# From Signal Processing to Machine Learning

Signal processing is a field at the intersection of systems engineering, electrical engineering, and applied mathematics. The field analyzes both analog and digitized signals that represent physical quantities. Signals include sound, electromagnetic radiation, images and videos, electrical signals acquired by a diversity of sensors, or waveforms generated by biological, control, or telecommunication systems, just to name a few. It is, nevertheless, the subject of this book to focus on *digital* signal processing (DSP), which deals with the analysis of digitized and discrete sampled signals. The word "digital" derives from the Latin word *digitus* for "finger," hence indicating everything ultimately related to a representation as integer countable numbers. DSP technologies are today pervasive in many fields of science and engineering, including communications, control, computing and economics, biology, or instrumentation. After all, signals are everywhere and can be processed in many ways: filtering, coding, estimation, detection, recognition, synthesis, or transmission, are some of the main tasks in DSP.

In the following sections we review the main landmarks of signal processing in the 20th century from the perspective of algorithmic developments. We will also pay attention to the cross-fertilization with the field of statistical (machine) learning in the last decades. In the 21st century, model and data assumptions as well as algorithmic constraints are no longer valid, and the field of *machine-learning signal processing* has erupted, with many successful stories to tell.

# 1.1 A New Science is Born: Signal Processing

## 1.1.1 Signal Processing Before Being Coined

One might argue that processing signals is as old as human perception of the nature, and you would probably be right. Processing signals is actually a fundamental problem in science. In ancient Egypt, the Greek civilization and the Roman Empire, the "men who knew" (nowadays called scientists and philosophers) measured and quantified river floods, sunny days, and exchange rates digitally. They also tried to predict them and model them empirically with simple "algorithms." One might say that system modeling, causal inference, and world phenomena prediction were matters that already existed at that time, yet were treated at a philosophical level only. Both the mathematical tools and the intense data exploitation came later. The principles of what we actually call

*Digital Signal Processing with Kernel Methods,* First Edition. José Luis Rojo-Álvarez, Manel Martínez-Ramón, Jordi Muñoz-Marí, and Gustau Camps-Valls.

© 2018 John Wiley & Sons Ltd. Published 2018 by John Wiley & Sons Ltd.

*signal processing* date back to the advances in classical numerical analysis techniques of the 17th and 18th centuries. Big names of European scientists, like Newton, Euler, Kirchhoff, Gauss, Cauchy, and Fourier, set up the basis for the latter development of sciences and engineering, and DSP is just the most obvious and particular case out of them. The roots of DSP can be found later in the digital control systems of the 1940s and 1950s, while their noticeable development and adoption by society took place later, in the 1980s and 1990s.

# 1.1.2 1948: Birth of the Information Age

The year 1948 may be regarded as the birth of modern signal processing. Shannon published his famous paper "A mathematical theory of communication" that established bounds for the capacity of a band-limited channel and created the discipline of information theory (Shannon, 1949). Hartley and Wiener fed Shannon's mind with their statistical viewpoint of communication; and others, like Gabor, developed the field enormously. In that year, Shannon also motivated the use of pulse code modulation in another paper. This year was also the year when modern digital methods were introduced: Bartlett and Tukey developed methods for spectrum estimation, while Hamming introduced error correcting codes for efficient signal transmission and recovery. These advances were most of the time motivated by particular applications: audio engineering promoted spectral estimation for signal analysis, and radar/sonar technologies dealt with discrete data during World War II that needed to be analyzed in the spectral domain. Another landmark in 1948 was the invention of the transistor at Bell Labs, which was still limited for commercial applications. The take-off of the signal processing field took place also because Shannon, Bode, and others discussed the possibility of using digital circuits to implement filters, but no appropriate hardware was available at that time.

# 1.1.3 1950s: Audio Engineering Catalyzes Signal Processing

DSP as we know it nowadays was still not possible at that time. Mathematical tools (e.g., the *z*-transform) were already available thanks to established disciplines like control theory, but technology was ready only to deal with low-frequency signal processing problems. Surprisingly, the field of audio engineering (boosted by the fever of rock 'n' roll in radio stations!) was the catalyst for new technological developments: automobile phonographs, radio transistors, magnetic recording, high-quality low-distortion loud-speakers and microphone design were important achievements.

The other important industry was telephony and the need for efficient repeaters (amplifiers, transistors): the transatlantic phone cable formed a huge low-pass filter introducing delays and intersymbol interferences in the communications, and time-assignment speech interpolators appeared as efficient techniques to exploit the pauses in speech during a phone conversation. Efficiency in communications took advantage of Shannon's theory of channel capacity, and introduced frequency-division multiplexing. Transmission capacity was alternatively improved by the invention of the coaxial cable and the *vocoder* in the 1930s by Dudley.

This decade is also memorable because of work on the theory of wave filters, mostly developed by Wagner, Campbell, Cauer, and Darlington. A new audio signal representation called *sound spectrograms*, which essentially shows the frequency content of speech as it varies through time, was introduced by Potter, Wigner, Ville, and other researchers. This time-frequency signal representation became widely used in signal processing some time later. Communications during World War II were quite noisy; hence, there was a notable effort in constructing a mathematical theory of signal and noise, notably by Wiener and Rice. The field of seismic data processing witnessed an important development in the early 1950s, when Robinson showed how to derive the desired reflection signals from seismic data carrying out one-dimensional deconvolution.

# 1.2 From Analog to Digital Signal Processing

## 1.2.1 1960s: Digital Signal Processing Begins

In the late 1950s, the introduction of the integrated circuit containing transistors revolutionized electrical engineering technology. The 1960s made technology ready for DSP. Silicon integrated circuits were ready, but still quite expensive compared with their analogical counterparts. The most remarkable contributions were the implementation of a digital filter using the bilinear transform by Kaiser, and the work of Cooley and Tukey in 1965 to compute the discrete Fourier transform efficiently, which is nowadays well known as the fast Fourier transform (FFT). DSP also witnessed the introduction of the Viterbi algorithm in 1967 (used especially in speech recognition), the chirp *z*-transform algorithm in 1968 (which widened the application range for the FFT), the maximum likelihood (ML) principle also in 1968 (for sensor-array signal processing), and adaptive delta modulation in 1970 (for speech encoding).

New and cheaper hardware made digital filters a reality. It was possible to efficiently implement long finite-impulse response (FIR) filters that were able to compete with analog infinite-impulse response (IIR) filters, offering better band-pass properties. But perhaps more crucial was that the 1960s were a time for numerical simulation. For instance, Tukey developed the concept of the *cepstrum* (the Fourier transform of the logarithm of the amplitude spectrum) for pitch extraction in a vocoder. In early 1961, Kaiser and Golden worked intensively to transfer filters from the analog to the digital domain. Digital filters also offered the possibility to synthesize time-varying, adaptive and nonlinear filters, something that was not possible with analog filters. Kalman filters (KFs) took advantage of the statistical properties of the signals for filtering, while Widrow invented the least mean squares (LMS) algorithm for adaptive filtering, which is the basis for neural networks (NNs) training. Bell Labs also developed adaptive equalizers and echo cancellers. Schroeder introduced the adaptive predictive coding algorithm for speech transmission of fair quality, while Atal invented linear predictive coding, which was so useful for speech compression, recognition, and synthesis.

In the 1960s, image processing stepped in the field of DSP through applications in space sciences. The topics of image coding, transmission, and reconstruction were in their infancy. In 1969, Anderson and Huang developed efficient coders and later introduced the famous discrete cosine transform (DCT) for image coding. Other twodimensional (2D) and three-dimensional (3D) signals entered the arena: computerized tomography scanning, interferometry for high-resolution astronomy and geodesy, and radar imaging contributed with improvements in multidimensional digital filters for image restoration, compression, and analysis. The wide range of applications, usually involving multidimensional and nonstationary data, exploited and adapted previous DSP techniques: Wiener filtering for radar/sonar tracking, Kalman filtering for control and signal-detection systems, and recursive time-variant implementation of closed-form filter solutions, just to name a few.

#### 1.2.2 1970s: Digital Signal Processing Becomes Popular

The 1970s was the time when video games and word processors appeared. DSP started to be everywhere. The speech processing community introduced adaptive differential pulse code modulation to achieve moderate savings in coding. Subband coding divided the signal spectrum into bands and adaptively quantized each independently. The technique was also used for image compression, with important implications in storage and processing.

Filter design continued introducing improvements, such as McClellan's and Parks's design of equiripple FIR filters, the analog-to-digital design procedures introduced by Burrus and Parks, Galand's quadrature mirror filters, and Darlington's multirate filters. Huang pioneered in developing filters for image processing. State-space methods and related mathematical techniques were developed, which were later introduced into fields such as filter design, array processing, image processing, and adaptive filtering. FFT theory was extended to finite fields and used in areas such as coding theory.

#### 1.2.3 1980s: Silicon Meets Digital Signal Processing

The 1980s will be remembered as the decade in which personal computers (PCs) became ubiquitous. IBM introduced its PC in 1981 and standardized the disk operating system. PC clones appeared soon after, together with the first Apple computers and the new IBM PC-XT, the first PC equipped with a hard-disk drive.

The most important fact in the 1980s for the signal processing point of view was the design and production of single-chip DSP. Compared with using general-purpose processors, specifically designed chips for signal processing made operations much faster, allowing parallel and real-time signal-processing systems. An important DSP achievement of the 1980s was JPEG, which essentially relies on the DCT, and still is the international standard for still pictures compression. The success of JPEG inspired efforts to reach standards for moving images, which was achieved in the 1990s in the form of MPEG1 and MPEG2. Automated image-recognition found its way into both military and civil applications, as well as for Earth observation and monitoring. Nowadays, these applications involve petabyte and multimillion ventures.

The introduction of NNs played a decisive role in many applications. In the 1950s, Frank Rosenblatt introduced the *perceptron* (a simple linear classifier), and Bernard Widrow the *adaline* (adaptive linear filter). Nevertheless, neural nets were not extensively used until the 1980s: new and more efficient architectures and training algorithms, capability to implement networks in very-large-scale integrated circuits, and the belief that massive parallelism was needed for speech and image recognition.<sup>1</sup> Besides the famous multilayer perceptron (MLP), other important developments are

<sup>1</sup> Interestingly, this is in current times a revived idea because of the concepts of "big data" and "deep learning."

worth mentioning: Hopfield's recurrent networks, radial basis function (RBF) networks, and Jordan's and Elman's dynamic and recurrent networks. NNs were implemented for automatic speech recognition, automatic target detection, biomedical engineering, and robotics.

# 1.3 Digital Signal Processing Meets Machine Learning

## 1.3.1 1990s: New Application Areas

The 1990s changed the rules with the effect of the Internet and PCs. More data to be transmitted, analyzed, and understood were present in our lives. On top of this, the continuing growth of the consumer electronics market impulsed DSP. New standards like MPEG1 and MPEG2 made efficient coding of audio signals widely used. Actually, the new era is "visual": image processing and digital photography became even more prominent branches of signal processing. New techniques for filtering and image enhancement and sharpening found application in many (in principle orthogonal) fields of science, like astronomy, ecology, or meteorology. As in the previous decade, space science introduced a challenging problem: the availability of multi- and hyperspectral images impulsed new algorithms for image compression, restoration, fusion, and object recognition with unprecedented accuracy and wide applicability.

New applications were now possible because of the new computer platforms and interfaces, the possibility to efficiently simulate systems, and the well-established mathematical and physical theories of the previous decades and centuries. People started to adopt DSP unconsciously when using voice-recognition software packages, accessing the Internet securely, compressing family photographs in JPEG, or trading in the stock markets using moving-average filters.

## 1.3.2 1990s: Neural Networks, Fuzzy Logic, and Genetic Optimization

NNs were originally developed for aircraft and automobile-engine control. They were also used in image restoration and, given its parallel nature, efficiently implemented on very-large-scale integrated architectures. Also exciting was the development and application of fractals, chaos, and wavelets. Fractal coding was extensively applied in image compression. Chaotic systems have been used to analyze and model complex systems in astronomy, biology, chemistry, and other sciences. Wavelets are a mathematical decomposition technique that can be cast as an extension of Fourier analysis, and intimately related to Gabor filters and time–frequency representations. Wavelets were considerably advanced in the mid-1980s and extensively developed by Daubechies and Mallat in the 1990s.

Other new departures for signal processing in the 1990s were related to fuzzy logic and genetic optimization. The so-called fuzzy algorithms use fuzzy logic, primarily developed by Zadeh, and they gave rise to a vast number of real-life applications. Genetic algorithms, based on laws of genetics and natural selection, also emerged. Since then they have been applied in different signal-processing areas. These are mere examples of the tight relation between DSP and computer science.

# 1.4 Recent Machine Learning in Digital Signal Processing

We are facing a totally new era in DSP. In this new scenario, the particular characteristics of signals and data are challenging the traditional signal-processing technologies. Signal and data streams are now massive, unreliable, unstructured, and barely fit standard statistical assumptions about the underlying system. The recent advances in interdisciplinary research are of paramount importance to develop new technologies able to deal with the new scenario. Powerful approaches have been designed for advanced signal processing, which can be implemented thanks to continuous advances in fast computing (which is becoming increasingly inexpensive) and algorithmic developments.

## 1.4.1 Traditional Signal Assumptions Are No Longer Valid

Standard signal-processing models have traditionally relied on the rather simplifying and strong assumptions of linearity, Gaussianity, stationarity, circularity, causality, and uniform sampling. These models provide mathematical tractability and simple and fast algorithms, but they also constrain their performance and applicability. Current approaches try to get rid of these approximations in a number of ways: by widely using models that are intrinsically nonlinear and nonparametric; by encoding the relations between the signal and noise (which are often modeled and no longer considered Gaussian independent and identically distributed (i.i.d.) noise); by using new approaches to treat the noncircularity and nonstationarity properties of signals; by learning in *anticausal systems*, which is an important topic of control theory; and, in some situations, since the acquired signals and data streams are fundamentally unstructured, by not assuming uniform sampling of the representation domain.

It is also important to take into account the increasing diversity of data. For example, large and unstructured text and multimedia datasets stored in the Internet and the increasing use of social network media produce masses of unstructured heterogeneous data streams. Techniques for document classification, part-of-speech tagging, multimedia tagging or classification, together with massive data-processing techniques (known as "big data" techniques) relying on machine-learning theory try to get rid of unjustified assumptions about the data-generation mechanisms.

## 1.4.2 Encoding Prior Knowledge

Methods and algorithms are designed to be specific to the target application, and most of the times they incorporate accurate prior and physical knowledge about the processes generating the data. The issue is two-sided. Nowadays, there is a strong need of constraining model capacity with proper priors. Inclusion of prior knowledge in the machines for signal processing is strongly related to the issue of encoding invariances, and this often requires the design of specific regularizers that constrain the space of *possible* solutions to be confined in a *plausible* space. Experts in the application field provide such knowledge (e.g., in the form of physical, biological, or psychophysical models), while engineers and computer scientists design the algorithm in order to fulfill the specifications. For instance, current advances in graphical models allow us to learn more about structure from data (i.e., the dynamics and relationships of each variable and their interactions), and multitask learning permits the design of models that tackle the problem of learning a task as a composition of modular subtask problems.

#### 1.4.3 Learning and Knowledge from Data

Machine learning is a powerful framework to reach the goal of processing a signal or data, turning it into information, and then trying to extract knowledge out of either new data or the learning machine itself. Understanding is much more important and difficult than fitting, and in machine learning we aim for this from empirical data. A naive example can be found in biological signal processing, where a learning machine can be trained from patients and control records, such as electrocardiograms (ECGs) and magnetic resonance imaging (MRI) spatio-temporal signals. The huge amount of data coming from the medical scanner needs to be processed to get rid of those features that are not likely to contain information. Knowledge is certainly acquired from the detection of a condition in a new patient, but there is also potentially important clinical knowledge in the analysis of the learning-machine parameters in order to unveil which characteristics or factors are actually relevant in order to detect a disease. For this fundamental goal, learning with hierarchical deep NNs has permitted increasingly complex and more abstract data representations. Similarly, cognitive information processing has allowed moving from low-level feature analysis to higher order data understanding. Finally, the field of causal inference and learning has irrupted with new refreshing algorithms to learn causal relations between variables.

The new era of DSP has an important constraint: the urgent need to deal with massive data streams. From images and videos to speech and text, new methods need to be designed. Halevy *et al.* (2009) raised the debate about how it becomes increasingly evident that machine learning achieves the most competitive results when confronted with massive datasets. Learning semantic representations of the data in such environments becomes a blessing rather than a curse. However, in order to deal with huge datasets, efficient automatic machines must be devised. Learning from massive data also poses strong concerns, as the space is never filled in and distributions reveal skewed and heavy tails.

## 1.4.4 From Machine Learning to Digital Signal Processing

Machine learning is a branch of computer science and artificial intelligence that enables computers to learn from data. Machine learning is intended to capture the necessary patterns in the observed data, such as accurately predicting the future or estimating hidden variables of new, previously unseen data. This property is known in general as *generalization*. Machine learning adequately fits the constraints and solution requirements posed by DSP problems: from computational efficiency, online adaptation, and learning with limited supervision, to their ability to combine heterogeneous information, to incorporate prior knowledge about the problem, or to interact with the user to achieve improved performance. Machine learning has been recognized as a very suitable technology in signal processing since the introduction of NNs. Since the 1980s, this particular model has been successfully exploited in many DSP applications, such as antennas, radar, sonar and speech processing, system identification and control, and time-series prediction (Camps-Valls and Bruzzone, 2009; Christodoulou and Georgiopoulos, 2000; Deng and Li, 2013; Vepa, 1993; Zhao and Principe, 2001).

The field of DSP revitalized in the 1990s with the advent of support vector machines (SVMs) in particular and of kernel methods in general (Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004; Vapnik, 1995). The framework of kernel machines

## 10 Digital Signal Processing with Kernel Methods

allowed the robust formulation of nonlinear versions of linear algorithms in a very simple way, such as the classical LMS (Liu *et al.*, 2008) or recursive least squares (Engel *et al.*, 2004) algorithms for adaptive filtering, Fisher's discriminants for signal classification and recognition, and kernel-based autoregressive and moving average (ARMA) models for system identification and time series prediction (Martínez-Ramón *et al.*, 2006). In the last decade, the fields of graphical models (Koller and Friedman, 2009), kernel methods (Shawe-Taylor and Cristianini, 2004), and Bayesian nonparametric inference (Lid Hjort *et al.*, 2010) have played an important role in modern signal processing. Not only have many signal-processing problems been tackled from a canonical machine-learning perspective, but the opposite direction has also been fruitful.

# 1.4.5 From Digital Signal Processing to Machine Learning

Machine learning is in constant cross-fertilization with signal processing, and thus the converse situation has also been satisfied in the last decade. New machine-learning developments have relied on achievements from the signal-processing community. Advances in signal processing and information processing have given rise to new machine-learning frameworks:

- *Sparsity-aware learning.* This field of signal processing takes advantage of the property of sparseness or compressibility observed in many natural signals. This allows one to determine the entire signal from relatively few scarce measurements. Interestingly, this topic originated from the image and signal processing fields, and rapidly extended to other problems, such as mobile communications and seismic signal forecasting. The field of sparse-aware models has recently influenced other fields in machine learning, such as target detection in strong noise regimes, image coding and restoration, and optimization, to name just a few.
- *Information-theoretic learning.* The field exploits fundamental concepts from information theory (e.g., entropy and divergences) estimated directly from the data to substitute the conventional statistical descriptors of variance and covariance. The field has encountered many applications in the adaptation of linear or nonlinear filters and also in unsupervised and supervised machine-learning applications. In the most recent years the framework has been interestingly related to the field of dependence estimation with kernels, and shown successful performance in kernel-based adaptive filter and feature extraction.
- Adaptive filtering. The urgent need for nonlinear adaptive algorithms in particular communications applications and web recommendation tools to stream databases has boosted the interest in a number of areas, including sequential and active learning. In this field, the introduction of online kernel adaptive filters is remarkable. Sequential and adaptive online learning algorithms are a fundamental tool in signal processing, and intelligent learning systems, mainly since they entail constraints such as accuracy, algorithmic simplicity, robustness, low latency, and fast implementation. In addition, by defining an instantaneous information measure on observations, kernel adaptive filters are able to actively select training data in online learning scenarios. This active learning mechanism provides a principled framework for knowledge discovery, redundancy removal, and anomaly detection.

Machine-learning methods in general and kernel methods in particular provide an excellent framework to deal with the jungle of algorithms and applications. Kernel methods are not only attractive for many of the traditional DSP applications, such as pattern recognition, speech, audio, and video processing. Nowadays, as will be treated in this book, kernel methods are also one of the primary candidates for emerging applications such as brain–computer interfacing, satellite image processing, modeling markets, antenna and communication network design, multimodal data fusion and processing, behavior and emotion recognition from speech and videos, control, forecasting, spectrum analysis, and learning in complex environments such as social networks.