ADVANCES IN

SPEECH, HEARING AND LANGUAGE PROCESSING

Editor: W. A. AINSWORTH

Volume 1 • 1990

ADVANCES IN SPEECH, HEARING AND LANGUAGE PROCESSING

A Research Annual

Editor: W.A. AINSWORTH
Department of Communication and Neuroscience
University of Keele

VOLUME 1 • 1990

JAI PRESS LTD

London, England *Greenwich, Connecticut*

JAI PRESS LTD
118 Pentonville Road
London N1 9JN

JAI PRESS INC.
55 Old Post Road No. 2
Greenwich, Connecticut 06836

ISBN: 1-55938-210-4

British Library Cataloguing in Publication Data

Advances in speech, hearing and language processing.
Vol. 1, 1990.
1. Man. Speech
I. Ainsworth, W. A. (William Anthony) *1939–*
612.78

Printed at the Alden Press
Oxford London and Northampton

LIST OF CONTRIBUTORS

G.J. Brown	Department of Computer Science University of Sheffield, England
Rolf Carlson	Department of Speech Communication and Music Acoustics Royal Institute of Technology Stockholm, Sweden
G. Chollet	Signal Department École Nationale Supérieure des Télécommunications Paris, France
K. Choukri	CAP SESA Innovation Paris, France
M.P. Cooke	Department of Computer Science University of Sheffield, England
M.D. Crawford	Department of Computer Science University of Sheffield, England
R.I. Damper	Department of Electronics and Computer Science University of Southampton, England
Christopher J. Darwin	Department of Experimental Psychology University of Sussex Brighton, England
Randy L. Deihl	Department of Psychology University of Texas Austin, U.S.A.

Björn Granström
Department of Speech Communication and Music Acoustics
Royal Institute of Technology
Stockholm, Sweden

P.D. Green
Department of Computer Science
University of Sheffield, England

Sheri Hunnicut
Department of Speech Communication and Music Acoustics
Royal Institute of Technology
Stockholm, Sweden

Keith R. Kluender
Department of Psychology
University of Wisconsin
Madison, U.S.A.

K.K. Paliwal
Tata Institute of Fundamental Research
Bombay, India

A.J.H. Simons
Department of Computer Science
University of Sheffield, England

Marcel Tatham
Centre for Cognitive Science/Linguistics
University of Essex
Colchester, England

G.D. Tattersall
School of Information Systems
University of East Anglia
Norwich, England

J.P. Tubach
Signal Department
École Nationale Supérieure des Télécommunications
Paris, France

Margaret A. Walsh
Department of Psychology
University of Texas
Austin, U.S.A.

INTRODUCTION

The aim of the Series is to present a survey of recent research in the fields of speech, hearing and language processing as carried out in various leading international laboratories. In the past these fields have been rather isolated from each other, but currently they appear to be drawing closer together. It is hoped that this continuing Series will encourage this trend.

This first volume, however, is concerned mainly with speech processing. Subsequent volumes will attempt to explore links between speech, hearing and language processing.

The signal processing facilities which are now available are enabling complex automatic speech recognition schemes to be tested, and more realistic models of speech perception to be investigated. Chapter 1 surveys the modern speech signal processing techniques. For speech recognition to be carried out the speech signal must be transformed into a symbolic representation. The next three chapters describe various techniques for performing this process: by time-dependent modelling to take account of speech variability (Chapter 2), by means of neural networks (Chapter 3), or via an intermediate representation called the speech sketch (Chapter 4).

One reason why speech recognition is so difficult both to understand and to automate is that language processing is different from speech processing. Chapter 5 attempts to reconcile these differences by presenting a unified theory of phonetics and phonology.

Another reason why speech recognition is difficult is that speech signals are corrupted by the environment in which they are produced. In Chapter 6

experiments are described which attempt to elucidate some of the ways in which the human perceptual system copes with such distortions. In Chapter 7 it is shown that many features of speech perception can be explained in terms of general auditory processing.

Signal processing is involved in speech production as well as recognition. Complex processing enables realistic synthetic speech to be generated, particularly multi-language synthesis as described in Chapter 8. Finally the application of speech aids for the handicapped is discussed in Chapter 9.

August 1989
Keele, Staffs.

W.A. Ainsworth
Series Editor

ADVANCES IN SPEECH, HEARING AND LANGUAGE PROCESSING

Volume 1 • 1990

CONTENTS

LIST OF CONTRIBUTORS vii

INTRODUCTION
W.A. Ainsworth ix

SPEECH PROCESSING TECHNIQUES
K.K. Paliwal 1

EXPERIMENTS IN SPEECH ANALYSIS AND RECOGNITION: TACKLING THE VARIABILITY OF SPEECH
G. Chollet, K. Choukri and J.P. Tubach 79

NEURAL NETWORKS AND SPEECH PROCESSING
G.D. Tattersall 107

BRIDGING THE GAP BETWEEN SIGNALS AND SYMBOLS IN SPEECH RECOGNITION
P.D. Green, G.J. Brown, M.P. Cooke, M.D. Crawford and A.J.H. Simons 149

COGNITIVE PHONETICS
Marcel Tatham 193

ENVIRONMENTAL INFLUENCES ON SPEECH PERCEPTION
Christopher J. Darwin 219

SOME AUDITORY BASES OF SPEECH PERCEPTION AND PRODUCTION
Randy L. Diehl, Keith R. Kluender and Margaret A. Walsh 243

MULTILINGUAL TEXT-TO-SPEECH DEVELOPMENT AND APPLICATIONS
Rolf Carlson, Björn Granström and Sheri Hunnicutt 269

SPEECH AIDS FOR THE HANDICAPPED
R.I. Damper 297

INDEX 333

SPEECH PROCESSING TECHNIQUES

K.K. Paliwal

OUTLINE

	ABSTRACT	2
1.	INTRODUCTION	2
2.	SPEECH PRODUCTION PROCESS	3
3.	SPEECH ANALYSIS TECHNIQUES	6
	3.1 Short-Time Fourier Analysis Technique	7
	3.2 Cepstral Analysis Technique	10
	3.3 Linear Prediction Analysis Technique	12
4.	PITCH EXTRACTION TECHNIQUES	23
	4.1 Autocorrelation Technique	24
	4.2 Average Magnitude Difference Function Technique	27
	4.3 Cepstrum Technique	30
	4.4 Maximum Likelihood Technique	30
	4.5 Harmonic-Peak-Based Techniques	31
	4.6 Analysis-by-Synthesis Technique	34
	4.7 Time-Domain Pattern Recognition Techniques	34
	4.8 Tracking and Smoothing	36
	4.9 Evaluation of Pitch Extraction Techniques	37
5.	FORMANT EXTRACTION TECHNIQUES	38

Advances in Speech, Hearing and Language Processing
Volume 1, pages 1–78.

ISBN: 1-55938-210-4

6. VECTOR QUANTIZATION 40
7. DYNAMIC TIME WARPING 45
8. HIDDEN MARKOV MODELLING 49
9. NEURAL-NET MODELLING 54
10. MULTIPULSE AND STOCHASTIC MODELLING 60
11. SUMMARY 67
REFERENCES 68
BIBLIOGRAPHY 78

ABSTRACT

Recently, significant advances have taken place in the development of powerful and efficient speech processing techniques. These techniques have been used to realize a number of ambitious speech processing application systems. The aim of this chapter is to provide a brief overview of these speech processing techniques and show their usefulness in different speech processing applications. The speech processing techniques described here include the short-time Fourier analysis technique, the cepstral analysis technique, the linear prediction analysis technique, the pitch and the formant extraction techniques, the vector quantization technique, the dynamic time warping technique, the hidden Markov modelling technique, the neural-net modelling technique, and the multipulse and the stochastic modelling techniques.

1. INTRODUCTION

Speech processing has been an area of interest for the last four decades; but the last decade has witnessed significant progress in this area of research. This progress has been possible mainly due to the recent advances made in the development of powerful and efficient speech processing techniques such as vector quantization, hidden Markov modelling, neural-net modelling, etc. Because of these techniques, it is now possible to develop ambitious speech processing application systems such as the 800 bits/s linear prediction (LP) vocoder (Wong *et al.*, 1982), the 4.8 kbits/s stochastic excited LP coder (Atal, 1986), the speaker-independent large-vocabulary continuous speech recognition system (Lee and Hon, 1988), and so on.

The aim of the present chapter is to provide a brief overview of the speech processing techniques and to show how these techniques are used in different speech processing applications. Speech processing applications considered in this chapter are speech coding, speech synthesis, speech recognition, speaker recognition and speech enhancement. It is not possible here to cover all the

details about speech processing techniques that have been reported in the literature. Instead, only those aspects of the techniques with which the author is familiar enough to comment confidently upon are presented. Thus, this chapter is not an exhaustive survey of the literature on speech processing techniques, but to supplement it, a selected bibliography for further reading, in addition to the references, is included.

Most speech processing applications require parametric modelling of the speech signal during the analysis phase. In order to select a proper parametric model for the speech signal, it is necessary to know how speech is produced. Therefore, Section 2 describes the speech production process and represents this process by a source-system model. In this model, the speech signal is generated as the output of a time-varying linear system which is excited either by a periodic pulse train (for voiced speech) or by a white random number sequence (for unvoiced speech). Section 3 describes three speech analysis techniques: (1) the short-time Fourier analysis technique, (2) the cepstral analysis technique, and (3) the LP analysis technique. The short-time Fourier analysis technique computes the spectrum of the speech signal, while the cepstral and the LP analysis techniques decompose this spectrum into two parts, one corresponding to the excitation source and the other to the linear system. These two parts can be characterized nicely in terms of the pitch and the formant parameters. The techniques for pitch and formant extraction are described in Sections 4 and 5, respectively. Vector quantization is powerful data compression technique and is described in Section 6. The dynamic time warping and the hidden Markov modelling techniques can align the speech events of two utterances through nonlinear time normalization and are mainly used for speech recognition. These techniques are described in Sections 7 and 8, respectively. The neural-net modelling technique is described in Section 9. This is a powerful technique for pattern classification as it can provide (arbitrarily shaped) nonlinear decision surfaces in the multidimensional pattern space. Section 10 discusses some limitations of the excitation-source part of the source-system model and describes the multipulse and the stochastic models for characterizing the excitation source better. These models are useful for very low bit-rate speech coding. Finally, the chapter is summarized in Section 11.

2. SPEECH PRODUCTION PROCESS

In order to perform efficient analysis of the speech signal at the acoustic level, it is advantageous to exploit the knowledge about the speech production process. This knowledge is useful in selecting a suitable parametric model for speech production. Once a speech production model is selected, the role of speech analysis techniques is to estimate the parameters of this model accurately and efficiently.

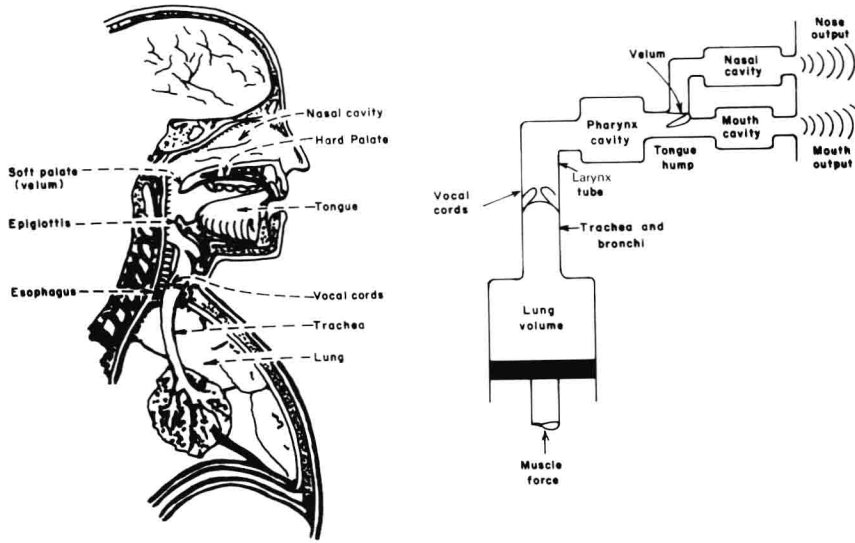

Figure 1. Speech production system and its schematic representation. (After Flanagan, 1972.)

Figure 1 shows the human speech production system along with its schematic representation. The speech production process can be decomposed into three components: (1) the generation of the excitation-source signal, (2) its modulation by the vocal-tract system, and (3) the radiation of the speech signal. These three components are shown in Figure 2. In order to generate the excitation-source signal, the lungs and the associated respiratory muscles constitute the source of power. This power is used to generate the quasi-periodic acoustic signal by means of the vibrating vocal cords for voiced sounds such as vowels. For fricative sounds (such as /f/ and /s/), it is converted into an aperiodic (noisy) signal due to the high velocity frictional flow of air through a narrow constriction formed in the mouth. For plosive sounds (such as /p/ and /t/), it is converted into short burst of noise by the sudden release of pressure which is built up by completely closing the vocal tract for short durations. Thus, all of the above mechanisms convert the more

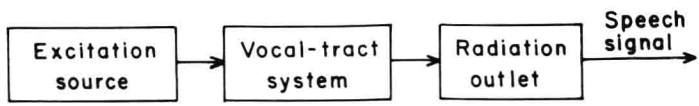

Figure 2. Three components of the speech production process: (1) the excitation source, (2) the vocal-tract system, and (3) the radiation outlet.

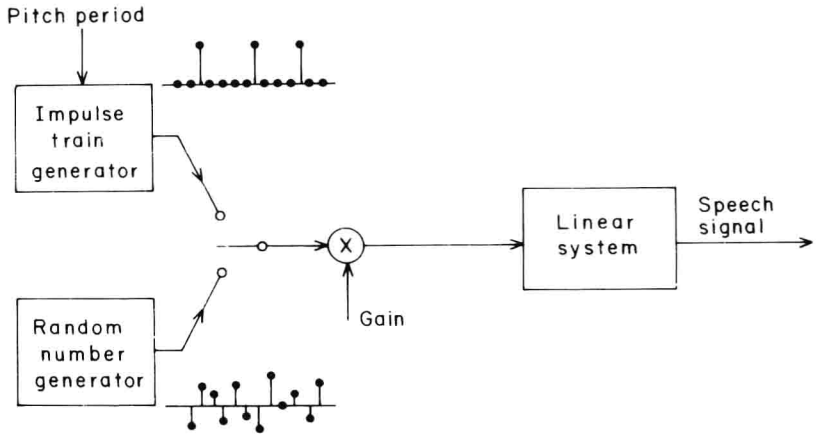

Figure 3. Source-system model of speech production.

or less steady pressure of the lungs (DC power) into an acoustic signal (AC power) which forms the excitation-source signal.

During the production of voiced speech, the vocal tract is excited by the periodic glottal waveform generated by vocal cords. The periodicity of this waveform is called the pitch period. The glottal waveform is triangular in shape. The excitation for the unvoiced speech sounds is random white noise. The shape of the vocal tract uniquely determines the sound that is produced. For a given speech sound, the vocal tract represents an acoustic cavity and, hence, it is usually characterized by natural frequencies (or formants) which correspond to the resonant frequencies of the acoustic cavity. Different speech sounds are produced by dynamically changing the shape of the vocal tract. This change is affected by the movement of the articulators: tongue, lips, jaws and velum. These speech sounds are radiated through the lips. For the production of nasal sounds (such as /m/ and /n/), the vocal tract is blocked at some point determined by the identity of the nasal consonant and the velum is moved to connect the nasal tract to the vocal tract. The nasal sounds are radiated through the nostrils.

The frequency-wise distribution of acoustic energy for a given speech sound depends on the excitation source, the vocal-tract system and the radiation impedance. As the excitation source, the vocal-tract system and the radiation impedance are relatively independent, the speech production process can be modelled as the source-system model shown in Figure 3. This model consists of the following two parts: the excitation source and the speech-generating linear system. These two parts work independently. The excitation-source generates the excitation signal either in the form of a periodic impulse train (for voiced speech) or in the form of a white random

number sequence (for unvoiced speech). The speech-generating linear system contains the combined spectral contributions of the glottal-wave shape (within a pitch period), the vocal-tract system and the radiation impedance. In both voiced and unvoiced speech cases, the gain factor controls the intensity of the excitation to the speech-generating linear system. The speech-generating linear system uses the excitation-source signal at its input and produces the speech signal at its output. In the time-domain, the output speech signal is the convolution of the excitation-source signal and the impulse response of the speech-generating linear system. In the frequency-domain, the spectrum of the output speech is the product of the source and system spectra. Different speech sounds are produced by this model by changing the excitation-source and the linear-system configurations.

3. SPEECH ANALYSIS TECHNIQUES

The aim of speech analysis techniques is to analyse the speech signal and estimate the parameters useful for the given speech processing application. Since the parameters used in most of the speech processing applications are derived from the frequency-domain representation of the speech signal, the main task of the speech analysis techniques is to compute the speech spectrum. There are some speech processing applications where time-domain parameters (such as energy and zero-crossing rate) are useful. However, these parameters can be estimated from the speech signal in a straightforward fashion. Therefore, their processing techniques are not elaborated here further. In this section, three speech spectrum analysis techniques are described: (1) the short-time Fourier analysis technique, (2) the cepstral analysis technique, and (3) the linear prediction (LP) analysis technique. The short-time Fourier analysis technique computes the spectrum of the speech signal, while the cepstral and the LP analysis techniques can decompose this spectrum into two components corresponding to the excitation-source and the linear-system parts of the speech production model (shown in Figure 3).

Before performing any type of digital processing on the speech signal, it is first necessary to digitize the analog speech signal. For this, the speech signal is filtered by a lowpass filter with a cutoff frequency of W Hz to avoid aliasing effects. It is then digitized using an analog-to-digital converter at a sampling frequency higher than the Nyquist rate of $2W$ Hz. It is preferable to select the cutoff frequency, W, to be high to get more information in the digitized speech signal which might be useful in a given speech processing application. But, this increases the computational load as the number of samples to be processed increases with W. Thus, there is a tradeoff involved in the selection of lowpass filter cutoff frequency, W. The value of the cutoff frequency, W, depends on the speech processing application and is typically in the range of 3–10 kHz.

The speech signal is nonstationary in nature, but it can be assumed to be stationary over short durations for the purpose of analysis. This assumption is not valid for regions where there are sharp transitions, as when the articulators are moving fast from the target positions of one sound to those of another. For the stationarity assumption to be valid, it is necessary to choose as short an analysis segment as possible. In pitch-synchronous analysis (Pinson, 1963), the pitch pulses mark the boundaries of the analysis segments; the analysis segments can then be quite short (usually less than one pitch period). Thus, the stationarity assumption is quite easily satisfied for pitch-synchronous analysis. However, it is not possible to reduce the analysis segment to that extent for pitch-asynchronous analysis. This is because arbitrary placement of analysis segments (with respect to pitch pulses) can cause large errors in spectral estimation if the analysis segment is too short. A reasonable compromise for pitch-asynchronous analysis is to use a segment duration which is two to four times the pitch period. Thus, in practice, the speech signal is analyzed frame-wise, with a frame-rate of 50–100 frames/s and for each frame the duration of speech segment is taken to be 20–40 ms.

3.1 Short-Time Fourier Analysis Technique

Fourier analysis is the traditional technique for computing the amplitude and phase spectra of the speech signal. Standard Fourier transform requires the speech signal to be available for all time (i.e. from minus infinity to plus infinity). Since speech is nonstationary in nature, it becomes necessary to perform short-time Fourier analysis using a window on the speech signal.

In the short-time Fourier analysis, the speech signal is multiplied by a suitable window function and the discrete Fourier transform (DFT) is computed using a fast Fourier transform (FFT) algorithm (Cooley and Tukey, 1965). The type of window function to be used in the analysis depends on the speech processing application. For example, in the spectral-subtraction method of speech enhancement (Boll, 1979; Paliwal, 1987a) where reconstruction of speech from the modified short-time Fourier transform is required, a triangular window with 50% overlap between adjacent frames is used to suppress block-edge effects. In adaptive transform coding application (Zelinski and Noll, 1977; Krishnan and Paliwal, 1986) where bit-reduction is important in addition to suppressing the block-edge effects, the trapezoidal windows are used. In other applications, the tapered cosine windows (such as the Hamming and the Hanning windows) are used in short-time Fourier analysis to avoid the spectral leakage effects (Harris, 1978).

For illustrating the performance of the short-time Fourier analysis technique, the voiced and the unvoiced speech signals of vowel /a/ and fricative /s/ are considered here. Each of these two signals has a duration of 32 ms. The