Foundations and Trends® in
Signal Processing
1:1-2 (2007)

Introduction to Digital Speech Processing

Lawrence R. Rabiner and Ronald W. Schafer

now
the essence of knowledge

Introduction to Digital Speech Processing

Lawrence R. Rabiner

Rutgers University and University of California Santa Barbara
USA

rabiner@ece.ucsb.edu

Ronald W. Schafer

Hewlett-Packard Laboratories
Palo Alto, CA
USA

E2008001330

now

the essence of knowledge

Boston – Delft

Foundations and Trends® in Signal Processing

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is L. R. Rabiner and R. W. Schafer, Introduction to Digital Speech Processing, Foundations and Trends® in Signal Processing, vol 1, no 1–2, pp 1–194, 2007

ISBN: 978-1-60198-070-0

Introduction to Digital Speech Processing

Foundations and Trends® in Signal Processing
Volume 1 Issue 1–2, 2007

Editorial Board

Editorial Scope

Foundations and Trends® in Signal Processing will publish survey and tutorial articles on the foundations, algorithms, methods, and applications of signal processing including the following topics:

- Adaptive signal processing
- Audio signal processing
- Biological and biomedical signal processing
- Complexity in signal processing
- Digital and multirate signal processing
- Distributed and network signal processing
- Image and video processing
- Linear and nonlinear filtering
- Multidimensional signal processing
- Multimodal signal processing
- Multiresolution signal processing
- Nonlinear signal processing
- Randomized algorithms in signal processing
- Sensor and multiple source signal processing, source separation
- Signal decompositions, subband and transform methods, sparse representations
- Signal processing for communications
- Signal processing for security and forensic analysis, biometric signal processing
- Signal quantization, sampling, analog-to-digital conversion, coding and compression
- Signal reconstruction, digital-to-analog conversion, enhancement, decoding and inverse problems
- Speech/audio/image/video compression
- Speech and spoken language processing
- Statistical/machine learning
- Statistical signal processing
- Classification and detection
- Estimation and regression
- Tree-structured methods

Information for Librarians
Foundations and Trends® in Signal Processing, 2007, Volume 1, 4 issues. ISSN paper version 1932-8346. ISSN online version 1932-8354. Also available as a combined paper and online subscription.

Foundations and Trends® in
Signal Processing
Vol. 1, Nos. 1–2 (2007) 1–194

DOI: 10.1561/2000000001

now
the essence of knowledge

Introduction to Digital Speech Processing

Lawrence R. Rabiner[1] and Ronald W. Schafer[2]

[1] *Rutgers University and University of California, Santa Barbara, USA, rabiner@ece.ucsb.edu*
[2] *Hewlett-Packard Laboratories, Palo Alto, CA, USA*

Abstract

Since even before the time of Alexander Graham Bell's revolutionary invention, engineers and scientists have studied the phenomenon of speech communication with an eye on creating more efficient and effective systems of human-to-human and human-to-machine communication. Starting in the 1960s, digital signal processing (DSP), assumed a central role in speech studies, and today DSP is the key to realizing the fruits of the knowledge that has been gained through decades of research. Concomitant advances in integrated circuit technology and computer architecture have aligned to create a technological environment with virtually limitless opportunities for innovation in speech communication applications. In this text, we highlight the central role of DSP techniques in modern speech communication research and applications. We present a comprehensive overview of digital speech processing that ranges from the basic nature of the speech signal, through a variety of methods of representing speech in digital form, to applications in voice communication and automatic synthesis and recognition of speech. The breadth of this subject does not allow us to discuss any

aspect of speech processing to great depth; hence our goal is to provide a useful introduction to the wide range of important concepts that comprise the field of digital speech processing. A more comprehensive treatment will appear in the forthcoming book, *Theory and Application of Digital Speech Processing* [101].

Contents

1

Introduction

The fundamental purpose of speech is communication, i.e., the transmission of messages. According to Shannon's information theory [116], a message represented as a sequence of discrete symbols can be quantified by its *information content* in bits, and the rate of transmission of information is measured in bits/second (bps). In speech production, as well as in many human-engineered electronic communication systems, the information to be transmitted is encoded in the form of a continuously varying (analog) waveform that can be transmitted, recorded, manipulated, and ultimately decoded by a human listener. In the case of speech, the fundamental analog form of the message is an acoustic waveform, which we call the *speech signal.* Speech signals, as illustrated in Figure 1.1, can be converted to an electrical waveform by a microphone, further manipulated by both analog and digital signal processing, and then converted back to acoustic form by a loudspeaker, a telephone handset or headphone, as desired. This form of speech processing is, of course, the basis for Bell's telephone invention as well as today's multitude of devices for recording, transmitting, and manipulating speech and audio signals. Although Bell made his invention without knowing the fundamentals of information theory, these ideas

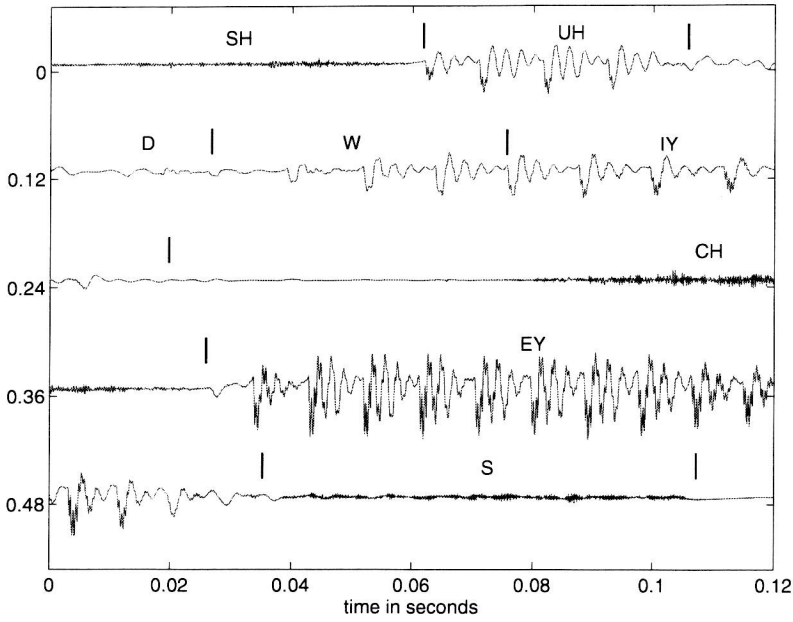

Fig. 1.1 A speech waveform with phonetic labels for the text message "Should we chase."

have assumed great importance in the design of sophisticated modern communications systems. Therefore, even though our main focus will be mostly on the speech waveform and its representation in the form of parametric models, it is nevertheless useful to begin with a discussion of how information is encoded in the speech waveform.

1.1 The Speech Chain

Figure 1.2 shows the complete process of producing and perceiving speech from the formulation of a message in the brain of a talker, to the creation of the speech signal, and finally to the understanding of the message by a listener. In their classic introduction to speech science, Denes and Pinson aptly referred to this process as the "speech chain" [29]. The process starts in the upper left as a message represented somehow in the brain of the speaker. The message information can be thought of as having a number of different representations during the process of speech production (the upper path in Figure 1.2).

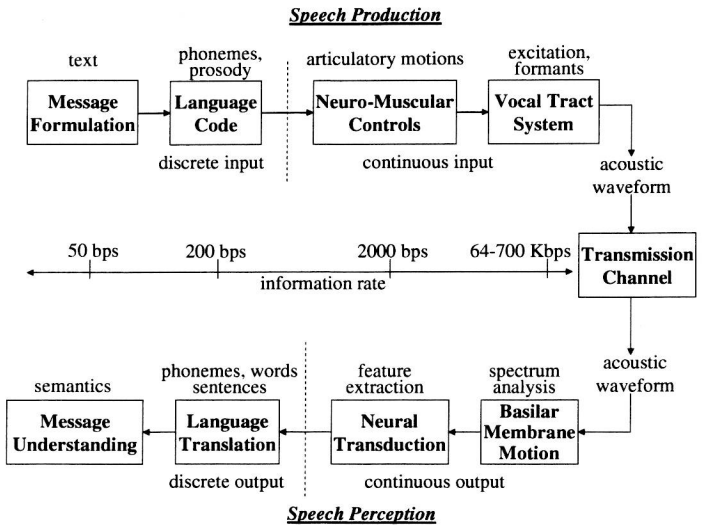

Fig. 1.2 The Speech Chain: from message, to speech signal, to understanding.

For example the message could be represented initially as English text. In order to "speak" the message, the talker implicitly converts the text into a symbolic representation of the sequence of sounds corresponding to the spoken version of the text. This step, called the language code generator in Figure 1.2, converts text symbols to phonetic symbols (along with stress and durational information) that describe the basic sounds of a spoken version of the message and the manner (i.e., the speed and emphasis) in which the sounds are intended to be produced. As an example, the segments of the waveform of Figure 1.1 are labeled with phonetic symbols using a computer-keyboard-friendly code called ARPAbet.[1] Thus, the text "should we chase" is represented phonetically (in ARPAbet symbols) as [SH UH D — W IY — CH EY S]. (See Chapter 2 for more discussion of phonetic transcription.) The third step in the speech production process is the conversion to "neuro-muscular controls," i.e., the set of control signals that direct the neuro-muscular system to move the speech articulators, namely the tongue, lips, teeth,

[1] The International Phonetic Association (IPA) provides a set of rules for phonetic transcription using an equivalent set of specialized symbols. The ARPAbet code does not require special fonts and is thus more convenient for computer applications.

jaw and velum, in a manner that is consistent with the sounds of the desired spoken message and with the desired degree of emphasis. The end result of the neuro-muscular controls step is a set of articulatory motions (continuous control) that cause the vocal tract articulators to move in a prescribed manner in order to create the desired sounds. Finally the last step in the Speech Production process is the "vocal tract system" that physically creates the necessary sound sources and the appropriate vocal tract shapes over time so as to create an acoustic waveform, such as the one shown in Figure 1.1, that encodes the information in the desired message into the speech signal.

To determine the rate of information flow during speech production, assume that there are about 32 symbols (letters) in the language (in English there are 26 letters, but if we include simple punctuation we get a count closer to $32 = 2^5$ symbols). Furthermore, the rate of speaking for most people is about 10 symbols per second (somewhat on the high side, but still acceptable for a rough information rate estimate). Hence, assuming independent letters as a simple approximation, we estimate the base information rate of the text message as about 50 bps (5 bits per symbol times 10 symbols per second). At the second stage of the process, where the text representation is converted into phonemes and prosody (e.g., pitch and stress) markers, the information rate is estimated to increase by a factor of 4 to about 200 bps. For example, the ARBAbet phonetic symbol set used to label the speech sounds in Figure 1.1 contains approximately $64 = 2^6$ symbols, or about 6 bits/phoneme (again a rough approximation assuming independence of phonemes). In Figure 1.1, there are 8 phonemes in approximately 600 ms. This leads to an estimate of $8 \times 6/0.6 = 80$ bps. Additional information required to describe prosodic features of the signal (e.g., duration, pitch, loudness) could easily add 100 bps to the total information rate for a message encoded as a speech signal.

The information representations for the first two stages in the speech chain are discrete so we can readily estimate the rate of information flow with some simple assumptions. For the next stage in the speech production part of the speech chain, the representation becomes continuous (in the form of control signals for articulatory motion). If they could be measured, we could estimate the spectral bandwidth of these

control signals and appropriately sample and quantize these signals to obtain equivalent digital signals for which the data rate could be estimated. The articulators move relatively slowly compared to the time variation of the resulting acoustic waveform. Estimates of bandwidth and required accuracy suggest that the total data rate of the sampled articulatory control signals is about 2000 bps [34]. Thus, the original text message is represented by a set of continuously varying signals whose digital representation requires a much higher data rate than the information rate that we estimated for transmission of the message as a speech signal.[2] Finally, as we will see later, the data rate of the digitized speech waveform at the end of the speech production part of the speech chain can be anywhere from 64,000 to more than 700,000 bps. We arrive at such numbers by examining the sampling rate and quantization required to represent the speech signal with a desired perceptual fidelity. For example, "telephone quality" requires that a bandwidth of 0–4 kHz be preserved, implying a sampling rate of 8000 samples/s. Each sample can be quantized with 8 bits on a log scale, resulting in a bit rate of 64,000 bps. This representation is highly intelligible (i.e., humans can readily extract the message from it) but to most listeners, it will sound different from the original speech signal uttered by the talker. On the other hand, the speech waveform can be represented with "CD quality" using a sampling rate of 44,100 samples/s with 16 bit samples, or a data rate of 705,600 bps. In this case, the reproduced acoustic signal will be virtually indistinguishable from the original speech signal.

As we move from text to speech waveform through the speech chain, the result is an encoding of the message that can be effectively transmitted by acoustic wave propagation and robustly decoded by the hearing mechanism of a listener. The above analysis of data rates shows that as we move from text to sampled speech waveform, the data rate can increase by a factor of 10,000. Part of this extra information represents characteristics of the talker such as emotional state, speech mannerisms, accent, etc., but much of it is due to the inefficiency

[2] Note that we introduce the term data rate for digital representations to distinguish from the inherent information content of the message represented by the speech signal.

of simply sampling and finely quantizing analog signals. Thus, motivated by an awareness of the low intrinsic information rate of speech, a central theme of much of digital speech processing is to obtain a digital representation with lower data rate than that of the sampled waveform.

The complete speech chain consists of a speech production/generation model, of the type discussed above, as well as a speech perception/recognition model, as shown progressing to the left in the bottom half of Figure 1.2. The speech perception model shows the series of steps from capturing speech at the ear to understanding the message encoded in the speech signal. The first step is the effective conversion of the acoustic waveform to a spectral representation. This is done within the inner ear by the basilar membrane, which acts as a non-uniform spectrum analyzer by spatially separating the spectral components of the incoming speech signal and thereby analyzing them by what amounts to a non-uniform filter bank. The next step in the speech perception process is a neural transduction of the spectral features into a set of sound features (or distinctive features as they are referred to in the area of linguistics) that can be decoded and processed by the brain. The next step in the process is a conversion of the sound features into the set of phonemes, words, and sentences associated with the in-coming message by a language translation process in the human brain. Finally, the last step in the speech perception model is the conversion of the phonemes, words and sentences of the message into an understanding of the meaning of the basic message in order to be able to respond to or take some appropriate action. Our fundamental understanding of the processes in most of the speech perception modules in Figure 1.2 is rudimentary at best, but it is generally agreed that some physical correlate of each of the steps in the speech perception model occur within the human brain, and thus the entire model is useful for thinking about the processes that occur.

There is one additional process shown in the diagram of the complete speech chain in Figure 1.2 that we have not discussed — namely the transmission channel between the speech generation and speech perception parts of the model. In its simplest embodiment, this transmission channel consists of just the acoustic wave connection between