great disadvantages to this biometric. It will still be a while until human DNA strands can be automatically recognized in a practical time interval. Although some work has been done on different aspects, including the classification problem [33]. Even when automatic DNA recognition matures, there are serious other limitations to DNA recognition. One problem is that people are not comfortable with giving up their DNA. Part of it may be stored while being replicated by the *PCR* process. This could seriously jeopardize a person's security and could possibly be misused if it came into the possession of the wrong people.

1.5.2 Ear

The folds in the *pinna* (the *auricle* of the ear) and the shape of the *ear canal* are different among individuals. These differences are quite pronounced and are easily

realizable by a visual inspection of the outer ear. The Ear has recently been used for establishing the identity of individuals through different approaches. We lump everything related to the ear together, in this section. To date, two separate branches of *ear recognition* have been studied. Aside from the *visual* approach, there is also an *acoustic* method for ear recognition.

The first branch of techniques uses images of the ear for recognition of the individual. This problem has several phases. Generally, a side image of the face is taken and the ear is segmented out. Then, depending on the algorithm, several processes may happen. To achieve invariance, most researchers use some flavor of the *Principal Component Analysis* (*PCA*) method – see Section 12.1.[22, 50] To handle rotational invariance, techniques such as conversion to polar coordinates and the usage of the polar coordinate version of the Fourier Transform to obtain *generic Fourier descriptor* (*GFD*) features have been successful.[22]

Some use a general 2-dimensional image of the ear.[80, 79, 50] Others use the more expensive apparatus of 3-dimensional scanning to obtain more detailed information about the contours.[18, 15] The 3-d scans usually require an additional 2-d reference image for color information. Therefore, the 3-d systems do not seem very practical. Some have used multiple view 2-d images to alleviate the expense and complexity associated with the 3-d systems.[81, 42] These systems do become more complex in the definition of the amount of rotation between different views and the possibility of repeating the same conditions with a practical apparatus.

To improve the accuracy of image-based ear recognition, many have fused this biometric with face recognition results to obtain a *multimodal biometric*.[31, 75, 78] Results of methods with these combinations and the best of breed seem to be in the order of about 2.5% error-rate for identification. The largest population seen in the 43 reviewed references was only in the order of 400 individuals. To date, no large-population study has been seen.

There are some major problems associated with image-based ear recognition approaches. Changes due to *illumination variations* plague this technique in a similar manner as in any other image-based biometric recognition system. These techniques usually have to work hard to attain rotation invariance, with some degree of success. The ear may be covered wholly or partially by hair, especially for individuals with long hair. This will create a gender bias since in general women have longer hair. Finally, there are issues with automatic segmentation of the image to extract the ear from the side-view image of the head. This problem is magnified with *illumination variations*.

The second ear recognition approach uses the acoustic properties of the pinna to establish the identity of an individual. In this approach, a small speaker device and a microphone, both point into the ear canal. The speaker sends out a wave (1.5-kHz

- 22-kHz) into the ear canal at an angle. Once the wave goes through the canal and reflects back from the ear drum and the wall of the canal, the microphone picks up the reflection wave. The wave, manipulated by this reflection, is related to the transfer function which is made up of the transfer functions of the speaker device, the *pinna*, the *ear canal* and the *microphone*. This transfer function is estimated based on the input and the reflected output.[2]

The phase difference between the emitted wave and the received wave is quite sensitive and is present even among within-class samples. To avoid these within-class variations and at the expense of losing of some biometric information, Reference [2] only uses the amplitude of the spectrum of the wave and throws away the phase information.

Unfortunately there are not that many researchers working on this branch of ear recognition and no test on large populations seems to be available for this method. The tests in [2] have only been done on $17-31$ subjects with best results of about $1.5\%-7\%$ equal error rate (EER) – see Section 19.1.1. Therefore, this method does seem to show some promise, but is mostly inconclusive. An interesting point in [2] is that several small earphone and cellular (mobile) phone installations were custom-made for the research. The earphones had much better performance than the cellular phones, as it would be expected, intuitively.

1.5.3 Face

Automatic face recognition has received quite a bit of attention in the last decade mostly due to the availability of the many video cameras in public locations for security purposes. Although, there has been active research in this field for more than 3 decades.[14] There have also been a handful of books written on the subject, in recent years.[41, 82] Face recognition may be achieved in two major forms, cooperative and passive.

The term *cooperative* is used to describe systems which basically work on mugshots of individuals. These are systems installed in airports, as well as systems for cataloging criminals. The airport versions are usually used in conjunction with other biometrics such as fingerprint and iris recognition. Normally, a frontal profile is captured using a digital camera and it is compared against a database of pictures. The lighting conditions are controlled in these systems and mostly take place at the desk of an official with a fixed apparatus. This type of recognition can be most effective, since at the time of capturing the photograph the officer will make sure that the target is not wearing any blocking attire such as glasses or a hat. Systems of this kind use an array of different types of features from geometric features and templates [14] to analogs of force field transforms [31]. Features are normally based on the locations,

shapes and distances of the different components of the face. A popular treatment uses *principal component analysis* (PCA) and goes by the name of *Eigenfaces*. This method uses normalization techniques to transform different frontal snapshots of faces into the same lighting condition as well as size and pixel distribution. Then *principal component analysis* is used to parametrize the faces.[38, 62]

The more challenging form is the passive one. In this method, usually a video camera constantly surveys an area and the individual is not necessarily cooperative in the data acquisition. In this case, the angle of the camera, the attire, lighting conditions, style of walking (the way the face is pointed) and many other variables dictate the quality of the outcome. Of all video-style face recognition applications, analyzing video from security cameras proves most challenging. Most modern cities such as New York and London have thousands of video cameras installed in public areas for surveillance. However, without an automatic face recognition system, the thousands of hours of video which are captured per day are not so useful. Another problem is the high bandwidth required for capturing video compared to audio. A speaker recognition system with segmentation and event classification/detection may be used to greatly reduce the work needed for searching these video streams. This type of speaker recognition may even be applied on-site at the location of the camera to selectively record high resolution video, based on islands of high audio activity. In lower audio activity situations, lower frame rates may be utilized to reduce the amount of storage and processing needed. There are, however, some legal challenges with the public recording of audio. Audio-related regulations are not always treated in the same way as video regulations.[56, 34]

There are simpler cases where the video is obtained in studio quality and under controlled lighting conditions. Broadcast news is one such example.[73] This type of video-face recognition really belongs with the cooperative kind even if the individual in the footage is not aware of face recognition being utilized on the video.

1.5.4 Fingerprint and Palm

Live scanning of fingerprints has been made possible in the last few decades. The minimum resolution required for scanning a finger is about 300 dots-per-inch (dpi), although the minimum required by the United States Federal Bureau of Investigation (FBI) is 500 dpi. There are many kinds of fingerprint scanners in the market, including optical, solid-state, and ultrasound.[44] These scanners started being mass produced years ago and are quite inexpensive. Some solid-state scanners are very small and use a sliding apparatus – the finger is slid on the scanner and the image is reconstructed from the data stream. Many notebook computers have these senors

built-in. This is a testament to the popularity of fingerprint recognition.

The fingerprint pattern is classically considered to include ridges, valleys, singularities (deltas, loops, and cores), and minutiae (local discontinuities and special features). Using these patterns, fingerprints are classified into classes starting from five major classes of patterns which encompass most prints (*left loop*, *right loop*, *whirl*, *arch*, and *tented arch*).[44] Once the main class categories are identified, the minutiae of the test prints are usually matched against those of the target templates on file.

Liveness is one of the major issues with fingerprint recognition. Fingerprints are left behind when an individual touches any hard surface. In the same way that forensic techniques have been collecting finger printers for more than a century, the prints of an individual may easily be lifted and with today's advanced latex molding techniques, a replica of the target person's finger may be created out of latex or similar materials. Imagine a latex fingertip, which may be worn over anyone's finger, that has been molded from the target individual's fingerprint. To be able to access anything that the target individual is allowed, all the impostor has to do is to wear the latex replica and pass the fingerprint recognition test. These latex impressions can be made so thin so that they are not easily visible to the naked eye.

Another problem is that close to 2% of the population do not have usable fingerprints. This is mostly due to damage caused by years of manual labor. Depending on the application, this percentage could be much higher. An example is the use of fingerprints for setting off dynamites for construction purposes. Most construction foremen get to that position after having years of hands-on experience as laborers. The percentage of construction workers that have non-usable prints is quite high. For this population, the use of fingerprint recognition would simply not work. Therefore, the dynamite detonation security systems would have to use other types of biometrics.

The whole palm may also be used in pretty much the same way as a fingerprint is used to identify individuals. The patterns on the palm of the hand possess a uniqueness based on the random generation of the patterns while the hands is formed. Recognition algorithms try to match the pattern of the ridges and shapes of the different zones of the hand to a database. Palm recognition suffers from most of the same problems as stated for fingerprints, including the liveness issue and the fact that palms could be somewhat damaged due to problems such as engagement in manual labor. New fingerprint sensors with light that penetrates the skin have been proposed to alleviate the population with damaged fingerprints. These techniques resemble the techniques of Section 1.5.9 far more than those discussed in this section.

1.5.5 Hand and Finger Geometry

Hand and finger geometry techniques usually use the back of the hand. An example is the technique used by [60] and [39] which is based on a special apparatus with a surface containing pegs designed to keep the fingers and the hand in a specific configuration. Envision a hand placed flat on a surface with the fingers kept wide apart. The amount of distance between the fingers is dictated by the pivot pegs on the specialized apparatus. Then, the shape of the fingers in that standard orientation is photographed and studied to utilize the different sizes of fingers and knuckle locations, and other features for establishing a unique identity. The geometry may of course be obtained by taking a photograph of the top of the hand once it is kept in the constrained position [60, 39] or just to scan the palm for the same information [59]. This is quite limited, due to the specialized apparatus which is certainly not very portable and not designed for everyday remote recognition of individuals the way speaker recognition can be utilized.

There are also techniques that concentrate on the fingers and do simple imaging of the upper surface of the hand to establish the locations of the knuckles. They use the lines in the knuckles just to come up with the locations of the knuckles and then the finger geometry is used to identify the individual.[40] It is unclear how unique these distances are since no large population study seems to be available like it is for some other biometrics.

1.5.6 Iris

After Fingerprint recognition, Iris has received the most attention in the biometrics field. It is partly because of the fact that based on large studies involving millions of different people, the uniqueness of the iris pattern has been established. In 2001, Reference [20] studied 2.3 million pairs of irides and based on the information in the patterns of the examined images, it concluded that the chance of two identical irides extrapolates to 1 in 7-billion. The same paper discusses a smaller study based on comparing the left and right eyes of 324 individuals and shows that the left eye and the right eye have the same *cross entropy* (see Definition 7.17) as the eyes of different people. This suggests that the iris pattern enjoys an epigenetic randomness which makes it an ideal candidate for a biometric measure.

As seen with other biometrics, iris recognition also has its downfall. One of the most serious hurdles is the problem of obtaining usable images from the eye. Illumination changes the shape of the iris as the pupil size changes. Also, non-cooperative subjects can pose problems in obtaining usable images at the proper angle. In addition, for a good image, the eye has to be still and should be close to the camera.

Another issue is that to reduce reflectivity from the cornea, an infrared leaning light should be used. The lighting conditions are paramount in the success of iris recognition. Again, these issues make the simplicity of obtaining audio samples for speaker recognition stand out in competition.

Note that the same issue as mentioned for most other biometrics regarding defective or missing body parts is also true for the iris. There is no known statistic on this issue, but there are certainly a number of people with damaged irides, either congenitally or through an accident or an illness. Another problem is the security of keeping around data which has been obtained directly from body parts. Speech samples do not have direct consequences on the person's identity and by changing the request for the audio different types of samples may be requested. Most iris recognition systems can easily be fooled by using a high quality image of the face of the person, so identity theft could become a real issue. Of course like any other problem, these also have solutions such as trying to purposefully change the pupil size of the individual while obtaining the images to test for liveness. However, it adds to the complexity of the problem and the problems stemming from different shapes of the iris, discussed earlier.

1.5.7 Retina

The idea for retinal scanning for biometric identification was first introduced by a New York Mount Sinai Hospital ophthalmologist in 1935.[68, 64] It was used to catalog criminals. It was shown that the tissue and vein patterns of the retina are almost unique to each individual, even more than the iris pattern, due to the larger surface area.

Realizing retinal patterns, has been achieved using several different signal processing techniques, mainly through the usage of Fourier and Mellin transforms.[13, 67] Most of the time, the color information in the image is discarded and the gray-scale image is used to conduct the pattern analysis.[13] There does not seem to be any real technical challenge with the image capture and the pattern recognition. However, retinal recognition has never been accepted for many other reasons, although it has been around longer than most other biometrics.

The failure of being widely accepted is mostly due to the invasive nature of the imaging as well as the difficulty of obtaining an image. Traditionally, a near infrared light was shone into the subject's eye to obtain the retinal image. This worries many users for the possible risk of damage to their retina. Also, the retina does not remain unchanged through the years. Illnesses such as *diabetes* and *glaucoma* can change the patterns of the retina. Furthermore, there are degenerative diseases such as *retinitis pigmentosa* and other retinal dystrophies which can change this pattern. In

addition, the image may be distorted by advanced *astigmatism* and *cataracts*. All of these problems attribute to its lack of popularity.

The special optical systems used for retinal image acquisition have been quite expensive in the past. To remedy this issue, there are new techniques for using normal cameras at the expense of some accuracy.[13]

1.5.8 Thermography

Thermographic biometrics utilize the distribution of thermal energy on the skin (generally of the face). As seen in Section 1.5.3, face recognition is plagued by the illumination curse. Namely, the within class variability due to illumination discrepancies could easily exceed those across different classes. The main motivation behind using thermographic imaging is to utilize light at a wavelength which is not abundantly available in normal lighting conditions. Therefore, variation in the lighting would not cause as many problems. This is an advantage of thermal imaging, however, like any other biometric, there are disadvantages.

Thermographic imaging inherits all other disadvantages pointed out for face recognition in Section 1.5.3 with the exception of the remedied illumination issue. However, as a trade-off which offsets the illumination solution, the cost of the photography increases substantially. The wavelength of the infrared light used in these operations is in the range of $8\mu m - 12\mu m$.[61] It cannot be detected using normal Charge-Coupled Device (CCD) cameras. For this purpose, costly *Microbolometer* technology has to be used which adds to the cost significantly. Prices of such cameras at the time of writing this book were in excess of $US\$26,000.00$!

1.5.9 Vein

The vein pattern (mostly of the hands) is used to identify individuals. It works by using near-infrared light which penetrates through the skin, but gets absorbed by the iron-rich red blood cells in the veins. The result is a pattern in which the veins look like dark lines and the rest of the light is reflected off the subcutaneous tissues of the hand. The vein pattern seems to be quite unique to the individual and great results are seen in the identification results.

Like any other biometric, vein recognition has many advantages as well as disadvantages. One of the advantages of the vein pattern is that it does not change readily, so it seems to be less susceptible to time lapse effects (see Chapter 20). Also, ev-

eryone has veins in contrast with fingerprints which are not present in about 2% of the population. Of course, we are considering people with limbs in this argument. Missing limbs would affect both fingerprint and vein recognition in the same way. That is more analogous to the mute population for speaker recognition.

One of the disadvantages of this technique is its need for specialized, expensive and typically large camera scanners. This makes it unsuitable for deployment in mobile applications such as in cellular (mobile) phones and notebook computer applications. A second problem which makes it hard for it to compete with speaker recognition is that the technology may seem somewhat invasive to the layperson. Although there is no indication that the light used in the process is harmful, it is not as well understood by the general public as talking.

Another important disadvantage, which may be addressed in time, is the unavailability of large independent studies for this biometric. Most results have been published by vendors and this makes them biased. Also, it seems like there is no standards development for this biometric at the time of writing of this book.

1.5.10 Gait

Some researchers have tried to recognize people based on their style of walking. The length of a person's *stride* and his/her *cadence* are somewhat behavioral, but they also possess some physiological aspects. They are affected by the person's height, weight, and gender among other factors.[9] Cadence is a function of the periodicity of the walk and by knowing the distance the person travels, his/her stride length can be estimated.

The sensor for gait recognition is usually a camera which has to first try to decipher the elements of walking from a video. Of course, the same issues as seen in other image related biometrics (lighting and angle consistencies) come up again. In addition there are specific problems related to video techniques in contrast with still techniques such as speed and direction of movement and their effects on recognition.

There have been very limited studies with this biometric, with small number of trials. Results seem to show some tangible correlation, but are not good enough to be used in the same sense as the more popular biometrics listed here. Reference [9] shows identification rates close to 40% for a handful of individuals in the database. Also, the verification equal error rates (EER) are higher than 10%.

Furthermore, it seems like the behavioral aspect of gait is also influenced by the company which is kept. For example, when a group of people walk together they affect one another in the style of walking for that session. This short-term behavioral

influence reduces the usability of this technique for serious recognition activities.

1.5.11 Handwriting

Handwriting may also be used in a similar fashion as speech. In fact it is quite similar in the sense that it also possesses behavioral (learned style of writing) and physiological (motor control) aspects.[6, 30] Although at a first glance it may seem like handwriting is more behavioral, it is quite apparent that we are constrained by our motor control in how we move a pen – if we are to do it in a natural way.[6, 30] This is very similar to the natural characteristics of our vocal tract versus our ability to somewhat affect the resulting audio which is uttered.

Signature verification greatly resembles speaker verification. Also, in the same way that the speech of a person may be used to assess his/her identity, handwriting may also be used. Of course, it is a bit vague to simply talk about handwriting. The handwriting we normally talk about is the trace of a pen on paper. However, what we are discussing mostly in this section is not limited to this trace. To truly be able to use handwriting for recognition purposes, a digitizer tablet should be used to capture not only the trace, but also the dynamics of the pen on the surface of the tablet. The tip of the pen is digitized at a fixed sampling rate. The resulting points will provide the location and speed of the pen motion. Sometimes a pressure sensitive pen is used to be able to use the pressure differentials as well. An offline assessment of handwriting loses most essential information about the physiological aspects of handwriting, hence losing most of the biometric capacity. This is not as pronounced when the objective is simply to read the content of a handwritten text. That problem is more related to speech recognition and may show relative success without the existence of the online features – this is the optical character recognition (OCR) problem. Although, in general, the online features of handwriting always help in both scenarios.

Therefore, since an online tablet seems to be a necessary device in using handwriting as a biometric, it becomes much harder to implement – if only for the infrastructural problems. Of course, if the will is present, it may be used in a wider sense. Take the ever-growing number of signature capture devices at department stores, your mail courier, etc. Unfortunately, at this point, these devices have been designed to perform a very simple capture with very little performance. Relaxed legal definitions of signature have not pushed these devices to their full potentials. You may have noticed that the signatures captured on most of these devices generate many spurious points and have a great miss rate which result in illegible traces of the signatures. High quality tablets do exist which allow for high quality capture of the online signature. If the topic receives more attention, the better quality tablets may be placed in service. Until that time, it is no competition for speaker recognition and

even then it will only tend to capture a specialized part of the market. Handwriting recognition, due to its independence from speech is a great candidate for fusion and multimodal use with speaker recognition.

1.5.12 Keystroke

Much in the same way that handwriting has its physiological angle, typing is governed by our motor control. The speed of transition between keys is not only behavioral, but is limited by the physiology of our fingers and hand movements. Also, the behavioral part will present a bias toward words we have typed more often in the past. If the text is original (of our own cognition), then another aspect becomes essential and that is the frequency of word usage and their relative placement. In other words, every one of us has his/her own personal built-in language model with word frequencies and personalized *n*-grams [26, 5].

Most software applications do not register the speed of transition between keys, although it is not so hard to accomplish. In a training stage, the statistics of the key transitions and the speed of transition may be learned by an enrollment system which may create a model in the same way as a speaker model is generated for speaker recognition. The model may then be used to recognize the typist at the time of testing.

1.5.13 Multimodal

Section 1.4.4 described an indexing application which can operate on the audio track of a video stream as well as plain audio. Imagine the example of using the audio track from a video clip. For each segment whose boundary is realized based on a speaker change, it is possible to produce a sorted list of speaker labels – sorted by a likelihood score. This list may be combined with other biometrics such as face recognition for a video stream as well as textual information either in the form of results from a natural language understanding system or raw and in the form of a search list. Such combinations have been shown to portray promising results.[73, 12, 77]

Figures 1.4 and 1.5 show the two parallel audio and video indexing processes applied on a video stream. Results are combined and indexed for future searches on the text, speaker by voice and speaker by face. The results of the face and voice may be combined for a more seamless search where the search query will ask for a combination of a text string being spoken by a certain speaker. The speaker tag-

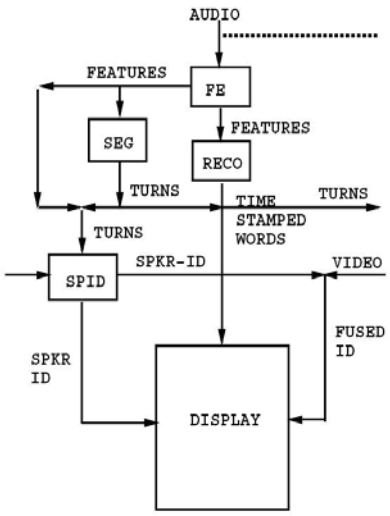

Fig. 1.4: Indexing Based on Audio

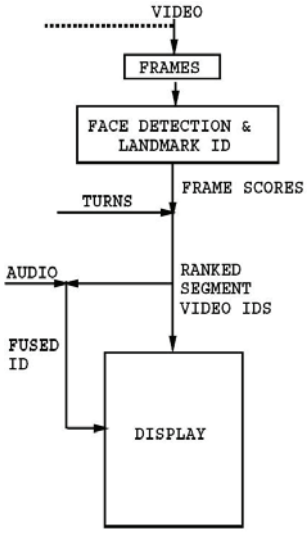

Fig. 1.5: Indexing Based on Video

ging includes results from matching the audio and the face of the target individual. Table 1.2 shows the ranking results based on Audio, Video and Fusion scores for a sample clip. In this case, the ID of the correct individual, OH, is returned by the fused system.[73]

Rank	Audio		Video		Fused	
1	UM	1.000	JW	1.000	OH	0.990
2	OH	0.997	OH	0.988	AK	0.966
3	UF	0.989	AK	0.961	GB	0.941
4	AK	0.988	SF	0.939	JM	0.936
5	JM	0.986	GB	0.932	JW	0.808
6	GB	0.980	JM	0.925	SF	0.759

Table 1.2: Sample Audio/Video and Fusion Results For Multimodal Speaker Recognition [73]

Since the basis for speaker recognition is the vocal tract characteristics of individuals and their uniqueness, it makes sense to combine this source of information with other presumably unrelated sources of biometric information. In fact speaker recognition may be combined with any of the biometrics listed here to improve results. Another example of such a multimodal system is the fusion of *speaker recognition* with *fingerprint recognition* [69].

Also, there have been studies which use microphone arrays in fixed settings such as conference rooms to identify the active speaker by the position of the speaker in the room. It works by comparing the strength of the signal being inputted into the different elements of the array. This information is then combined with more classical speaker segmentation and identification techniques to come up with a better performance.[58]

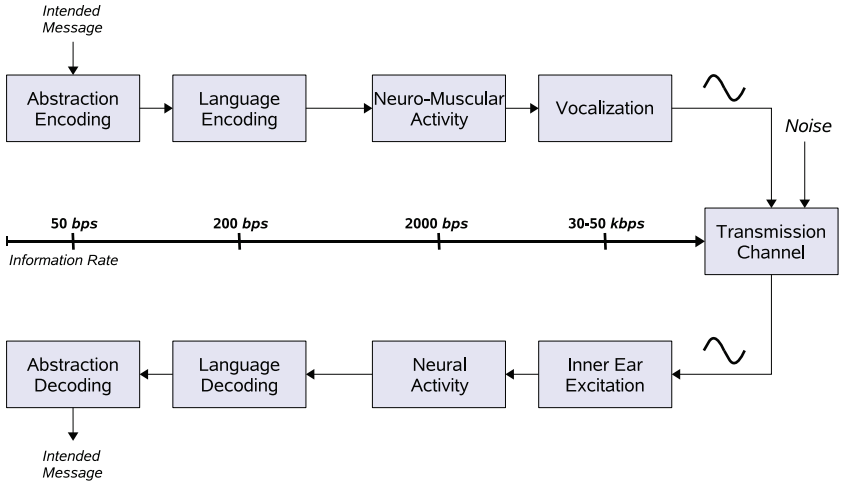

Fig. 1.6: Speech Generation Model after Rabiner [55]

1.5.14 Summary of Speaker Biometric Characteristics

Telephones and microphones have been around for a long time and have prepared the public for their acceptance. Usage of a biometric system based on microphones seems to be much better tolerated than newer systems using unfamiliar means. The negative aspect of this existing infrastructure is the presence of somewhat antiquated systems in use which degrade the quality of speaker recognition systems. Examples are band-limited analog networks which are still in use in many countries. This is true with all established infrastructures, but the positive aspects of the existence of the infrastructure outweigh the difficulties with legacy networks.

Major advantages of using speech as a biometric are its the non-intrusive nature (especially based on the modern culture revolving around the telephone) and the possibility of remote processing, again based on the telephone and the Internet in-

frastructures.

Notable disadvantages on the other hand are its variability, channel effects, and background noise susceptibilities. The variability aspect may be due to illnesses such as nasal congestion, laryngitis, behavioral variations, and variations due to lack of coverage. Let us discuss the lack of coverage in some more detail. In most biometrics, the samples are quite repeatable and one or at most a handful of samples would be enough for acceptable recognition. In general, the speech signal is a high capacity signal conveying a small amount of information – See Figure 1.6. In certain cases, long samples of speech may be obtained with a small coverage of all possible speech sequences. This means that the data seen in training and enrollment stages does not necessarily possess sufficient coverage of the data being seen at the recognition time.

The trouble with channel effects, mentioned above, may be present in different forms such as noise on the channel, channel variability, and compression effects. The noise present on a channel is usually well modulated into the speech signal with almost impossible separation. The characteristics of the noise can manifest themselves by modifying the properties of the speaker characteristics observed in the signal. This is also true about the channel characteristics. Since telephony networks are quite complex with unpredictable paths, these characteristics may change with each instance of communication. Also, due to the small amount of information present in the high-capacity speech signal, most of the time very aggressive compression schemes are utilized, which still allow intelligibility of the content of speech, but may modify the speaker characteristics significantly by eliminating some of the dynamics of the signal.

References

1. Abu-El-Quran, A., Gammal, J., Goubran, R., Chan, A.a.: Talker Identification Using Reverberation Sensing System. In: Sensors, 2007 IEEE, pp. 970–973 (2007)
2. Akkermans, A., Kevenaar, T., Schobben, D.a.: Acoustic ear recognition for person identification. In: Automatic Identification Advanced Technologies, 2005. Fourth IEEE Workshop on, pp. 219–223 (2005)
3. ARPA: Proceedings of the DARPA Speech Recognition Workshop (1996)
4. ARPA: Proceedings of the DARPA Speech Recognition Workshop (1997)
5. Beigi, H.S.: Character Prediction for On-line Handwriting Recognition. In: Proceedings of the Canadian Conference on Electrical and Computer Engineering, vol. II, pp. TM10.3.1–TM10.3.4 (1992)
6. Beigi, H.S.: Pre-Processing the Dynamics of On-Line Handwriting Data, Feature Extraction and Recognition. In: A. Downton, S. Impedovo (eds.) Progress in Handwriting Recognition, pp. 191–198. World Scientific Publishers, New Jersey (1997). ISBN: 981-02-3084-2
7. Beigi, H.S., Maes, S.H., Chaudhari, U.V., Sorensen, J.S.: A Hierarchical Approach to Large-Scale Speaker Recognition. In: EuroSpeech 1999, vol. 5, pp. 2203–2206 (1999)

8. Beigi, H.S.M., Maes, S.H., Chaudhari, U.V., Sorensen, J.S.: IBM Model-Based and Frame-By-Frame Speaker Recognition. In: Speaker Recognition and its Commercial and Forensic Appications (1998)
9. BenAbdelkader, C., Cutler, R., Davis, L.a.: Stride and cadence as a biometric in automatic person identification and verification. In: Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on, pp. 372–377 (2002)
10. Bengherabi, M., Tounsi, B., Bessalah, H., Harizi, F.: Forensic Identification Reporting Using A GMM Based Speaker Recognition System Dedicated to Algerian Arabic Dialect Speakers. In: 3rd International Conference on Information and Communication Technologies: From Theory to Applications (ICTTA 2008), pp. 1–5 (2008)
11. Bennani, Y., Gallinari, P.a.: A modular connectionist architecture for text-independent talker identification. In: Neural Networks, 1991., IJCNN-91-Seattle International Joint Conference on, vol. ii, pp. 857–860 (1991)
12. Besson, P., Popovici, V., Vesin, J.M., Thiran, J.P., Kunt, M.a.: Extraction of Audio Features Specific to Speech Production for Multimodal Speaker Detection. Multimedia, IEEE Transactions on **10**(1), 63–73 (2008)
13. Borgen, H., Bours, P., Wolthusen, S.: Visible-Spectrum Biometric Retina Recognition. In: International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP2008), pp. 1056–1062 (2008)
14. Brunelli, R., Poggio, T.: Face Recognition: Features versus Templates. IEEE Transactions on Pattern Analysis and Machine Intelligence **15**(10), 1042–1052 (1993)
15. Cadavid, S., Abdel-Mottaleb, M.a.: Human Identification based on 3D Ear Models. In: Biometrics: Theory, Applications, and Systems, 2007. BTAS 2007. First IEEE International Conference on, pp. 1–6 (2007)
16. Cetingul, H., Yemez, Y., Erzin, E., Tekalp, A.a.: Discriminative lip-motion features for biometric speaker identification. In: Image Processing, 2004. ICIP '04. 2004 International Conference on, vol. 3, pp. 2023–2026 (2004)
17. Chau, C.K., Lai, C.S., Shi, B.E.: Feature vs. Model Based Vocal Tract Length Normalization for a Speech Recognition-Based Interactive Toy. In: Active Media Technology, Lecture Notes in Computer Science, pp. 134–143. Springer, Berlin/Heidelberg (2001). ISBN: 978-3-540-43035-3
18. Chen, H., Bhanu, B.a.: Human Ear Recognition in 3D. Pattern Analysis and Machine Intelligence, IEEE Transactions on **29**(4), 718–737 (2007)
19. Chunrong, X., Jianhuan, Z., and, L.F.: A Dynamic Feature Extraction Based on Wavelet Transforms for Speaker Recognition. In: Electronic Measurement and Instruments, 2007. ICEMI '07. 8th International Conference on, pp. 1–595–1–598 (2007)
20. Daugman, J., Downing, C.: Epigenetic Randomness, Complexity and Singularity of Human Iris Patterns. Biological Sciences **268**(1477), 1737–1740 (2001)
21. Drygajlo, A.a.: Forensic Automatic Speaker Recognition [Exploratory DSP]. Signal Processing Magazine, IEEE **24**(2), 132–135 (2007)
22. Fabate, A., Nappi, M., Riccio, D., Ricciardi, S.a.: Ear Recognition by means of a Rotation Invariant Descriptor. In: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, vol. 4, pp. 437–440 (2006)
23. Fermentas Nucleotides Catalog. Website (2009). URL http://www.fermentas.com/catalog/nucleotides
24. Flanagan, J.L.: Speech Analysis, Synthesis and Perception, 2nd edn. Springer-Verlag, New York (1972). ISBN: 0-387-05561-4
25. Foote, J., Silverman, H.a.: A model distance measure for talker clustering and identification. In: Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on, vol. i, pp. I/317–I/320 (1994)
26. Fujisaki, T., Beigi, H., Tappert, C., Ukelson, M., Wolf, C.: Online Recognition of Unconstrained Handprinting: A Stroke-based System and Its Evaluation. In: S. Impedovo, J. Simon (eds.) From Pixels to Features III: Frontiers in Handwriting, pp. 297–312. North Holland, Amsterdam (1992). ISBN: 0-444-89665-1

27. Gonzalez-Rodriguez, J., Fierrez-Aguilar, J., Ortega-Garcia, J.a.: Forensic identification reporting using automatic speaker recognition systems. In: Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on, vol. 2, pp. II–93–6 (2003)
28. Hairer, G., Vellekoop, M., Mansfeld, M., Nohammer, C.: Biochip for DNA Amplification and Label-free DNA Detection. In: IEEE Conference on Sensors, pp. 724–727 (2007)
29. Hodges, N.D.C.: Census of the Defective Classes. Science **VIII** (1889). URL http://www.census.gov
30. Hollerbach, J.M.: An Oscillation Theory of Handwriting. MIT Press (1980). PhD Thesis
31. Hurley, D., Nixon, M., Carter, J.a.: A new force field transform for ear and face recognition. In: Image Processing, 2000. Proceedings. 2000 International Conference on, vol. 1, pp. 25–28 (2000)
32. J., A.: Effect of age and gender on LP smoothed spectral envelope. In: The IEEE Odyssey Speaker and Language Recognition Workshop, pp. 1–4 (2006)
33. Jaakkola, T., Haussler, D.: Exploiting Generative Models in Discriminative Classifiers. In: Advances in Neural Information Processing Systems, vol. 11, pp. 487–493. MIT Press (1998)
34. Kablenet: No Snooping on the Public. World Wide Web (2007). URL http://www.theregister.co.uk/2007/08/03/cctv_audio_recording_consultation
35. Kanak, A., Erzin, E., Yemez, Y., Tekalp, A.a.: Joint audio-video processing for biometric speaker identification. In: Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on, vol. 2, pp. II–377–80 (2003)
36. Kersta, L.G.: Voiceprint Identification. Nature **196**, 1253–1257 (1962)
37. Kijima, Y., Nara, Y., Kobayashi, A., Kimura, S.a.: Speaker adaptation in large-vocabulary voice recognition. In: Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84., vol. 9, pp. 405–408 (1984)
38. Kirby, M., Sirovich, L.: Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces. IEEE Transactions on Pattern Analysis and Machine Intelligence **12**(1), 103–108 (1990)
39. Kukula, E., Elliott, S.: Implementation of hand geometry: an analysis of user perspectives and system performance. IEEE Aerospace and Electronic Systems Magazine **21**(3), 3–9 (2006)
40. Kumar, A., Ravikanth, C.: Personal Authentication Using Finger Knuckle Surface. IEEE Transactions on Information Forensics and Security **4**(1), 1–13 (2009)
41. Li, S.Z., Jain, A.K. (eds.): Handbook of Face Recognition. Springer, New York (2005). ISBN: 978-0-387-40595-7
42. Liu, H., and, J.Y.: Multi-view Ear Shape Feature Extraction and Reconstruction. In: Signal-Image Technologies and Internet-Based System, 2007. SITIS '07. Third International IEEE Conference on, pp. 652–658 (2007)
43. Maes, S.H., Beigi, H.S.: Open SESAME! Speech, Password or Key to Secure Your Door? In: Asian Conference on Computer Vision (1998)
44. Maltoni, D.: A Tutorial on Fingerprint Recognition. In: Advanced Studies in Biometrics, *Lecture Notes in Computer Science*, vol. 3161, pp. 43–68. Springer, Berlin/Heidelberg (2005). ISBN: 978-3-540-26204-6
45. Miller, R.L.: Nature of the Vocal Cord Wave. Journal of the Acoustical Society of America **31**, 667–677 (1959)
46. Morito, M., Yamada, K., Fujisawa, A., Takeuchi, M.a.: A single-chip speaker independent voice recognition system. In: Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86., vol. 11, pp. 377–380 (1986)
47. Naini, A.S., Homayounpour, M.M.: Speaker age interval and sex identification based on Jitters, Shimmers and Mean MFCC using supervised and unsupervised discriminative classification methods. In: The 8th International Conference on Signal Processing, vol. 1 (2006)
48. Nava, P., Taylor, J.a.: Speaker independent voice recognition with a fuzzy neural network. In: Fuzzy Systems, 1996., Proceedings of the Fifth IEEE International Conference on, vol. 3, pp. 2049–2052 (1996)
49. Nolan, F.: The Phonetic Bases of Speaker Recognition. Cambridge University Press, New York (1983). ISBN: 0-521-24486-2

50. Nosrati, M.S., Faez, K., Faradji, F.a.: Using 2D wavelet and principal component analysis for personal identification based On 2D ear structure. In: Intelligent and Advanced Systems, 2007. ICIAS 2007. International Conference on, pp. 616–620 (2007)
51. Ortega-Garcia, J., Cruz-Llanas, S., Gonzalez-Rodriguez, J.a.: Speech variability in automatic speaker recognition systems for forensic purposes. In: Security Technology, 1999. Proceedings. IEEE 33rd Annual 1999 International Carnahan Conference on, pp. 327–331 (1999)
52. Parris, E., Carey, M.a.: Language independent gender identification. In: Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on, vol. 2, pp. 685–688 (1996)
53. Peterson, G., Barney, H.L.: Control Methods Used in a Study of the Vowels. The Journal of the Acoustical Society of America (JASA) **24**(2), 175–185 (1952)
54. Pollack, I., Pickett, J.M., Sumby, W.: On the Identification of Speakers by Voice. Journal of the Acoustical Society of America **26**, 403–406 (1954)
55. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Prentice Hall Signal Processing Series. PTR Prentice Hall, New Jersey (1990). ISBN: 0-130-15157-2
56. RCFP: Can We Tape? World Wide Web (2008). URL http://www.rcfp.org/taping
57. Reynolds, D.A., Quatieri, T.F., , Dunn, R.B.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing **10**, 19–41 (2000)
58. Rozgic, V., Busso, C., Georgiou, P., Narayanan, S.a.: Multimodal Meeting Monitoring: Improvements on Speaker Tracking and Segmentation through a Modified Mixture Particle Filter. In: Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on, pp. 60–65 (2007)
59. Saeed, K., Werdoni, M.: A New Approach for hand-palm recognition. In: Enhanced Methods in Computer Security, Biometric and Artificial Interlligence Systems, Lecture Notes in Computer Science, pp. 185–194. Springer, London (2005). ISBN: 1-4020-7776-9
60. Sanchez-Reillo, R., Sanchez-Avila, C., Gonzalez-Marcos, A.: Biometric identification through hand geometry measurements. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(10), 1168–1171 (2000)
61. Selinger, A., Socolinsky, D.A.: Appearance-Based Facial Recognition Using Visible and Thermal Imagery: A Comparative Study. Computer Vision and Image Understanding **91**(1–2), 72–114 (2003)
62. Sharkas, M., Elenien, M.A.: Eigenfaces vs. fisherfaces vs. ICA for face recognition; a comparative study. In: IEEE 9th International Conference on Signal Procesing (ICSP2008), pp. 914–919 (2008)
63. Shearme, J.N., Holmes, J.N.: An Experiment Concerning the Recognition of Voices. Language and Speech **2**, 123–131 (1959)
64. Simon, C., Goldstein, I.: Retinal Method of Identification. New York State Journal of Medicine **15** (1936)
65. Slomka, S., Sridharan, S.a.: Automatic gender identification optimised for language independence. In: TENCON '97. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications'., Proceedings of IEEE, vol. 1, pp. 145–148 (1997)
66. Stagni, C., Guiducci, C., Benini, L., Ricco, B., Carrara, S., Paulus, C., Schienle, M., Thewes, R.: A Fully Electronic Label-Free DNA Sensor Chip. IEEE Sensors Journal **7**(4), 577–585 (2007)
67. Tabatabaee, H., Fard, A., Jafariani, H.: A Novel Human Identifier System using Retinal Image and Fuzzy Clustering. In: International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP2008), vol. 1, pp. 1031–1036 (2006)
68. Eye Prints. Time Magazine (1935)
69. Toh, K.A., and, W.Y.Y.: Fingerprint and speaker verification decisions fusion using a functional link network. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on **35**(3), 357–370 (2005)
70. Tosi, O.I.: Voice Identification: Theory and Legal Applications. University Park Press, Baltimore (1979). ISBN: 978-0-839-11294-5

71. Disability Census Results for 2005. World Wide Web (2005). URL http://www.census.gov
72. Viswanathan, M., Beigi, H., Tritschler, A., Maali, F.a.: Information access using speech, speaker and face recognition. In: Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on, vol. 1, pp. 493–496 (2000)
73. Viswanathan, M., Beigi, H.S., Maali, F.: Information Access Using Speech, Speaker and Face Recognition. In: IEEE International Conference on Multimedia and Expo (ICME2000) (2000)
74. Voice Biometrics. Meeting (2008). URL http://www.voicebiocon.com
75. Wang, Y., Mu, Z.C., Liu, K., and, J.F.: Multimodal recognition based on pose transformation of ear and face images. In: Wavelet Analysis and Pattern Recognition, 2007. ICWAPR '07. International Conference on, vol. 3, pp. 1350–1355 (2007)
76. Wikipedia: A Taste of Freedom. Website. URL http://en.wikipedia.org/wiki/A_Taste_Of_Freedom
77. Xiong, Z., Chen, Y., Wang, R., Huang, T.a.: A real time automatic access control system based on face and eye corners detection, face recognition and speaker identification. In: Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on, vol. 3, pp. III–233–6 (2003)
78. Xu, X., and, Z.M.: Feature Fusion Method Based on KCCA for Ear and Profile Face Based Multimodal Recognition. In: Automation and Logistics, 2007 IEEE International Conference on, pp. 620–623 (2007)
79. Yuan, L., Mu, Z.C., and, X.N.X.: Multimodal recognition based on face and ear. In: Wavelet Analysis and Pattern Recognition, 2007. ICWAPR '07. International Conference on, vol. 3, pp. 1203–1207 (2007)
80. Zhang, H.J., Mu, Z.C., Qu, W., Liu, L.M., and, C.Y.Z.: A novel approach for ear recognition based on ICA and RBF network. In: Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on, vol. 7, pp. 4511–4515 (2005)
81. Zhang, Z., and, H.L.: Multi-view ear recognition based on B-Spline pose manifold construction. In: Intelligent Control and Automation, 2008. WCICA 2008. 7th World Congress on, pp. 2416–2421 (2008)
82. Zhou, S.K., Chellappa, R., Zhao, W.: Unconstrained Face Recognition, *International Series on Biometrics*, vol. 5. Springer, New York (2008). ISBN: 978-0-387-26407-3

http://www.springer.com/978-0-387-77591-3

Fundamentals of Speaker Recognition

Beigi, H.

2011, LXI, 942 p., Hardcover
ISBN: 978-0-387-77591-3