Signal Processing for Multimedia
J.S. Byrnes (Ed.)
IOS Press, 1999

Speech Processing

Manfred R. Schroeder

University of Göttingen, Germany
and AT&T Bell Laboratories (ret.)

ABSTRACT. The processing of speech signals has a long and venerable history. As early as 1770 Wolfgang von Kempelen demonstrated his mechanical talking machine to the courts of Europe. In 1928 Homer Dudley invented the "vocoder" (voice coder) arguing that speech is specified by a few slowly varying parameters requiring only a fraction of the telephone bandwidth for transmission. A digital vocoder was first put into service in World War II for a secure telephone link connecting Roosevelt, Churchill and major military commands around the world.

Exploiting properties of the human ear, such as "phase deafness" and auditory masking, perceptual coders have been demonstrated that transmit speech (and even high-quality music!) at fractional bits per sample. Applications for mobile radio, voice-email, and Internet radio abound.

The success of speech recognition depends on the size of the vocabulary and the quality of the speech signal. The zero-error recognition of unrestricted, continuous speech from a noisy environment (the "electronic secretary"), however, is still in the future.

Speaker identification has helped solve several disasters (mid-air collision over the Grand Canyon, burning-up of three astronauts), but its forensic applications are limited if the pool of potential speakers is large. Speaker *verification* is of increasing importance in limiting access to restricted data (financial, medical, military).

Text-to-speech (TTS), although still suffering from an "electronic accent", has innumerable applications from "talking books" for the blind to a wide variety of spoken-language information services.

1. Speech Compression

Speech compression, once an esoteric preoccupation of a few speech enthusiasts, has taken on a practical significance of undreamed-of proportion. It all began in 1928 when Homer Dudley, a telephone engineer at Bell Laboratories, had a brilliant idea of how to compress a speech signal with a bandwidth of over 3000 Hz into the 100 -Hz bandwidth of a new *transatlantic telegraph* cable. Instead of sending the speech signal itself, he thought it would suffice to transmit a *description* of the signal to the far end of the link. This basic idea of substituting for the signal a sufficient specification from which it could be recreated is still with us in the latest linear prediction standards and other methods of speech compression for cell phones, secure digital voice channels, compressed-speech storage for multimedia, and, last but not least, Internet broadcasting and Internet telephony.

Dudley's original idea was to transmit information about the motions of the speaker's articulatory organs, like the tongue and the lips. When this turned out to be impossible (in fact, it is still difficult to extract these parameters from a running speech signal), Dudley suggested sending a succession of *short-time spectra* instead. This led to the *channel vocoder* in which the speech spectrum is described by its smooth *envelope* and, in the case of voiced sounds, the spectral *fine structure*, that is the spacing of the harmonics of the fundamental frequency or voice "pitch".

The first important application of the vocoder occurred in World War II when it was used to encrypt the telephone link between Churchill in London and Roosevelt in Washington. The compression made possible by the vocoder permitted the speech signal to be encoded by as little as the 1500 bits per second that fitted into existing transatlantic radio channels.

Vocoder work for civilian applications was resumed after the war, but it had to start almost from scratch because the progress made during the war, along with numerous patents, were classified "top secret" and kept under wraps. For general telephony, the principal difficulty was the "pitch problem", the need to track the exact fundamental frequency in real time. To make matters worse, the fundamental frequency of much telephone speech is actually missing in the telephone signal that does not transmit frequencies below 200 or 300 Hz.

When the author joined these efforts at Bell Telephone Laboratories in 1954, the most promising approach to the pitch problem was autocorrelation analysis. Of course, for a steady voiced speech sound, the first maximum of the autocorrelation function (at nonzero delay) occurs at a delay corresponding to one pitch period. But, unfortunately, most speech signals are not "steady" and the highest maximum in the delay region of interest corresponds to a delay of one pitch period plus or minus one *formant* period. A vocoder driven from such a pitch signal sounds horrible – drunk, to be more precise. To overcome these difficulties dozens of schemes were floated and tried – and found wanting for various reasons. The pitch problem was finally laid to rest with the invention of cepstrum pitch detectors.

However, even the cepstrum had problems with tracking the pitch of two different voices on the same line. For such cases, a better solution than the cepstrum was the Fourier transform of the *magnitude* of the Fourier transform, in other words, replacing the logarithm of the power spectrum (as in the cepstrum) by the square root of the power spectrum. Another method that sometimes outperformed the cepstrum is the "harmonic product spectrum", in which each harmonic frequency is considered, in a probabilistic manner, an integer multiple of the fundamental frequency.

The frequency-channel vocoder was soon joined by numerous other parametric compression schemes such as formant vocoders, harmonic compressors, correlation vocoders and phase vocoders.

Great strides were also made in speech and audio *waveform* compression, beginning with simple delta-modulation. Another early attempt on the waveform was made by M. V. Mathews by means of *extremal coding*, in which only the positions and amplitudes of the maxima and minima of the waveform are maintained, while intermediate values are approximated by a spline function. The digital simulation of extremal coding was the first successful digital simulation of a signal processor on a general purpose computer at Bell Labs. Digital simulation had been transplanted from MIT to Bell by Mathews in 1955, when he joined Bell's acoustics research. Extremal coding was tailor-made for digital simulation because computer-running times were quite reasonable. (The author later took digital simulation to "unreasonable" lengths by simulating signal processors containing hundreds of sharp bandpass filters and simulating sound transmission in full-size concert halls.)

Later, several refinements were made to waveform coding, leading to adaptive differential pulse code modulation (ADPCM), which became an international standard. Although requiring higher bit rates than parametric compressors, waveform coding, combined with proper *subjective* error criteria, is still much in the running for high-quality audio coding for high-definition television (HDTV) and motion-picture industry standards (MPEG). But for the highest compression factors and relaxed quality requirements, parametric compressors, especially vocoders based on linear prediction, reign supreme.

1.1. *Compression by Linear Prediction*

Linear prediction, especially in the form of *code-excited linear prediction* (CELP), has become the method of choice for speech compression.

Linear prediction came to the fore in speech research in 1966. The author, after 12 years of work on vocoders, had become impatient with their unsatisfactory speech quality. The idea was to encode speech signals not in a rigid vocoder-like fashion but to leave room for "error" in the coding process. Thus was born linear predictive coding (LPC) for speech signals with a prediction residual or "error" signal to take up the slack from the prediction. Since speech is a highly variable signal, B. S. Atal and the author opted for an *adaptive* predictor. Having taken up research in hearing in the 1950s (with the aim to design better sounding speech codes), the author proposed replacing the r.m.s. error criterion in linear prediction by a *subjective* measure, namely the perceived *loudness of the quantizing noise*. If given a proper spectral shape, the quantizing noise becomes less audible or is even completely *masked* by the speech signal itself. Beginning in 1972 J. L. Hall, Jr., and the author measured the *masking of noise by signals* (as opposed to the customary masking of signals by noise). The result was linear prediction with a perceptual error criterion, first implemented by Atal. The method of perceptual audio coding (PAC) has now found wide application in both speech and general audio (music) coding. Together with an excitation signal derived from a code book (CELP), bit rates for the prediction residual of 1/4 bit per sample for high-quality speech were realized by Atal and the author in 1980.

For audio compression, rates as low as 16 kilobits per second were demonstrated by D. Sinha, J. D. Johnston, S. Dorward and S. R. Quackenbush. Near compact disk (CD) quality was achieved at 64 kbps!

1.2. *Waveform Coding*

The history of waveform coding is almost as old as that of vocoders. It all started with delta-modulation at Philips Research in the Netherlands in the 1940s. Later *adaptive* delta-modulation became a method of choice. After the advent of pulse-code modulation (PCM), adaptive pulse-code modulation (APCM) was introduced.

Subband coding, advanced by N. S. Jayant and P. Noll, and wavelets lead to a near revolution in waveform coding.

2. Speech Synthesis from Text

Speech synthesis from written text has been a long-standing goal of linguists and engineers alike. One of the early incentives for "Talking Machines" came from the desire to permit the blind to "read" books and newspapers. True, tape-recorded texts give the blind access to some books, magazines and other written information, but tape recordings are often not available. Here a "reading machine" might come in handy, a machine that could transform letters on the printed page into intelligible speech. Scanning a page and optically recognizing the printed characters is no longer a great problem – witness the plethora of optical text scanners available in computer stores today. The real problem is the conversion of strings of letter, the "graphemes", to phonetic symbols and finally the properly concatenated sequence of speech sounds.

In speech synthesis from written material, one of the first steps is usually the identification of whole words in the text and their pronunciation, as given by a string of phonetic symbols. But this string is only a guide to pronouncing the word in isolation and not as embedded in a meaningful grammatical sentence. Speech is decidedly not, as had long been surmised, a succession of separate speech sounds strung together like a string of pearls. Rather, the ultimate pronunciation is determined by the syntactical function of the word within its sentence and the *meaning* of the text. This meaning can often be inferred only by inspecting several sentences. Thus, proper speech synthesis from general text requires lexicographical, syntactical, and semantic analyses. These prerequisites are the same as for automatic translation from one language to another and they are one reason why translation by machine remains difficult.

(Another reason of course is that some utterances in one language are literally untranslatable into certain other languages.) No wonder then that good, natural-sounding automatic speech synthesis from unrestricted texts is anything but easy!

Beyond reading machines for the blind, there is an ever-increasing need to convert text, on the printed page or in computer memory, into audible form as speech. With the ever-spreading Internet, a huge store of information is only a mouse click away for ever more people. While much of this information is best absorbed by looking at a printed document, in many cases an oral readout would be preferable: think of a driver in a moving car, the surgeon bent over the operating table or any other operator of machinery who has his hands and eyes already fully occupied by other tasks. Or think of receiving text information via a modem over cable or over the air (by mobile phone, say). In such cases a voice output of text would be a good option to have. This is particularly true for people on the go who could receive their text e-mail by listening to the output of a text-to-speech synthesizer. Such *voice e-mail* would obviate the need for lugging a portable printer around the country (or the world). Finally, many people on our globe cannot read; they have to rely on the spoken word.

Still other applications of speech synthesis from text result from the great bit compression it permits. Waveform and parameter coding of speech signals allow compression down to a few thousand bits per second. By contrast, the corresponding written text, albeit lacking intonation, requires only a few hundred bits per second at normal read-out rates (and even less with proper entropy coding). Thus text-to-speech synthesis (in connection with automatic speech recognition) would allow the ultimate in bit compression.

Synthesis of natural speech from unrestricted text also requires proper *prosody*: word and sentence intonation, segment durations, and stress pattern. All three aspects of prosody have inherent ("default") values, which govern the word when spoken in isolation. But necessary modifications from these standards depend on the structure of the sentence and, again, the intended meaning and mode of speaking: is the utterance a question, an order, a neutral statement, or what?

A related aspect of human speech is its "style": is the speaker shouting or preaching? Is he or she reading from a newspaper or a detective novel? How fast is he speaking? Does the speaker feel anxiety? How confident is she? All these different styles affect not only the prosody but reach into the articulatory domain and influence the course of the formant frequencies. (For example, for fast speech, vowels tend to be "neutralized", i.e. the formant frequencies migrate to those of the uniform-area vocal tract.) There are also many interesting interactions. The pause structure, for instance, influences the intonation. The beginning of a talk sounds subtly different from its ending. (This writer, apparently on the basis of such subtle linguistic cues, is almost always aware – he may even wake up in time to applaud – when "the end is near".)

2.1. *Model-Based Speech Synthesis*

Most synthetic speech is "manufactured" by speech synthesizers such as linear predictive coders (LPC), formant vocoders or "terminal analogs" of the vocal tract. These synthesizers may exists either as hardware or, more commonly, as software. The ("low level") parameters (predictor coefficients, formant frequencies, samples of the area functions) that control these synthesizers are computed from a few "high level" parameters (such as tongue position, lip rounding etc.). These parameters are obtained from articulatory models that incorporate the physical and linguistic constraints of human speech production.

Needless to say, the algorithms necessary for these conversions are not exactly simple. A vast body of research has been expended on the study of the human speaking process, including high-speed motion pictures of the human vocal chords, x-ray movies of the articulators, electrical contacts on the palate, hot-wire flow meters, magnetic field probes to track the motions of various articulators, and myographic recordings from the muscles that activate the

articulators. In addition, neural networks have been trained to speak in an attempt to learn more about the human speaking facility.

One of the several areas in which more research is required is the functioning of the vocal chords. Future high-quality speech synthesizers may also have to forego the fiction that vocal chords and vocal tract are completely decoupled mechanical systems. There is no dearth of research topics in speech synthesis!

2.2. *Synthesis by Concatenation*

One of the most seductive methods of synthesizing speech from text is by stringing together, or *concatenating*, prerecorded words, syllables, or other speech segments. This avoids many of problems encountered in phoneme-by-phoneme synthesis, such as the coarticulatory effects between neighboring speech sounds. Still, even words do not usually occur in isolation: the words immediately preceding or following a given word influence its articulation, its pitch, duration and stress, and the *meaning* of the entire utterance. (You just can't get away from meaning in speech, be it synthesis, recognition and, perforce, translation.)

Another problem of word concatenation is the large dictionary required for general-purpose texts. (I once gave a talk in Philadelphia and had the computer deliver the introduction by text-to speech synthesis using word concatenation. I wanted the machine to say "I just arrived from New Jersey", but, alas, the word *Jersey* wasn't in the dictionary. What to do? Well, Philadelphia isn't Brooklyn, but, as I had hoped, *Joy-See* was readily understood.)

The size-of-the-dictionary problem is, of course, alleviated if one concatenates *syllables* rather than whole words. But then coarticulation effects become more complex again. To minimize the more difficult coarticulation effects, it is best to base the dictionary on consonant-vowel-consonant (CVC) strings and to cut these strings in the center of the steady-state vowel, yielding *demisyllables*. Another approach to divide and conquer syllables are *diphones* (vowel to postvocalic consonant transitions).

For many languages, demisyllables minimize the coarticulation effects at syllable boundaries because the demisyllables are obtained from natural utterances by "cutting" in the middle of a steady-state vowel. Thus only relatively simple concatenation rules might be required – in the best of all worlds. But the reality of human speech is more complex and a successful concatenation system may have to rely on a combination of demisyllables, diphones and suffixes (postvocalic consonant clusters).

2.3. *Prosody*

For some time now, text-to-speech (TTS) systems have produced intelligible, if unpleasant sounding, speech. Much synthetic speech still has an unnatural ("electronic") accent and the fault lies largely at the door of prosody: voice pitch, segment durations, loudness fluctuations and other aspects of speech that go beyond the sequence of phonemes of the utterance. It has been shown that proper prosody is also crucial in ease of understanding. For example, subjects who have to perform a "competing" task do so more reliably while listening to high-quality speech and they tire later compared to subjects listening to speech with improper prosody.

Prosody is also heavily (and heavenly) dependent on the gender of the speaker. And there is more to the gender difference than pitch height. (B. S. Atal and the writer once tried to change a male into a female voice by just raising the fundamental frequency. The resulting "hermaphrodite" was a linguistic calamity. Even changing the formant frequencies and bandwidths in accordance with female vocal tract physiology did not help much: the voice of the gynandroid never sounded very attractive.)

But there is considerable commercial interest in changing voices and accents, not only from male to female (and vice versa), but from, say, "Deep South" to Oxford English (and vice versa?).

In spite of persisting difficulties, progress toward more human sounding, intelligible speech has been made during the last several decades. In his inaugural lecture at the University of Göttingen in 1970, the writer demonstrated the then current standard of TTS by playing a German poem by Heinrich Heine, synthesized on an American computer (slightly modified for the occasion, courtesy of Noriko Umeda and Cecil Coker). It seemed that nobody understood more than a few words. Then the wily lecturer played the same tape once more, this time around with a simultaneous (but unannounced) projection of the text. Suddenly everybody understood. But most listeners were not aware why they understood the second playing. Thus, 30 years ago anyhow, providing visual cues (preferably the complete text) was a great help in rendering TTS intelligible.

3. Speech Recognition and Speaker Identification

3.1. *Speech Recognition*

The automatic recognition of spoken language and its transcription into readable text has long been a dream for people in the word business. I wish I could dictate this paper into an automatic speech recognizer rather than laboriously tapping it out with my two index fingers. Of course, once "voice typewriters" were widely available, people would miss the human typist brightening up the office as an intelligent working partner. In addition to the voice typewriter, there are many other useful applications of automatic speech recognition.

The success of automatic speech recognition depends critically on the specific task to which it is put. The recognition of the few words from a small vocabulary, spoken in isolation, preferably by a "master's voice", has been within reach for decades: witness the toy dog "Rex" of yore who wagged his tail in recognition (!) when addressed as "Rex". According to John Pierce, my longtime boss at Bell, this was – and he believed would remain – the only useful application of speech recognition. (Actually, the mechanical marvel responded in the same manner to any loud enough sound or noise – so much for early "speech recognition".)

Another early speech recognition system was Bell's Automatic Digit Recognizer, dubbed "Audrey," intended for voice dialing. After a brief training session on a new voice, it would dial correctly *most* of the numbers much of the time – but rarely a complete seven-digit number correctly. In the meantime, voice dialing has become very important in car phones, allowing the driver to keep his hands on the steering wheel and his eyes on the road. But instead of calling out telephone numbers, the driver would simply say "office", "home", and so forth (or "broker" if he isn't broke already).

As the size of the vocabulary grows and the pool of talkers is widened, reliable speech recognition becomes more difficult. If, in addition, words are not pronounced in isolation but strung together into fluent, conversational speech, if there are background noises, echos and reverberation, useful speech recognition soon strains present capabilities. However, interest in speech recognition and speaker identification refuses to wither and research continues vigorously. Progress might be slow, but progress there will be. After all, the automobile, originally conceived as an ersatz horse, runs triumphant, although it still balks at leaping fences. And the airplane, too, inspired by the swift bird, has grown in size and importance without ever having mastered the fine art of alighting on a tree.

3.2. *Dialogue Systems*

From among the numerous applications for automatic speech recognition, perhaps the most (im)pertinent are the task-oriented dialogue systems, such as travel information services. These require, of course, speech synthesis as well as recognition. One prominent example is the Philips Train Timetable Information System, which has been running in Switzerland and the Netherlands since 1996. Another (earlier) example is the AT&T Bell Laboratories

"Conversational Mode Speech Understanding System" by S. E. Levinson and L. R. Rabiner. Here is a brief dialogue with the Bell system (not the long-defunct Bell System):

Customer: I want one first class seat on flight number 31 to Denver on Sunday.

Machine: I am sorry the flight is not available on Sunday.

Customer: I want to go on Saturday.

Machine: Flight number 31 leaves New York at 09:45 and arrives in Denver at 13:10. Etc.

The machine can, reportedly, also cope with some non-sequiturs and nonsensical questions by the customer (without losing its temper).

3.3. *Speaker Identification*

Speech recognition's sibling, automatic speaker identification, also, has many potential applications. I remember a visit, in 1995, by a delegation from the American Bankers Association at my Murray Hill office who wanted to know the chances of replacing payment by paper check by voice-actuated money transfer. The customer's voice was to take the place of the signature on the check. When I pointed out the unreliability of automatic speaker verification, they wanted none of it: North American banks were losing (I forget how many) millions of dollars every year owing to forged or illegible signatures – or no signatures at all. So a certain "false accept" rate was quite acceptable to the associated bankers.

Speaker identification or verification could also be of crucial importance in allowing (or denying) access to sensitive data or restricted facilities. Think of confidential medical reports or bank statements. A crazy colonel could conceivably start a war by pretending to be someone much higher up in the chain of command. In World War II, speaker identification (by visual inspection of sound spectrograms) was used to track the movements of German radio communicators, thereby allowing the Allies to anticipate forthcoming enemy forays. This was the first "field" application of the sound-spectrograph and "visible speech".

Beyond verifying a given speaker, identifying his accent or *dialect* is sometimes the goal. Again, I remember a visit, this time by a pair of "spooks" from Virginia. They were eager to learn whether it was possible to build a machine that could identify the dialect of an unknown voice. They had a secret recording, taped in a bar in Rio de Janeiro, of a Russian-speaking voice and they wanted to know whether a machine could tell that it had an Odessa accent. (I'll spare you my answer, which is "top secret" anyhow.)

3.4. *Pinpointing Disasters by Speaker Identification*

My first encounter with the usefulness of voice recognition was in 1956, when two airliners collided over the Grand Canyon. There was a last message, just before the crash, from one of the planes ending in the words "We are going in..." After that: silence. The Federal Aviation Authority surmised that the speaker had just seen the other plane and was crying out his fateful discovery. But who was the speaker? The answer would identify the position of the speaker in the cockpit and therefore the probable direction of the other plane.

Careful analysis of the spectrogram of the unknown voice by Larry Kersta revealed that it matched the characteristics of the flight engineer. This modest piece of information helped the Authority to reconstruct the course of the collision. Subsequently the FAA issued orders aimed at forestalling future accidents of this type.

Another tragedy that caught the world's attention was the burning up of three U.S. astronauts on the ground during a training session. Again I was at the receiving end of a horrible tape recording: the last words of a human being engulfed by flames. The voice screamed, at a

pitch exceeding 400 Hz, “Fire! We’re burning up!!”. Whoever said those words probably saw the fire first, implying that it had started on his side. But whose voice was it? The screaming had distorted it beyond human identification. But spectral analysis identified the screamer and helped NASA to take corrective action. (This included replacing the highly flammable pure-oxygen breathing atmosphere by a safer mixture of oxygen and nitrogen. The Russians had been capable of doing this much earlier in their space program because their rockets were more powerful and could carry the required greater payload.)

3.5. *Speaker Identification for Forensic Purposes*

With these successes in voice identification it is not surprising that linguists soon thought of enlisting spectral analysis for forensic purposes. The main interest was in identifying the voice of an extortionist or suspected criminal. Before long, sound spectrograms were christened “voice prints” by those eager to sell the new “art”, the implication being that they were as reliable as fingerprints.

To keep the discussion on safe scientific ground, the Acoustical Society of America formed a committee of speech experts to look into these claims. The main conclusions of the committee’s report emphasized that a suspect’s voice could sometimes be *excluded* with certainty on the basis of incompatible spectral data. In other words, the suspect, given his or her vocal apparatus, could never have produced all the features of the given utterance. Furthermore, an unknown voice could sometimes be “identified” with some probability from a limited pool of potential candidates. But, according to the committee, all bets were off for the identification of a voice from an *open* ensemble of speakers. Voiceprints are just not as uniquely characteristic of a person’s identity as fingerprints – notwithstanding the entry in the American Heritage Dictionary of the English Language (Third Edition, 1992): voice print (noun), an electronically recorded graphic representation of a person’s voice, in which the configuration for any utterance is uniquely characteristic of the individual speaker.

3.6. *Conclusion*

With still faster computers and ever wider applications, speech processing can look to a healthy future, bringing benefits to business and individuals alike.