

Jacob Benesty

## About the Editors:

### Jacob Benesty

Jacob Benesty received the Masters degree from Pierre & Marie Curie University, France, in 1987, and the Ph.D. degree in control and signal processing from Orsay University, France, in 1991. After positions at Telecom Paris University as a consultant and a member of the technical staff at Bell Laboratories, Murray Hill, Dr. Benesty joined the University of Quebec (INRS-EMT) in Montreal, Canada in May 2003 as a professor. His research interests are in signal processing, acoustic signal processing, and multimedia communications.

### M. M. Sondhi

M. Mohan Sondhi is a consultant at Avaya Research Labs, Basking Ridge, New Jersey. Prior to joining Avaya he spent 39 years at Bell Labs, from where he retired in 2001. He holds undergraduate degrees in Physics and Elec-

trical Communication Engineering, and M.S. and Ph.D. degrees in Electrical Engineering. At Bell Labs he conducted research in speech signal processing, echo cancellation, acoustical inverse problems, speech recognition, articulatory models for analysis and synthesis of speech and modeling of auditory and visual processing by humans.

### Y. Huang

Yiteng (Arden) Huang received the B.S. degree from the Tsinghua University in 1994, the M.S. and Ph.D. degrees from the Georgia Institute of Technology (Georgia Tech) in 1998 and 2001, respectively, all in electrical and computer engineering. Upon graduation, he joined Bell Laboratories as a member of technical staff in March 2001. His current research interests are in acoustic signal processing and multimedia communications.

## Authors:

A. Acero, Redmond, WA, USA

Jont Allen, Urbana, IL, USA

Jacob Benesty, Montreal, QC, Canada

R. Bimbot, Rennes, France

Nick Campbell, Kyoto, Japan

William Campbell, Lexington, MA, USA

Rolf Carlson, Stockholm, Sweden

Jingdong Chen, Murray Hill, NJ, USA

Juin-Hwey Chen, Irvine, CA, USA

Israel Cohen, Haifa, Israel

Jordan Cohen, Woburn, MA, USA

Corinna Cortes, New York, NY, USA

Eric Diethorn, Basking Ridge, NJ, USA

T. Dutoit, Mons, Belgium

Simon Doclo, Leuven-Heverlee, Belgium

Jasha Droppo, Redmond, WA, USA

Gary Elko, Summit, NJ, USA

Sadaoki Furui, Tokyo, Japan

Sharon Gannot, Ramat-Gan, Israel

Mazin Gilbert, Florham Park, NJ, USA

Michael Goodwin, Scotts Valley, CA, USA

B. Granstrom, Stockholm, Sweden

Volodya Grancharov, Stockholm, Sweden

Patrick Haffner, Florham Park, NJ, USA

Roar Hagen, Stockholm, Sweden

Mary Harper, College Park, MD, USA

Larry Heck, Sunnyvale, CA, USA

Jürgen Herre, Erlangen, Germany

Wolfgang Hess, Bonn, Germany

Kiyoshi Honda, Kyoto, Japan

Yiteng Huang, Murray Hill, NJ, USA

Juang, B.-H. (Fred), Atlanta, GA, USA

Tatsuya Kawahara, Kyoto, Japan

Esther Klabbbers, Beaverton, OR, USA

Bastiaan Kleijn, Stockholm, Sweden

Birger Kollmeier, Oldenburg, Germany

Sen Kuo, DeKalb, IL, USA

Chin-Hui Lee, Atlanta, GA, USA

Haizhou Li, Singapore, Singapore

Jan Linden, San Francisco, CA, USA

Bin Ma, Singapore, Singapore

Michael Maxwell, College Park, MD, USA

Alan McCree, Lexington, MA, USA

Mehryer Mohri, New York, NY, USA

Marc Moonen, Leuven-Heverlee, Belgium

Dennis Morgan, Murray Hill, NJ, USA

David Nahamoo, Yorktown Heights, NY, USA

Shri Narayanan, Los Angeles, CA, USA

Douglas O'Shaughnessy, Montreal, QC, Canada

Lucas Parra, New York, NY, USA

S. Parthasarathy, Florham Park, NJ, USA

Fernando Pereira, Philadelphia, PA, USA

Michael Picheny, Yorktown Heights, NY, USA

USA

Rudolf Rabenstein, Erlangen, Germany

Lawrence Rabiner, Piscataway, NJ, USA

Douglas Reynolds, Lexington, MA, USA

Mike Riley, New York, NY, USA

Aaron Rosenberg, Piscataway, NJ, USA

Salim Roukos, Yorktown Heights, NY, USA

Ronald Schafer, Palo Alto, CA, USA

Juergen Schroeter, Florham Park, NJ, USA

Stephanie Seneff, Cambridge, MA, USA

Wade Shen, Lexington, MA, USA

Elliot Singer, Lexington, MA, USA

Jan Skoglund, San Francisco, CA, USA

Mohan Sondhi, Basking Ridge, NJ, USA

Sascha Spors, Berlin, Germany

Ann Spriet, Leuven-Heverlee, Belgium

Richard Sproat, Urbana, IL, USA

Yannis Stylianou, Heraklion, Greece

Jes Thyssen, Irvine, CA, USA

Jan Van Santen, Beaverton, OR, USA

Jay Wilpon, Florham Park, NJ, USA

Jan Wouters, Leuven, Belgium

Arie Yeredor, Tel-Aviv, Israel

Steve Young, Cambridge, USA

Victor Zue, Cambridge, MA, USA

# Springer eBook Collection

Introducing The World's Most Comprehensive Online Scientific Book Collection.

Springer, the world's largest international publisher of scientific books introduces the world's most comprehensive digitized scientific, technical and medical book collection. The Springer eBook collection offers the first online book collection especially made for the requirements of researchers and scientists. The collection includes online access to more than 10,000 books adding 3,000 new research book titles each year.

For more information visit [springer.com/ebooks](http://springer.com/ebooks)

recommend to  
YOUR LIBRARY

012504x

## Order Now!

Springer Handbook of Speech Processing

Yes, please send me \_\_\_\_\_ copies

- Print \_\_\_\_\_ ISBN 978-3-540-49125-5 ▶ € 249,00 | £191.50 Prepublication price, valid until February 29, 2008 ▶ € 199,95 | £154.00  
 eReference \_\_\_\_\_ ISBN 978-3-540-49127-9 ▶ € 249,00 | £191.50 Prepublication price, valid until February 29, 2008 ▶ € 199,95 | £154.00  
 Print + eReference \_\_\_\_\_ ISBN 978-3-540-49128-6 ▶ € 311,00 | £239.00 Prepublication price, valid until February 29, 2008 ▶ € 250,00 | £192.50

- Please bill me \_\_\_\_\_  
 Please charge my credit card:  Eurocard/Access/Mastercard  Visa/Barclaycard/Bank/Americard  AmericanExpress

Number                 Valid until

Available from	Name
	Dept.
	Institution
	Street
	City / ZIP-Code
	Country
	Email
	Date ✕
	Signature ✕

Springer Distribution Center GmbH, Haberstrasse 7, 69126 Heidelberg, Germany

▶ Call: + 49 (0) 6221-345-4301 ▶ Fax: +49 (0) 6221-345-4229

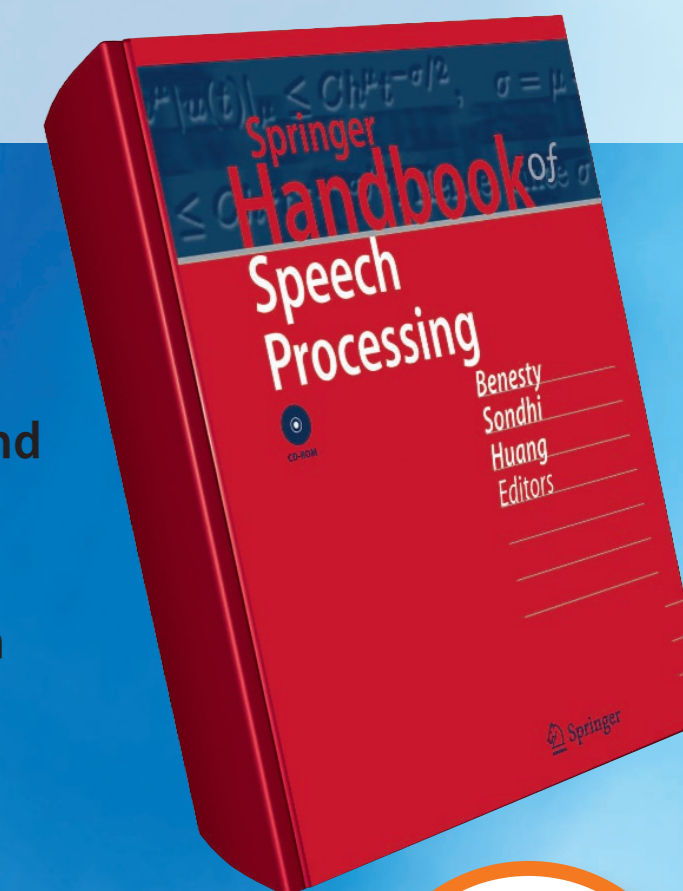
▶ Email: [SDC-bookorder@springer.com](mailto:SDC-bookorder@springer.com) ▶ Web: [springer.com](http://springer.com)

All € and £ prices are net prices subject to local VAT, e.g. in Germany 7% VAT for books and 19% VAT for electronic products. Pre-publication pricing: Unless otherwise stated, pre-pub prices are valid through the end of the third month following publication, and therefore are subject to change. All prices exclusive of carriage charges. Prices and other details are subject to change without notice. All errors and omissions excepted. M0254

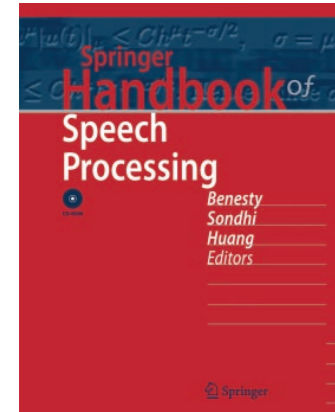
# Springer Handbook of Speech Processing

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

- ▶ Authoritative desk reference
- ▶ Quick access to applicable, reliable, and comprehensive knowledge
- ▶ Electronic contents on accompanying DVD



WITH  
DVD



# Springer Handbook of Speech Processing

**J. Benesty**, Université de Québec, Montréal, QC, Canada; **M. M. Sondhi**, Avayalabs Research, Basking Ridge, NJ, USA; **Y. Huang**, Bell Labs, Murray Hill, NJ, USA (Eds.)

these technologies, this work combines the established knowledge derived from research in such fast evolving disciplines as Signal Processing and Communications, Acoustics, Computer Science and Linguistics.

- ▶ Authoritative desk reference of one of tomorrow's breakthrough technologies
- ▶ Provides quick access to applicable, reliable, and comprehensive knowledge
- ▶ Electronic contents on accompanying DVD
- ▶ Handy format, lucid two-color layout, uniformly styled figures
- ▶ For PC. For the complete system requirements see: [springer.com](http://springer.com)

This handbook is designed to play a fundamental role in sustainable progress in speech research and development. With an accessible format and with accompanying DVD-RoM, it targets three categories of readers: graduate students, professors and active researchers in academia, and engineers in industry who need to understand or implement some specific algorithms for their speech-related products. It is a superb source of application-oriented, authoritative and comprehensive information about

### Key topics

- ▶ Introduction to Speech Processing.
- ▶ Production, Perception, and Modeling of Speech.
- ▶ Signal Processing for Speech.
- ▶ Speech Coding.
- ▶ Text-to-Speech Synthesis.
- ▶ Speech Recognition.
- ▶ Speaker Recognition.
- ▶ Language Recognition.
- ▶ Speech Enhancement.
- ▶ Multichannel Speech Processing

SPRINGER REFERENCE

Prepublication price, valid until February 29, 2008

▶ € 199,95 | £154.00

With DVD

sample chapter online at [springer.com](http://springer.com)

### Print

2008. Approx. 1500 p. 560 illus.

With DVD. Hardcover

ISBN 978-3-540-49125-5

▶ € 249,00 | £191.50

Prepublication price, valid until

February 29, 2008

▶ € 199,95 | £154.00

### eReference

2008. eReference.

ISBN 978-3-540-49127-9

▶ € 249,00 | £191.50

Prepublication price, valid until

February 29, 2008

▶ € 199,95 | £154.00

### Print +eReference

2008. Approx. 1500 p.

With DVD. Print + eReference.

ISBN 978-3-540-49128-6

▶ € 311,00 | £239.00

Prepublication price, valid until

February 29, 2008

▶ € 250,00 | £192.50

## Perception of

### 4. Perception of Speech and Sound

The transformation of acoustical signals into auditory sensations can be characterized by psychophysical quantities like loudness, tonality, or perceived pitch. The resolution limits of the auditory system produce spectral and temporal masking phenomena and impose constraints on the perception of amplitude modulations. Binaural hearing (i. e., utilizing the acoustical difference across both ears) employs interaural time and intensity differences to produce localization and binaural unmasking phenomena such as the binaural intelligibility level difference, i. e., the speech reception threshold difference between listening to speech in noise monaurally versus listening with both ears.

The acoustical information available to the listener for perceiving speech even under adverse conditions can be characterized using the Articulation Index, the Speech Transmission Index, and the Speech Intelligibility Index. They can objectively predict speech reception thresholds as a function of spectral content, signal-to-noise ratio and preservation of amplitude modulations in the speech waveform that enter the listener's ear. The articulatory or phonetic information available to and received by the listener can be characterized by speech feature sets. Transformation analysis allows to detect the relative transmission error connected with each of these speech features. The comparison across man and machine

4.1 Basic Psychoacoustic Quantities	2
4.1.1 Mapping of Intensity into Loudness	2
4.1.2 Pitch	4
4.1.3 Perception	5
4.1.4 Binaural Hearing	7
4.1.5 Binaural Noise Suppression	8
4.2 Acoustical Information Required for Speech Perception	10
4.2.1 Speech Intelligibility and Speech Reception Threshold (SRT)	10
4.2.2 Measurement Methods	11
4.2.3 Factors Influencing Speech Intelligibility	12
4.2.4 Prediction Methods	12
4.3 Speech Feature Perception	14
4.3.1 Formant Features	15
4.3.2 Phonetic and Distinctive Feature Sets	15
4.3.3 Internal Representation Approach and Higher-Order Temporal-Spectral Features	16
4.3.4 Man-Machine Comparison	20
References	21

in speech recognition allows to test hypotheses and models of human speech perception. Conversely, automatic speech recognition may be improved by introducing human signal processing principles into machine processing algorithms.

Acoustically produced speech is a very special sound to our ears and our brain. Humans are able to extract the information contained in a spoken message extremely efficiently even if the speech energy is lower than any competing background sound. Hence, humans are able to communicate acoustically even under adverse listening conditions, e.g., in a cafeteria. The process of understanding speech can be subdivided into two stages. First, an auditory pre-processing stage where the speech sound is transformed into its internal representation in the brain and special speech features are extracted (such

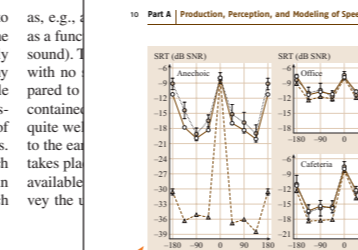


Fig. 4.11 SRT data (filled symbols) and predictions for three different acoustic conditions and normal listeners. The convex dome model predictions without introducing appropriate processing errors, whereas the open symbols denote predictions employing internal processing errors, that have been taken from average values in other psychophysical tasks (after 4.4.1)

Note that in each frequency band only one EC circuit is employed in the model. This reflects the empirical evidence that the brain is only able to cancel out one direction for each frequency band at each instant of time. Hence, the processing strategy adopted will use appropriate compromises for any given real situation.

#### 4.2 Acoustical Information Required for Speech Perception

4.2.1 Speech Intelligibility and Speech Reception Threshold (SRT) Speech intelligibility (SI) is important for various fields of research, engineering, and diagnosis for quantifying very different phenomena such as the quality of recordings, communication and playback devices, the rehabilitation of audiology, characteristics of hearing impairment, benefit using hearing aids, or combinations of these topics. The most useful way to define SI is: speech intelligibility, SI is the proportion of speech items (i.e., syllables, words, or sentences) correctly repeated by (a) listener(s) for a given speech intelligibility test. This operative definition makes SI directly and quantitatively measurable.

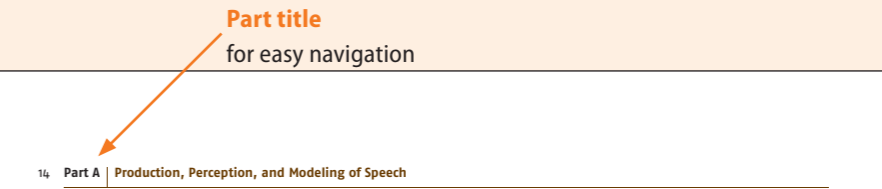
$$SI(L) = \frac{1}{2} \left( 1 + \frac{A-1}{1 + \exp\left(-\frac{L-L_{50}}{s}\right)} \right) \quad (4.4)$$

with  $L_{50}$ : speech level of the midpoint of the intelligibility function;  $s$ : slope parameter, the slope at  $L_{50}$  is

as, e.g., as a function of the sound pressure level (measured in dB) of the speech signal or the speech-to-noise ratio (SNR) (measured in dB), if the test is performed with interfering noise.

Easy to read and use: includes more than 500 diagrams and illustrations drawn to scale

One-column section headings



### 4.3 Speech Feature Perception

The information-theoretic approach to describing speech perception assumes that human speech recognition is based on the combined, parallel recognition of several acoustical cues that are characteristic for certain speech elements. While a *phoneme* represents the smallest unit of speech information, its acoustic realization (denoted as *phone*) can be quite variable in its acoustical properties. Such a phone is produced in order to deliver a number of acoustical speech cues to the listener who should be able to deduce from it the underlying phoneme. Each speech cue represents one feature value of more- or less-complex speech features like *voicing*, *fricative*, or *duration*, that are linked to phonetics and to perception. These speech feature values are decoded by the listener independently of each other and are used for recognizing the underlying speech element (such as, e.g., the represented phoneme). Speech perception can therefore be interpreted as reception of certain values of several speech features in parallel and in discrete time steps.

Each phoneme is characterized by a unique combination of the underlying speech feature values. The articulation of words and sentences produces (in the sense of information theory) a discrete stream of information via a number of simultaneously active channels (Fig. 4.13).

The spoken realization of a given phoneme causes a certain speech feature to assume one out of several different possible values. For example, the speech feature *voicing* can assume the value one (i. e., voiced sound) or the value zero (unvoiced speech sound). Each of these features is transmitted via its own, specific transmission channel to the speech recognition system of the listener.

The *channel* consists of the acoustical transmission channel to the listener's ear and the subsequent decoding of the signal in the central auditory system of the receiver (which can be hampered by a hearing impairment or a speech pathology). The listener recognizes the actually assumed values of certain speech features and combines these features to yield the recognized phoneme.

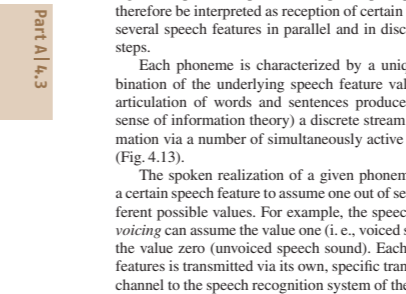
If  $p(i)$  gives the probability (or relative frequency) that a specific speech feature assumes the value  $i$  and  $p'(j)$  gives the probability (or relative frequency, respectively) that the receiver receives the feature value  $j$ , and  $p(i, j)$  gives the joint probability that the value  $j$  is recognized if the value  $i$  is transmitted, then the so-called transformation  $T$  is defined as

$$T = - \sum_{i=1}^N \sum_{j=1}^N p(i, j) \log_2 \left( \frac{p(i)p'(j)}{p(i, j)} \right) \quad (4.10)$$

The transformation  $T$  assumes its maximal value for perfect transmission of the input values to the output values, i. e., if  $p(i, j)$  takes the diagonal form or any permutation thereof.  $T$  equals 0 if the distribution of received feature values is independent of the distribution of input feature values, i. e., if  $p(i, j) = p(i)p'(j)$ . The maximum value of  $T$  for perfect transmission (i. e.,  $p(i, j) = p'(j)$ ) equals the amount of information (in bits) included in the distribution of input feature values  $H$ , i. e.,

$$H = - \sum_{i=1}^N p(i) \log_2 p(i) = - \sum_{i=1}^N p(i) \log_2 [p(i)] \quad (4.11)$$

In order to normalize  $T$  to give values between 0 and 1, the so-called transformation index (TI) is often used.



### 4.3.3 Internal Representation Approach and Higher-Order Temporal-Spectral Features

Such an internal representation model puts most of the peculiarities and limitations of the speech recognition process into the nonlinear, destructive transformation process from the acoustical speech waveform into its internal representation, assuming that all transformation steps are due to physiological processes that can be characterized completely physiologically or by psychophysical means (Fig. 4.14 for illustration).

Several concepts and models to describe such an internal representation have been developed so far. Some of the basic ideas are as follows:

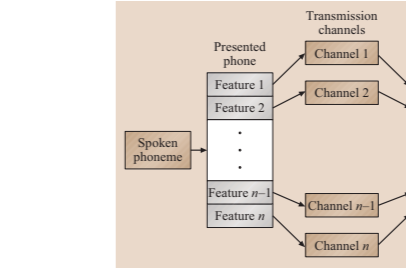
1. **Auditory spectrogram:** The basic internal representation assumes that the speech sound is separated into a number of frequency bands distributed unevenly across a psychoacoustically-based frequency scale like the Bark- or ERB-scale and that the compressed frequency-channel-specific intensity is represented

noise can be estimated quite well from psychoacoustical experiments.

Such an internal representation model puts most of the peculiarities and limitations of the speech recognition process into the nonlinear, destructive transformation process from the acoustical speech waveform into its internal representation, assuming that all transformation steps are due to physiological processes that can be characterized completely physiologically or by psychophysical means (Fig. 4.14 for illustration).

Several concepts and models to describe such an internal representation have been developed so far. Some of the basic ideas are as follows:

1. **Auditory spectrogram:** The basic internal representation assumes that the speech sound is separated into a number of frequency bands distributed unevenly across a psychoacoustically-based frequency scale like the Bark- or ERB-scale and that the compressed frequency-channel-specific intensity is represented



Chapter title and section heading

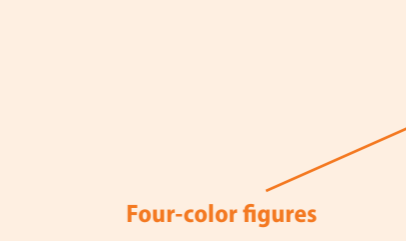


Fig. 4.16a-c Auditory spectrogram representation of the German word 'Stall'. It can be represented by a spectrogram, a bark spectrogram on a log-loudness scale(b) or as a contrast-enhanced version using nonlinear feedback loops (after 4.4.14)(c)

Four-color figures

Clearly displayed math

Chapter and section title for easy navigation

Thumb indices identify the part and chapter section