1. Introduction to Speech Processing

J. Benesty, M. M. Sondhi, Y. Huang

In this brief introduction we outline some major highlights in the history of speech processing. We briefly describe some of the important applications of speech processing. Finally, we introduce the reader to the various parts of this handbook.

1.1 A Brief History of Speech Processing

Human beings have long been motivated to create machines that can talk. Early attempts at understanding speech production consisted of building mechanical models to mimic the human vocal apparatus. Two such examples date back to the 13th century, when the German philosopher Albertus Magnus and the English scientist Roger Bacon are reputed to have constructed metal talking heads. However, no documentation of these devices is known to exist. The first documented attempts at making speaking machines came some five hundred years later. In 1769 Kratzenstein constructed resonant cavities which, when he excited them by a vibrating reed, produced the sounds of the five vowels a, e, i, o, and u. Around the same time, and independently of this work, Wolfgang von Kempelen constructed a mechanical speech synthesizer that could generate recognizable consonants, vowels, and some connected utterances. His book on his research, published in 1791, may be regarded as marking the beginnings of speech processing. Some 40 years later, Charles Wheatstone constructed a machine based essentially on von Kempelen's specifications [1.1–3].

Interest in mechanical analogs of the human vocal apparatus continued well into the 20th century. Mimics of the type of von Kempelen's machine were constructed by several people besides Wheatstone, e.g., Joseph Faber, Richard Paget, R. R. Riesz, et al.

It is known that as a young man Alexander Graham Bell had the opportunity to see Wheatstone's implementation. He too made a speaking machine of that general nature. However, it was his other invention – the telephone – that provided a major impetus to modern speech processing. Nobody could have guessed at that time the impact the telephone would have, not only on the way people communicate with each other but also on research in speech processing as a science in its own right. The availability of the speech waveform as an electrical signal shifted interest from mechanical to electrical machines for synthesizing and processing speech.

Some attempts were made in the 1920s and 1930s to synthesize speech electrically. However it is Homer Dudley's work in the 1930s that ushered in the modern era of speech processing. His most important contribution was the clear understanding of the *carrier* nature of speech [1.4]. He developed the analogy between speech signals and modulated-carrier radio signals that are used, for instance, for the transmission or broadcast of audio signals. In the case of the radio broadcast, the message to be transmitted is the audio signal which has frequencies in the range of 0–20 kHz. Analogously, the message to be transmitted in the case of speech is carried mainly by the time-varying shape of the vocal tract, which in turn is a representation of the *thoughts* the speaker wishes to convey to the listener. The movements of the vocal tract are at syllabic rates, i. e., at frequencies between 0 and 20 Hz. In each case – electromagnetic and acoustic – the message is in a frequency range unsuitable for transmission. The solution in each case is to imprint the message on a carrier. In the electromagnetic case the carrier is usually a high-frequency sinusoidal wave. In the acoustic case the carrier can be one of several signals. It is the quasi periodic signal provided by the vocal cords for voiced speech, and a noise-like signal provided by turbulence at a constriction for fricative and aspirated sounds. Or it can be a combination of these for voiced fricative sounds. Indeed, the selection of the carrier as well as the changes in intensity and fundamental frequency of the

vocal cords may be conveniently regarded as additional parts of the message.

Being an electrical engineer himself, Dudley proceeded to exploit this insight to construct an *electrical* speech synthesizer which dispensed with all the mechanical devices of von Kempelen's machine. Electrical circuits were used to generate the carriers. And the message (i. e., the characteristics of the vocal tract) was imprinted on the carrier by passing it through a time-varying filter whose frequency response was adjusted to simulate the transfer characteristics of the vocal tract.

With the collaboration of Riesz and Watkins, Dudley implemented two highly acclaimed devices based on this principle – the Voder and the Vocoder. The Voder was the first versatile talking machine able to produce arbitrary sentences. It was a system in which an operator manipulated a keyboard to control the sound source and the filter bank. This system was displayed with great success at the New York World Fair of 1939. It could produce speech of much better quality than had been possible with the mechanical devices, but remained essentially a curiosity. The Vocoder, on the other hand had a much more serious purpose. It was the first attempt at *compressing* speech. Dudley estimated that since the message in a speech signal is carried by the slowly time-varying filters, it should be possible to send adequate information for the receiver to be able to reconstruct a telephone speech signal using a bandwidth of only about 150 Hz, which is about 1/20 the bandwidth required to send the speech signal. Since bandwidth was very expensive in those days, this possibility was extremely attractive from a commercial point of view.

We have devoted so much space here to Dudley's work because his ideas were the basis of practically all the work on speech signal processing that followed. The description of speech in terms of a carrier (or excitation function) and its modulation (or the time-varying spectral envelope) is still – 70 years later – the basic representation. The parameters used to quantify these components, of course, have evolved in various ways. Besides the channel Vocoder (the modern name for Dudley's Vocoder) many other types of Vocoders have been invented, e.g., formant Vocoder, voice-excited Vocoder. Besides speech compression, Dudley's description was also considered for other applications such as secure voice systems, and the sound spectrograph and its use for communication with the deaf.

Unfortunately, the quality achieved by analog implementations of Vocoders never reached a level acceptable for commercial telephony. Nevertheless they found useful applications for military purposes where poor speech quality was tolerated. The Vocoder representation was also the basis of a speech secrecy system that found extensive use during World War II.

Another example of an analog implementation of Dudley's representation is the sound spectrograph. This is a device that displays the distribution of energy in a speech signal as a function of frequency, and the evolution of this distribution in time. This tool has been extremely useful for investigating properties of speech signals. A real time version of the spectrograph was intended for use as a device for communication with the deaf. That, however, was not very successful. A few people were able to identify about 300 words after 100 hours of training. However, it turned out to be too difficult a task to be practical.

During more than three decades following Dudley's pioneering work, a great amount of research was done on various aspects and properties of speech – properties of the speech production mechanisms, the auditory system, psychophysics, etc. However, except for the three applications mentioned above, little progress was made in speech signal processing and its applications. Exploitation of this research for practical applications had to wait for the general availability of digital hardware starting in the 1970s. Since then much progress has been made in speech coding for efficient transmission, speech synthesis, speech and speaker recognition, and hearing aids [1.5–7]. In the next section we discuss some of these developments.

Today, the area of speech processing is very vast and rich as can be seen from the contents of this Handbook. While we have made great progress since the invention of the telephone, research in the area of speech processing is still very active, and many challenging problems remain unsolved.

1.2 Applications of Speech Processing

As mentioned above, one of the earliest goals of speech processing was that of coding speech for efficient transmission. This was taken to be synonymous with reduction of the bandwidth required for transmitting speech. Several advances were needed before the modern success in speech coding was achieved. First, the

notions of information theory introduced during the late 1940s and 1950s brought the realization that the proper goal was the reduction of information rate rather than bandwidth. Second, hardware became available to utilize the sampling theorem to convert a continuous band-limited signal to a sequence of discrete samples. And quantization of the samples allowed digitization of a band-limited speech signal, thus making it usable for digital processing. Finally, the description of a speech signal in terms of linear prediction coefficients (LPC) provided a very convenient representation [1.8–11]. (The theory of predictive coding was in fact developed in 1955. However, its application to speech signals was not made until the late 1970s.)

A telephone speech signal, limited in frequency from 0 to 3.4 kHz, requires 64 kbps (kilobits per second) to be transmitted without further loss of quality. With modern speech compression techniques, the bit rate can be reduced to 13 kbps with little further degradation. For commercial telephony a remaining challenge is to reduce the required bit rate further but without sacrificing quality. Today, the rate can be lowered down to 2.4 kbps while maintaining very high intelligibility, but with a significant loss in quality. Some attempts have been made to reduce the bit rate down to 300 bps, e.g., for radio communication with a submarine. However the quality and intelligibility at these low bit rates are very poor.

Another highly successful application of speech processing is automatic speech recognition (ASR). Early attempts at ASR consisted of making deterministic models of whole words in a small vocabulary (say 100 words) and recognizing a given speech utterance as the word whose model comes closest to it. The introduction of hidden Markov models (HMMs) in the early 1980s provided a much more powerful tool for speech recognition [1.12–14]. Today many products have been developed that successfully utilize ASR for communication between humans and machines. And the recognition can be done for continuous speech using a large vocabulary, and in a speaker-independent manner. Performance of these devices, however, deteriorates in the presence of reverberation and even low levels of ambient noise. Robustness to noise, reverberation, and characteristics of the transducer, is still an unsolved problem.

The goal of ASR is to recognize speech accurately regardless of who the speaker is. The complementary problem is that of recognizing a speaker from his/her voice, regardless of what words he/she is speaking. At present this problem appears to be solvable only if the speaker is one of a small set of N known speakers. A variant of the problem is speaker *verification*, in which the aim is to automatically verify the claimed identity of a speaker. While speaker recognition requires the selection of one out of N possible outcomes, speaker verification requires just a yes/no answer. This problem can be solved with a high degree of accuracy for much larger populations. Speaker verification has application wherever access to data or facilities has to be controlled. Forensics is another area of application. The problem of reduced performance in the presence of noise, as mentioned above for ASR, applies also to speaker recognition and speaker verification.

A third application of speech processing is that of synthesizing speech corresponding to a given text. When used together with ASR, speech synthesis allows a complete two-way spoken interaction between humans and machines. Speech synthesis is also a way to communicate for persons unable to speak. Its use for this purpose by the famous physicist Stephen Hawking is well known.

Early attempts at speech synthesis consisted of deriving the time-varying spectrum for the sequence of phonemes of a given text sentence. From this the corresponding time variation of the vocal tract was estimated, and the speech was synthesized by exciting the time-varying vocal tract with periodic or noise-like excitation as appropriate. The quality of the synthesis was significantly improved by concatenating pre-stored units (i. e., short segments such as diphones, triphones) after modifying them to fit the context. Today the highest-quality speech is synthesized by the unit *selection* method in which the units are selected from a large amount of stored speech and concatenated with little or no modification.

Finally we might mention the application of speech processing to aids for the handicapped. Hearing aid technology has made considerable progress in the last two decades. Part of this progress is due to a slow but steady improvement in our knowledge of the human hearing mechanism. A large part is due to the availability of high-speed digital hardware. At present performance of hearing aids is still poor under noisy and reverberant conditions.

A potentially useful application of speech processing to aid the handicapped is to display the shape of one's vocal tract as one speaks. By trying to match one's vocal tract shape to a displayed shape, a deaf person can learn correct pronunciation. Some attempts to implement this idea have been made, but have still been only in the realm of research.

Another useful application is a reading aid for the blind. The idea is to have a device to scan printed text from a book, and synthesize speech from the

scanned text. Coupled with a device to change speaking rate, this forms a useful aid for the blind. Several products offering this application are available on the market.

Many other application examples are described in the various parts of this handbook. We invite the reader to browse this volume on speech processing to find topics relevant to his/her specific interests.

1.3 Organization of the Handbook

This handbook on speech processing is a comprehensive source of knowledge in speech technology and its applications. It is organized as follows. This volume is divided into nine parts. For each part we invited at least one associate editor (AE) to handle it. All the AEs are very well-known researchers in their respective area of research. Part A (AE: M. M. Sondhi) contains four chapters on production, perception, and modeling of speech signals. Part B (AEs: Y. Huang and J. Benesty) concerns signal processing tools for speech, in eight chapters. Part C (AE: B. Kleijn) covers five chapters on speech coding. In part D (AE: S. Narayanan), the areas of text-to-speech synthesis are presented in seven chapters. Part E (AEs: L. Rabiner and B.-H. Juang), with 10 chapters, is a comprehensive overview on speech recognition. Part F (AE: S. Parthasarathy) contains three chapters on speaker recognition. Part G (AE: C.-H. Lee) is about language identification and contains four chapters. In part H (AEs: J. Chen, S. Gannot, and J. Benesty), various aspects of speech enhancement are developed in seven chapters. Finally the last section, part I (AEs: J. Benesty, I. Cohen, and Y. Huang), presents the important aspects of multichannel speech processing in four chapters.

References

1.1 H. Dudley, T.H. Tarnoczy: The speaking machine of Wolfgang von Kempelen, J. Acoust. Soc. Am. **22**, 151–166 (1950)

1.2 G. Fant: *Acoustic Theory of Speech Production* (Mouton, 's-Gravenhage 1960)

1.3 J.L. Flanagan: *Speech Analysis, Synthesis and Perception* (Springer, New York 1972)

1.4 H. Dudley: The carrier nature of speech, Bell Syst. Tech. J. **19**(4), 495–515 (1940)

1.5 L.R. Rabiner, R.W. Schafer: *Digital Processing of Speech Signals* (Prentice Hall, Englewood Cliffs 1978)

1.6 S. Furui, M.M. Sondhi (Eds.): *Advances in Speech Signal Processing* (Marcel Dekker, New York 1992)

1.7 B. Gold, N. Morgan: *Speech and Audio Signal Processing* (Wiley, New York 2000)

1.8 P. Elias: Predictive coding I, IRE Trans. Inform. Theory **1**(1), 16–24 (1955)

1.9 P. Elias: Predictive coding II, IRE Trans. Inform. Theory **1**(1), 24–33 (1955)

1.10 B.S. Atal, M.R. Schroeder: Adaptive predictive coding of speech, Bell Syst. Tech. J. **49**(8), 1973–1986 (1970)

1.11 B.S. Atal: The history of linear prediction, IEEE Signal Proc. Mag. **23**(2), 154–161 (2006)

1.12 L.R. Rabiner: A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE **77**(2), 257–286 (1989)

1.13 L.R. Rabiner, B.-H. Juang: *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliff 1993)

1.14 F. Jelinek: *Statistical Methods for Speech Recognition* (MIT, Boston 1998)