James L. Flanagan

Parametric Representation of Speech Signals

EDITOR'S INTRODUCTION

Our guest in this column is Dr. James L. Flanagan. Dr. Flanagan holds the doctor of science degree in electrical engineering from the Massachusetts Institute of Technology (MIT), the master of science degree from MIT, and the bachelor of science degree from Mississippi State University. Dr. Flanagan is Professor Emeritus at Rutgers University, serving earlier as director of the Rutgers Center for Advanced Information Processing and Board of Governors Professor of Electrical and Computer Engineering. He was Rutgers University's vice president for research until retirement in 2005. Dr. Flanagan spent 33 years at Bell Laboratories before joining Rutgers University. At Bell Labs he led Acoustics Research and later served as director of Information Principles Research. Over the course of his impressive career, Dr. Flanagan has had a long list of inventions and contributions to the signal processing field in several areas including psychoacoustics, array microphone processing, and digital loudspeakers. Most notably, many of his pioneering achievements were reduced to practice with an impact on our current daily lives including speech coding in MP3 and speech recognition. Dr. Flanagan has published approximately 200 technical papers in scientific journals. He is the author of a research text *Speech Analysis, Synthesis and Perception* (Springer Verlag), which has appeared in five printings and two editions, and has been translated in Russian. He holds 50 U.S. patents.

Dr. Flanagan is an IEEE Life Fellow, a long-time member of the Signal Processing Society, which he served as president in the earlier formative stages. Among his awards are the IEEE Medal of Honor (2005) and the U.S. National Medal of Science (1996), presented at the White House by the President of the United States. A special pride is the Signal Processing Society's creation and sponsorship of the IEEE James L. Flanagan Speech and Audio Processing Technical Field Award. He was chosen as the 2005 recipient of the Research and Development Council of New Jersey's Science/Technology Medal. Dr. Flanagan is a member of the National Academy of Engineering and the National Academy of Sciences.

In the past, Dr. Flanagan has enjoyed deep-sea fishing, swimming, sailing, hiking, and flying as an instrument-rated pilot. He currently lives in New Jersey with his wife, Mildred, and they have three sons, all married and with families.

In October 2009, the Marconi Foundation in Italy combined with the Marconi Society based at Columbia University celebrated the centennial of the Nobel Prize to Guglielmo Marconi for his contribution in advancing wireless telegraphy. The occasion, in Bologna, Italy, was also the platform for the 2009 Marconi Fellowship Award. A main part of the program was a technical symposium, which additionally was joined by the Italian Federation of Industry Leaders. Several Marconi Fellows were asked to make presentations in the symposium. Dr. Flanagan chose to talk about efficient digital speech communication, one area favored in his research at AT&T Bell Labs. Specifically, Dr. Flanagan offered a perspective that highlighted junctures from conventional analog telephony to ambitions for the future.

In this article, Dr. Flanagan gives a condensed summary of his Marconi presentation, devoted to parametric representation of speech signals. We have arranged for his audio demonstrations to be available at http://www.signalprocessingsociety.org/publications/periodicals/spm/columns-resources/, as well as in IEEE *Xplore*. Regarding the future of speech coding, Dr. Flanagan says "The future is certain to prove interesting!" I am confident that you, our readers, will find this column interesting and you will enjoy reading this perspective from a long-term innovator and expert in the signal processing field.

Ghassan AlRegib

Telephony was conceived as the electrical transmission of a facsimile of the sound pressure waveform radiated from a talker's mouth. A microphone performed the acoustic to electrical conversion, and a low-pass filter typically confined the signal to a bandwidth adequate for intelligibility, about 3,000 Hz. Electrical noise might intrude in transmission. When needed, electronic amplification strengthened the signal to compensate for its attenuation over distance. But, accumulated noise would also be amplified along with the signal, hence signal-to-noise ratio could diminish with transmission distance.

Digital Object Identifier 10.1109/MSP.2010.936028

GENESIS

Even with these analog deficiencies, this principle has served voice communication, both by wire and by radio, for more than 100 years.

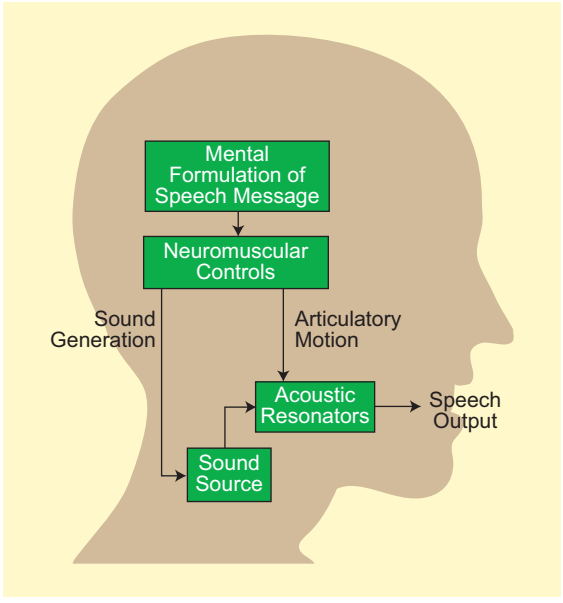
Despite the success and utility of this principle, it was recognized early that it was not efficient. Neural-activated vocal musculatures can exert only finite force, so the velocities and displacements of

the massive articulators are continuous functions of time. Further, the articulators change relatively slowly in producing a sequence of distinctive sounds—something at the rate of ten phonemes/s—not nearly at the rate of 3,000 cycles/s, typical of telephone bandwidth.

BANDWIDTH CONSERVATION

An early step towards bandwidth saving was the cogent observation that the vocal sound source, and the intelligence modulated upon it by the resonant vocal system, were largely linearly separable functions (Figure 1). This raised the possibility for parametric description of the radiated signal more in terms of the slowly changing vocal motions. This notion led to the Bell Labs Vocoder [1], where a frequency-modulated pulse generator and a broad-spectrum noise generator could approximate vocal-cord vibration and turbulent frication, and the modulated intelligence could be approximated by values of the short-time amplitude spectrum taken at ten frequencies over the audible frequency range. Implicitly, this development suggested that while waveform facsimile transmission was sufficient, it was not necessary. Rather, perceptually, preservation of the short-time amplitude spectrum was central to speech intelligibility.

The Vocoder was demonstrated in 1939. (And, a keyboard-operated version of the synthesizer, the Voder [2], provided a popular display at the New York World's Fair.) The time-varying parameters that described the source and resonant system occupied a bandwidth less than 300 Hz, or one-tenth that of the telephone channel. This was almost small enough to transmit speech over the transatlantic telegraph cable, laid in 1866! (The first transatlantic telephone cable had to await development of submersible amplifiers, and didn't become a reality until 1956.) But the parametric description was too coarse to provide good speech quality when synthesized at the receiver. Additionally, the analog parameters were susceptible to noise interference. The issues of how to compress speech bandwidth and resist interference continued to command attention.



[FIG1] **Source-resonator representation of the speech process.**

> CONTINUED PROGRESS AIMED TO EXPLOIT THE SLOWLY CHANGING NATURE OF THE SPEECH SIGNAL AND ITS LOW-PASS CHARACTER.

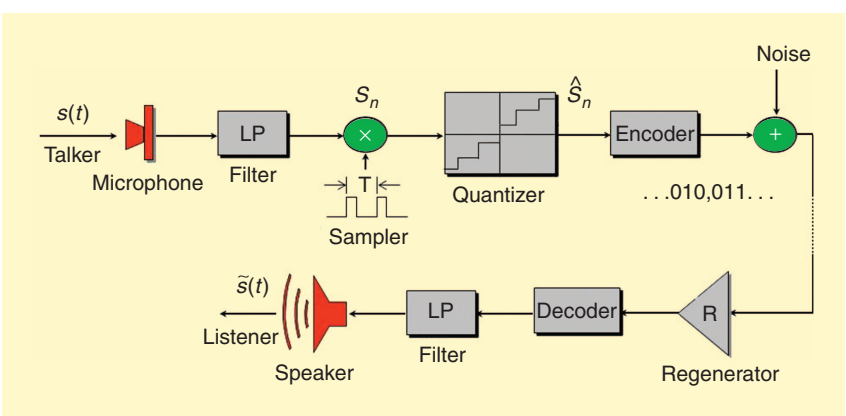
QUALITY INDEPENDENT OF DISTANCE

Resistance to analog noise was dramatically impacted by expanded understanding of sampled-data theory and by the advent of digital technology. An initial step, pulse code modulation (PCM), was simply the conversion of the 3 kHz sound waveform into digital form (Figure 2). This entailed sampling a band-limited signal, quantizing the amplitude samples, and converting the quantized values into time-framed binary "words" by an encoder. Any noise accumulated in transmission could be "stripped away" by detecting the binary pulses and regenerating them before they were overwhelmed by interference. At the receiver, the binary words were decoded, converted to pulse amplitudes, and low-pass filtered to recover the original signal (along with quantizing noise, which could be made negligible with enough steps in the quantizer, or enough binary digits, i.e., bits per word).

Although conceived by Rainey in 1926 and rediscovered independently by Reeves in 1937 [3], PCM had to await electronic progress. The first commercial deployment was in 1962, when Illinois Bell introduced the T1 carrier, employing 8 kHz sampling and 8-bit log-amplitude quantization. This process was still a waveform transmission system. But it gave the world noise-free telephonic transmission whose quality was essentially independent of the transmission distance.

DIFFERENTIAL CODING

Continued progress aimed to exploit the slowly changing nature of the speech

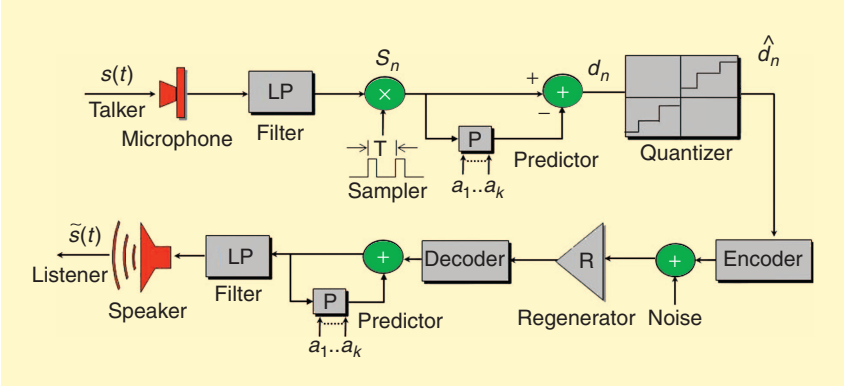


[FIG2] **An example of PCM.**

signal and its low-pass character. (The ratio of the frequency of the upper band edge to the centroid of the speech spectrum is about six, with the bulk of spectral energy in the lower frequencies—by virtue of the characteristics of vocal sound generation and radiation.) Adjacent sample values of the waveform are consequently similar. Differential PCM (DPCM) (Figure 3) was therefore proposed, whereby, at the transmitter, a local estimate of each signal sample is made based upon correlation statistics of past values [4], [5]. This parameterized estimate, or prediction, is subtracted from the input signal. If the estimate is good, the difference signal is greatly reduced in power and requires fewer bits of quantization. After transmission, with regeneration and then decoding to pulse amplitude form, the signal is recovered by an accumulator with the same predictor, and finally desampled by a low-pass filter.

The nature of the typical predictor is a weighted linear sum of some number, k, of past samples. This essentially is a transversal filter, whose time domain impulse response is the sum of weighted delta functions of delay equal to k times the sampling interval T. The weights are the coefficients $\{a_k\}$. In the sampled-data frequency domain, the filter response, P(z), is the sum over k of the product of the predictor coefficients and their corresponding delay operator z^{-k}. The transmitter operates on the spectral input as $[1-P(z)]$ and the receiver operates on the difference spectrum as $1/[1-P(z)]$ which, in the absence of quantizing, exactly recovers the input.

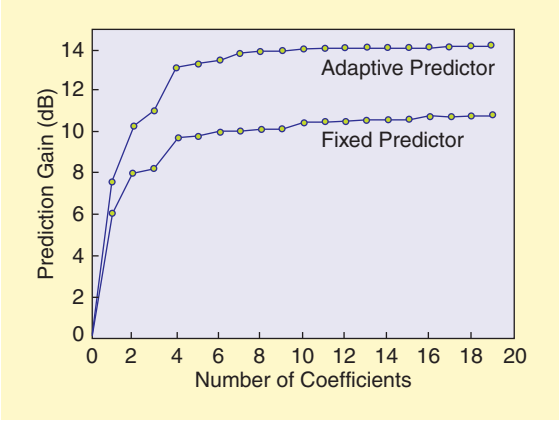
A body of mathematics provides a closed-form computation of the predictor coefficients, $\{a_k\}$, that minimizes the power of the difference signal [6]. The computation requires inversion of a matrix of correlation values, and hence comprises the main processing requirement. If long-term statistics are used, the coefficients of the predictor can



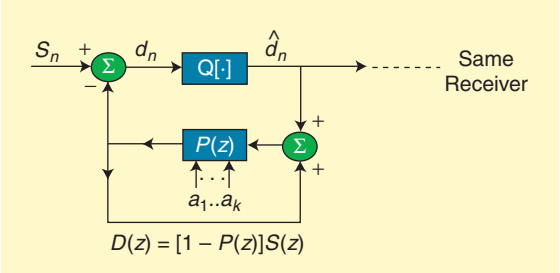
[FIG3] **DPCM: open-loop quantizing.**

be fixed at both the transmitter and receiver, and only the difference signal is transmitted. Its power is reduced by about 10 dB (Figure 4). If short-term statistics are used to make the predictor adaptive (and hence achieve better prediction), the coefficients are typically computed every 20–30 ms. The time-varying coefficients are then separately transmitted to the receiver along with the difference signal. Refinements in prediction techniques permit encoding delays significantly shorter than 20–30 ms. For predictors of order greater than about six, adaptive prediction reduces the power of the difference signal by about another 4 dB, requiring still fewer bits for quantization [7].

LINEAR PREDICTION CAN BE EXTENDED TO CHARACTERIZE THE VOCAL SOUND SOURCE AS WELL AS THE RESONATOR SPECTRUM.



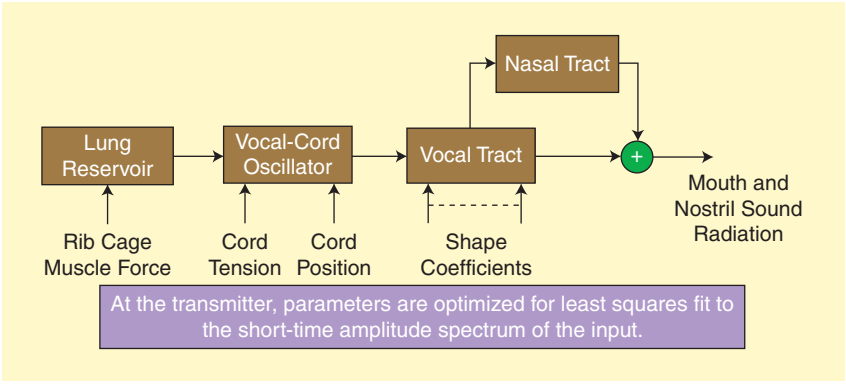
[FIG4] **Prediction gain as a function of predictor order [7].**



[FIG5] **DPCM: closed-loop quantizing.**

Commonly, traditional DPCM employs fixed predictors that accommodate the low-pass nature of the speech spectrum. Linear predictive coding (LPC) additionally transmits time-varying predictor coefficients that follow the slow changes in the amplitude spectrum of the signal [8], [9]. The difference signal retains much of the characteristics of the vocal sound source.

While shown in Figure 3 as open loop and fixed quantizing at the transmitter, there are advantages to closed-loop quantization (Figure 5). In this arrangement, the predicted signal derives from the quantized difference, and is the same as that generated at the receiver. Quantizing noise at transmitter and receiver are the same and do not accumulate [10]. (One can confirm that the closed-loop transmitter operation, in absence of quantizing, is [1-P(z)], as for the open-loop case.) Further, arranging for the



[FIG6] Components of the "speech mimic" system.

quantizer step size to be adaptive in time additionally reduces the number of bits required in the quantizer. This involves specifying the desired number of bits and simple processing logic to constantly examine the code words issued from the encoder, and expand or contract the step size in accordance with whether the signal is consistently occupying the highest or lowest quantal value. In the absence of transmission error, the receiver "sees" the same code words and has the same logic to modify step size for digital-to-analog recovery [11].

Linear prediction can be extended to characterize the vocal sound source as well as the resonator spectrum [12]. This parameterization of vocal excitation allows further reduction in transmission rate, especially into the Vocoder range. Various approaches have been established for this, embracing pitch and voiced/unvoiced estimation at one extreme, and "code books" of excitation at the other. If employed, parameters for these analyses must be additionally transmitted to the receiver. The price of this progressive reduction in transmission rate is increased complexity.

At this time, ITU standards have been promulgated for a range of coding rates from 64 kb/s down to cell phone speed 8 kb/s, based upon differential predictive coding. Military standards have been established for the low rates of 4.8 and 2.4 kb/s. The latter are greatly refined and more complex derivatives of the original concept of the Vocoder. Complexity and cost are essentially in inverse relation to the coding rate. If complexity is expressed in millions of instructions per second (MIPS), processing grows from a fractional MIP for 64 kb/s PCM up to the order of 100 MIPS for 2.4 kb/s Vocoders. For error-free transmission, good quality and talker recognition can be maintained by increased complexity down to cell phone speeds of about 8 kb/s (requiring about 20 MIPS). In the Vocoder range, even with greater complexity, some degradation in quality and talker recognition typically remains evident.

PRACTICAL SPEECH COMPRESSION HAS ADVANCED A DISTANCE, BUT LIKELY HAS NOT REACHED THE LIMIT OF EFFICIENCY.

RESEARCH OUTLOOK

What is the outlook for ultimately achieving good performance into the very low coding rates? The most refined present-day Vocoders at 2.4 kb/s provide useful intelligibility, but do not achieve good talker recognition. Is there a fundamental limit to the minimum information rates that meet the joint objectives of speech intelligibility and high quality? Informal experiments have indeed demonstrated an "existence proof" of transparency (where the synthesis is indistinguishable from the natural input) for coding at rates as low as 2,000 b/s. But, this compression has required lengthy and laborious human intervention in the analysis. No practical solution yet exists for this goal.

If interest focuses solely upon the written equivalent of the information in a speech signal, a greater gap is evident. For a given language, the number of distinctive speech sounds is of the order of 32–64. Typically, about ten phonemes are uttered per second, corresponding to an information rate of only 50–60 b/s. Some perceptual experiments suggest that the human is capable of processing and making decisions on pattern features at information rates only of the order of 100 b/s [13], [14]. This seems incredibly low, but the task of making as many as 100 yes/no decisions per second could indeed be taxing.

These observations provoke attempts to look at information representation higher up the chain of acoustic, muscular, and neural processes. That is, not focus on the radiated sound, but on the factors that produce the sound. A small step in this direction is to examine speech-sound generation from first principles of fluid flow, incorporating the physiology and dynamic constraints of the vocal mechanism (Figure 6). Here, the control factors relate to the subglottal air reservoir, the vocal-cord source, turbulence generation and the time-varying vocal resonator system. With enough computation, controls for this formulation can be sought by having a physiological model "mimic" a continuous speech input. Control parameters can be computed by gradient decent to minimize the difference between the short time amplitude spectra of the original speech and that of the "mimic." In simplest form, excitation information is separately measured [15].

ARTICULATORY REPRESENTATION

A main intelligence-bearing component here is the shape of the vocal tract, which can be parameterized [16]. Some of the inherent constraints and dynamics can be incorporated into a model of the sagittal plane cross-sectional area of the vocal conduit. A model of the whole system allows computation of sound for one-dimensional wave propagation in the conduit, when it is excited by nonlinear valving of air flow by the vocal cords and turbulence generation at

positions where the Reynolds number exceeds a critical value. The acoustic volume velocities can be obtained for the glottal excitation and for the mouth and nostril radiation. The output volume currents act through radiation impedances and encounter atmospheric pressure. The sound pressure in front of the speaker's mouth is determined as the superposition of the radiation of pistons (mouth and nostril) set in a spherical baffle (the head). All controls are related to the dynamic physiology. Initial implementations are exceedingly primitive. But, by appealing to an articulatory domain for parameterization, we are able to focus on the speech-producing mechanism, rather than on the sound output itself.

Preliminary experiments with such a "mimic" suggest that information rates in the range of 1,000–2,000 b/s may preserve quality and personal characteristics. But, so far deep studies of the fluid flow approach have not been made (hampered in part by the fact that even a "stripped down" model runs over 100 times real time on a mainframe computer, mainly to compute solutions of the Navier-Stokes fluid flow equations).

Because this discussion has focused primarily on applied commercial voice transmission, it has not touched on a variety of related topics that partake of common fundamental components. Automatic speech recognition (What was said?) and talker verification (Who said it?) are cases in point, and are being brought into commercial telecom services. The continuing interests in formant analysis/synthesis seek automatic extraction of the time-varying eigen frequencies of the vocal system. These contribute the prominent maxima in the short-time amplitude spectrum and, perceptually, promise even more parsimonious description of speech information. All these factors underlie the transmission techniques emphasized here.

EPILOGUE

So, practical speech compression has advanced a distance, but likely has not reached the limit of efficiency. Implied is even the possibility for obtaining fundamental speech coding parameters at the neural level. That is, just *think* what you want to say! And, there are ambitious studies commencing in this sector. The future is certain to prove interesting!

ACKNOWLEDGMENTS

This review is an abbreviated form of a presentation to the Marconi Foundation Symposium honoring the centennial of G. Marconi's Nobel Prize for radio telegraphy in Bologna, Italy, 9 October 2009. I am indebted to Prof. Lawrence Rabiner, Dr. Richard Cox, Dr. Joseph Hall, and Ann-Marie Flanagan for their advice and assistance in preparing this article.

AUTHOR

James L. Flanagan (jlf@caip.rutgers.edu) is a Professor Emeritus at Rutgers University.

REFERENCES

[1] H. Dudley, "The vocoder," *Bell Labs Rec.*, vol. 17, pp. 122–126, 1939.

[2] H. Dudley, R. Riesz, and S. Watkins, "A synthetic speaker," *J. Franklin Inst.*, vol. 227, pp. 739–764, 1939.

[3] E. O'Neill, Ed., in *A History of Engineering and Science in the Bell System: Transmission Technology (1925–1975)*. AT&T Bell Laboratories, 1985, ch. 18, p. 527.

[4] C. Cutler, "Differential quantization of communications," U.S. Patent 2 605 361, July 1952.

[5] F. de Jager, "Delta modulation, a method of PCM transmission using a 1-unit code," *Philips Res. Rep.*, vol. 7, pp. 442–466, 1952.

[6] P. Elias, "Predictive coding," *IRE Trans. Inform. Theory*, vol. IT-1, pp. 16–33, 1955.

[7] P. Noll, "A comparative study of various schemes for speech encoding," *Bell Syst. Tech. J.*, vol. 54, pp. 1597–1611, 1975.

[8] B. Atal and S. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637–655, 1971.

[9] F. Itakura and S. Saito, "An analysis-synthesis telephony based on maximum likelihood method," in *Proc. Int. Congr. Acoustics*, Tokyo, Japan, 1968, Paper C-5-5.

[10] R. McDonald, "Signal-to-noise and idle channel performance of differential pulse code modulation systems," *Bell Syst. Tech. J.*, vol. 45, pp. 1123–1151, 1966.

[11] P. Cummiskey, N. Jayant, and J. Flanagan, "Adative quantization in differential PCM coding of speech," *Bell Syst. Tech. J.*, vol. 52, pp. 1105–1118, 1973.

[12] B. Atal and M. Schroeder, "Predictive coding of speech signals," in *Proc. Int. Congr. Acoustics*, Tokyo, Japan, 1968, Paper C-5-4.

[13] J. Pierce and J. Karlin, "Information rate of the human channel," *Proc. IRE*, vol. 45, p. 368, 1957.

[14] W. Keidel, "Information processing by sensory modalities in man," in *Cybernetic Problems in Bionics*, H. Oestreicher and D. Moore, Eds. New York: Gordon and Breach, 1968, pp. 277–300.

[15] J. Flanagan, K. Ishizaka, and K. Shipley, "Signal models for low bit-rate coding of speech," *J. Acoust. Soc. Amer.*, vol. 68, pp. 780–791, 1980.

[16] C. Coker, "Speech synthesis with a parametric articulatory model," in *Proc. Kyoto Speech Symp.*, Kyoto, Japan, 1968, pp. A-4-1–A-4-6.

[SP]

[from the **GUEST EDITORS**] continued from page 19

an important role in system design. The article by Zhang et al. gives a broad overview of the spectrum sharing approach for cognitive radio networks and describes in detail various convex optimization formulations and solutions for the design of cognitive radio systems.

Finally, the article by Jiang and Li focuses on the applications of convex optimization to discriminative training in speech and language processing. For many widely used statistical models, discriminative training for speech processing normally leads to nonconvex optimization problems. This article shows how convex relaxation techniques (such as linear programming relaxation or SDR) can be used in this context.

In closing, we would like to thank all of our colleagues who have contributed to this special issue, including the authors of submitted papers. We also thank the reviewers for their quality work, and the editorial board for their support, without which this special issue would not have been possible.

[SP]