Application Areas

2

2.1 Speech

The interaction by spoken language is the dominant modality of communication between humans. By means of speech emotions can be conveyed, irony can be expressed, simply "small talk" can be made, or information can be transmitted. The last of these aspects is by far the most important for the automatic processing of speech even though approaches for recognizing emotions in spoken utterances are pursued, too. Spoken language makes it possible to transmit information without hardly any effort—at least for healthy humans—and with a rather high "data rate" of up to 250 words per minute. Thus with respect to ease of use and efficiency this modality principally outperforms all other means of communication used by humans as, for example, gesture, handwriting, or typing on a keyboard. In the literature it is, therefore, often concluded that speech would also be the best solution for the communication with technical systems. This may well be doubted, however, as an open-plan office where all employees talk to their computers or a coffee machine that can be controlled by spoken language only and not by simply pushing a button might not seem to be the best ideas.

There are, however, a number of scenarios in which man–machine communication by spoken language makes sense—if necessary including additional modalities—and can be applied successfully. In such scenarios the goal is either to control a certain device or to acquire information from an automatic system. Examples for the latter are information systems from which time-table or event information can be queried over the telephone and also the respective train, cinema, or theater tickets can be ordered, if necessary. Among the control applications are the operation of mobile phones, which make the respective connection when the appropriate name or phone number is called, the operation of machinery in an industrial context, where the use of other modalities besides speech is not possible, and also the control of so-called non-safety relevant functions in vehicles as, e.g., the car stereo or the air condition. As a very special case of device control the automatic transcription of texts by a dictation system can be viewed. Though automatic

G.A. Fink, *Markov Models for Pattern Recognition*,
Advances in Computer Vision and Pattern Recognition,
DOI 10.1007/978-1-4471-6308-4_2, © Springer-Verlag London 2014

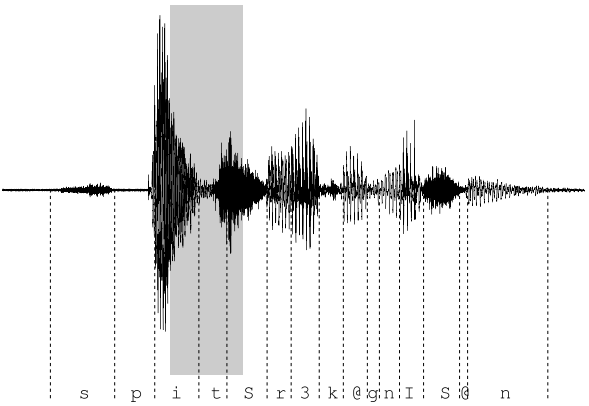


Fig. 2.1 Example of a digitized speech signal of the phrase "`speech recognition`" with manually marked phone segments

dictation did not make it to become the "killer application" of speech technology, it has had a crucial influence on the developments in the field.

In order to make spoken language man–machine communication possible, spoken utterances need to be mapped onto a suitable computer-internal symbolic representation. This internal representation later serves as the basis for determining the actions of the system. For this purpose first the physical correlate of speech—i.e. the minute changes in air pressure caused by the radiation of sound—needs to be represented digitally. Microphones convert the sound pressure level into a measurable electrical quantity. The temporal progression of these measurements corresponds to the acoustic signal. In order to represent this signal with sufficient accuracy in digital form, it is sampled, i.e., the analog values are measured at certain regular time intervals, and subsequently quantized, i.e., the analog quantities are mapped onto a finite discrete domain of values. The combination of sampling and quantization is referred to as digitization. For speech one usually works with sampling rates of 11 to 16 kHz and stores the quantized measurements with a precision of 8 to 16 bits.[1]

Figure 2.1 exemplarily shows a digitized speech signal of the utterance "`speech recognition`". For extremely simple applications of speech processing this information sometimes is already sufficient. Thus voice dialing in a mobile phone can be achieved by the direct comparison of the current speech signal with a small set of stored reference signals. In complex applications of spoken language processing, however, it is indispensable to first create a suitable intermediate symbolic representation before an interpretation of the data in the context of the application is attempted.

Besides the realization of language as an acoustic signal, there also exists the dual representation in written form. Though a number of characteristics of speech as, for example, loudness, speed, or timbre, cannot be represented in writing, still the central information content can be specified orthographically. This "encoding"

[1] Representing speech sampled at 16 kHz with 16 bits per sample surely is not the best possible digital representation of speech signals. However, by and large it is sufficiently accurate for the automatic processing methods applied and has become standard in the field.

of the acoustic signal can also easily be represented in and manipulated by digital computers. Therefore, it is standard to first map speech signals onto a textual representation in more complex systems for spoken language processing. This processing step is referred to as *speech recognition*. The process of *speech understanding* starts from the results of speech recognition which usually consist of a sequence of word hypotheses. On this basis methods for speech understanding try to derive a representation of the meaning for the utterance considered. In information systems the intention of the user is determined in this step and the relevant parameters of his query are extracted. An automatic dialog system of an airline, for example, needs to distinguish between a request for flight schedules and the actual order of a ticket. In both cases the airport of departure, the destination, and the desired travel time must be determined. For the syntactic-semantic analysis of so-called *natural language*, i.e., language input encoded in textual form, a multitude of different approaches were proposed in the literature (cf. e.g. [318]). The interpretation of utterances is almost exclusively achieved by applying rule-based methods which either build directly on linguistic theories or are motivated by these.

However, the mapping of a speech signal onto its textual representation, as it is the goal of automatic speech recognition, cannot be achieved with purely symbolic methods. The main reason for this is the large variability in the realization of principally identical spoken utterances by different speakers or in different environments. Furthermore, boundaries between acoustic units are generally *not* marked at all within the speech signal which makes the segmentation of spoken language extremely problematic.

The elementary unit for describing speech events is the so-called *phone* which denotes a single speech sound. In contrast to a *phoneme*, i.e., the smallest unit of speech used to distinguish meaning, phones define "building blocks" of spoken utterances that can be discriminated perceptually by listeners. The categories used for describing phones were developed on the basis of the articulation of the respective speech units (cf. e.g. [9, 47]). The articulation of speech events can be explained by a model of the speech production process, the principles of which will be described briefly in the following.

First a stream of air from the lungs is created—usually by exhalation—which passes the phonation mechanism in the larynx formed by the vocal folds. If this so-called *glottis* is closed, a series of periodic impulses of air pressure is generated. In contrast, with the glottis opened the air passing by only produces something similar to white noise. This voiced or un-voiced *excitation signal* is then modified in its spectral content in the so-called *vocal tract*, and a certain speech sound is formed. The vocal tract consists of the oral and nasal cavities and the pharynx. It can be modified in shape depending on the opening of the jaw and the position of the tongue and the soft palate. When putting it in simple terms, the two coarse classes of speech sounds *vowels* and *consonants* can be distinguished by the gross type of sound modification effected by the vocal tract. For vowels the excitation is always voiced and the vocal tract merely forms a resonant space. For example, with the largest possible opening of the vocal tract one obtains a vowel as in the word "`start`" (in phonetic

transcription[2] `[stArt]`). In contrast, consonants result from a sort of constriction formed in the vocal tract being combined with either a voiced or un-voiced excitation signal. For example, if the tip of the tongue touches the back of the lower teeth, either a voiced or un-voiced S sound as in "*raise*" or "*race*" is generated, respectively (`[reIz]` vs. `[reIs]`).

Spoken language utterances always develop as a sequence of such elementary sounds. However, the units are not represented in isolation within these and, therefore, are by no means easy to segment. As the articulatory organs cannot change their positions instantaneously from one sound to the next, this is achieved by continuous movements. Therefore, speech signals reflect the smooth transition between the characteristic features of subsequent sounds. In strong idealization of the real situation one may assume that in deliberately articulated, slow speech the typical properties of a speech sound are expressed at the center of its total duration. The border regions of the respective signal segment, however, are influenced by the neighboring sounds. This mutual influencing among sounds in the speech current is referred to as coarticulation. In reality its effects can also extend across multiple neighboring sounds. In Fig. 2.1 the segmentation of an example signal is shown. However, the discrimination between the individual phones is not uniquely defined, in general, as even by experts the segment boundaries cannot be specified beyond any doubt.

The inherent continuity of speech makes a purely data-driven segmentation without any model knowledge virtually impossible. Therefore, today exclusively so-called "segmentation-free" methods on the basis of hidden Markov models are applied for the purpose of automatic speech recognition. Though the digitized speech signal itself is already a linear sequence of samples, the statistical models of speech always start from a suitable feature representation. This aims at numerically describing the characteristic properties of speech units which are mainly defined by the local spectral composition of the signal. This feature extraction needs to be performed on sections of speech where the properties in question vary as little as possible over time as no segmentation information is available in this early stage of processing. Therefore, on the one hand the respective sections should be quite short. On the other hand, they also need to be sufficiently long in order to make the computation of useful spectral characteristics possible. Therefore, the speech signal is subdivided into sections of constant length of approximately 16 to 25 ms, which are called *frames*. In order to avoid losing important information at the boundaries created by this elementary segmentation of the signal, the frames usually overlap. A *frame rate* of 10 ms has virtually become a standard in the field. With a frame length of 20 ms the signal sections would overlap by 50 percent. Figure 2.2 shows the subdivision of a speech signal into frames of 16 ms length for a part of the example signal known from Fig. 2.1.

[2]The phonetic transcriptions of spoken utterances given in this book use the symbol inventory defined by SAMPA which was developed as a machine-readable version of the International Phonetic Alphabet (IPA) especially for the automated processing in digital computers [303].

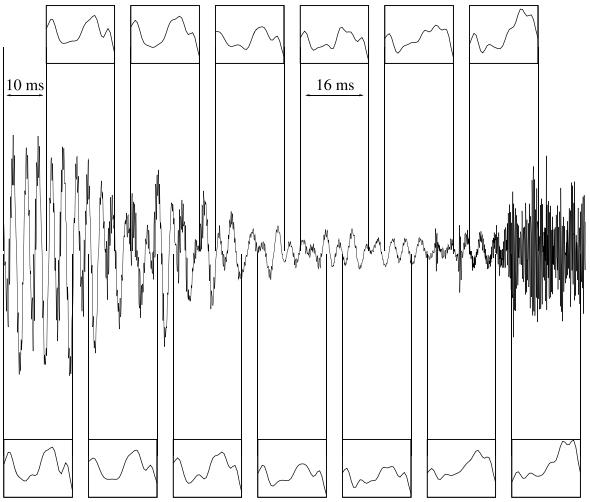


Fig. 2.2 Frame segmentation for the short example section marked in the speech signal known from Fig. 2.1, which represents the transition from the vowel [i] through the plosive [t] to the fricative [S]. The model spectrum which was created by cepstral smoothing is shown as the hypothetical feature representation

For every frame features are calculated. Thus one obtains a sequence of high-dimensional, continuous feature vectors which are identified with the outputs of a hidden Markov model. All feature extraction methods have in common that they use a measure for the signal energy and generate an abstract representation of the spectral composition of the respective section of the signal. Originally developed for the analysis of seismic data, the so-called *cepstral*[3] analysis has become the standard feature extraction method in the field of automatic speech recognition ([26], cf. also [123, pp. 306–318]). The so-called model spectrum implicitly characterizes the shape of the vocal tract during the formation of speech and thus allows to draw inferences on the speech sound articulated.[4] The combination of subdividing a speech signal into frames and carrying out a local feature extraction is referred to as *short-time analysis*. In Fig. 2.2 results of such a procedure are exemplarily shown. There the computation of a model spectrum created by cepstral smoothing was used as the hypothetical feature extraction method.

The training of hidden Markov models for acoustic units attempts to reproduce the statistical properties of the feature vector sequences that were generated by the short-term analysis. Usually, for this purpose a modular approach is applied for the description of complex structures of speech. Models for words are constructed by concatenation on the basis of models for elementary units as, e.g., phones. An arbitrary sequence of word models from a given lexicon then defines a model for spoken utterances from a certain application domain. The overall model is again a hidden Markov model. On the basis of a sequence of feature vectors that represents a spoken language utterance, its segmentation into the respective word sequence can be

[3] Terms as *cepstrum*, *saphe*, and also *alanysis* were artificially derived by the authors from their "equivalent" terms in the frequency domain, i.e., *spectrum*, *phase*, and *analysis*.

[4] A detailed explanation of different methods for feature extraction is, e.g., given in [123, Chap. 6 pp. 275–336].

obtained by computing the optimal state sequence through the model. This state sequence passes through certain word models which by construction are part of the utterance model. Therefore, the corresponding optimal textual representation can be derived easily. However, in general this solution will only represent an approximation of what was really said.

In addition to the modeling on the acoustic level, statistical restrictions on a symbolic level can be introduced by means of Markov chain models. This extension of the model helps to avoid the consideration of arbitrary word sequences during the search for the solution, which might be quite implausible in the context of the respective application. Therefore, an additional Markov chain model statistically describes regularities of the language fragment considered and, therefore, is usually referred to as a so-called *language model*. Principally, also purely symbolic methods can be used for this purpose as, e.g., formal grammars. However, the combination of two statistical techniques, in general, leads to more powerful integrated systems. Therefore, the combination of hidden Markov models for the acoustic modeling and Markov chain models for the language modeling has become the standard procedure within the field of automatic speech recognition.

The difficulty of an actual speech recognition task can be estimated from the restrictions which apply for the spoken utterances to be expected. The more constrained, the simpler and the more diverse, the more difficult the necessary modeling will be. The problem of speech recognition is considerably simplified if the actual speech data originates from only a single speaker. This is then said to be a *speaker-dependent* recognition task. In contrast, systems are referred to as *speaker independent* which are approximately capable of processing utterances from a wide range of different persons. The recognition problem can also be simplified by limiting the vocabulary considered. If only a simple set of command words is to be recognized, this can be achieved comparably easily and robustly. The decoding of a large vocabulary of several 10 000 words, however, requires a dramatically increased effort. Therefore, off-the-shelf dictation systems work in speaker-dependent mode in order to achieve an acceptable recognition accuracy even for very large vocabularies. Information systems accessible over the telephone, in contrast, need to be speaker independent and generally use vocabularies which are constrained to the actual task.

But not only the size of the lexicon influences the recognition performance achievable. Usually, in large vocabulary speech recognition systems one uses statistical models for the restriction of the potential or probable word sequences. Depending on how well these constraints can be brought to bear during model decoding, the recognition problem is simplified. This can be achieved especially well for utterances from clearly defined application areas in which possibly even formalized language structures might be used. Therefore, the first dictation systems that were commercially available were aimed at offices of attorneys and medical doctors.

In addition to the size of the space of potential solutions, also the speaking style critically influences the quality of the recognition results. In experiments in the laboratory one often works with speech signals that were read from given text prompts and which, therefore, exhibit considerably less variability in their acoustic realization than can be observed in spontaneous speech. In general, when speaking spon-

taneously, the care taken in the articulation decreases, and coarticulation effects between neighboring speech sounds increase. In certain contexts individual phones or complete phone groups are potentially not realized at all. Furthermore, spontaneous speech effects as, e.g., hesitations or false starts may occur which, of course, need to be taken into account when building the models required.

A further severe difficulty for speech recognition systems are changes in the environmental conditions in which the signal data is captured. On the one hand, these might affect the recording channel itself which is defined by the technical solutions used for signal recording and the acoustic properties of the recording environment. Therefore, in early off-the-shelf dictation systems often the special microphone required was sold together with the software package. Even today dictation systems work in a quiet office environment only and not in a large exhibition hall. In such public spaces two additional effects make the task of automatic speech recognition extremely challenging. First, interfering noises appear which adversely affect the system performance in two respects. On the one hand, they overlay the actual speech signal and thus influence the feature representations extracted. On the other hand, also the speaker himself perceives the interfering noises and, in general, modifies his articulation in consequence. This phenomenon according to its discoverer is referred to as the *Lombard effect* [181]. Therefore, the robust recognition of utterances spoken in uncontrolled acoustic environments as, for example, in a driving vehicle or in crowded public spaces where severe and unpredictable noises occur is a considerable challenge for current speech recognition systems. Secondly, speech signals recorded in public spaces are distorted by a considerable amount of reverberation not present in quiet in-door environments. Even though human listeners show a remarkable capability to understand speech in reverberant environments, automatic techniques are still not able to achieve a comparable performance.

Over decades of speech recognition research, respectable achievements were made in counteracting the manifold difficulties of automatically recognizing speech input by further and further refinements in the statistical methods as well as by special application specific techniques. Nevertheless, even today no system exists that is able to recognize arbitrary utterances of an arbitrary person on an arbitrary subject in an arbitrary environment—but these requirements are not even met by the human example. But also for less ambitious goals as, for example, the building of speaker-dependent dictation systems the actual system performances often fall short of the promises made by the manufacturers [97]. The problem of automatic speech recognition, therefore, is not solved at all, and also in the future considerable research efforts and potentially radically new methods will be required for creating systems that, at least approximately, achieve the capabilities of a human listener.

2.2 Writing

When considering writing, in the Western world first character alphabets come to mind and especially the widely used Roman alphabet, which was also used for printing this book. Alphabetic writing systems, in general, follow the phonographic principle. According to this principle the phonetic structure of the respective language is

represented by the comparably few symbols of the alphabet.[5] Even though a Western reader might consider this procedure to be the only one making sense due to his cultural background still completely different approaches for writing down spoken language exist today.[6] The most prominent example is the Chinese writing system which mostly works according to the logograhic principle. In Chinese script each of the complex symbols represents a certain word of the language. Therefore, at least a few thousand symbols are required in order to write texts in everyday Mandarin, e.g., in newspapers. Such a form of writing might appear overly complex in the eyes of Western people. Nevertheless, even in the computer age such writing systems are not replaced by alphabetic ones. However, the typing of Chinese or Japanese texts on today's computer keyboards is usually facilitated by the use of romanized versions of the respective scripts, i.e., systems for writing Chinese or Japanese that use Roman characters.

Japanese texts are written in a mixture of Chinese symbols (so-called *kanji*) for representing word stems and two syllabic writing systems that were derived from it by simplification and which again follow the phonographic principle. *Hiragana* symbols are used for writing grammatical elements. Names and foreign words are written with the symbols of the *katagana* writing system.

Independently from the actual writing system written texts can either be produced by machine, i.e., printed, or by handwriting on paper by using a pen or brush. In machine print the form of the individual symbols or characters is not constrained in principle. In contrast, in handwriting one aims at bringing the symbols of a certain script to paper as easily and as fluently as possible. Therefore, for most writing systems also a cursive version adapted for handwriting exists in addition to the symbol set used for producing machine-printed texts.

In contrast to the recognition of spoken language, the automatic processing of writing is only partly carried out in the context of man–machine interaction. Rather, the origins of this technology lie to a major extent in industrially relevant applications in the field of automation technology.

This book does not aim at treating the automatic processing of writing thoroughly for all existing writing systems. Therefore, we will limit the considerations to the widely used alphabetic writing systems in the following and there to the Roman alphabet as a typical representative. The interested reader is referred to the respective specialized technical literature for a presentation of techniques, which are used for processing, e.g., Japanese, Chinese, or Arabic texts (cf. e.g. [36, Chap. 10 to 15]). With the exception of highly logographic writing systems, as, e.g., Chinese, the same principal approaches are applied. The main differences result from taking into account the respective special appearance of the image of the writing in the selection of methods and the combination of processing steps.

[5] The relationship between characters of a certain alphabet and the respective pronunciation is more or less obviously preserved in their current writing due to the historical development of languages.

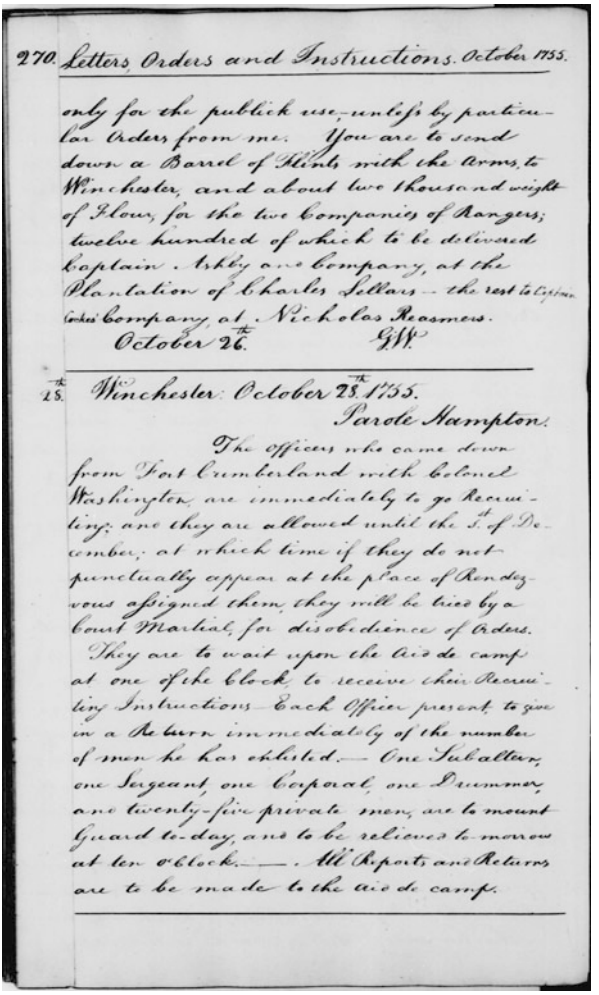
[6] In [51] a thorough overview is given over ancient writing systems and those still in use today together with their historical development and relationships amongst each other. A brief overview of the major scripts with examples of their appearance can be found in [112].

The classical application of automatic processing of writing is the so-called *optical character recognition* (OCR, cf. e.g. [200]). There the goal is to automatically "read" machine-printed texts which were captured optically and digitized afterwards. In other words, one aims at transforming the image of the writing into a computer-internal symbolic representation of the text. Thus the underlying data is images of document pages, as, for example, the one at hand, which are converted into a digital image by means of a scanner with a typical resolution of 300 to 2400 dots per inch. Therefore, methods for automatic image processing are predominant in the field of OCR due to the very nature of the input data itself.

Before the actual text recognition can be started, an analysis of the document layout is necessary in almost any document image analysis task. This layout analysis attempts to identify text areas and other elements of the document structure as, e.g., headlines or graphics, within the available page image. Afterwards, the text areas can be segmented into paragraphs, individual lines, and, in general, also single characters due to the usually high precision in the production of machine-printed texts. As soon as the images of the written symbols are isolated, they can be mapped onto a symbolic representation by arbitrary techniques from the field of pattern classification (cf. e.g. [36, 275]). The results of the classification are generally subject to one or more post-processing steps. These post-processing operations attempt to correct errors on the character level as far as possible by incorporating context restrictions, e.g., in the form of a lexicon (cf. e.g. [55]).

As in the field of automatic speech recognition, the complexity of the processing task is defined by the variability of the input data to be expected. The size of the lexicon used is of subordinate importance as the processing mainly is performed on the character level and the word and document context is only taken into account during post-processing. On the one hand, the variability of the data results from differences in the respective image of the writing as it is created by the printing process. On the other hand, distortions in the optical capturing of the documents may severely affect the appearance of individual characters and text sections. In the printing process itself the type face (e.g. Times, Helvetica or `Courier`), the font family (e.g. regular, **bold**, or *italic*), and the character size (e.g. tiny, normal, large) may vary. The difficulty of the automatic processing is increased considerably if the image of the writing could only be captured with poor quality. This can be due to aging or a comparable degradation of the original document that might be caused, for example, by contamination. A similar effect is caused by the repeated reproduction or transmission of a document, e.g., by fax or by means of copying machines, which severely reduces the quality of the source data for automatic character recognition. In such cases a segmentation on character level is generally no longer possible with sufficient reliability. Therefore, segmentation-free methods on the basis of Markov models offer substantial advantages in the processing of such "degraded" documents as opposed to classical OCR methods (cf. e.g. [75]).

The problem of automatic character recognition also becomes considerably more challenging if the texts considered were not printed by machine but written by hand. Especially the use of cursive writing which, in general, links individual characters with connecting strokes in alphabetic writing systems makes a reliable segmentation


270. Letters, Orders and Instructions. October 1755.

only for the publick use, unless by particular Orders from me. You are to send down a Barrel of Flints with the Arms, to Winchester, and about two thousand weight of Flour, for the two Companies of Rangers; twelve hundred of which to be delivered Captain Ashby's one Company, at the Plantation of Charles Sellars — the rest to Captain Cocks Company, at Nicholas Reasmers.

October 26th GW.

28th Winchester October 28th 1755.

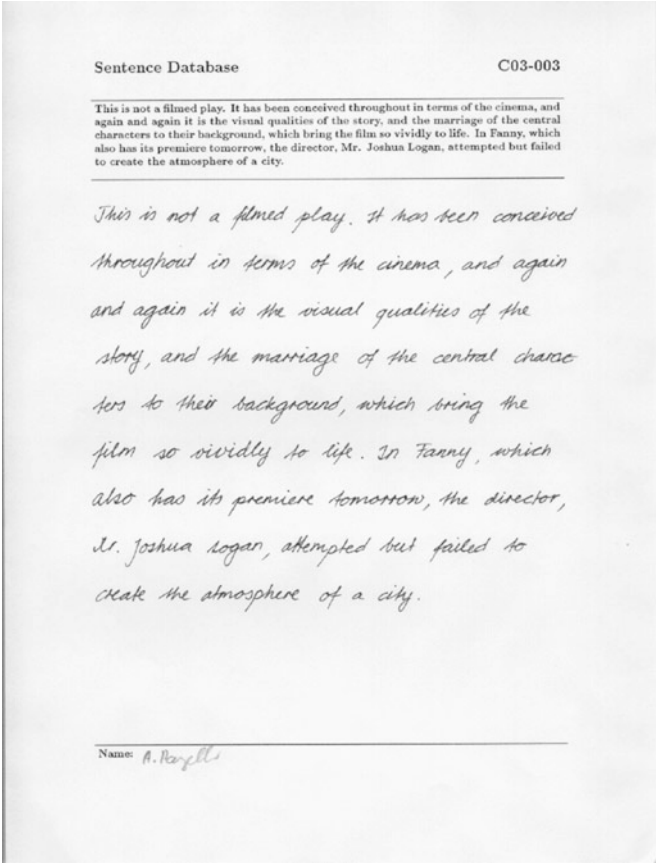
Parole Hampton.

The officers who came down from Fort Cumberland with Colonel Washington are immediately to go Recruiting: and they are allowed until the 5th of December; at which time if they do not punctually appear at the place of Rendezvous assigned them, they will be tried by a Court Martial, for disobedience of Orders.

They are to wait upon the Aid de camp at one of the Clock to receive their Recruiting Instructions. Each Officer present to give in a Return immediately of the number of men he has enlisted. — One Subaltern, one Sergeant, one Corporal, one Drummer, and twenty-five private men, are to mount Guard to-day, and to be relieved to-morrow at ten o'Clock. — All Reports and Returns are to be made to the Aid de camp.

Fig. 2.3 Example of a digitized handwritten document from *The George Washington Papers at the Library of Congress, 1741–1799*, Series 2, Letterbook 1, page 270, reprinted with permission of Library of Congress, Manuscript Division

on character level virtually impossible. Therefore, segmentation-free methods on the basis of Markov models are predominant in the field of handwriting recognition today. However, the automatic reading of larger handwritten documents is currently not a relevant application possibly due to the enormous difficulty of the task. Even in the scientific area today only a few related research efforts exist. An example of a historic document in quite neatly written cursive script is shown in Fig. 2.3. In contrast, Fig. 2.4 shows a contemporary document as it was used for building a large corpus of handwritten texts within a research initiative of the University of Bern, Switzerland (cf. [192]). The challenges of automatically reading every-day texts are exemplified by the example of a historical postcard shown in Fig. 2.5.



Sentence Database C03-003

This is not a filmed play. It has been conceived throughout in terms of the cinema, and again and again it is the visual qualities of the story, and the marriage of the central characters to their background, which bring the film so vividly to life. In Fanny, which also has its premiere tomorrow, the director, Mr. Joshua Logan, attempted but failed to create the atmosphere of a city.

This is not a filmed play. It has been conceived throughout in terms of the cinema, and again and again it is the visual qualities of the story, and the marriage of the central characters to their background, which bring the film so vividly to life. In Fanny, which also has its premiere tomorrow, the director, Mr. Joshua Logan, attempted but failed to create the atmosphere of a city.

Name: A. Perelli

Fig. 2.4 Example page from the corpus of handwritten documents created by the Institute for Informatics and Applied Mathematics, University of Bern, Switzerland [192]. In the *upper part* the text section to be written together with an identification code is printed. *Below* the respective handwritten version was filled in by one of the subjects (reprinted with permission)

The complexity in the processing of handwritten documents results from the considerable larger variability in the image of the writing as opposed to machine-printed texts. Similarly to sounds in spoken language, the individual characters differ even when repeatedly realized by the same person, depending on their context, and the most severely between different writers. Without suitable restrictions on the potential character sequences, no satisfactory results are achieved in the automatic recognition of handwritten texts. Therefore, the vocabulary used and comparable contextual knowledge is of great importance for this task.

An especially important application of the automatic recognition of writing is the automated reading of postal addresses in mail sorting machines. Here the essential difficulty results from the part of handwritten addresses which is large even today. Therefore, in practice an automatic sorting of the mail pieces can be achieved only

Fig. 2.5 Example of a historical postcard written in German Kurrent script during the First World War and delivered by the German Military Postal Service (Private Collection Dr. Britta Bley, Dortmund, Germany, reprinted with permission). Please note how the writer used every possible writing space by adding paragraphs with different script orientations and with ever decreasing character size—even on the image side of the postcard

partially as recognition errors by wrongly dispatched mail cause enormous costs. In order to be able to keep the error rate of such systems as low as possible, a quite large rejection rate has to be accepted. The addresses on rejected mail pieces then need to be transcribed manually.

Powerful methods for the automatic analysis of addresses—especially for machine-printed postal addresses—have been applied in practice for many years. In postal address recognition the quality of the results not only depends on the classification accuracy on character level which should be as high as possible. Additionally, it is extremely important to intelligently exploit relations in structure and content

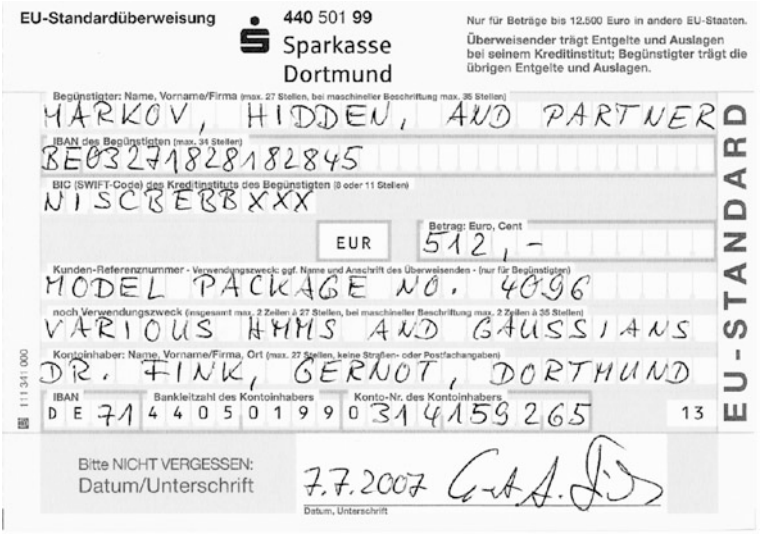


Fig. 2.6 Example of a form for bank transfers in the European Union filled out manually. Handwritten capital characters need to be written into the appropriate fields given

between the individual elements of a postal address, e.g., city name, zip-code, street name, and house number. After the introduction of a new automatic address reading system at the end of the last century in the U.S. more than half of the handwritten addresses were analyzed automatically while the estimated error rate was below 3 percent [50].[7]

Almost exclusively devoted to the processing of handwritten input are methods for the automatic processing of forms. When automatically reading forms, the complexity of the recognition task can be considerably reduced by suitable technical measures for limiting the variability of the character images. Special fields for writing individual words or even short phrases are used in order to facilitate text segmentation. Additionally, the writing style is often restricted to the exclusive use of hand-printed characters which frequently also means that only capital letters may be used. The most restricted writing style is obtained if every character has to be written within an individual field of the form. Figure 2.6 shows this principle with the example of a form for bank transfers used in many countries sharing the Euro as currency in the European Union.

In the U.S., in contrast to Europe, checks play an especially important role in financial transactions. A multitude of approaches exist for automatically analyzing the handwritten legal amount, its numeric equivalent, the so-called courtesy amount, and the date of issue (cf. e.g. [165]). In this recognition task the lexicon

[7] Today, the processing of handwritten addresses will be even more reliable. Unfortunately, manufacturers of postal automation equipment or postal service providers hardly ever publish up-to-date performance figures.

Fig. 2.7 Example of a pen trajectory of the handwritten word "*handwriting*" captured online. The size of the dots representing the pen positions encodes the pen pressure in pen-down strokes (black) and the shading represents the distance to the writing surface during pen-up movements

is restricted to only a few entries consisting of numerals and digits. However, from these basic units longer sequences may be constructed without appreciable sequencing restrictions. The writing style used also satisfies no constraints. Thus the use of hand-printed characters, cursive writing, and an arbitrary mixture of those styles is possible. This mixed writing style is usually referred to as so-called unconstrained handwriting.

All methods for processing writing presented so far have in common that documents are captured digitally *after* they were completed and are then processed further independently from their generation process. This type of analysis is referred to as *offline* handwriting or character recognition. In contrast to that, also so-called *online* methods exist for the processing of handwritten documents where the motion of the pen is already captured during the process of writing. For this purpose special sensors as, e.g., pressure sensitive tablets or LC displays, are required and, in general, also the use of specialized tools for writing. In essence a sequence of two-dimensional measurements of the pen position is obtained as the digital representation of the written texts. This sequence encodes the trajectory of the pen as observed during the writing process. In simple devices as they are, e.g., used in so-called personal digital assistants (PDAs), position measurements are only obtained if the pen touches the writing surface resulting in so-called *pen-down* strokes. However, if the pen is lifted, only more sophisticated graphics tablets can still track the pen in so-called *pen-up* movements in the close vicinity of the writing surface by using inductive techniques and specialized pens. Such devices usually also provide the pen pressure and the inclination of the pen with respect to the writing surface in addition to the position measurements. Figure 2.7 shows a potential pen trajectory by the example of the word "`handwriting`". There the pen pressure and the distance of the pen tip to the writing surface are encoded in the size and shading, respectively, of the dots used to represent the pen positions.

Compared to the offline processing of handwriting, online methods have the advantage that they can take into account the additional information about the temporal organization of the writing process during recognition. Thus it is, for example, not possible that neighboring characters overlap in the image of the writing and, therefore, can hardly be separated in the automatic segmentation. The dynamic information is also essential for the verification of signatures. It represents a highly

writer specific peculiarity of the signature which even by experts cannot be forged on the basis of an available image of the signature.

However, the main application area for online handwriting recognition is man–machine interaction. Especially for operating extremely small, portable computing devices, which would not be reasonably possible with a keyboard, this form of text entry has become very attractive for device control.[8] Usually the problem is simplified as much as possible in order to achieve satisfactory results with the limited resources of a PDA, organizer, or smart phone while at the same time reaching sufficiently high reaction times. In the well known PalmPilot and its descendants only isolated characters were captured in special input fields. Additionally, a special writing style optimized for the purposes of automatic recognition needed to be used. On today's more powerful devices usually also the input of complete handwritten words or phrases is possible.

Inspired by the success of Markov model-based methods in the field of automatic speech recognition, the principal approach of this technique was also transferred to problems of automatic character and handwriting recognition in recent years. These segmentation-free methods are mostly applied where the "classical" OCR approaches, which first segment on the level of characters and later classify, are either not reliably enough or fail completely. Therefore, Markov models are mainly used for the recognition of handwritten documents in both online and offline mode and only rarely for the processing of machine-printed texts. Similarly to automatic speech recognition, hidden Markov models are applied for modeling the appearance in writing of individual characters or whole words, and Markov chain models are used for restricting potential sequences of elementary units on character or word level.

The fundamental prerequisite for the applicability of these methods is that the signal data considered can be represented as a linear sequence. This is rather easily possible in the field of online handwriting recognition. The temporal progress of the writing process itself defines a chronological order of the position measurements which are provided by the respective sensors. The time-line of the signal thus virtually runs along the trajectory of the pen. Similarly to the short-time analysis of speech signals, local characteristic properties of the pen trajectory can be described by feature vectors. For this purpose mainly shape properties are evaluated as, for example, the writing direction or the curvature. In contrast, the writing speed is usually normalized in pure recognition systems in order to avoid variations of the signal characteristics due to different writers.

It is considerably more difficult to define a comparable procedure for the serialization of offline documents as these are principally two-dimensional images. However, in general a segmentation of the document considered into individual text lines

[8] With the introduction of capacitative touch-sensitive displays, pen-based input methods have often been replaced by so-called soft keyboards. Apple's former CEO Steve Jobs severely influenced the turning away from pen-based interaction. Technically, it is also more challenging to use pen input on a capacitative display. Just recently pens with conductive tips were introduced that make writing on these interactive surfaces possible. Interestingly, Apple itself may be reviving pen-based interaction in the future with its ideas of an "active stylus".

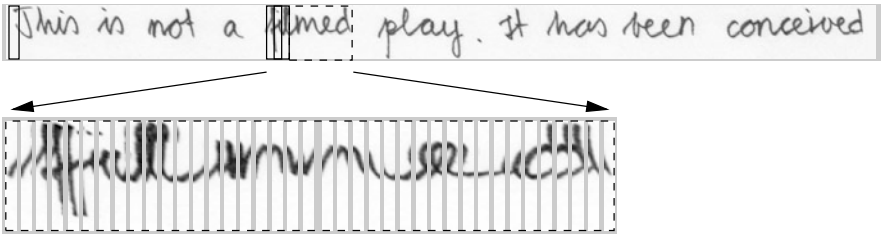


Fig. 2.8 Example for the serialization of offline handwriting: *Above*, the text line to be analyzed is shown with some of the overlapping analysis windows superimposed. *Below*, the extracted image stripes are shown for part of the text line (inspired by [241]; example based on the document image known from Fig. 2.4, used with permission; pre-processed by the author)

can be generated with sufficient reliability. When analyzing forms, which usually comprise only isolated words or phrases, the segmentation of the field contents is possible even more easily. Then a hypothetical time-line can be defined in parallel to the direction of the text. Following this direction the changing local properties of the image of the writing can be described by a sequence of feature vectors. For this purpose the text line is usually subdivided into a sequence of narrow overlapping image windows which then in principle correspond to the frames known from automatic speech recognition. Figure 2.8 shows an example of this so-called *sliding-window approach* applied to a text-line image. For each of the analysis windows extracted such a feature vector is computed. Unfortunately, in the field of offline handwriting or character recognition no generally accepted method exists for extraction of feature representations from images of writing. In some cases local structural properties are computed as, e.g., the number of line ends or arcs, which lie within a certain frame of writing. Most recent feature extraction approaches rely on the computation of statistical image descriptors as, e.g., moments or histograms. The sequence of the feature vectors generated thus is then—similarly to automatic speech recognition—identified with the outputs of hidden Markov models for characters or words.

2.3 Biological Sequences

The genetic information of all living organisms, which influences their growth, their metabolism, and to a major extent also their behavior, is encoded in a symbolic sequence. In the majority of cases this is the macro molecule *deoxyribonucleic acid* (DNA) which consists of two strands intertwined in the form of a double helix. The strands are built as sequences of so-called *bases*. There exist four different types of bases (adenine, cytosine, guanine, and thymine) which are pairwise complementary and, therefore, in addition to the chemical bonds within a DNA strand also establish pair bonds to bases from the other strand. Thus one obtains the "ladder-type" structure of the double-stranded DNA. As the pairwise bonds are unique, already a single DNA strand contains the complete genetic information. Therefore, in the double strand it is encoded redundantly.

In higher developed organisms as, for example, mammals, the DNA is not available as a single complete sequence but distributed across so-called *chromosomes*. Human cells contain 23 pairs of those, which represent the genetic information from maternal and paternal side, respectively. The entirety of the DNA strands in all chromosomes is referred to as the *genome*. Every cell of a living being contains an identical copy of this total genetic information. The size of the genome is coarsely connected to the complexity of the respective organism. While the genetic information of bacteria contains only a few million bases, the human genome comprises approximately 3 billion base pairs.

However, the majority of the DNA sequence has no cell-biological function or none that has been understood so far. In this additional "junk" material the elementary units of genetic information—the so-called *genes*—are embedded. Their relevant information that encodes the function of a gene—the so-called coding region—is generally split up into multiple *exons* which are interrupted by *introns*. A few years ago it was still assumed that the human genome contains approximately 30 000 to 40 000 genes [131, 295] while more recent estimates propose a total number of only approximately 25 000 coding regions [132].

For controlling most cell-biological functions, genes are "transformed" into *proteins* in a process called *expression*. The proteins created then influence the metabolism and the growth of the cell and control its reproduction during cell division.

In order to express a certain gene, first the genetic information available on the double-stranded DNA is read and transformed into the equivalent representation of the single-stranded *ribonucleic acid* (RNA) which also represents a sequence of bases. This process, which is referred to as *transcription*, begins in a so-called *promotor* region before the actual DNA sequence which contains the information of a specific gene. In the resulting raw version of the RNA, the coding region of a gene, in general, is still interrupted by introns without known function. Therefore, the RNA is "cleaned" in a subsequent modification process where the introns are discarded. The cleaned RNA is called *messenger* RNA or for short mRNA.

Finally, a certain protein is generated from mRNA in an additional transformation process which is referred to as *translation*. This protein realizes the functionality of the underlying gene. In contrast to DNA and RNA, proteins consist of a sequence of 20 different *amino acids*. Within the mRNA sequence a triple of bases—a so-called *codon*—encodes a certain amino acid.[9] Special start and stop codons control which area of the mRNA is covered by the translation process. After the generation of the amino acid sequence, proteins form a characteristic three-dimensional structure by folding which makes up a substantial part of their functionality. Figure 2.9 shows a part of the amino acid sequence of a protein for the example of hemoglobin as well as its representation as a sequence of codons on the level of DNA.

In contrast to sensor data, which are always affected by measurement noise, the symbolic representations of DNA sequences or proteins can in principle be given exactly. Therefore, one might assume that symbolic and rule-based methods are

[9] The relationship between codons and amino acids is not unique as with 4 bases there exist $4^3 = 64$ potential triples.

Amino Acid Sequence

```
MALSAEDRALVRALWKKLGSNVGVYTTEALERTFLAFPATKTYFSHLDLS
PGSSQVRAHGQKVADALSLAVERLDDLPHALSALSHLHACQLRVDPASFQ
LLGHCLLVTLARHYPGDFSPALQASLDKFLSHVISALVSEYR
```

DNA Sequence

```
atggcgctgt ccgcggagga ccgggcgctg gtgcgcgccc
tgtggaagaa gctgggcagc aacgtcggcg tctacacgac
agaggccctg gaaaggacct tcctggcttt ccccgccacg
aagacctact tctcccacct ggacctgagc cccggctcct
cacaagtcag agcccacggc cagaaggtgg cggacgcgct
gagcctcgcc gtggagcgcc tggacgacct accccacgcg
ctgtccgcgc tgagccacct gcacgcgtgc cagctgcgag
tggacccggc cagcttccag ctcctgggcc actgcctgct
ggtaaccctc gcccggcact accccggaga cttcagcccc
gcgctgcagg cgtcgctgga caagttcctg agccacgtta
tctcggcgct ggtttccgag taccgctga
```

Fig. 2.9 Part of the amino acid sequence and the underlying DNA sequence of the protein hemoglobin according the SWISS-PROT database [12]. Individual amino acids are encoded by capital letters and bases by lower case letters

completely sufficient for genome analysis. However, the sequencing of a genome poses considerable difficulties in practice (cf. [76, Chap. 5]). Therefore, even after the completion of the "*Human Genome Project*" [129], the genetic information of the human cells investigated can still not be given completely and with absolute certainty. Furthermore, genetic information is not encoded uniquely in the sequence of base pairs and is subject to a wide range of random variations within a family of organisms and also within the same species. Therefore, for complex genomes the actual number and position of the individual genes can still be estimated only even for the extensively studied human genome. In order to understand the function of the proteins expressed, it is additionally essential to consider the so-called expression pattern, i.e., under what conditions they are created from the respective genes, and to investigate the three-dimensional structure that is formed. The latter can result in functional equivalent form from different amino acid sequences.

The variations within biological sequences pointed out above, which—according to current scientific knowledge—are largely random, have helped statistical methods to become predominant for their investigation and modeling.

Depending on which data basis the analysis of genetic material starts from, different processing steps are relevant. When starting from so-called genomic DNA, i.e., virtually raw genetic information, first the coding regions of genes need to be found and the DNA sequences present there subsequently need to be cleaned from introns in the same way as in the creation of mRNA.

When applying Markov model-based methods for the analysis of genomic DNA, individual HMMs for promotor regions as well as for exons and introns need to be created. A segmentation of the DNA sequence considered then allows to localize genes and to extract their coding region in a cleaned-up representation (cf. [118, 160]). However, genes can also be identified within DNA sequences on the basis

of Markov chain models which define restrictions for the occurrence of individual bases within different genetic contexts (cf. [222, 266]).

When starting directly from mRNA, this first processing step is not necessary as only a single gene is transcribed at a time and the final mRNA was already cleaned. However, depending on the life cycle of a cell, only a limited set of genes is expressed so that the investigation of a complete genome is virtually impossible on the basis of mRNA only.

Often only the final product of the transcription and translation process itself is considered, namely the proteins. When analyzing proteins, the goal is not a segmentation but the finding of similar sequences. The comparison of proteins is the most simple if they are only considered pairwise. Long before the application of hidden Markov models, probabilities for the mapping between amino acids at certain positions of the sequence as well as for their insertion and deletion were defined in order to be able to capture statistical variations of proteins. Such a statistical model can be used to associate an amino acid sequence with another one position by position. From such a position-wise pairing one obtains a so-called *alignment*. The logical positions within this mapping between two proteins are mostly directly connected to the three-dimensional structure formed.

It is considerably more demanding to apply the sequence alignment to multiple proteins of a certain family. The results of such efforts are represented in the form of so-called *multiple alignments*. From pre-existing multiple alignments the statistical properties of the respective groups of similar sequences can be derived and then be described by hidden Markov models (cf. [67, 69, 70, 161]). These so-called *profiles* can then be used to search the respective databases automatically for further similar proteins. Figure 2.10 shows a multiple alignment created by an expert for the example of different goblins.

Of course, the detection of new genes by the segmentation of a genome or the extension of a family of proteins with new members by means of statistical comparisons cannot show the cell-biological functions of the newly found structures. In the end this needs to be proven in biological experiments. However, from the structural comparison of biological sequences and the similarities found, hypotheses about the function of genes and proteins can be derived which can then be verified experimentally in a considerably more goal directed manner.

Such efforts are embedded in the endeavor of biologist and bioinformatics researchers to be able to explain the function of biological organisms. Especially an exact understanding of the human metabolism is of fundamental interest to the pharmaceutical industry. Substances constructed on the genetic level and especially adapted to a certain individual—such is the hope of the researchers—might make a substantially improved treatment of diseases possible without at the same time causing the often dramatic side effects of classical drugs. Therefore, the sequencing of more and more genetic material and its detailed analysis with respect to structure and cell-biological function is especially pushed by pharmaceutical companies.

```
Helix                AAAAAAAAAAAAAAA   BBBBBBBBBBBBBBBBCCCCCCCCCCC    DDDDDDDEE
HBA_HUMAN  ---------VLSPADKTNVKAAWGKVGA--HAGEYGAEALERMFLSFPTTKTYFPHF-DLS-----HGSA
HBB_HUMAN  --------VHLTPEEKSAVTALWGKV----NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP
MYG_PHYCA  ---------VLSEGEWQLVLHVWAKVEA--DVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASE
GLB3_CHITP ----------LSADQISTVqASFDKVKG------DPVGILYAVFKADPSIMAKFTQFAG-KDLESIKGTA
GLB5_PETMA PIVDTGSVAPLSAAEKTKIRSAWAPVYS--TYETSGVDILVKFFTSTPAAQEFFPKFKGLTTADQLKKSA
LGB2_LUPLU --------GALTESQAALVKSSWEEFNA--NIPKHTHRFFILVLEIAPAAKDLFS-FLK-GTSEVPQNNP
GLB1_GLYDI ---------GLSAAQRQVIAATWKDIAGADNGAGVGKDCLIKFLSAHPQMAAVFG-FSG----AS---DP

Helix      EEEEEEEEEEEEEEEEEE            FFFFFFFFFFFF  FFGGGGGGGGGGGGGGGGGGG
HBA_HUMAN  QVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL--RVDPVNFKLLSHCLLVTLAAHLPAE
HBB_HUMAN  KVKAHGKKVLGAFSDGLAHL---D--NLKGTFATLSELHCDKL--HVDPENFRLLGNVLVCVLAHHFGKE
MYG_PHYCA  DLKKHGVTVLTÄLGAILKK----K-GHHEAELKPLAQSHATKH--KIPIKYLEFISEAIIHVLHSRHPGD
GLB3_CHITP PFETHANRIVGFFSKIIGEL--P---NIEADVNTFVASHKPRG---VTHDQLNNFRAGFVSYMKAHT--D
GLB5_PETMA DVRWHAERIINAVNDAVASM--DDTEKMSMKLRDLSGKHAKSF--QVDPQYFKVLAAVIADTVAAG----
LGB2_LUPLU ELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG---VADAHFPVVKEAILKTIKEVVGAK
GLB1_GLYDI GVAALGAKVLAQIGVAVSHL--GDEGKMVAQMKAVGVRHKGYGNKRIKAQYFEPLGASLLSAMEHRIGGK

Helix       HHHHHHHHHHHHHHHHHHHHHHHHH
HBA_HUMAN  FTPAVHASLDKFLASVSTVLTSKYR------
HBB_HUMAN  FTPPVQAAYQKVVAGVANALAHKYH------
MYG_PHYCA  FGADAQGAMNKALELFRKDIAAKYKELGYQG
GLB3_CHITP FA-GAEAAWGATLDTFFGMIFSKM-------
GLB5_PETMA -----DAGFEKLMSMICILLRSAY-------
LGB2_LUPLU WSEELNSAWTIAYDELAIVIKKEMNDAA---
GLB1_GLYDI MNAAAKDAWAAAYADISGALISGLQS-----
```

Fig. 2.10 Multiple alignment of the amino acid sequence of seven goblins of different organisms after [161] with the respective identifiers used in the SWISS-PROT database [12]. The line designated by `Helix` defines the mapping onto the three-dimensional structure of the proteins. Deletions of amino acids at certain positions are designated by –

2.4 Outlook

Markov models represent a formalism which has received substantial attention in the field of pattern recognition and beyond due to the success of the technique in the area of automatic speech recognition. Therefore, it would be a pointless endeavor trying to list all problems that were ever tackled by applying these methods. However, in the following we will give an overview of the most important topics for which Markov models were used to a larger extent and which do not fall into their main application areas of automatic speech recognition, character and handwriting recognition, and the analysis of biological sequences.

Alternative recognition tasks on the basis of speech signals that have been tackled by Markov model-based systems are, for example, the recognition of prosodic structures (cf. [33]) or the recognition of emotions conveyed (cf. e.g. [219]). As HMMs are generative models, they can even be successfully applied to the approximately inverse problem of speech recognition, namely speech synthesis (see [293] for a recent survey).

Similarly to the processing of speech signals, which merely represent a sequence of measurements of the sound pressure level, hidden Markov models can be applied for the analysis of other series of measurements as they are, e.g., obtained in material testing (cf. e.g. [270, 300]) in biomedical applications (cf. e.g. [4, 127, 220]), or in remote sensing (cf. e.g. [104, 173]).

Comparable to online handwriting recognition is the automatic recognition of human gestures (cf. e.g. [29, 39, 72, 74, 149, 202, 213, 253]) which includes the recognition of sign-language as a special case (cf. e.g. [62, 288]). However, the trajectories of the hands and arms of a person and, if necessary, also the respective hand postures need to be extracted from the respective image sequences with costly image processing methods before the statistical analysis of the motion sequences. In this respect these methods are comparable in their structure to a video-based online handwriting recognition system [92, 313] developed on the basis of a method for tracking pen movements during writing in image sequences [201].

The recognition of human actions or behavior can be regarded as a generalization of gesture recognition. Thus in, e.g., [323] motion sequences of tennis players and in [31] of people walking are analyzed. As human gait is quite characteristic for individuals, can easily be observed from a distance, and is hard to conceal, hidden Markov models are increasingly applied to gait recognition in the context of surveillance applications (cf. e.g. [41, 144, 179, 291]). In [121] human motion patterns learned are used for mimicking them by a robot and thus having the demonstrated action carried out by the machine. Extremely special human actions are changes in facial expressions as they are, e.g., analyzed in [120, 177]. Related to action recognition and surveillance applications is the problem of detecting unusual—and, therefore, interesting—events (cf. e.g. [249, 329]).

In all these methods, first chronologically organized sequences of feature vectors are created from the input image sequences. The Markov model-based techniques then start from the serialized feature representations. However, in the literature also methods were proposed which extend the formalism of hidden Markov models such that a modeling of two- or even three-dimensional input data is directly possible (cf. e.g. [73, 141, 175, 176, 267]).

In virtually all approaches mentioned so far, hidden Markov models are used in isolation and not in combination with Markov chain models. This is mainly due to the fact that for applications as, e.g., gesture or action recognition only a rather small inventory of segmentation units is used. Therefore, probabilistic restrictions on the respective symbol sequences are not of immediate importance.

On the purely symbolic level, Markov chain models, in contrast, are applied for describing state sequences without being complemented by a hidden Markov model. An important application area is the field of information retrieval where statistical models of texts are described by Markov chain models (cf. e.g. [244]). As a compact representation of documents in principle corresponds to a compression of their content, the same principles also form the foundation of different methods for text compression (cf. [17]). Markov chain models are also applied in slightly modified form as so-called Markov decision processes for, e.g., the solution of planning tasks (cf. e.g. [310]).

http://www.springer.com/978-1-4471-6307-7

Markov Models for Pattern Recognition
From Theory to Applications
Fink, G.A.
2014, XIII, 276 p. 45 illus., Hardcover
ISBN: 978-1-4471-6307-7