LOW-BIT-RATE CODING OF SPEECH SIGNALS

by

Abraham Alcaim

February 1981

A thesis submitted for the degree of Doctor of Philosophy
in the Faculty of Engineering, University of London

Department of Electrical Engineering
Imperial College of Science and Technology
University of London

ABSTRACT

The possibilities of exploiting two different approaches to the low-bit-rate coding of speech signals are examined. The possibilities relate to the dominant coefficient Hadamard transform coding technique and to a time-domain coding technique. The aim is to develop simple coding techniques for the transmission of speech at rates below 1 bit per sample, whilst maintaining intelligibility and naturalness of the coded speech.

In the first part of the thesis, an extensive review of previous work in the field of speech coding is presented.

The Hadamard transform coding approach is examined and two techniques are investigated which are aimed at improving the system performance at low-bit-rates (below 6 kbits/sec). One technique considers the possibility of further exploiting the residual correlation between the dominant coefficients of speech segments in the Hadamard domain, and the second technique is concerned with reducing the effects of spectrum aliasing and discontinuities in the Hadamard transform coded speech.

A new time-domain coding technique is then developed. The technique has many of the basic characteristics of waveform coding, but information is not transmitted every sampling instant. The work on time-domain coding is divided into four parts. In the first part, a new time-domain pitch period detection algorithm is developed, and

the second part consists of a series of experiments on the 'voiced' portions of the speech waveform in which the quasi-periodic nature of the waveform and a small number of features of the waveform are studied with a view to effecting bit-rate reduction. In the third part, a new method of clipped unvoiced speech regeneration is examined, and in the fourth and final part, a complete time-domain coding technique is developed, which operates at an average bit-rate of 7.6 kbits/sec. Waveforms of original and processed speech, spectrograms, and recorded utterances are presented and discussed.

The final part of the thesis is devoted to consideration of aspects relating to the buffering of variable-rate data sources. Estimates of communication delays and buffer sizes are obtained for zero-crossings-based speech coding schemes and for the time-domain coding technique developed in the thesis.

ACKNOWLEDGEMENTS

I wish first and foremost to express my gratitude to Professor L.F. Turner, whom I am privileged to have had as supervisor. I am indebted not only for Professor Turner's technical guidance and criticism but also for his constant encouragement during the course of my research and the preparation of this thesis.

I am very grateful to Dr. E. Frangoulis for his close collaboration throughout the course of the research. His comments were most helpful and are greatly appreciated. Another debt is owed to my colleagues Behrouz Farhang and Orlando Loques for discussions on microprocessor-based implementation of systems.

I would also like to thank Mr. R. Sulley who assisted me with the computer facilities during this work, and Mike Perdios for his help in the preparation of the recordings. To Judy Perle, who greatly helped to improve my English, I express my gratitude. Thanks are also due to Mr. R.W. Puddy who executed the drawings with outstanding accuracy, and to Moira, for her skillful typing of the manuscript.

Special thanks are due to Ivonice for her support and friendship.

The study presented in this thesis was financially supported by CAPES (Coordination for the Upgrading of University Graduate Level Personnel), an agency of the Brazilian Ministry of Education and Culture. Financial assistance was also received from PUC/RJ (Pontifical Catholic University of Rio de Janeiro). I would like to express my thanks to both CAPES and PUC/RJ.

TABLE OF CONTENTS

ABSTRACT 2

ACKNOWLEDGEMENTS 4

TABLE OF CONTENTS 6

LIST OF FIGURES 11

LIST OF TABLES 18

CHAPTER 1: INTRODUCTION 19

1.1. The Speech Signal 21

1.2. The Low-Bit-Rate Coding of Speech 26

1.3.. Organisation of the Thesis 31

CHAPTER 2: SPEECH CODERS : A REVIEW 36

2.1. Waveform Coding 38

2.1.1. PCM and APCM Coders 39

2.1.2. DPCM and ADPCM Coders 44

2.1.3. DM and ADM Coders 49

2.2. Sub-Band Coding 55

2.3. Adaptive Predictive Coding 57

2.4. Vobanc and Analytic Rooter 61

2.5. Channel Vocoder 63

2.6. Formant Vocoder 67

2.7. Linear Prediction Vocoder 72

2.7.1. The Covariance and Auto-Correlation Methods 73

2.7.2. Maximum Likelihood Vocoder 74

2.7.3. PARCOR Vocoder 75

2.7.4. Other Developments in the LPC Vocoder Method of Speech Coding 77

2.8. Voice-Excited Vocoder 81

2.9. Homomorphic Vocoder 83

2.10. Pattern-Matching Vocoder 87

2.11. Correlation Vocoder 89

2.12. Harmonic Vocoder 91

2.13. Laguerre and Chebyshev Vocoders 92

2.14. Phase Vocoder 94

2.15. Transform Coding 96

2.15.1. The Karhunen-Loève, Walsh-Hadamard, Discrete Fourier, and Discrete Cosine Transforms as Applied to Speech Coding 98

2.15.2. Dominant Coefficient Walsh-Hadamard Transform Coding 108

2.15.3. Hadamard-Haar Transform Coding 110

2.15.4. Adaptive Transform Coding 111

2.16. Time Coding 114

2.16.1. Clipped Speech 115

2.16.2. Time-Shape Pattern Matching Coder 118

2.16.3. Extremal Coding 119

2.16.4. Pitch-Synchronous Time Coding 120

2.16.5. Adaptive Aperture Coding 122

2.17. Discussion 124

CHAPTER 3 : NEW APPROACHES TO THE HADAMARD TRANSFORM CODING OF SPEECH 139

3.1. Double Hadamard Transform Coding Technique 141

3.1.1. Motivation 141

3.1.2. The Average Power Density Function in the Double Hadamard Transform Domain 143

3.1.3. Bit Allocation for Primary and Secondary Labelling 146

3.1.4. Discussion of Experimental Results. 149

3.2. The Combined Use of Oversampling and Smoothing in Decoding 152

3.2.1. Motivation 152

3.2.2. Proposed Modifications 155

3.2.3. Speech Quality Obtained with Modified Hadamard Coding 157

CHAPTER 4 : TIME CODING OF SPEECH SIGNALS 177

4.1. Introduction 178

4.2. A Time-Domain Pitch Period Detector 180

4.2.1. General Comments 180

4.2.2. Description of the New Pitch Detection Algorithm 183

4.2.3. Voiced-Unvoiced Classification 189

4.2.4. Experimental Results 192

4.3. Experiments on Pitch-Synchronous Time Coding of Voiced Speech 197

4.3.1. Experiment I 199

4.3.2. Experiment II 200

4.3.3. Experiment III 202

4.3.4. Experiment IV 204

4.3.5. Experiment V 205

4.3.6. Discussion 209

4.4. A Method of Clipped Unvoiced Speech Regeneration 209

4.5. Low-Bit-Rate Time Coding of Speech Signals: A New Scheme 213

4.5.1. System Description 213

4.5.2. Experimental Results 218

4.6. Considerations Relating to Speech Quality and System Complexity 220

CHAPTER 5 : DELAYS IN VARIABLE-INFORMATION-RATE TO CONSTANT-TRANSMISSION-RATE SYSTEMS 269

5.1. Introduction 270

5.2. Theoretical Results 271

5.2.1. Delay-Operated Systems 271

5.2.2. The Symbol-Injection Procedure 275

5.2.3. Distortion Intervals in Systems with Insufficient Delays 282

5.3. Applications to Zero-Crossings-Based Speech Coders 283

5.3.1. Statistical Properties of the Zero-crossings of Low-Pass Filtered Speech Signals. 283

5.3.2. Estimates of Communication Delay and Buffer Storage 287

5.4. Applications to the Proposed Time Coding Technique for Speech Signals 289

5.5. Discussion 290

CHAPTER 6 : CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH 310

6.1. Conclusions 311

6.2. Suggestions for Further Research 318

REFERENCES 321

APPENDIX 335

LIST OF FIGURES

Figure 1.1.	The human vocal mechanism.	34
Figure 1.2.	Illustration of voiced and unvoiced speech waveforms (time interval: 64 msec).	35
Figure 2.1.	DPCM system.	126
Figure 2.2.	DM system.	127
Figure 2.3.	Continuous DM.	128
Figure 2.4.	Step-size adaptation logic in DCDM.	129
Figure 2.5.	Step-size adaptation logic in discrete adaptive DM.	129
Figure 2.6.	Step-size adaptation logic in one-bit memory DM.	129
Figure 2.7.	Sub-band coder using low-pass translation technique.	130
Figure 2.8.	Technique of integer-band sampling.	130
Figure 2.9.	Predictor used in adaptive predictive coding.	131
Figure 2.10.	Adaptive predictive coder with adjustable noise spectrum.	131
Figure 2.11.	Channel vocoder.	132
Figure 2.12.	Channel synthesis using spectral flattening.	133
Figure 2.13.	Discrete all-pole model in the time domain.	133
Figure 2.14.	Synthesis procedure in the maximum likelihood vocoder.	133
Figure 2.15.	PARCOR vocoder.	134
Figure 2.16.	Voice-excited vocoder.	135
Figure 2.17.	Homomorphic vocoder.	135

Figure 2.18.	Autocorrelation vocoder.	136
Figure 2.19.	Laguerre vocoder.	137
Figure 2.20.	Transform coding.	137
Figure 2.21.	Adaptive transform coder.	138
Figure 2.22.	Illustration of aperture coding of length L = 3.	138
Figure 3.1.	Average power density in the double Hadamard transform domain (sentence no. 1).	160
Figure 3.2.	Average power density in the double Hadamard transform domain (sentence no. 2).	161
Figure 3.3.	Average power density in the double Hadamard transform domain (sentence no. 3).	162
Figure 3.4.	Average power density in the double Hadamard transform domain (sentence no. 4).	163
Figure 3.5.	Double Hadamard transform coding.	164
Figure 3.6.	Original (a) and Hadamard processed (b) speech waveform (16 msec), using 3 dominant coefficients out of 64.	165
Figure 3.7.	Original (a) and Hadamard processed (b) speech waveform (16 msec), using 3 dominant coefficients out of 64.	166
Figure 3.8.	Original (a) and Hadamard processed (b) speech waveform (16 msec), using 4 dominant coefficients out of 64.	167
Figure 3.9.	Original (a) and Hadamard processed (b) speech waveform (16 msec), using 4 dominant coefficients out of 64.	168
Figure 3.10.	Short-time spectrum magnitude of a 32 msec segment of voiced speech.	169
Figure 3.11.	Short-time spectrum magnitude of the same speech segment as used in Figure 3.10 reconstructed by the 3 dominant coefficient system.	170

Figure 3.12.	Short-time spectrum magnitude of the same speech segment as used in Figure 3.10 reconstructed by the 4 dominant coefficient system.	171
Figure 3.13.	Short-time spectrum magnitude of a 32 msec segment of voiced speech.	172
Figure 3.14.	Short-time spectrum magnitude of the same speech segment as used in Figure 3.13 reconstructed by the 3 dominant coefficient system.	173
Figure 3.15.	Short-time spectrum magnitude of the same speech segment as used in Figure 3.13 reconstructed by the 4 dominant coefficient system.	174
Figure 3.16.	Hadamard processed waveform and smoothed equivalent of the speech segment used in Figure 3.7.	175
Figure 3.17.	Short-time spectrum magnitude of the same original speech used in Figune 3.15 after Hadamard processing, oversampling (16 KHz) and smoothing.	176
Figure 4.1.	Non-linear smoothing of pitch period contours (after Rabiner, Sambur, and Schmidt[156]).	227
Figure 4.2.	Illustration of pitch peaks polarity (time interval: 32 msec).	228
Figure 4.3.	TDPD algorithm : selection of NP1 and the N peaks (N = 6) according to step 1.	229
Figure 4.4.	TDPD algorithm : result of step 1 for the speech segment used in Figure 4.3.	230
Figure 4.5.	TDPD algorithm : illustration of an incorrect pitch estimation to be corrected by step 3.	231
Figure 4.6.	TDPD algorithm : illustration of an incorrect pitch estimation to be corrected by step 4.	232
Figure 4.7.	TDPD algorithm ; illustration of a typical speech segment to which step 5 is to be applied.	233
Figure 4.8.	Speech waveform segment (32 msec) and pitch markers as given by the TDPD algorithm.	234
Figure 4.9.	Speech waveform segment (32 msec) and pitch markers as given by the TDPD algorithm.	235
Figure 4.10.	Speech waveform segment (32 msec) and pitch markers as given by the TDPD algorithm.	236
Figure 4.11.	Speech waveform segment (32 msec) and pitch markers as given by the TDPD algorithm.	237

Figure 4.12.	Speech waveform segment (32 msec) and pitch markers as given by the TDPD algorithm.	238
Figure 4.13.	Speech waveform segment (32 msec) and pitch markers as given by the TDPD algorithm.	239
Figure 4.14.	Speech waveform segment (32 msec) and pitch markers as given by the TDPD algorithm.	240
Figure 4.15.	Speech waveform segment (32 msec) and pitch markers as given by the TDPD algorithm.	241
Figure 4.16.	Pitch period contours for utterance no..1.	242
Figure 4.17.	Pitch period contours for utterance no. 5.	243
Figure 4.18.	Experiment I: (a) original waveform; (b) reconstructed waveform. Time interval: 128 msec.	244
Figure 4.19.	Experiment I: (a) original waveform; (b) reconstructed waveform. Time interval: 128 msec.	245
Figure 4.20.	Experiment I: (a) original waveform; (b) reconstructed waveform. Time interval: 32 msec.	246
Figure 4.21.	Experiment II: illustration of intrapitch waveform reconstruction.	247
Figure 4.22.	Experiment III: (a) original waveform; (b) reconstructed waveform. Time interval: 32 msec.	248
Figure 4.23.	Experiment IV: illustration of features extracted from the waveform intrapitch.	249
Figure 4.24.	Experiment IV: (a) original, and (b) processed waveform using only the absolute extreme between zero-crossings.	250
Figure 4.25.	Experiment IV: (a) original, and (b) processed waveform using all peaks and valleys.	251
Figure 4.26.	Experiment V: extraction of intrapitch parameters.	252
Figure 4.27.	Experiment V: extraction of other pitch peaks.	252
Figure 4.28.	Experiment V: (a) original, and (b) processed waveform (NUP = 128 samples). Time interval: 128 msec.	253

Figure 4.29.	Time occurrence of alternate extremes of a 16 msec segment of unvoiced speech waveform.	254
Figure 4.30.	Histogram relating $D_i - \bar{D}$, where D_i is the time interval between alternate unvoiced waveform extremes and $\bar{D}$ is the average over each 32 msec segment (male speech)	255
Figure 4.31.	Histogram relating $D_i - \bar{D}$, where D_i is the time interval between alternate unvoiced waveform extremes and $\bar{D}$ is the average over each 32 msec segment (female speech).	256
Figure 4.32.	Time coding technique: (a) original, and (b) processed waveform. Time interval: 128 msec.	257
Figure 4.33.	Time coding technique: (a) original, and (b) processed waveform. Time interval: 128 msec.	258
Figure 4.34.	Time coding technique: (a) original, and (b) processed waveform. Time interval: 128 msec.	259
Figure 4.35.	Time coding technique: (a) original, and (b) processed waveform. Time interval: 128 msec.	260
Figure 4.36.	Time coding technique: (a) original, and (b) processed waveform. Time interval: 128 msec.	261
Figure 4.37.	Time coding technique: (a) original, and (b) processed waveform. Time interval: 128 msec.	262
Figure 4.38.	Spectrogram of original speech for the utterance " The box was thrown beside the parked tru(ck)" spoken by a male speaker.	263
Figure 4.39.	Spectrograms of processed speech, (a) unquantised and (b) quantised, for the utterance " The box was thrown beside the parked tru(ck)" spoken by a male speaker.	264
Figure 4.40.	Spectrogram of original speech for the utterance " Glue the sheet to the dark blue backgroun(d)" spoken by a male speaker.	265
Figure 4.41.	Spectrograms of processed speech, (a) unquantised and (b) quantised, for the utterance " Glue the sheet to the dark blue backgroun(d)" spoken by a male speaker.	266

Figure 4.42. Spectrogram of (a) original, and (b) processed (quantised) speech for the utterance " It's easy to tell the depth of the well" spoken by a male speaker. 267

Figure 4.43. Spectrogram of (a) original, and (b) processed (quantised) speech for the utterance " It's easy to tell the depth of the well" spoken by a female speaker. 268

Figure 5.1. Delay at transmit buffer to avoid underflow. 292

Figure 5.2. Illustration of output distortion due to insufficient decoding delay. 293

Figure 5.3. Delay at receiver to avoid output distortion. 294

Figure 5.4. Illustration of symbol injection at transmitter. 295

Figure 5.5. Illustration of delay necessary to avoid distortion in a symbol-injection system. 295

Figure 5.6. Illustration of symbol injection. 296

Figure 5.7. Illustration of overall system delays involved in the delay (γ) and symbol-injection (β) systems. (a) $\beta \neq \gamma$, (b) $\beta = \gamma$. 297

Figure 5.8. Illustration of distortion in systems with insufficient delays. 298

Figure 5.9. Histogram relating number of crossings in 2 sec intervals of speech (male speaker; zero offset voltage). 299

Figure 5.10. Histogram relating number of crossings in 2 sec intervals of speech (male speaker; 10 mV offset voltage). 300

Figure 5.11. Histogram relating number of crossings in 2 sec intervals of speech (male speaker; 20 mV offset voltage). 301

Figure 5.12. Histogram relating number of crossings in 2 sec intervals of speech (female speaker; zero offset voltage). 302

Figure 5.13. Histogram relating number of crossings in 2 sec intervals of speech (female speaker; 10 mV offset voltage). 303

Figure 5.14. Histogram relating number of crossings in 2 sec intervals of speech (female speaker; 20 mV offset voltage). 304

Figure 5.15. Variations in cumulative function (number of zero-crossings) over 7 sec interval of speech (male). 305

Figure 5.16. Variations in cumulative function (number of zero-crossings) over 7 sec interval of speech (female 1). 306

Figure 5.17. Variations in cumulative function (number of zero-crossings) over 7 sec interval of speech (female 2). 307

Figure 5.18. Variations in cumulative function (number of bits) for the proposed time coding scheme over 7 sec interval of speech (male). 308

Figure 5.19. Variations in cumulative function (number of bits) for the proposed time coding scheme over 7 sec interval of speech (female). 309

LIST OF TABLES

Table 2.1.

Objective and Subjective Performance of ADPCM versus Log-PCM. 49

Table 2.2.

Consonant Articulation Scores of Channel Vocoders 65

Table 3.1.

Percentage Power Contained in Hadamard Coefficients after Second Transformation. 145

Table 4.1

Error Parameters for MAUT (M) and TDPD (T) Pitch Period Detection Algorithms. 195

Table 4.2.

Quantisation of Some Parameters Used in the Time Coding Technique. 218

Table 4.3.

Time Coding Technique: Average Bit-Rate per Sentence. 221

Table 4.4.

Lower and Upper Limits of the Number of Operations Required for Voiced-Unvoiced Classification and Pitch Detection. 223

Table 5.1.

Zero-Crossings Measurements of Low-Pass Filtered Speech Signals (Zero Offset Voltage). 285

Table 5.2.

Zero-Crossings Measurements of Low-Pass Filtered Speech Signals (Offset Voltages of 10 mV, 20 mV, and 30 mV; 2 sec Sections of Speech). 286

CHAPTER 1

INTRODUCTION

CHAPTER 1

INTRODUCTION

Speech is a remarkable and unique communication means used by man - the most effective mode of conveying his thoughts and ideas to other human beings. It results from longitudinal vibrations in the air medium induced by specialised movements of man's respiratory and masticatory apparatus. Such acoustic waves are perfectly adequate as a means of communication between humans as far as face-to-face conversation is concerned. For long-distance communication, however, an effective representation of the naturally produced acoustic waves had to be invented. This was accomplished by transforming acoustic waves into electromagnetic waves, an invention which marked the birth of telephony.

The growth of population, and an increasing need for more efficient means of speech communication over long distances stimulated investigation into analogue techniques which would make it possible to transmit speech information using a much reduced spectrum. With the advent of digital techniques research has mainly concentrated on the digital coding of speech for a variety of reasons, to be explained later in this Chapter.

When using digital facilities, the transmission of the speech information is described in terms of the number of bits per second necessary for the information to be transmitted. This is referred to as the necessary transmission bit-rate. The problem of finding

efficient means for coding speech constitutes a particularly complex problem since a compromise between bit-rate (or bandwidth) requirements, the complexity of the coding and decoding operations, and the quality of the signal thus produced must be reached. Clearly, therefore, there exists no single answer - each possible solution offers different advantages.

This introductory Chapter is divided in three parts. Firstly, some of the essential basic characteristics of the speech signal are considered. This is then followed by a brief discussion of those problems related to the low-bit-rate coding of speech, on which the motivation and scope of the research work reported on in this thesis was based. Finally, in the last part of the chapter, a brief outline of the organisation of the thesis is presented.

1.1. The Speech Signal.

The vocal tract is a non-uniform acoustic tube, approximately 17 cm in length in the case of adult males, which extends from the glottis (opening between the vocal folds) to the lips. Its shape varies with time and is determined by the lips, jaw, tongue and velum. The cross-sectional areas of the vocal tract vary in the range from 0 cm^2 to 20 cm^2. During the production of non-nasal sounds, the nasal cavity, which extends from the nostrils to the velum, is closed off by the velum. A schematic diagram of the anatomic structures involved is presented in Figure 1.1.

Speech sounds are produced by air expelled from the lungs into the trachea. Depending on the speech sound to be generated, three basic mechanisms of excitation may be involved. In the case of voiced sounds (such as /æ/ in cat), the air flow from the lungs causes the vocal folds to vibrate, and this action, in turn, generates a sequence of quasi-periodic pulses of air pressure which serves as the excitation source to the vocal tract. For unvoiced fricative sounds (such as /s/ in side), turbulence is created at narrow constrictions in the vocal tract, producing a source of continuous noise. The third mechanism of excitation consists of a sudden release of excess air pressure after a complete closure at some point in the vocal tract, and is characteristic of plosive sounds (such as /p/ in park).

Natural languages comprise a number of basic sound elements, termed phonemes. In the literature relating to and describing speech, phonemes are usually bracketed between two slanting lines - /./. These elements are characterised by the fact that two words differ if only one of their basic elements differ. Some characteristics of the different sounds which occur in the English language will now be outlined.

A. Vowels (e.g., /u/ as in boot, /ɛ/ as in met) are relatively long in duration, a factor which depends, however, on the phonetic environment. For unvoiced and voiced consonant environments, average durations of 170 msec and 310 msec, respectively, have been measured.[1] Vowel sounds are produced

by excitation of the vocal tract by quasi-periodic pulses of air pressure caused by vocal fold vibrations, and contain a number of resonances (or formants) associated with the vocal tract shapes.

B. Glides (e.g., /j/ in you, /w/ in want) and Semivowels (e.g., /r/ in run, /ℓ/ in let) are vowel-like sounds, which differ from vowels in that they exhibit a continuous formant motion.

C. Diphthongs (e.g., /aI/ as in five, /eI/ as in say) are also vowel-like sounds, which result from combinations of two vowel units. They also present characteristics similar to glides and semivowels.

D. Nasals (e.g., /m/ in mouth, /n/ in no) are produced by opening the velum and coupling the nasal cavity with the vocal tract. These sounds are voiced since excitation is also provided by vocal fold vibrations, as in vowels, but are radiated mainly through the nostrils. Their spectra generally present a low first formant, a spectral zero, and uniform characteristics for high frequencies.

E. Fricatives can be either voiced (as /v/ in seven) or unvoiced (as / ʃ / in she), depending on the presence or absence of vocal fold vibrations. Both, however, are typified by a source of continuous noise created by turbulence at narrow constrictions in the vocal tract.

F. Plosives or stops can also be either voiced (as /b/ in boy) or unvoiced (as /p/ in pound), and are characterised by an increased pressure at a point in the vocal tract followed by an abrupt release of air. The pressure build up is observed as a silence interval, and in the case of voiced stops vocal fold vibrations are usually present.

G. Affricates (e.g., /t ʃ/ as in chain , /dz/ as in jar) are combinations of stops and fricatives, and exhibit characteristics similar to both types of sounds.

Broadly speaking, then, it can be seen that speech sounds are divided into two main categories - voiced or unvoiced - depending on the presence or absence of vocal fold vibrations. In Figure 1.2, 64 msec - long speech samples of actual speech waveform are shown to illustrate the two categories. It can be seen that for the voiced speech sound the waveform is nearly periodic. The interval between the major peaks gives a measure of the pitch period for that particular speaker, and the inverse is the pitch frequency, which can range over about one octave[2]

during the course of a sentence spoken by an individual. Typical average pitch frequencies are 120 Hz for men and 220Hz for women. The vocal tract geometry, on the other hand, is controlled by the speaker in order to obtain the distinction between different sounds, which is reflected in their spectra by resonances (or formants). Below 3 KHz, three formants are usually found. From Figure 1.2 it can be seen that each pitch period consists of approximately damped oscillations, which are actually a consequence of the spectral resonances of the tract. The oscillation period is roughly related to the major formant position and the rate of damping to the half-power bandwidth of the formant. It should be noted, however, that the waveform in a single pitch period results from a superimposition of decaying oscillations due to the multiplicity of formants.

The unvoiced speech waveform shown in Figure 1.2 exhibits a noise-like characteristic. This waveform is typical of this class of sounds, and shows a concentration of energy at high frequencies, which is evidenced by the high number of zero-crossings. In general, it can also be said that these sounds are of lower intensity than voiced sounds.

Speech signals are highly complex and the aspects which have been considered in this section relate only to certain basic characteristics. An excellent study which deals with the subject in depth is contained in a book by Flanagan[3] which, in addition, contains an extensive bibliography.

1.2. The Low-Bit-Rate Coding of Speech.

There are many reasons for the attractiveness of the digital representation of speech. In the case of long-distance communication, digital links are very reliable. Noise effects are not accumulative as in the case with analogue transmission. Digital representation of signals also offers the advantage of a uniform format for different signals. This means that a digital communication system can handle speech information together with other signals such as facsimile signals, video signals, telegraph, computer data, etc. Another advantage is the possibility of combining transmission and switching functions. Furthermore, the ease with which encryption can be accomplished (by bit permutation, for instance), once the signal is in digital form is an attractive property of the digital representation of signals.

As mentioned earlier, an important parameter associated with the digital transmission of speech is the transmission bit-rate. At high bit-rates, such as 56 Kbits/sec, pulse-code-modulation (PCM) is a standard technique of speech coding. In general, 7 bits are used to represent each sample of low-pass filtered speech, and the speech is sampled at a rate of 8 Ksamples/sec. At 56 Kbits/sec the decoded PCM speech signal is virtually indistinguishable from the original.

A question arises, however, with respect to the efficiency of the above mentioned PCM system - this concerns the need to use such high bit-rates to transmit speech information. On considering, for instance, the written equivalent of speech in terms of the 42 English

phonemes, it can be seen that only a very low bit-rate (of the order of 50 bits/sec) is required.[3] Transmission of speech as text however, does not include important characteristics which are particular to the speaker. Although it is not as yet possible to define what is meant by the information rate of speech, experiments have shown that much of the information contained in the speech signal is not essential to its understanding and, indeed, transmission bit-rates much lower than 56 Kbits/sec can be used.

Bit-rate reduction is closely associated with efficiency. If the required data rate could be reduced by a factor of 10 then ten times as many voice channels would be available. In the coding of speech to remove redundancies and lower the bit-rate, however, a number of factors have to be taken into account. These include the perceptual quality of the resulting speech and the complexity of the encoding and decoding equipment. The number of variables involved has led to the development of a large number of alternative digital coding schemes.

A system which has given rise to a large number of low-bit-rate speech coding techniques is the channel vocoder invented over 40 years ago by Dudley.[4,5] The idea behind this, and other related techniques developed later, was largely the exploitation of important features of the source modelled on the human vocal mechanism. This made it possible to obtain an efficient representation of speech in terms of a small number of slowly varying parameters, conveying information about excitation source and vocal tract transmission characteristics for the generation of speech sounds.

The original aims of research into channel vocoders were largely concerned with fundamental speech research.[6] It was only during World War II, when requirements for secure speech communication arose, that attention was focused on the channel vocoder as a means of obtaining digital speech. Later developments of this and other related techniques have made possible the digital transmission of speech at remarkably low-bit-rates (1000 bits/sec, or even lower). However, the high cost of the coding operation, and a certain lack of naturalness in the decoded speech, limit the uses of the channel vocoder. More recent developments of vocoder-type techniques have made it possible to obtain acceptably natural speech reproduction at low-bit-rates. The techniques involved, however, require complex methods of processing and analysis as part of the estimation of the parameters of the speech model.

In the case of higher bit-rate systems (16 Kbits/sec and above), waveform coding techniques, which typically include pulse-code-modulation (PCM), differential pulse-code-modulation (DPCM), delta-modulation (DM), and their adaptive configurations,[7] are among the most attractive, since good speech reproduction can be attained with very simple coder/decoder structures. These techniques involve direct digital coding of the speech waveform with various strategies being employed to reduce the amount of information to be sent, without, however, any use being made of the principles involved in the human mechanism of speech generation and/or perception.

Sub-band coding (SBC),[8] adaptive predictive coding (APC),[9] and adaptive transform coding (ATC)[10] represent important developments in speech coding techniques at bit-rates below 16 Kbits/sec since they enable good quality speech to be obtained with equipment which is less complex than that associated with vocoder-type devices. However, the performance is considerably diminished if the bit-rate is reduced below approximately 4.8 Kbits/sec although intelligibility can still be largely maintained. It should be noted that with vocoder-type systems, increasing the bit-rate above 4.8 Kbits/sec does little to improve the quality of the speech.

Another approach to the low-bit-rate coding of speech signals consists of utilising a more efficient representation of speech in the Hadamard transform domain. This is achieved by transmitting only a small sub-set of dominant (largest absolute valued) coefficients.[11,12] The merits of this method relate to the fact that, for bit-rates of 6 Kbits/sec and above, highly intelligible speech can be attained with computationally simple operations, since only additions and subtractions are required in order to obtain the transform coefficients. An additional advantage is that the direct and inverse transform operations are essentially identical. At lower bit-rates (below 6 Kbits/sec), however, the performance of Hadamand coding schemes is noticeably reduced, although the resulting decoded speech remains intelligible.

The advantages associated with the above coding procedure motivated part of the research work reported in this thesis. The aim of this particular aspect of the study was to examine the possibilities of improving the quality of decoded speech for the lower bit-rates (less than 6 Kbits/sec) without increasing the complexity of the coding and decoding procedures.

Another approach to the low-bit-rate coding of speech comprises methods in which information is extracted from the speech waveform only at certain time instants or intervals,[13-20] and will be referred to as 'time-domain coding' or, more simply, as 'time coding'. A major part of the research presented in this thesis was devoted to the investigation of the possibilities offered by the 'time coding' approach. The aim was to develop a relatively simple scheme capable of giving intelligible speech, without sacrificing its naturalness, at transmission bit-rates below 8 Kbits/sec.

The decision to study time coding systems was based on several factors. Firstly, since one of the major aims was to develop simple schemes, vocoder related techniques were ruled out because of their complexity. Secondly, waveform coders, despite their simplicity and low cost, require at least 1 bit per sample and, thus, do not have the potential to operate at bit-rates below 8 Kbits/sec. Time coding techniques, on the other hand, do not transmit information at every sampling instant and are generally based on the determination of important characteristics of the speech waveform in relation to the perception of the sounds. The methods employed are usually simple and can potentially operate at bit-rates below 1 bit per sample. The time coding approach has not been extensively explored in the area of low-bit-rate speech coding, a factor which further contributed to the decision to study its possibilities.

One of the disadvantages of time coding techniques involves the fact that they generate information at non-constant rates, which means that buffering is required if information is to be communicated (as is usually the case) using a constant-rate transmission system. A consequence of the buffering is that delays are introduced into the communication process. For this reason, a separate study was undertaken in which theoretical aspects of the delay problem were considered. In addition, an experimental evaluation of the delays involved was also carried out.

1.3. Organisation of the Thesis.

A very large number of schemes have been developed for the coding of speech signals. Chapter 2 of this thesis is devoted to a review of existing schemes developed during the past 40 years, which represent the range of research interests in the field. An attempt is made to unify the material related to numerous past and current techniques, and to discuss some of the relative merits and difficulties associated with the various methods.

Because of the simplicity of the basic scheme of dominant-coefficient Hadamard transform coding of speech, two investigations were carried out in an attempt to improve the system performance for low-bit-rate transmission. Chapter 3 contains a detailed description of these investigations. The first is concerned with the possibility of further exploiting the residual redundancy contained in a limited number (m) of dominant coefficients, by applying a second Hadamard transformation of length m. The technique was simulated for bit-rates

below 6 Kbits/sec and is discussed in relation to the Hadamard and Hadamard-Haar schemes. The second investigation examines the effects of two factors, namely spectrum aliasing and discontinuities in the waveform, which are likely to introduce distortions in the Hadamard transform coded speech. Some modifications are then proposed which involve no bit-rate penalty and which are aimed at improving speech quality. Experimental results are presented and discussed.

Chapter 4 focuses on the time coding approach. It was clear from the outset that in order to obtain an efficient signal representation, the time-domain coding technique should exploit the quasi-periodic nature of voiced speech. With this in mind, effort was first directed towards the search for a simple pitch period detection method which would be sufficiently accurate but less complex than existing methods. Chapter 4 begins with a description and some illustrative examples of the performance of a time-domain pitch estimation algorithm which was developed during this initial phase. This is followed by a series of experiments concerned with the time coding of voiced speech in which different types of distortion are introduced into the speech waveform by eliminating certain characteristics of the original signal, and observing the effects on the speech regenerated from the remaining information. From these experiments, a relatively simple technique for coding the voiced segments of the waveform was developed. Then, a method for regenerating clipped unvoiced speech is presented, and in the final part of the chapter a complete time coding technique, which operates at an average bit-rate of 7.6 Kbits/sec, is proposed. The results obtained from a computer simulation of the proposed technique are presented and discussed.

Chapter 5 is devoted to the study of certain factors related to the performance of variable-source-rate to constant-transmission-rate schemes of data compression, in which the time coding technique proposed in Chapter 4 is included. The question of buffering at the trasmitter and the receiver is considered, and expressions are derived for the necessary minimum delay which must be introduced at the transmitter and the receiver if distortion is to be avoided. In addition, a symbol-injection procedure is considered and an expression is derived for the system delay necessary to avoid distortion. The results of a statistical analysis of low-pass filtered speech are presented, and these results are then used to obtain estimates of the necessary minimum overall system delay, and the necessary buffer sizes, in zero-crossings-based speech coding schemes. In addition, estimates are provided of the period of time over which the output will contain distortions if a delay below the minimum necessary to avoid distortion is used. Finally, in the last part of the chapter, estimates of communication delay and buffer storage associated with the proposed time coding scheme presented in Chapter 4 are provided for illustrative samples of speech.

Finally, in Chapter 6, an attempt is made to provide an overall view of the research work presented in this thesis, together with an objective evaluation of the main results achieved. Some lines of further research in the field are also suggested.

Most of the typical results presented in Chapters 3 and 4 are illustrated by recorded utterances accompanying this thesis.* An appendix to the thesis describes and catalogues these recordings.

---

*After the preparation of this thesis it was learnt that University rules do not permit recordings to be appended to theses. However, a copy of the recordings referred at various points in the thesis can be obtained, upon request, from Professor L.F. Turner, Department of Electrical Engineering, Imperial College of Science & Technology.

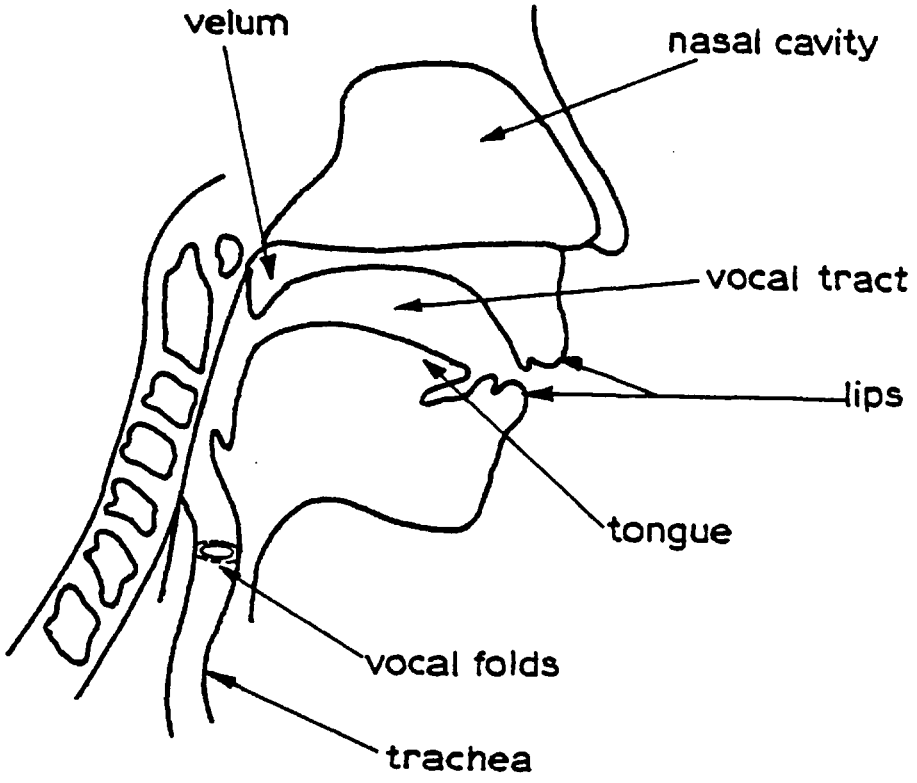


Figure 1.1. The human vocal mechanism

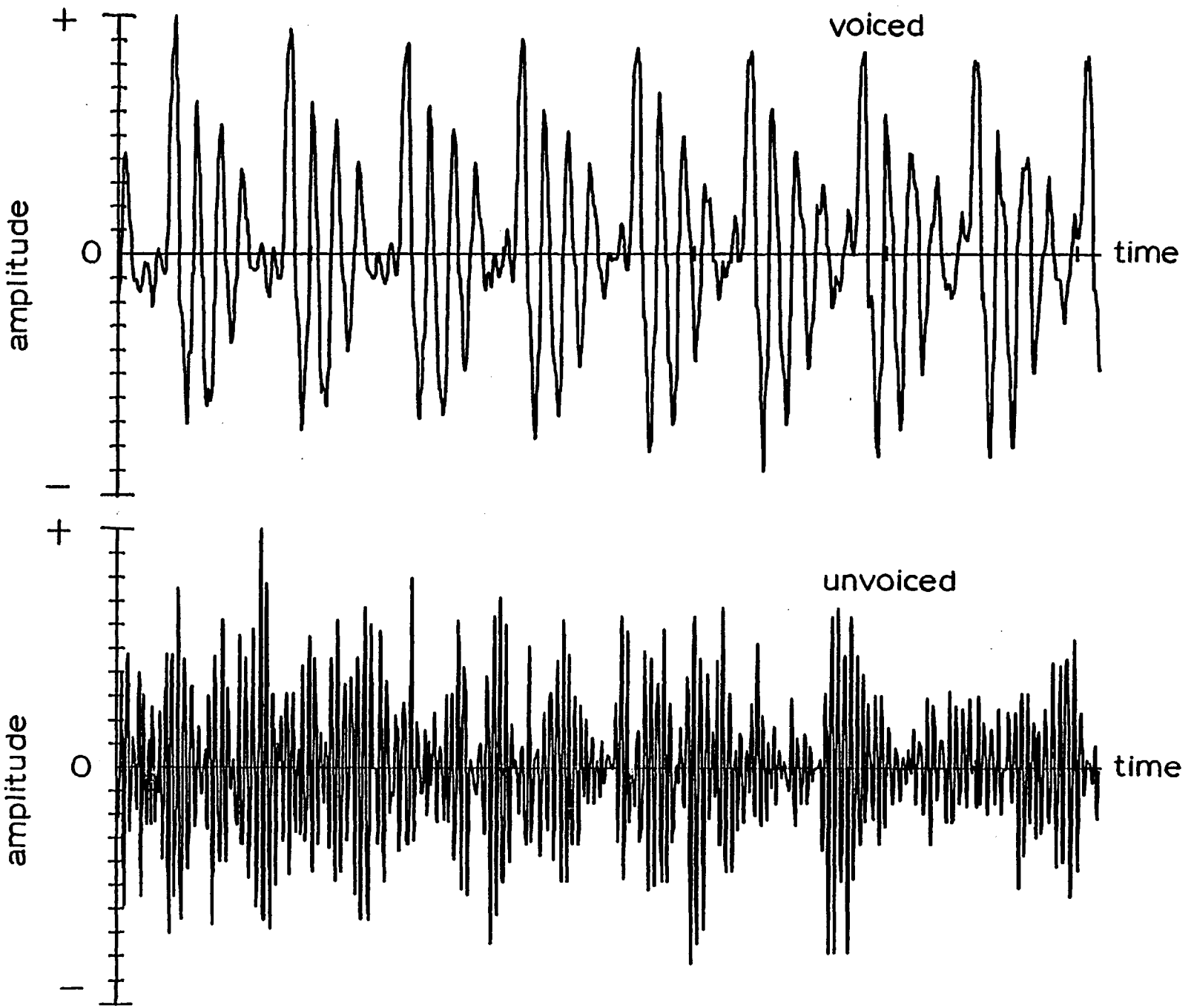


Figure 1.2. Illustration of voiced and unvoiced speech waveforms (time interval: 64 msec).

CHAPTER 2

SPEECH CODERS : A REVIEW

CHAPTER 2

SPEECH CODERS : A REVIEW

The development of speech coders in a variety of categories of coding philosophy has been growing rapidly in past years, especially following Dudley's invention of the Vocoder[4,5] in the late 1930's. Accordingly, many important review works have appeared in the literature, which are either focused on a few particular methods of speech coding or consider a wider range of approaches.

In a 1966 paper, Schroeder[21] reviews a great variety of vocoder-type techniques, whereas Schafer,[22] on the other hand, considers only those schemes which are important to the field of aids to the handicapped. A comprehensive review on the topic of speech coding is presented in a book by Flanagan, [3] which includes important developments on the subject prior to 1972, when the second edition of the book was issued. An interesting collection of papers addressed to the problem of speech synthesis[23] was published in 1973. Transform coding, particularly in the Walsh domain, is briefly reviewed in a paper by Elliott,[24] and the class of waveform coders is considered extensively by Jayant[7] in a 1974 work. A survey including a number of important methods of speech coding can also be found in Reference 6, and more recently, in Reference 25, a wider range of coders and their relative complexity and performance is considered at some length.

The aim of this chapter is not to present a complete study of the subject, but to give a general view of the many different types of coding philosophy which have appeared during the past 40 years. Thus, an attempt is made to unite the material of past and present techniques, with particular attention being focused on the main characteristics and problems involved.

In the review are included the techniques of waveform coding (PCM, DPCM, DM and adaptive configurations), sub-band coding and adaptive predictive coding, methods of frequency-division, a variety of vocoder-type schemes, different techniques of transform coding, and finally, other developments which are considered as elements of another class of techniques, which will be referred to as time coding.

2.1. Waveform Coding.

Techniques of speech coding in which an attempt is made to send information about the acoustic time waveform, without taking into account the production and/or the perception of the signal, are known as waveform coding. Many of these techniques make use, however, of statistical properties of the signal in order to design certain strategies to reduce the amount of information to be transmitted. Broadly speaking, these techniques are the simplest coding techniques which operate at high bit-rates and reproduce high-quality speech.

Waveform coding techniques include pulse-code-modulation (PCM), differential pulse-code-modulation (DPCM), delta-modulation (DM), and the adaptive versions of these.

2.1.1. PCM and APCM Coders.

Pulse-code-modulation[7,26] consists of the basic operations of sampling, quantisation, and representation of discrete (quantised) amplitudes by groups of binary digits. The signal is first band-limited to a frequency F Hertz and then sampled at intervals of at least 1/2F seconds, since these values uniquely specify the band-limited signal. Each sample amplitude is represented by a value chosen from a set of 2^n discrete amplitudes, as the nearest one to the signal value. This operation is called n-bit quantisation, since each of those discrete values can be represented by a group of n bits. Each group is called a code word, and the representation of the discrete amplitudes in code words is the coding operation. At the receiving end of the system there exists a sequence of inverse operations which consists of generating, for each code word, a quantised sample of amplitude represented by that code. The approximation of the original analogue signal is obtained by low-pass filtering the sequence of samples using a filter of bandwidth F Hertz.

In PCM the information is transmitted at a rate of 2Fn bits per second, where n is the number of bits transmitted per sample and 2F is the number of samples generated per second (sampling rate).

With error-free channels the error introduced by representing the samples by discrete amplitudes, the quantisation error, is the most important disturbance acting on the system. For fine quantisation ($n > 6$) and no overload of the quantiser, this error can be described statistically in terms of a uniform distribution[7] and can be considered as an additive noise. For coarse quantisation this error is correlated with the speech waveform and it cannot be modeled as additive noise. The signal-to-quantisation noise ratio SNR (in dB), under fine quantisation and no overload conditions, can be shown[7] to be a linear function of the number of bits, n, per sample.

The presence of frequently occurring low-amplitude values and infrequent high-amplitude values is a well-known property of speech waveforms.[27] Consequently, the use of a non-uniform quantiser, in which smaller step-sizes are associated with the lower amplitudes of the speech and higher step-sizes for higher speech amplitudes results in an improvement in the signal-to-noise performance of the system for a given bit-rate. In general, the error criterion adopted is the minimum mean-square-error. The structure and properties of both uniform and non-uniform quantisers have been reported on in many works appearing in the literature.[28-31] An extensive study of non-uniform quantisation using logarithmic compandors (compressors/expandors) is presented in a classical work by Smith.[29] Non-uniform logarithmic quantisation can be regarded as uniform quantisation of logarithmically compressed input, and the so-called μ-law is defined[29] by

$$C(x) = \frac{V\ell_n[1 + \mu x/V]}{\ell_n(1 + \mu)}, \qquad 0 \leq x \leq V, \tag{2.1}$$

where x is the input signal, $\mu > 0$, and $C(-x) = -C(x)$. For speech signals, Smith observes that a good choice for μ is around 100. A value of μ equal to 255 is, however, usually adopted.

The effectiveness of logarithmic quantisation is well established. In terms of SNR performance, a 7-bit log-quantiser, for example, is approximately equivalent to a 12-bit uniform quantiser,[25] and this clearly represents a considerable improvement.

The problem of designing an optimum quantiser for a given class of input signals is directly related to the problem of finding an appropriate method for matching the quantiser step-size to the input signal variance. This gives rise to what is referred to as adaptive quantisation. Two basically different classes of adaptation algorithm exist. One of them is adequate for input signals with unknown but constant variance. The methods, which may be referred to as " learning" algorithms, attempt to learn the constant value of the signal variance.[32,33] " Tracking-mode" algorithms,[34,35] on the other hand, are suitable for tracking changes in the variance of the input signal, and are, therefore, the most appropriate ones for speech signals which, of course, do have significant changes in variance with respect to time. These tracking-mode algorithms can operate, depending on the particular design, at a syllabic[34] (long-term) or instantaneous[35] (short-term) rate.

The syllabic adaptive PCM of Wilkinson[34] involves the following operations: the input speech is band-limited to the range 250-2400 Hz, 2-bit quantised using threshold voltages +V, 0, -V, and then sampled

at a rate of 4.8 Ksamples/sec. The amplitude digits are applied to a 10 msec time constant integrator, and the output of this integrator is used to control the threshold value V. In this scheme, the least significant digit (the amplitude digit) is 1 for signal amplitudes which are either more positive than +V or more negative than -V. Hence, a large number of 'ones' appearing on the amplitude digital output indicates that V should be increased. At the receiving end, a similar process is involved in order to convert the digits to 4-level amplitude samples, which are then low-pass filtered at 2.4 KHz. The total transmission bit-rate is 9.6 Kbits/sec, and the speech quality is comparable with the 2.4 Kbits/sec channel vocoder and some adaptive delta systems operating at 9.6 Kbits/sec. The syllabic APCM of Wilkinson has the advantage that it can be implemented with simple equipment, but at 9.6 Kbits/sec the speech quality obtained with this technique is still very low with respect to the commercial quality standard.

Jayant's method of instantaneous adaptive quantisation[35] can be interpreted in the following manner. At each instant i + 1 the input standard deviation σ_{i+1} is estimated by multiplying its last estimate, σ_i, by a factor $M(I_i)$ which depends only upon the last quantiser output, i.e.,

$$\sigma_{i+1} = [M(I_i)]\,\sigma_i \,, \tag{2.2}$$

where $I_i = k$ if $Z_i = \pm \sigma_i y_k$, $k = 1, \ldots, N$, Z_i is the quantiser output at instant i, and $\{y_k;\ k = 1, \ldots, N\}$ are the quantising levels for

unitary input variance. An initial value σ_{START} is assumed for the sequence of estimates. Note that if the input signal level is high, which is evidenced by a quantising level $\pm \sigma_i y_k$ with k near to N, then this indicates that σ_i should be increased. On the other hand, if the level is low, σ_i should be decreased. These observations imply the following constraints on the multipliers:

$$M(N) > 1 \tag{2.3a}$$

$$M(1) < 1 \tag{2.3b}$$

$$M(1) \leq M(2) \leq \ldots \leq M(N). \tag{2.3c}$$

For low-pass filtered input speech, desirable values of these multipliers have been obtained by Jayant,[35] for the case of 2, 3, 4, and 5-bit quantisers. The criterion adopted was to maximise the signal-to-noise ratio, as measured over the entire duration of the input data. The SNR gain achieved with this technique over μ-law log-PCM was 6 dB, 7 dB, and 4 dB, respectively, for 2, 3, and 4-bit quantisation. Owing to the coarseness of quantisation, these results should be taken, however, only as indicative and not as adequate evidence of the system performance.

In general, adaptive PCM systems are useful for speech transmission at bit-rates which are usually higher than 20 Kbits/sec. At these bit-rates, good quality speech can be obtained at a low-cost of coding and decoding operations. For lower bit-rates, however, the resulting speech quality becomes noticeably deteriorated.

2.1.2. DPCM and ADPCM Coders.

Differential pulse-code-modulation (DPCM)[36,37] is a more sophisticated class of waveform coding, which is tailored to transmit, more efficiently, information signals exhibiting high correlation between adjacent amplitude values. Therefore, speech is a particularly suitable class of signals for this type of coding.

The basic approach in DPCM is to code the difference e_i between the real input value s_i and the estimated, or predicted, value $\hat{s}_i$ (based on previous values) of each sample amplitude, rather than coding s_i itself as in PCM. A block diagram of the system is shown in Figure 2.1, in which the quantity q_i, the quantisation error, denotes the difference between the input e_i and the output Z_i of the quantiser. The process of prediction is simplified if a linear prediction is used[38] in which $\hat{s}_i$ is obtained as a linear combination of a number m of past values. It is important to note that in a DPCM system these are not the actual values, but the values obtained from the quantised versions of the error sequence. The reason is that since at the receiver, only these values are available, the placing of the feedback loop before the quantiser would result in the reconstructed signal containing not only the quantisation noise of the error, but also an error introduced by using estimates which are different from those obtained at the transmitter.

For the DPCM coder (see Figure 2.1)

$$e_i = s_i - \hat{s}_i \tag{2.4}$$

and

$$\hat{s}_i = \sum_{j=1}^{m} a_j \, (s_{i-j} - q_{i-j}), \tag{2.5}$$

where $\{a_j \; ; \; j = 1, \ldots, m\}$ are the time-invariant coefficients of an m th-order linear predictor. The aim is to maximise the signal-to-noise ratio as given by

$$SNR = \frac{1}{Q(N)} \cdot \frac{E[s_i^2]}{E[e_i^2]} \quad , \tag{2.6}$$

where $Q(N) = E[q_i^2]/E[e_i^2]$, and $E[\cdot]$ denotes the expected value, or expectation.

The values of q_i are, in general, much smaller than the values of s_i, and for this reason, it is usual to neglect them in equation (2.5). This implies that if $\{s_i\}$ is a stationary process then the probability density function $p(e)$ of e_i is independent of i, and $E[Z_i^2]$ is also independent of i. It should be observed that the probabilities $P\{Z_i = \sigma_e y_j\}$ and $P\{\sigma_e x_j \leq e_i < \sigma_e x_{j+1}\}$ are identical, where $\sigma_e = \{E[e_i^2]\}^{\frac{1}{2}}$, and $\{y_j; \; j = 1, \ldots, N\}$ are the quantising levels for normalised input intervals $[x_j, \; x_{j+1})$, and hence, it can be seen that $K^2 = E[Z_i^2]/E[e_i^2]$ depends only on the p.d.f. of the normalised variable e_i/σ_e and the design of the quantiser. Assuming negligible correlation between e_i and q_i, this conclusion is also true for Q(N), since in this case it depends only on K^2. For Gaussian inputs it follows that e_i will also be Gaussian, irrespective of the choice of the predictor coefficients. Thus, K^2 is independent of the predictor and SNR can be maximised by designing a quantiser which minimises Q(N) and a predictor which minimises $E[e_i^2]$, separately.

Since 1/Q(N) represents the SNR in a PCM system, the gain achieved over PCM by the DPCM system will be given by

$$G = E[s_i^2]/E[e_i^2]. \tag{2.7}$$

The predictor characteristics result, by minimising $E[e_i^2]$, in the well-known " orthogonality principle" ,[38] i.e.,

$$E[(s_i - \sum_{j=1}^{m} a_j s_{i-j}) s_{i-k}] = 0, \tag{2.8}$$

$k = 1, \ldots, m.$

From this equation, the optimum coefficients are given by

$$\underline{A}_{opt} = \underline{R}^{-1} \underline{\rho} \tag{2.9}$$

and the minimum mean-square-prediction-error by

$$\sigma_{e_{min}}^2 = \sigma^2 - \underline{\rho}^t \underline{R}^{-1} \underline{\rho}\ , \tag{2.10}$$

where

$$\sigma^2 = E[s_i^2],$$

$$\underline{A} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \qquad \underline{\rho} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_m \end{bmatrix} \qquad \underline{R} = \begin{bmatrix} \rho_0 & \rho_1 \ldots \rho_{m-1} \\ \rho_1 & \rho_0 \ldots \rho_{m-2} \\ \vdots & \vdots \\ \rho_{m-1} & \rho_{m-2} \ldots \rho_0 \end{bmatrix}$$

and $\rho_{|i-j|} = E[s_i s_j]$.

At this point, it should be noted that DPCM coders are strongly dependent on the input signal statistics, which, for speech varies from speaker to speaker and from one type of sound to another. Nevertheless, adjacent amplitudes in speech are, in general, highly correlated and this makes DPCM an attractive waveform coding technique for this class of signals, even though the basic theoretical results relating to the overall maximisation of SNR, previously discussed, are not strictly valid for speech signals.

The gain G given by equation (2.7) has been computed by McDonald[36] as approximately 6 dB and 10 dB, respectively, for 1st and 2nd-order predictors, and has been shown to saturate at about this point. The performance with respect to one type of signal, such as speech, has also been shown to be poor[37] when the DPCM coder was designed for another class of signals.

In DPCM systems, the changes in input speech variance will be reflected in changes of prediction error variance. Since the latter is the input to the DPCM quantiser, adaptation techniques are also required in this case. The one-word memory adaptive quantisation of Jayant,[35] described previously in connection with APCM systems, was also used in adaptive DPCM. The difference is that a different set of multipliers $\{M(1), \ldots, M(2^n)\}$ should be obtained, since the quantiser input is now the prediction error sequence $\{e_i\}$ and not the speech amplitude sequence $\{s_i\}$. These multipliers have, in fact, been obtained by Jayant,[35] and the ADPCM coder has yielded 10 dB, 10 dB, and 7 dB gains over μ-law log-PCM,

when using 2, 3, and 4-bit quantisers, respectively. A more extensive study of this algorithm as applied to ADPCM coding of speech has been reported by Cummiskey, Jayant, and Flanagan.[39] Using band-pass (200-3200 Hz) speech as input, and sampling at 8 KHz, the multipliers obtained for a 3-bit Max[30] non-uniform quantiser were M(1) = M(2) = 7/8, M(3) = 5/4, and M(4) = 2. The resulting 24 Kbits/sec ADPCM coder, using a simple 1st-order predictor, has been found to be 8 dB better in SNR than a corresponding μ-law (μ = 100) companded PCM. The effectiveness of this system can be better assessed in terms of the quality of reconstructed speech. It has, in fact, proved quite successful[39] in this respect, with, for example, a 3-bit ADPCM giving better results than 5-bit log-PCM. Subjective tests[39] have in general indicated a better performance of ADPCM coders than the one predicted by the SNR objective measurements. These findings are summarised in Table 2.1.

Adaptive DPCM, thus, represents an attractive technique of speech coding at bit-rates between 24 Kbits/sec and 32 Kbits/sec. The technique is simple, and at these bit-rates good quality speech (better than 5-bit log-PCM quality) is obtained. The method, is, however, not useful for lower transmission bit-rates, except for special applications (e.g., mobile radio) at about 16 Kbits/sec.

Table 2.1. Objective and Subjective Performance of ADPCM versus Log-PCM

Objective Rating (SNR)		Subjective Rating (Preference)
7- bit PCM	(high)	7-bit PCM
6-bit PCM		4-bit ADPCM
4-bit ADPCM		6-bit PCM
5-bit PCM		3-bit ADPCM
3-bit ADPCM		5-bit PCM
4-bit PCM	(low)	4-bit PCM

2.1.3. DM and ADM Coders.

Delta-modulation (DM)[40,41] is another type of waveform coding, in which the quantisation process has been reduced to its simplest form. In DM only two quantisation levels are used. The basic idea is to approximate the input signal $\{s_i\}$ by a staircase function $\{\hat{s}_i\}$ and to transmit information only about the sign of the difference $\{e_i\}$ between $\{s_i\}$ and $\{\hat{s}_i\}$. The approximation $\{\hat{s}_i\}$ is obtained by means of a local decoder which consists, in its simplest configuration, of a perfect integrator inserted in a feedback loop around the quantiser, suggesting that delta-modulation may be regarded as a DPCM system making use of a very coarse quantisation. The one-bit quantisation is somewhat compensated for by having a sampling rate, f_s, which is

considerably higher than the Nyquist rate. At the receiver, $\{\hat{s}_i\}$ is regenerated from the transmitted symbols $\{b_i\}$ and low-pass filtered to obtain the final decoded signal. It should be observed that the transmission bit-rate is identical to the sampling rate, f_s, since each symbol uses only 1 bit per sample.

The block diagram shown in Figure 2.2 represents a single-integration linear delta-modulation (LDM) system, where Δ is the quantiser step-size, assumed to be a constant in non-adaptive DM. Under idling conditions (zero input level) the approximation $\{\hat{s}_i\}$ is a series of alternating +1 and -1 pulses, which, however, exhibit no increase in magnitude and no instability arises. This is not true for the case of double integration LDM, where a second integration is inserted in the feedback loop in order to obtain an approximation signal which will be a closer representation of the input. In this case, each symbol b_i changes the slope rather than the amplitude of $\hat{s}_i$. In the idling condition however, depending on initial conditions, the oscillation period of $\{\hat{s}_i\}$ can assume different values, and a d.c. component is generally present.

The types of encoding error in DM systems are known[41] as 'granular' noise and 'slope-overload' distortion. The first is inherent to the quantisation process, and occurs when the approximation function is correctly tracking the input signal but oscillates around it due to the discrete amplitude changes of $\{\hat{s}_i\}$. In this case the error is within the range of the step-size value. The second type of error, the slope-overload distortion, is characterised by the inability

of the approximation function to track the input signal in situations where the slope of the input is greater than the slope of the approximation. In this case the error is greater than the step-size value.

For a fixed sampling frequency and a constant value of input slope variance there is a point at which the SNR is maximum and the corresponding step-size is optimum. However, if the input slope variance changes, the performance will deteriorate and a change in step-size will then be required in order to bring the performance closer to the optimum. The idea is to increase the value of the step-size when slope-overload distortion is present, and to decrease step-size in situations in which the distortion is of the granular-noise type. This basic idea gives rise to a number of different types of adaptive delta-modulators (ADM), which use either syllabic (long-term) adaptation, such as continuous DM[43] and digitally controlled DM,[44] or instantaneous (short-term) adaptation, such as discrete adaptive DM[42] and one-bit memory DM.[45]

In continuous delta-modulation[43] the rectified derivative of the band-limited (200-3400 Hz) input speech, s(t), is low-pass filtered at 100 Hz producing a signal d(t) which is added to s(t) as shown in Figure 2.3. The transmitted binary symbols $\{b_i\}$ are controlled in amplitude by a signal, c(t), which is obtained by low-pass filtering $\{b_i\}$ at 100 Hz, and represents the mean number m of 1's in $\{b_i\}$. For high input signals m is set equal to 1/2 and for very small input signals to 1/3 (i.e., 1.5 times lower). The control c(t) is then used

to change the amplitude of the pulses $\{p_i\}$ (before integration) by a factor of 20, resulting in a change of the d.c. voltage of $\hat{s}(t)$ by a factor of 30. Thus, $\hat{s}(t)$ attempts to follow s(t) and a slowly varying signal which is related to the mean changes in input derivative. At 40 Kbits/sec continuous DM is reported[43] to give very good performance, and a lower but still fair quality of reproduced speech is obtained at 16 Kbits/sec.

Digitially controlled delta-modulation (DCDM), another technique which uses syllabic adaptation, has also been judged[44] to give fairly good results at 16 Kbits/sec. The adaptation logic consists of the following operations. A threshold circuit generates a signal $V = V_o$ as long as there are four or more consecutive bits of same polarity in the coder output, otherwise V is set equal to zero. This signal passes through an integrator with 5 msec time constant, resulting in a control signal V_c which varies between zero and V_{cm}, and is related to the mean value of the input signal derivative. This control signal, as shown in Figure 2.4, is finally used to provide the adaptation of the step-size Δ between a minimum value, fixed by means of δV_{cm} (a small constant added to V_c), and a maximum value set to Δ_m. The SNR performance at 56 Kbits/sec is better than log-PCM. At 32 Kbits/sec the performance of DCDM is still very good as indicated[44] by the CCITT specification (for good voice transmission by telephone) achieved by this technique. Even at 7.2 Kbits/sec, intelligible, though quite noisy, reproduction of speech can be attained with this method. At bit-rates below 7.2 Kbits/sec, however, the DCDM technique does not

provide satisfactory performance in terms of speech quality and intelligibility.

Because the variation in step-size is based on the long-term characteristics of the input signal derivative, syllabic adaptation techniques are quite resistant to bit errors. Instantaneous adaptation methods, on the other hand, are generally of simpler structures. One of these methods has been proposed by Abate[42] and is illustrated in Figure 2.5. In this system, known as discrete adaptive DM, the step-size Δ is chosen from a discrete set $\{\Delta_o, G_2\Delta_o, G_3\Delta_o, \ldots, G_m\Delta_o\}$, where $G_i < G_{i+1}$ for all i. The idea is that if the present and previous bit in the output of the quantiser are of same polarity, this is interpreted as slope-overload condition, and the previous gain G_i is switched now to the first greater neighbour G_{i+1}; otherwise it is switched to G_{i-1}. The same principle is used in the decoding process. This scheme has been simulated with different selections of gains. Particularly, for $G_i = i$, $i = 1, \ldots, 8$, the SNR performance of this system, when using a sampling frequency equal to four times the Nyquist rate, has been determined to be nearly the same as the one obtained with a 4-bit log-PCM. Thus, this scheme proves to be inferior to the DCDM technique, since at 32 Kbits/sec DCDM can provide better results than log-PCM.

Another important suggestion in ADM techniques consists of a very simple scheme using instantaneous adaptation which is referred to as one-bit memory DM.[45] The final part of this section will be devoted to a brief description of this system and the main results achieved for speech signal input.

The step-size adaptation logic in one-bit memory DM is based on the two most recent channel symbols b_i and b_{i-1}, as shown in Figure 2.6, and defined by

$$\Delta_i = P.\Delta_{i-1} \text{ if } b_i = b_{i-1},$$
$$= -Q.\Delta_{i-1} \text{ if } b_i \neq b_{i-1}. \quad (2.11)$$

In (2.11), Δ_i and Δ_{i-1} are present and past step-size values, respectively, and P and Q are parameters of the adaptation rule, with the requirement that $P > 1$ to deal with slope-overload, and $Q < 1$ to deal with granular noise. It has been verified[45] that the condition $P.Q = 1$ is the condition of optimality with regard to SNR when dealing with speech signals, and that $P = 1.5$ maximises the SNR for sampling frequencies of 20, 40, and 60 KHz.

At 60 Kbits/sec an improvement of more than 10 dB is achieved in the maximum SNR over linear delta- modulation. At bit-rates of about 40 Kbits/sec and below, the performance of one-bit memory DM is better than that achieved by log-PCM, but at 20 Kbits/sec a noticeable degradation in quality is observed, although satisfactorily intelligible speech can be attained with this very simple coding structure.

In a separate study[46] it has been observed that the value of P which gives perceptually better speech in the one-bit memory DM scheme is $P = 1.2$, which is different from the value $P = 1.5$ based on SNR performance. Experiments have been carried out at 20 and 40 Kbits/sec

and a further interesting conclusion has also been drawn. The indications are that even though the contribution of slope-overload distortion to the total noise power is significantly greater than that of granular noise, the latter seems to be more important with respect to the resulting subjective quality of speech signals.

As can be observed from the multiplicity of techniques which have appeared in recent years, the effectiveness of waveform coding techniques is restricted, in general, to bit-rates above 16 Kbits/sec. The disadvantages of high-bit-rates are, however, offset by the fact that at these higher bit-rates the waveform coded speech is much less vulnerable to different types of degradations, such as digital transmission errors.- Furthermore, they are very attractive in terms of the simplicity of their structure.

2.2. Sub-Band Coding.

In sub-band coding (SBC)[8,47,48] different bands of the speech frequency spectrum are separately coded, such that the effects of the quantisation distortion can be more appropriately controlled according to perceptual characteristics of the considered band. One possibility concerning the choice of these bands is to separate the speech spectrum into bands which make equal contributions to the articulation index.

A block diagram of a sub-band coder using low-pass translation technique is shown in Figure 2.7. The speech signal spectrum is partitioned into sub-bands (f_{1n}, f_{2n}), n = 1, ..., N, which are then modulated by $\cos(2\pi f_{1n} t)$ and low-pass filtered at f_n Hz, where $f_n = f_{2n} - f_{1n}$. Each sub-band signal is sampled at a rate of $2f_n$, n = 1, ..., N, and coded by an APCM system using the one-word memory adaptation rule[35] described in section 2.1. A number, N, of digital signals are obtained in this manner, and multiplexed for transmission. At the receiver, the N signals are demultiplexed, demodulated to their original bands, and band-limited. The approximation of the original signal is finally obtained by adding these N resulting sub-bands.

A more efficient implementation of sub-band coding in terms of complexity, uses the technique of integer-band sampling which is illustrated in Figure 2.8. Each sub-band n is obtained by band-pass filtering, and is located between mf_n Hz and $(m + 1)f_n$ Hz, where m is an integer. Then, a sampling at a rate of $2f_n$ is used, and the regeneration of the original sub-band is obtained by another band-pass filtering to the range $[mf_n, (m + 1)f_n]$. The advantage is that no modulators are necessary.

In terms of performance, as judged by informal listening tests,[8] the sub-band coder has been considered to be better than ADPCM at 16 Kbits/sec. The results are slightly worse for SBC, however, if the ADPCM bit-rate is increased to 24 Kbits/sec. Comparisons of speech quality between 9.6 Kbits/sec SBC and 10.3, 12.9 and 17.2 Kbits/sec ADM have also shown superior performance for the sub-band coder. A small

reduction in quality has been observed when the bit-rate is decreased from 9.6 to 7.2 Kbits/sec, and a much inferior speech reproduction results at 4.8 Kbits/sec. A more extensive set of subjective tests[47] has also indicated that SBC at 9.6, 16 and 24 Kbits/sec can give results which are similar to ADPCM with variable predictor at bit-rates of 9.6, 18, and 26 Kibts/sec, respectively, despite of the higher complexity of the latter, which requires the computation of auto-correlation coefficients and the solution of a set of linear equations every adaptation period.

Sub-band coding, thus, can give acceptable results at bit-rates greater than 16 Kbits/sec. This technique is more complex than waveform coding schemes (e.g., ADM), due mainly to the filter bank required in SBC. In addition, it does not provide acceptable speech quality at bit-rates below 16 Kbits/sec, and at 4.8 Kbits/sec a poor reproduction of speech is obtained with this technique.

2.3. Adaptive Predictive Coding.

Adaptive predictive coding (APC)[9] is a refined form of differential pulse-code-modulation, in which the characteristics of the speech production mechanism are exploited to a large extent. The idea arises by considering an approximate model of voiced speech production as the output $\{\acute{s}_i\}$ of a discrete linear time-varying filter with a quasi-periodic (period M) pulse train input $\{u_i\}$, with the amplitude adjusted from period to period according to the rule

$$u_i = \beta\, u_{i-M}\,, \tag{2.12}$$

where β is an amplitude adjusting parameter. Each sample s_i is then represented in this model by the following expression:

$$s_i = \sum_{j=1}^{p} a_j \, s_{i-j} + \beta \, u_{i-M} , \tag{2.13}$$

where $\{a_j;\ j = 1, \ldots, p\}$ are the coefficients of the filter, which are related to the short-time spectral envelope of the speech signal, assumed to be time-invariant in the duration of a pitch period (excitation period M). This implies, together with equation (2.13), that

$$s_i = \beta \, s_{i-M} + \sum_{j=1}^{p} a_j (s_{i-j} - \beta \, s_{i-j-M}), \tag{2.14}$$

an expression which characterises the prediction P(Z) used with this method. As shown in Figure 2.9, this predictor can be represented by a combination of a predictor $P_1(Z)$ given by

$$P_1(Z) = \beta Z^{-M} , \tag{2.15}$$

and a predictor $P_2(Z)$ given by

$$P_2(Z) = \sum_{j=1}^{p} a_j Z^{-j}, \tag{2.16}$$

where the first exploits the quasi-periodicity of voiced speech sounds, and the second exploits the non-uniform characteristics of the short-time spectral envelope, which is determined by the vocal tract configuration. From (2.14), (2.15) and (2.16), it can also be seen that $[1 - P(Z)]$ can be obtained by a serial combination of $[1 - P_1(Z)]$ and $[1 - P_2(Z)]$, i.e.,

$$[1 - P(Z)] = [1 - P_1(Z)][1 - P_2(Z)] \tag{2.17}$$

The approximate model of voiced speech production is changed only with respect to the excitation function $\{u_i\}$ to obtain the unvoiced speech production model. In this case, $\{u_i\}$ is a noise-like excitation, implying that only the non-uniformity of the short-time spectral envelope is exploited and $P(Z) = P_2(Z)$.

Parameters β and M are obtained by minimising the mean-square-error between $\{s_i\}$ and $\{\beta s_{i-M}\}$. It can be shown[9] that

$$\beta = E[s_i s_{i-M}]/E[s_i^2 - M], \tag{2.18}$$

and $M(> 0)$ is given by the value which maximises

$$\rho = E[s_i s_{i-M}]/\{E[s_i^2]E[s_i^2 - M]\}^{\frac{1}{2}} \tag{2.19}$$

where the averages are taken over 5 msec intervals of speech. The parameters $\{a_j;\ j = 1, \ldots, p\}$ are obtained by minimising the total mean-square prediction error with the values of β and M as previously computed. The resulting expression is similar to the one obtained for DPCM [see equation (2.9)].

In a simulation of the APC technique,[9] M was chosen between 20 and 150, and p set to 8. Using a sampling rate of 6.67 Ksamples/sec for speech signals low-pass filtered at 3.1 KHz, the system has shown a superior performance when compared with 5-bit log-PCM. In this simulation 1-bit adaptive quantisation was employed, and the transmission bit-rate estimated to be approximately 10 Kbits/sec. As was seen in section 2.1, similar results were obtained with a simple ADPCM scheme, but at bit-rates more than twice as high. Adaptive

predictive coding can give fairly good regeneration of speech using transmission bit-rates as low as 7.2 Kbits/sec. The system complexity is, however, considerably higher, and for bit-rates below 7.2 Kbits/sec it does not provide satisfactory results in terms of speech quality.

In a recent paper by Atal and Schroeder[49] a further improvement in the performance of the APC is achieved by exploiting perceptual characteristics of speech. The idea arises from the observation that most of the perceived noise originates from those parts of the spectrum associated with low-signal levels. From a perception point of view, noise in regions of high-signal level (formants) is, in turn, masked by the speech signal . This implies that the control of the noise spectrum appears as an efficient approach to improving the coder performance.

Referring to equation (2.5), it can be seen that the last term represents the quantising noise filtered through a filter P defined by the coefficients $\{a_j\}$. If a different filter F is used then the shape of the output noise spectrum can be appropriately controlled. It has been found[49] that a good selection of F is given by

$$F = P(\alpha\, Z^{-1}),\ 0 < \alpha < 1, \tag{2.20}$$

which allows reduction of the noise in frequency regions associated with low-signal levels at the expense of a corresponding increase in the formant regions. The application of this technique to APC provides a better quality of reconstructed speech. However, it makes the system still more complex, and for low-bit-rates (e.g., 4.8 Kbits/sec) the adaptive predictive coder with adjustable noise spectrum does not give acceptable speech quality. The system is shown in Figure 2.10.

2.4. Vobanc and Analytic Rooter.

In this section two techniques which are designed to achieve a 2-to-1 reduction in the bandwidth of speech by using frequency division of formant bands will be illustrated briefly.

One of these techniques, known as Vobanc, was proposed in 1956 by Bogert.[50] In this system, the input speech is first separated into three bands, each normally containing one formant region, by means of three band-pass filters denoted as " A" filters. The strongest components in each of these bands are then halved, and the adjacent components translated to lower frequencies correspondingly, by using regenerative modulators, where the carrier is derived from the modulator output. By passing these signals through filters (denoted as " B" filters) with bandwidths equal to one-half of the corresponding " A" filters, the strongest components are then transmitted using only one-half of the original band. At the receiver, the combined signal is separated again, by " B" filters, into three bands which are 'frequency doubled' and recombined.

The Vobanc is a fairly complex system which does not give a significant saving in transmission bandwidth. The technique can achieve a 2-to-1 reduction in transmission bandwidth, and some improvement[50] in speech intelligibility when compared with the unprocessed band-limited (200-1700 Hz) speech.

In 1967 a method of frequency division was proposed,[51] which is called an " Analytic Rooter" . The approach is based on the fact that frequency division of a signal s(t) can be obtained by the rooting of the analytic signal

$$\sigma(t) = s(t) + j\hat{s}(t) \ , \tag{2.21}$$

where $\hat{s}(t)$ is the Hilbert transform of $s(t)$ (phase shift of $\pi/2$). This can be easily seen by writing $\sigma(t)$ in the form

$$\sigma(t) = a(t)\ e^{j\phi(t)} \ , \tag{2.22}$$

and defining $s_{1/n}(t)$ by

$$s_{1/n}(t) = \mathrm{Re}\{[\sigma(t)]^{1/n}\} = [a(t)]^{1/n}\cos[\phi(t)/n], \tag{2.23}$$

where

$$a(t) = [s^2(t) + \hat{s}^2(t)]^{1/2}. \tag{2.24}$$

From these equations it can be shown[51] that for $n = 2$, $s_{1/2}(t)$ is given by

$$s_{1/2}(t) = (1/2)^{1/2}[a(t) + s(t)]^{1/2}, \tag{2.25}$$

and since $s_{1/2}(t)$ is related to $\cos[\phi(t)/2]$, the sign in this equation should be changed when $\phi(t) = 2k\pi$ with k integer.

Equations (2.24) and (2.25) represent the operations used for frequency dividing by two in the analytic rooter. The corresponding frequency doubling at the receiver is obtained[51] by

$$s(t) = s^2_{1/2}(t) - \hat{s}^2_{1/2}(t). \tag{2.26}$$

The number of frequency bands used in this case is equal to four. Certain degradations found in the Vobanc processed speech are much reduced by this method.

Both the Vobanc and the analytic rooter avoid the problem of pitch extraction found in vocoder-type techniques (see sections on vocoders), but still remain fairly complex, and only modest savings in transmission bandwidth are achieved by these schemes.

2.5. Channel Vocoder.

The origin of a large number of important developments in the low-bit-rate coding of speech signals dates from the discovery by Dudley,[4,5] more than 40 years ago, of the Vocoder. This is a system which exploits the major features of the speech production and perception mechanisms.

The basic idea involved in this system, which is nowadays referred to as the 'channel vocoder', is to use a simplified model of speech production, in which the vocal source and vocal tract are assumed as two separate systems. The excitation is represented either by a quasi-periodic sequence of pulses (voiced source), or by a noise-like signal (unvoiced excitation), which drives a time-varying linear filter representing the slowly varying characteristics of the vocal tract transfer function. Since speech is non-stationary, both the excitation and the spectral envelope determined by the vocal tract shape vary with time, which implies that the analysis must be performed on a short-time basis.

In the channel vocoder, the vocal tract spectral envelope is estimated by a bank of N band-pass filters covering the speech frequency spectrum. The output of each band-pass filter is rectified and low-pass

filtered, which results in the average amplitude of each band. The extraction of source information is governed by an algorithm which decides between unvoiced or voiced excitation, and which, in the case of the latter, also provides an estimate of the period of the vocal fold vibrations (the pitch period). At the receiver, these excitation signals are generated and used as input to a bank of N band-pass filters (see Figure 2.11) whose outputs are amplitude modulated by the corresponding channel signals, and recombined to form the synthesised speech.

The early vocoder of Dudley [4,5] consisted of a 10-channel system with one filter covering the range from 0 to 250 Hz and the other nine, each of 300 Hz bandwidth, covering the frequency range from 250 Hz to 2950 Hz. Low-pass filters, each with a 25 Hz cutoff have been used for both the spectrum channel signals and the pitch signal. In another implementation, reported on in a 1956 paper by Vilbig and Haase,[52] 100 filters operating at high frequencies (130 to 133 KHz) have been used in the analysis, and scanned at a rate of 30 sec^{-1} for transmission by time multiplexing. In that same year, consonant articulation scores were reported[53] for channel vocoders using digital transmission of the channel signals. Some of these results are shown in Table 2.2 for different numbers of channels and different bit-rates (not including the pitch channel). From the table it can be seen that the variation in the scores is quite small over the range of bit-rates. In this same study a 90% articulation score was attained by increasing the number of channels to 16, and by using a logarithmic spacing of

filters above 1 KHz. Another implementation[54] considered the cases of 14 and 18-channel signals at total bit-rates of 1200 and 2400 bits/sec.

Although a highly intelligible speech can be reproduced with a vocoder operating at low-bit-rates (e.g., 2.4 Kbits/sec),[55] the quality is noticeably deteriorated in most situations.

Table 2.2. Consonant Articulation Scores of Channel Vocoders.

No. of Channels N	Bit-rate (bits/sec) without including pitch channel	Consonant Articulation Score (%)
10	1300	82
10	1160	79
10	1000	79
8	1040	79
6	780	79
5	650	78

A considerable inprovement in this respect has been obtained by using a special configuration of the synthesiser which is known as spectral flattening.[55-57] This is illustrated in Figure 2.12. The idea is to insert a hard limiter between the excitation and the modulation, such that square-waves with amplitudes +1 and -1 are created at the input to all the modulators. Improvements in quality have been

explained by the experimental observations that rapid pitch changes produce fluctuations on the speech spectral envelope and result in audible distortion of synthetic speech.

Some work on channel vocoders has been directed towards reducing the possible remaining redundancy in the representation of channel signals and, thus, providing a means for transmission at lower bit-rates. One technique[58] transmits only a small number, m, of channels which represent local maxima. Implementations using an 18-channel vocoder with m equal to 3, 4, or 5, and sampling rates of 30 or 60 sec^{-1}, resulted in high sentence intelligibility (60% or 70% nonsense syllable articulation) at bit-rates as low as 1000 bits/sec. Another method, proposed by Kramer and Mathews,[59] considers the use of the Karhunen-Loève transformation (minimum mean-square error) in order to reduce the correlation between adjacent channels. The transformed signals consist of $m < N$ signals, where N is the total number of channels. It has been found[59] that for $N = 16$ and $m = 10$ the quality of synthesised speech is superior to that obtained with a direct 10-channel vocoder. A similar approach, proposed some 10 years later,[60] considers a simpler realisation where a 16-point Hadamard transform is applied to the logarithm of the channel signals using a sampling rate equal to 50 sec^{-1} and different number of bits for quantisation of the transformed signals. At the resulting bit-rate of 1650 bits/sec the speech quality has been found to be comparable to a direct coding of the channel signals at 4 Kbits/sec, thus providing a considerable reduction in bit-rate.

A somewhat simpler structure of the channel vocoder synthesiser has been proposed[61] which is known as vocal-response synthesiser. It avoids the use of band-pass filters, and gives rise to the possibility of microminiaturising the synthesiser. With regard to digital implementation, one possible approach is to use a bank of digital filters, with design characteristics as found in reference 62.

Although channel vocoders are capable of reproducing speech signals having high intelligibility whilst operating at very low transmission bit-rates, the filtering (or complex digital signal processing) necessary to obtain the narrow-band spectral information involves a high implementation cost. The complex problem of pitch extraction, which is general to all vocoder-type techniques, also contributes substantially to the high cost of these techniques. The pitch problem is also, at least partially, responsible for the speaker-dependent quality of the processed speech and, in general, speech of commercial telephone quality standard is not achieved.

2.6. Formant Vocoder.

As was seen in the previous section, spectral envelope information is obtained in channel vocoders by means of a bank of contiguous filters, rectifiers, and smoothers, the outputs of which, in each case, represent the average amplitude of the signals in each band. However, the characteristics of the vocal tract can be approximately specified by its spectral peaks (or formants),[3] which correspond to the poles of the vocal tract transmission function. For non-nasalised voiced speech,

for instance, only three formants are usually found below 3 KHz. Thus, the use of formant information only represents a more efficient way of transmitting the major characteristics of the spectrum envelope, since a reduced number of control signals is used.

The above considerations are, in fact, the basis of one of the most important low-bit-rate schemes of speech coding. This is called the formant vocoder. One of the earliest forms of the formant vocoder appeared in 1950. This used a configuration in which formant measurements were used as the average number of zero-crossings in each of three bands.[63] In this early implementation, the energy in each of these bands, the pitch frequency, and the energy in the band 40-400 Hz was also extracted in the analysis. The synthesiser used three resonant networks with controlled peak frequencies, excited by one of the two source signals (a pulse train or a wide-band noise), and modulated by the energy information signals. The outputs were then combined to produce the synthetic speech. Approximate intelligibility scores of 100% for vowels and 70% for consonants have been reported.[3] The formant measurements used in this early system were, however, inaccurate, due mainly to the difficulty associated with overlapping of formant frequency ranges. For non-nasalised voiced speech, and a male speaker, typical ranges for the first three formants are 150-850 Hz, 500-2500 Hz, and 1700-3500 Hz, respectively.[2]

The configuration of formant (or "terminal-analog") synthesisers is quite varied, depending on the degree of refinements included. In an early work published by Lawrence[64] in 1953 the synthesis of speech

from three formant frequencies and an excitation function was considered, and resulted in good intelligibility of synthesised speech. The information rate was estimated to be 550 bits/sec.

Formant synthesisers can be classified as series or parallel, depending on their mode of connection. In a 1957 paper by Flanagan[65] a study is presented of these two types of synthesiser. The series synthesiser has the advantage of automatically adjusting the relative formant amplitudes and of reproducing non-nasalised vowels better than parallel connection. The latter, however, introduces zeros into the transfer function, making it more appropriate for reproducing consonant sounds.

A more complex configuration of the formant synthesiser has been simulated in a digital computer.[66] It uses one branch to generate voiced sounds, and one for unvoiced sounds. The first contains three variable resonators for formants F1, F2, F3, a fixed fourth formant F4, and a pole-zero pair for nasalised speech. The latter consists of one or two poles and one zero. A more sophisticated version was later proposed by Rabiner.[67] It included a new design for the higher pole correction network, an amplitude control to shaped pitch-impulses with the purpose to obtain low-frequency low-energy waveforms for the closure interval of voiced stops, and a network to model the generation of the unvoiced component of voiced fricatives. Good quality speech was claimed to be obtained with a digital hardware implementation using these same principles.[68] More recently, Holmes[69] has shown, from experiments with a parallel formant synthesiser, that by using a synthetic glottal pulse obtained by inverse filtering a typical vowel from the same talker, synthetic and natural speech are very similar.

The analysis problem in formant vocoders has proved to be a complex task. This is due mainly to the difficulty involved in making accurate formant measurements. As has previously been mentioned, formant frequency ranges overlap, and also there is the possibility of masking or elimination of a pole by a source zero or a side branch (such as nasal cavity) zero.

Many techniques of formant analysis have been developed, with variations in terms of accuracy and complexity. Comments have already been made about the zero-crossings method[63] presented in 1950. Two other methods have since been proposed by Flanagan.[70] They involve peak-picking of the short-time amplitude spectrum. In one of them, a bank of 36 band-pass filters is used, and the maxima are picked at points of zero derivative. In the other, the spectrum is segmented into formant ranges, and the frequency of the maximum in each segment is measured. Although this method has the disadvantage of limiting formant frequencies to fixed ranges, it is simpler, and has been found to provide more accurate measurements.[71] In both cases, the highest error rate occurred for nasalised sounds.

Another technique of formant analysis uses a digital computer to analyse the rectified and smoothed outputs from a bank of 24 band-pass filters.[72] The method first locates a peak in the whole spectrum. Then, a resonance curve of appropriate frequency and average bandwidth is subtracted, and the process is repeated for the next formant, thus, avoiding the problem of fixed segmentation.

A more recent analysis approach considers the use of digital inverse filtering.[73] The results show that 90% of the time, the first three peaks in the reciprocal of the inverse filter spectrum define the first three formants. However, the algorithm involves rather complex operations such as the computation of the autocorrelation sequence, and the solution of a set of autocorrelation equations.

Transmission bit-rates less than 1 Kbit/sec have been reported in the literature[74] for a formant vocoder using three formants only, and the effect of quantisation did not have significant influence on the synthetic speech when compared to the analogue mode of operation of the vocoder. A further formant vocoder system has also been implemented, which uses analogue equipment for synthesis and spectrum filtering operations, and a digital computer for formant analysis.[75] Good performance has resulted.[3]

As can be observed, formant vocoders provide a very efficient description of speech information. They are, perhaps, the most effective means for obtaining very low-bit-rate coding of speech signals, but, on the other hand, the complexity, resulting particularly from the sophisticated analysis procedure involved is even higher than for the channel vocoders. In addition, formant vocoded speech is not of commercial quality and, as in the case of channel vocoders, the problem of pitch extraction also contributes to the general complexity of the system.

2.7. Linear Prediction Vocoder.

An alternative method of extracting the spectral envelope of speech makes use of linear prediction techniques. The method uses an all-pole model of the vocal tract, in which the effects of zeros are approximated by multiple poles, and are included in the representation by a single recursive time-varying filter with p coefficients given by

$$H(Z) = \frac{1}{1 + \sum_{j=1}^{p} a_j Z^{-j}} \quad . \tag{2.27}$$

A time-domain representation is shown in Figure 2.13, where the speech sample s_i is obtained as a linear combination of past values and an excitation e_i, i.e.,

$$s_i = - \sum_{j=1}^{p} a_j s_{i-j} + e_i \tag{2.28}$$

The first term in this equation is the predicted sample $\hat{s}_i$, and the error with respect to the actual value, which is also known as the residual, is exactly the excitation function of the model. The coefficients $\{a_j;\ j = 1, \ldots, p\}$ are obtained by minimising

$$\alpha = \sum_{i=n_o}^{n_1} e_i^2 \ , \tag{2.29}$$

which leads to the set of linear equations

$$\sum_{i=1}^{p} a_i c_{ij} = -c_{oj} \ , \ j = 1, \ldots, p, \tag{2.30}$$

where

$$c_{ij} = \sum_{n=n_o}^{n_1} s_{n-i}\, s_{n-j}. \tag{2.31}$$

In the linear prediction vocoder (LPC vocoder) the time-varying filter is driven, as in channel or formant vocoders, by a series of quasi-periodic pulses in the case of voiced sounds, or a noise-like excitation in the case of unvoiced sounds. Thus, the LPC vocoder and the APC technique are, in their basic features, very similar schemes with the main difference being that the residual in the APC is replaced by noise or pulse excitation in the LPC vocoder.

Linear prediction as applied to the analysis of discrete signals, with special attention being given to all-pole modelling, is presented in a review paper by Makhoul.[76] Some historical notes can be found in a work by Sorenson,[77] and the major aspects of the theory and application of linear prediction techniques to speech signals are contained in a book published in 1976 by Markel and Gray.[78] In this section the aim is to present briefly some of the important developments achieved with this approach.

2.7.1. The Covariance and Autocorrelation Methods.

In the covariance method of linear prediction, the error α in equation (2.29) is minimised over a finite interval defined by $n_0 = p$ and $n_1 = N - 1$, where N is the length of the speech sample sequence. The method consists in solving[78] (2.30) with error e_n, $n = p, p + 1, \ldots, N - 1$, given by

$$e_n = \sum_{i=0}^{p} a_i s_{n-i}, \quad a_0 = 1. \qquad (2.32).$$

The autocorrelation method, on the other hand, assumes a minimisation over the infinite interval ($n_0 = -\infty$, $n_1 = \infty$), and defines $s_i = 0$ outside [0, N - 1], so that[78]

$$c_{ij} = \sum_{n=0}^{N-1-|i-j|} s_n s_{n+|i-j|} \quad . \tag{2.33}$$

Equation (2.30) then becomes

$$\sum_{i=1}^{p} a_i \, \rho_{|i-j|} = -\rho_j \, , \; j = 1, \ldots, p, \tag{2.34}$$

where

$$\rho_j = \sum_{n=0}^{N-i-j} s_n s_{n+j}, \quad j \geq 0, \tag{2.35}$$

with the error e_n, $n = 0, \ldots, N + p - 1$, being given by (2.32). The elements $\rho_{|i-j|}$ in equation (2.34) form a symmetric matrix with equal values along each diagonal (i.e., a Toeplitz matrix[76]).

2.7.2. Maximum Likelihood Vocoder.

The first linear prediction vocoder appeared in the literature in 1968, in a work by Itakura and Saito.[79] In this scheme an all-pole model is assumed, and the sequence of speech samples $\underline{s} = \{s_0, s_1, \ldots s_{N-1}\}$ is considered as samples of a random Gaussian process, with a joint probability density function which is a function of the set of parameters $\underline{\omega} = \{\sigma_e^2, a_1, a_2, \ldots, a_p\}$, where σ_e^2 is a gain factor for matching the speech spectral density. Denoting $p(\underline{s}|\underline{\omega})$ as the

probability density of $\underline{s}$ for a given set $\underline{\omega}$, the maximum likelihood method states that the value of $\underline{\omega}$ which makes the observation $\underline{s}$ most likely is the one which maximises $p(\underline{s}|\underline{\omega})$. This formulation, using approximations for N >> p leads to results identical to those obtained with the autocorrelation method.

In the maximum likelihood vocoder, pitch extraction is accomplished by a procedure of peak-picking the autocorrelation sequence of the error signal, which also controls the proportion of voiced excitation V. Unvoiced excitation is mixed with the voiced excitation in a proportion of $UV = \sqrt{1 - V^2}$, as shown in Figure 2.14, and the number of coefficients $\{a_j\}$ is set equal to 10.

At 5.4 Kbits/sec the maximum likelihood vocoder has been considered to perform better than digitised channel vocoders,[3] and also to provide a more accurate extraction of speech spectral envelope when compared with formant vocoders. The system, however, is still complex, despite the advantage of completely eliminating the need for filter banks for the analysis and synthesis of speech spectra. Also, the speech quality obtained with this technique does not satisfy commercial telephone quality standard, and the mixing of voiced and unvoiced excitation does not provide any significant improvement of the processed speech.[78]

2.7.3. PARCOR Vocoder.

Consider a sequence $\{s_{n-m-1}, s_{n-m}, \ldots, s_n\}$ of speech samples, and let $e_{fn}^{(m)}$ and $e_{bn}^{(m)}$ be, respectively, the forward and backward prediction errors at time n, given by

$$e_{fn}^{(m)} = s_n - \left(-\sum_{i=1}^{m} a_{mi}\, s_{n-i}\right), \tag{2.36}$$

and

$$e_{bn}^{(m)} = s_{n-m-1} - \left(-\sum_{i=1}^{m} b_{mi}\, s_{n-i}\right), \tag{2.37}$$

which can be regarded as the outputs of two filters, $A_m(Z)$ and $B_m(Z)$, with coefficients $\{a_{mi}\}$ and $\{b_{mi}\}$. The minimisation of the total squared errors for m = 1, 2, ..., p, leads to[78]

$$A_m(Z) = A_{m-1}(Z) - K_m B_{m-1}(Z), \tag{2.38}$$

and

$$B_m(Z) = Z^{-1}[B_{m-1}(Z) - K_m A_{m-1}(Z)], \tag{2.39}$$

where the coefficients

$$K_m = \frac{\sum_{n=-\infty}^{\infty} e_{fn}^{(m-1)}\, e_{bn}^{(m-1)}}{\left\{\sum_{n=-\infty}^{\infty}[e_{fn}^{(m-1)}]^2 \sum_{n=-\infty}^{\infty}[e_{bn}^{(m-1)}]^2\right\}^{1/2}} \tag{2.40}$$

are known as partial correlation (PARCOR) coefficients.[80] These coefficients are also the same set of intermediate parameters which appear in the solution of the autocorrelation method equations (differing only in sign), which are known as reflection coefficients,[78] and are given by $K_m' = a_{mm}$, with final solution for $\{a_j\}$ resulting in $\{a_j = a_{pj}\ ;$ $j = 1, \ldots, p\}$. The PARCOR coefficients have the property that if $|K_m| < 1$, m = 1, ..., p, then the stability of the linear prediction filter H(Z), as given by (2.27), is guaranteed.[78]

From (2.38) and (2.39) it can be seen that

$$e_{fn}^{(m)} = e_{fn}^{(m-1)} - K_m \, e_{bn}^{(m-1)} \tag{2.41}$$

and

$$e_{bn}^{(m)} = e_{b,\,n-1}^{(m-1)} - K_m \, e_{f,\,n-1}^{(m-1)} \tag{2.42}$$

where $e_{fn}^{(o)} = s_n$ and $e_{bn}^{(o)} = s_{n-1}$. These expressions define the lattice structure of the PARCOR vocoder analyser, another technique proposed by Itakura and Saito,[80] which is shown in Figure 2.15. The coefficients $\{K_i \; ; \; i = 1, \ldots, p\}$ are calculated by the cross-correlators COR, and the synthesiser is composed of p sections, with transfer function inverse of $A_p(Z)$. Listening tests indicated a minimum p equal to 8, and no improvement for p greater than 12, at a sampling rate of 8 Ksamples/sec.

The PARCOR formulation, like the recursive solution of the autocorrelation method, has the advantage of computing the transmission parameters (PARCOR coefficients) on a sample-by-sample basis. The system has also been considered[78] more attractive in terms of hardware implementation as compared to the recursive solution of the autocorrelation method, although less advantageous in terms of computer simulation. The general complexity of the system is still high, and the technique is not capable of giving good quality speech which is speaker-independent.

2.7.4. Other Developments in the LPC Vocoder Method of Speech Coding.

Many other developments on the subject of LPC vocoders have appeared in the literature in the last decade, and in the past few years it has actually become one of the most popular topics in the area of low-bit-rate speech coding.

A classical paper on the LPC vocoder,[81] published in 1971, used the covariance method of linear prediction, and presented the basic ideas contained in the LPC analysis - synthesis of speech signals. In this work, Atal and Hanauer found that p = 12 was the most desirable number of coefficients $\{a_j; j = 1, \ldots, p\}$ to be used. Direct quantisation of these coefficients, however, is not efficient because high accuracy is necessary to avoid possible large errors in the pole frequencies of the filter , which may cause it to become unstable. Other equivalent sets of parameters, such as the frequencies and bandwidths of the poles or the areas of an acoustic tube model of p cylindrical sections of equal length, have been found to be more efficient with respect to stability.[81]

In the above development, a total of 72 bits per frame are used, with 60 being used for the filter parameters. At 100, 67 and 33 frames/sec, corresponding bit-rates of 7.2, 4.8, and 2.4 Kbits/sec have been achieved, giving fairly good speech reproduction (though not of commercial quality), as indicated by the recorded utterances of synthesised speech which were enclosed with the paper of Atal and Hanauer. The analysis frame used in this scheme was adjusted to 10 msec for unvoiced sounds and one pitch period for voiced sounds. In a study of this system, reported two years later by Haskew et al.,[82] a constant analysis frame of N samples was considered and the results indicated that N = 128 was the best choice from the values of N = 64, 128 and 256. Frame rates above 200 sec^{-1} produced no difference in synthetic speech quality, and below 33 sec^{-1} resulted in rapid loss of quality. With respect to selection and quantisation of filter

parameters it has been found that the most suitable arrangement was the logarithmic quantisation of the ratios of adjacent areas of the acoustic tube model as previously mentioned. It has also been found that there is no significant difference in speech quality for the 3.6 and 7.2 Kbits/sec systems and, again, p = 12 was considered a good choice, although there were talker and text sensitivities.

In an attempt to attain a better quality speech from the LPC vocoder, a technique which uses a non-impulsive source has been suggested[83] and has been found to be important in effectively masking the characteristic buzz existing in LPC synthetic speech. Another work,[84] directed towards obtaining a more efficient transmission by means of DPCM encoding of the log-area ratios, has resulted in bit-rates of 2 Kbits/sec and lower. In this same work, the suggestion of LPC vocoder with orthogonalised coefficients was also made. The idea exploits the redundancy still existing in the coefficients by means of the Karhunen-Loève transformation of the log-area ratio parameters. The transmission of only the 6 most significant orthogonal coefficients out of 12 has been observed[84] to be sufficient to produce synthetic speech without significant degradation in quality. This quality is however, talker-dependent and naturalness is, in general, not well preserved. Associated with DPCM encoding, bit-rates as low as 600 bits/sec (a function of utterance and talker) have been achieved. However, this low bit-rate is obtained by a considerable increase in system complexity, which is introduced by the eigenvector analysis for the computation of the orthogonal coefficients.

In a study of intelligibility and consonant attributes of LPC vocoded speech[85] it has been found that the difference between real and 12-predictor coefficient LPC vocoded speech is most apparent for the feature sustention which distinguishes interrupted from continuous consonants (such as pence from fence, and choose from shoes). Modifications to improve this feature, without any significant cost in respect of the bit-rate have been proposed.[86] They consist of using only a reduced number of coefficients to code unvoiced sounds combined with a faster unvoiced frame updating to better preserve the onset of energy (gradual or abrupt).

A study has been carried out[85] in order to evaluate the performance of LPC and formant coded speech in terms of intelligibility. The conclusions, although based only on one male speaker, indicated that 12-predictor coefficient, 5-formant, and 3-formant coded speech give equivalent results.

The class of LPC vocoder techniques involves a high implementation cost on account of the computational difficulties associated with the extraction of linear prediction filter coefficients and the pitch frequency. The system performance in terms of speech quality is usually good at bit-rates as low as 2.4 Kbits/sec, but it is talker and text-dependent, and the system is not satisfactory, at least at this time, for commercial application.

2.8. Voice-Excited Vocoder.

Extraction of excitation function in vocoder techniques of speech coding has proved to be a difficult problem. Voiced excitation is not periodic, and varies not only in period but in shape and amplitude. Also, the effect of the formant structure on the measurements of the pitch period contributes to make the task of designing a reliable pitch detector a difficult one. Furthermore, the modelling of the excitation into either voiced or unvoiced intervals is a simplification which is not always satisfactory. These problems stimulated the development of a vocoder system in which voicing decision and pitch extraction have been eliminated. This technique results, however, in some increase in the necessary transmission bandwidth.

The basic principles involved in the voice-excited vocoder are shown in the block diagram presented in Figure 2.16. The scheme was suggested by David et al.[87] in 1962. The excitation signal is generated at the receiving end from a sub-band of the original speech, which is transmitted in place of voicing decision and pitch information. This avoids the simplification of separating the excitation into voiced and unvoiced intervals and avoids the problem of pitch detection. This technique is relatively insensitive to the input speech quality. Speech quality is a factor of importance for reliable pitch extraction. The transmitted sub-band is flattened in the receiver by applying it to a spectral flattener, which consists of a rectifier, a bank of band-pass filters, and peak-clippers to remove spectrum amplitude fluctuations.

In one implementation of the voice-excited vocoder,[87] the baseband (sub-band carrying excitation information) occupies the range between 250 Hz and 940 Hz. The spectral analysis uses 17 vocoder channels in the band from 940 Hz to 3650 Hz, and the total transmission bandwidth is between 1 KHz and 1.2 KHz. Intelligibility tests revealed a greater degradation for the female than for the male speaker, and with respect to quality of synthetic speech, the voice-excited vocoder performed much better than a conventional 18-channel vocoder, removing much of the unnaturalness of the vocoded speech.

A digitised voice-excited channel vocoder was proposed some years later,[88] which used band-pass sampling of the baseband signal in order to improve efficiency in terms of transmission bit-rate. The following specification was adopted: - baseband: 300 to 900 Hz, sampled at 1.2 Ksamples/sec, and quantised with 5 or 6 bits, resulting in a baseband transmission rate of 6 or 7.2 Kbits/sec; - spectral analysis: 10 channel signals and one normalisation parameter at a frame rate of 50 sec^{-1} quantised with 4 bits each, resulting in 2.2 Kbits/sec. The total bit-rate was 8.2 or 9.4 Kbits/se, and informal listening tests indicated that the 9.4 Kbits/sec system could reproduce speech with quality similar to the analogue mode of operation of the vocoder, which still contains some degree of unnaturalness, and cannot be applied in any other than special circumstances. Although the voice-excited vocoder is not as difficult to implement as pitch-excited vocoder, it is still much more complex than waveform coders such as ADM and ADPCM.

Voice excitation has found application not only in channel vocoders. At least two other vocoder methods, namely, the formant vocoder,[89] and more recently the LPC vocoder,[90] have also been developed using this technique. The latter, when operating at a bit-rate of 8 Kbits/sec, has been found to have a level of performance comparable with an adaptive predictive coder operating at same bit-rate, but was less degraded by telephone-quality speech.

2.9. Homomorphic Vocoder.

The application of homomorphic filtering[91,92] to separate the effects of excitation and vocal tract, with the purpose of efficient speech coding, has been shown[93] to be effective.

Homomorphic systems[91], H, are systems which satisfy a principle of superposition for some operation o. With these systems, if $s(t) = s_1(t) \circ s_2(t)$, then

$$H[s(t)] = H[s_1(t)] \circ H[s_2(t)]. \tag{2.43}$$

In the case that o is normal addition, H is a linear system and hence it follows that the concept represents a generalisation of the concept of linear system. A homomorphic system H can be represented as a cascade of three systems,[91,92] namely, a system D, a linear system, and the inverse of D, where D satisfies

$$D[s_1(t) \circ s_2(t)] = D[s_1(t)] + D[s_2(t)]. \tag{2.44}$$

Now, consider the speech signal s(t) to be modelled by the convolution of an excitation function e(t) and the vocal tract impulse response v(t), i.e.,

$$s(t) = e(t) * v(t). \tag{2.45}$$

The system D, in this case, is realised by inverse Fourier transforming the logarithm of the magnitude of the Fourier transform. Thus,

$$\begin{aligned} D[s(t)] &= F^{-1}[\log|S(\omega)|] \\ &= F^{-1}[\log|E(\omega)|] + F^{-1}[\log|V(\omega)|], \end{aligned} \tag{2.46}$$

where $S(\omega)$, $E(\omega)$, and $V(\omega)$ are the Fourier transforms of s(t), e(t), and v(t) respectively, and F^{-1} is the inverse Fourier transform operator. From expressions (2.45) and (2.46) it can be seen that

$$D[e(t)*v(t)] = D[e(t)] + D[v(t)], \tag{2.47}$$

which means that the effects of excitation and vocal tract can be separated by the operator D into two components which are added. The speech coding technique which is based on this principle is known as homomorphic vocoder. The operator D, when applied to s(t), results in what is known as the 'cepstrum' of s(t), and on account of this the homomorphic vocoder is sometimes called cepstrum vocoder.

The cepstrum contains two distinct portions: a low one (near the origin), which is representative of the slow variation with frequency of the vocal tract features, and a high-time portion, which represents the contribution of the rapid variation with frequency of the excitation. In the homomorphic vocoder, voiced-unvoiced detection and pitch extraction are obtained by selecting the peak in the cepstrum which occurs in an interval equal to the pitch period. Absence of this peak indicates an unvoiced segment.

In the synthesiser, the vocal tract impulse response approximation is obtained by applying the operator D^{-1} to the low-time portion of the cepstrum, resulting in a signal which is then convolved with an appropriate excitation function, as in other vocoder configurations.

The analysis and synthesis principles of the homomorphic (or cepstrum) vocoder are illustrated in the block diagram of Figure 2.17. The cepstrum is obtained by 512-point transforms taking the first 32 points as the low-time portion. It is recalculated every 20 msec, and the synthetic speech is claimed[93] to be natural sounding and of good quality. Little degradation has been observed[93] when the low-time portion is reduced to 26 points and quantised to 6 bits each (bit-rate: 7.8 Kbits/sec).

The homomorphic vocoder has the advantage that pitch estimation can be obtained directly from the cepstrum and no further analysis is required. The system is conceptually sophisticated and important, but its complexity and performance are not sufficiently attractive to guarantee a practical application of the scheme.

The transmission rate of the cepstrum low-time portion can be reduced if linear prediction techniques are used.[94] The idea consists in using the all-pole model given by.

$$\hat{V}(Z) = \frac{a_o}{1 - \sum_{i=1}^{p} a_i Z^{-i}} \quad , \tag{2.48}$$

in order to represent the effect of the vocal tract. Now, let v_n be the impulse response as obtained in the synthesiser of the original homomorphic vocoder. The method computes the coefficients $\{a_j\}$ by minimising the mean-square-error between the output of $1/\hat{V}(Z)$ when driven by v_n, and an impulse function. Of course, now the analyser, in addition to performing the cepstrum analysis, also performs the operations of direct and inverse Fourier transformation, exponentiation, and the linear prediction analysis. This makes the system considerably more complex. The synthesiser, on the other hand, uses a recursive filter, given by (2.48), excited by an impulse function in order to obtain the approximation to the vocal tract impulse response.

Using 10 coefficients, 8 bits per coefficient, 50 frames/sec, and 800 bits/sec for excitation parameters, a resultant bit-rate of 4.8 Kbits/sec is involved and the quality of decoded speech is found to be comparable with that obtained with the original homomorphic vocoder. The overall system complexity is, however, too high, and although the bit-rate was reduced to 4.8 Kbits/sec the speech quality obtained is not good enough for commercial telephone application.

2.10. Pattern-Matching Vocoder.

The aim in this type of vocoder is to avoid the transmission of the infinity of possible spectra which can be transmitted by conventional vocoders and thereby permit only a limited number of possibilities , i.e., by classifying spectral patterns into a finite alphabet. In this way, the whole spectrum can be described (with a number of limitations which will be discussed later in this section) by a single signal (or symbol).

The pattern-matching vocoder reported by Dudley[95] in 1958 uses an analysis system[96] in which the speech signal is first filtered by a bank of ten band-pass filters, rectified, and then low-pass filtered, as in the spectrum analyser of a 10-channel vocoder. However, in this case, the spectral patterns of six vowels, two semivowels, and two fricative consonants are stored in a 10 x 10 matrix $\underline{M}$, and a best spectral match is chosen in terms of a cross-correlation criterium. Thus, if ν_{bs} is the output of the b-th band-pass channel of a spoken sound s, and r_{pb} is the element of the matrix $\underline{M}$ corresponding to the p-th pattern and the b-th band-pass channel then the value of p, for which

$$\nu_{ps} = \sum_{b=1}^{10} r_{pb} \nu_{bs} \tag{2.49}$$

is maximum, corresponds to the selected pattern. This pattern, together with conventional pitch and voicing parameters, is then transmitted to a 10-channel vocoder synthesiser.

Two listeners, in an articulation test,[95] scored 97% and 99% respectively, of correct recognition of digits uttered by one speaker, whose speech material was used in the matrix of stored patterns, but significantly lower scores (45% and 66%) were obtained when common monosyllabic words were used. This shows one of the limitations of the system, namely, the failure to handle a wider vocabulary. Also, it should be noted that in the tests the stored spectral patterns were taken from the same speaker who uttered the material for the tests. Thus, from the basic characteristics of the system, it would be expected that considerable reductions in the performance would result with respect to different talkers. Besides vocabulary limitations, the system thus seems to be limited with respect to speaker population.

The problem of size and composition of the matrix $\underline{M}$ has been considered further and a new strategy has been used, in which patterns with higher probability are correspondingly matched with higher accuracy.[97,98] The number of patterns used varies from 497 (328 voiced, 169 unvoiced) to 4006 (728 voiced, 3338 unvoiced), with a corresponding bit-rate variation in spectrum data which ranges from 450 to 600 Kbits/sec, at a 50 Hz sampling frequency. Intelligibility tests have shown that with respect to consonant discriminability no significant reduction in performance results when compared with the conventional channel vocoder.

The pattern-matching vocoder appears to be an interesting attempt to obtain very low transmission bit-rates. However, the characteristics of the method are such as to indicate that coder performance is

speaker-dependent and speaker recognition is considerably degraded. Furthermore, the system is even more complex than channel vocoders, since it involves additional number of computation to determine the best spectral match of a spoken sound to the stored patterns.

2.11. Correlation Vocoder.

The short-time autocorrelation function $\phi_t(\tau)$ of a signal $s(t)$ can be measured[99] by averaging the product $s(t)\ s(t-\tau)$ by means of a low-pass filter with impulse response given by

$$h(t) = 2\alpha\ e^{-2\alpha t}\ u(t), \tag{2.50}$$

where $u(t) = 1$ for $t > 0$, and $u(t) = 0$ for $t < 0$.

Thus,

$$\phi_t(\tau) = 2\alpha \int_{-\infty}^{t} s(x)\ s(x-\tau) e^{-2\alpha(t-x)} dx \tag{2.51}$$

The short-time power spectrum, $\Psi_t(\omega)$, of the signal $s(t)$, on the other hand, is measured[99] by passing $s(t)$ through band-pass filters

$$h_1(t) = \sqrt{2\alpha}\ e^{-\alpha t}.\sin\omega t.\ u(t), \tag{2.52a}$$

and

$$h_2(t) = \sqrt{2\alpha}\ e^{-\alpha t}.\cos\omega t.\ u(t), \tag{2.52b}$$

and adding the squares of the output to obtain

$$\Psi_t(\omega) = |S_t(\omega)|^2 \tag{2.53}$$

where

$$S_t(\omega) = \int_{-\infty}^{t} s(x).[h(t-x)]^{1/2}\ e^{-j\omega x}\ dx. \tag{2.54}$$

It has been shown[99] that (2.51) and (2.53) are related by a Fourier transformation, i.e.,

$$\phi_t(\tau) = \frac{e^{\alpha|\tau|}}{2\pi} \int_{-\infty}^{\infty} \Psi_t(\omega)\ e^{j\omega\tau}\ d\omega \tag{2.55}$$

and

$$\Psi_t(\omega) = \int_{-\infty}^{\infty} e^{-\alpha|\tau|}\ \phi_t(\tau)\ e^{-j\omega\tau}\ d\tau\ . \tag{2.56}$$

In 1954, two studies were reported[100,101] with regard to the autocorrelation analysis of speech sounds. Later, time-domain vocoders were developed[102-104] which made use of the relationship between short-time power-spectrum and short-time autocorrelation function of a signal, as expressed by (2.55) and (2.56).

The analyser portion of an autocorrelation vocoder computes the short-time autocorrelation function at discrete delays separated by an interval $\Delta\tau \leq 1/2\ F$, where F is the maximum frequency contained in the speech input signal. When 18 delay channels are used, a total bandwidth of 360 Hz results,[104] which is similar to that achieved with channel vocoders. Values of $\Delta\tau = 0.167$ msec and $\tau_{max} \cong 3$ msec were employed. At the synthesiser, pitch and voiced-unvoiced parameters are used to generate a proper excitation which drives a transversal filter having an impulse response given by the short-time autocorrelation function for delays between $-\tau_{max}$ and τ_{max}. It should be noted, however, that the output spectrum envelope is the power spectrum of the input signal. This squaring effect is compensated by a spectral equaliser, which is introduced before the autocorrelation

analysis in order to reduce the three formant regions to square roots of the original amplitudes.

A complete block diagram of the autocorrelation vocoder is shown in Figure 2.18. This system can yield highly intelligible speech, but of quality comparable with that of a channel vocoder, which is not satisfactory when compared with the commercial quality standard. Also, the general complexity of the autocorrelation vocoder is still substantially high. The technique involves the computation of the autocorrelation function for a number of discrete delays, spectral flattening,, and narrow-band filtering to obtain an equivalent description of the short-time power spectrum and pitch extraction is also required.

2.12. Harmonic Vocoder.

Another development in the class of vocoder devices involves the representation of the short-time spectral envelope of speech by a small set of coefficients taken from its Fourier series expansion. For this reason it is known as harmonic vocoder.[105]

The analyser part of the device consists in obtaining a time function s(t) at time intervals $T_a = 1/F_a$, where t is actually representing instantaneous frequency. This can be accomplished using the principle of the " scan vocoder ",[52] which uses a scanning operation on the rectified outputs of 100 analysing filters, and a low-pass filter to remove fine structure. Variations of F_a between 25 Hz and 50 Hz, and of spectral envelope transmission bandwidths between 75 Hz and 250Hz

have been considered, which result in a total band between 100 Hz and 300 Hz when pitch and voicing information is added.

The synthesiser used in the harmonic vocoder first determines a number $(2N + 1)$ of Fourier coefficients from the series expansion of $s(t)$, and these are then used to control a variable frequency response network, representing the vocal tract characteristics within the accuracy of the Fourier analysis. The characteristics of this system, however, indicate that acceptable quality speech cannot be obtained for the reduced transmission bandwidth mentioned above, in view of the limitations involved in the spectral representation and the simple model used to analyse the signal.

2.13. Laguerre and Chebyshev Vocoders.

As has been seen from equations (2.55) and (2.56), the short-time autocorrelation function $\phi_t(\tau)$ and the short-time power spectrum $\Psi_t(\omega)$ are related by a Fourier transformation when the weighting function $h(t)$ is given by (2.50). This property has served as the basis for the correlation vocoder. The Laguerre vocoder,[106] on the other hand, considers the expansion of $\phi_t(\tau)$ in a series of orthonormal functions over the interval $0 \leqslant \tau \leqslant \infty$. Particular attention has been given to the set of Laguerre functions $\{\ell_j(\tau)\}$

$$\ell_j(\tau) = e^{\lambda\tau/2} \frac{d^j}{d\tau^j} \left\{ \frac{\tau^j}{j!} e^{-\lambda\tau} \right\} \tag{2.57}$$

on account of the possibility of using RC elements only.[106] Thus, $\phi_t(\tau)$ is expressed by

$$\phi_t(\tau) = \sum_{j=0}^{\infty} a_j(t)\ell_j(\tau), \qquad \tau \geq 0 \tag{2.58}$$

where

$$a_j(t) = \int_{-\infty}^{t} s(\gamma)h(t - \gamma)d\gamma \int_{-\infty}^{\gamma} s(\xi)\ell_j(\gamma - \xi)d\xi, \tag{2.59}$$

and s(t) denotes the speech signal. The latter equation defines the analyser of the Laguerre vocoder, which is shown in the block diagram of Figure 2.19.

For all τ, $\phi_t(\tau)$ can be written as

$$\phi_t(\tau) = \sum_{j=0}^{\infty} a_j(t)[\ell_j(\tau) + \ell_j(-\tau)], \tag{2.60}$$

and the spectrum involves the terms $[L_j(\omega) + L_j^*(\omega)]$, where * denotes complex conjugate. However, due to difficulties in implementation of (2.60),[3,106] and since the spectrum magnitudes of $[\ell_j(\tau) + \ell_j(-\tau)]$ and $[\ell_{m-j}(\tau) + \ell_{m+j+1}(\tau)]$ are the same, the latter term has been used in the synthesis operation, as shown in Figure 2.19.

The minimum number of coefficients required in the Laguerre vocoder has been found to be equal to seven (N = 6 in Figure 2.19). The same basic principles have also been considered in a separate paper[107] by Kulya, in which Chebyshev polynomials were used.

Laguerre and Chebyshev vocoders, though conceptually interesting techniques of speech bandwidth compression, suffer from the basic disadvantages of vocoder-type techniques. The indications are that speech quality is still not good enough, and at present the system is not sufficiently simple to justify practical application.

2.14. Phase Vocoder.

Another possibility for transmission bandwidth compression is based upon an approximation of the speech signal s(t) as a combination of the output signals $s_n(t)$, n = 1, ..., N, of N band-pass filters with impulse responses as given by

$$g_n(t) = h(t) \cos\omega_n t, \tag{2.61}$$

where h(t) is the impulse response of a low-pass filter. The method has been proposed by Flanagan and Golden[108] and is known as the phase vocoder.

The band-pass signals $s_n(t)$ can be written as

$$s_n(t) = |S_t(\omega_n)| \cos [\omega_n t + \phi_t(\omega_n)], \tag{2.62}$$

where $S_t(\omega_n)$ is the short-time spectrum of s(t) at $\omega = \omega_n$, i.e.,

$$S_t(\omega_n) = \int_{-\infty}^{t} s(\lambda)\, h(t - \lambda) e^{-j\omega_n \lambda} \, d\lambda, \tag{2.63}$$

with amplitude $|S_t(\omega_n)|$ and phase $\phi_t(\omega_n)$.

Because $\phi_t(\omega_n)$ is not always bounded, it has not been used as a transmission parameter, and has been replaced by its time derivative $\dot{\phi}_t(\omega_n)$. This, however, assumes that the additive phase constant is not important. Thus, in place of (2.62) we have

$$\tilde{s}_n(t) = |S_t(\omega_n)| \cos[\omega_n t + \tilde{\phi}_t(\omega_n)], \tag{2.64}$$

where

$$\tilde{\phi}_t(\omega_n) = \int_0^t \dot{\phi}_t(\omega_n)\, dt. \tag{2.65}$$

An approximation to the original speech signal is obtained by

$$\tilde{s}(t) = \sum_{n=1}^{N} \tilde{s}_n(t), \tag{2.66}$$

where $\tilde{s}_n(t)$, given by (2.64), is the output of an oscillator with frequency ω_n, which is modulated in amplitude by $|S_t(\omega_n)|$ and in phase by $\tilde{\phi}_t(\omega_n)$.

A computer simulation of the phase vocoder has been carried out[108] with N = 30 and $\omega_n = 2\pi n$. 100 rad/sec, covering a frequency range from 50 to 3050 Hz. In this simulation, each of the phase and amplitude information signals was low-pass filtered at 25 Hz, resulting in a bandwidth compression of 2 to 1. A performance superior to that obtained with a channel vocoder has been attained in terms of quality of decoded signal. It should be noted, however, that the compression achieved by the latter is typically of the order of 10 to 1.

Parameters of the phase vocoder have also been quantised and coded for transmission at bit-rates of 7.2 Kbits/sec and 9.6 Kbits/sec, yielding results which are comparable with log-PCM at bit-rates of the order of three times that of the phase vocoder.[3] More recently, a computationally efficient implementation which uses the fast Fourier transform (FFT) algorithm to perform analysis and synthesis procedures, has been suggested.[109]

The phase vocoder, thus, produces fairly good results in terms of speech quality and transmission bit-rates. However, it does not give commercial quality speech at low-bit-rates (e.g., 7.2 Kbits/sec and below), and although this technique avoids the problem of pitch estimation (and is simpler than the channel vocoder) it is substantially more complex than waveform coders.

2.15. Transform Coding.

A class of speech coding techniques, which have been widely explored in the literature (see, for example, references 10, 11, 12, 24, 47, 48, 59, 60 and 110 to 121), involves the application of orthogonal transforms in order to obtain a more efficient representation of the signal. This class of coding techniques is known as transform coding and is illustrated in the simplified block diagram shown in Figure 2.20. The idea consists in applying an N x N transform matrix $\underline{T}$ to a vector $\underline{s}$ representing a sequence of N speech samples $\{s_0, \ldots, s_{N-1}\}$, resulting in a vector $\underline{c}$ of N transform coefficients given by

$$\underline{c} = \underline{T}\,\underline{s}, \tag{2.67}$$

which are separately quantised to $\hat{\underline{c}}$. This vector is then transmitted, and in the decoding process is transformed by the inverse matrix, $\underline{T}^{-1}$, of the matrix $\underline{T}$ into a vector $\hat{\underline{s}}$, which represents an approximation to the original segment of speech samples arranged into the vector $\underline{s}$. Thus,

$$\hat{\underline{s}} = \underline{T}^{-1}\,\hat{\underline{c}}. \tag{2.68}$$

The orthogonal transform $\underline{T}$ can be written as

$$\underline{T} = \begin{bmatrix} \underline{T_o}^t \\ \vdots \\ \underline{T_{N-1}}^t \end{bmatrix} \tag{2.69}$$

where the basis vectors $\{\underline{T_o}, \ldots, \underline{T_{N-1}}\}$ are orthogonal to each other. This implies that the transpose, $\underline{T}^t$, and the inverse, $\underline{T}^{-1}$, of $\underline{T}$ are identical if $\underline{T_j}$, $j = 0, \ldots, N - 1$, are normalised. Expression (2.68) can then be written as

$$\hat{\underline{s}} = \sum_{j=0}^{N-1} \hat{c}_j \, \underline{T_j}. \tag{2.70}$$

From the above equation it can be seen that if $\underline{T}$ is chosen such that the distribution of coefficients $\hat{c}_j$, $j = 0, \ldots, N - 1$, can be efficiently exploited by independent quantisation and/or by discarding some of them, then the method offers the possibility of bit-rate reduction.

This section begins with a brief discussion of the Karhunen-Loève, Walsh-Hadamard, Discrete Fourier, and Discrete Cosine transforms as possibilities for digital coding of speech signals. Then, different approaches in the design of speech transform coders are presented, including some of the most recent developments in this field. One of them, the dominant coefficient Hadamard transform coding technique has been further investigated on account of the simplicity of the implementation and the possibility of operating at bit-rates which are lower than those associated with waveform coding techniques. The study of this method forms part of the research work reported on in this thesis (see Chapter 3).

2.15.1. The Karhunen-Loève, Walsh-Hadamard, Discrete Fourier, and Discrete Cosine Transforms as Applied to Speech Coding.

Before discussing the possibilities of the application of the transforms mentioned above, the definition and some important aspects of these transforms will be considered. A more detailed discussion and extensive bibliography can be found in the books of Harmuth,[122] and Ahmed and Rao.[123]

The transformation which minimises the total mean-square-error given by

$$\bar{\varepsilon} = E[(\underline{s} - \hat{\underline{s}})^t . (\underline{s} - \hat{\underline{s}})] \tag{2.71}$$

is the transformation $\underline{K}$ which is comprised of the eigenvectors $\underline{K}_j$ of the covariance matrix $\underline{R}_s$ of the input $\underline{s}$. It is known as the Karhunen-Loève transformation[10,59,112,123-125] and is given by

$$\underline{K} = \begin{bmatrix} \underline{K_o}^t \\ \vdots \\ \underline{K_{N-1}}^t \end{bmatrix} \tag{2.72}$$

where

$$\underline{R}_s \, \underline{K}_j = \lambda_j \, \underline{K}_j \, , \tag{2.73}$$

and $\underline{K}_j$ is the eigenvector of $\underline{R}_s$ corresponding to the eigenvalue λ_j.

The covariance matrix of the transform coefficients $\{c_j ; j = 0, \ldots, N - 1\}$ is given by

$$\begin{aligned} \underline{R}_c &= \underline{K} \, \underline{R}_s \, \underline{K}^t \\ &= \text{diag}(\lambda_0, \lambda_1, \ldots, \lambda_{N-1}), \end{aligned} \tag{2.74}$$

which means that these coefficients are uncorrelated with variances $\{\sigma_j^2 \; ; \; j = 0, \ldots, N - 1\}$ which are equal to the eigenvalues of $\underline{R}_s$, i.e.,

$$\sigma_j^2 = \lambda_j \text{ for all } j. \tag{2.75}$$

Since speech is a non-stationary process, the transformation $\underline{K}$ will be different in each segment of samples, implying that eigenvalues and eigenvectors must be calculated for each of these segments. However, this results in a complex and time consuming set of operations.

An interesting set of functions, which take only the values +1 and -1 and are ordered so that each function has one more zero-crossing than the preceding one, was described in 1923 by J.L. Walsh.[126] In the present case, however, of more interest is the set of N-length discrete Walsh functions which is obtained by sampling the first N continuous Walsh functions at N equidistant points k = 0, 1, ..., N - 1. These can be expressed by an N x N matrix $\underline{W}_N$.[127]

The discrete Walsh function, wal(n, k), of order n, where n = 0, 1, ..., N - 1, can be represented for $N = 2^p$ by[127]

$$\text{wal}(n, k) = \prod_{r=0}^{p-1} (-1)^{(n_{p-r} + n_{p-r-1})k_r} \tag{2.76}$$

$$k = 0, 1, \ldots, N - 1,$$

where n_r and k_r denote the r-th bit in the binary representation of n and k, i.e.,

$$n = (n_{p-1} \cdots n_1 \; n_0)_{\text{binary}}$$

and

$$k = (k_{p-1} \cdots k_1 \; k_0)_{\text{binary}}.$$

The finite Walsh transform of the sequence $\{s_j;\ j = 0, \ldots, N-1\}$ is difined by $\underline{T} = (1/N)\underline{W}_N$, so that the transform coefficients are given by

$$c_n = \frac{1}{N} \sum_{k=0}^{N-1} s_k \, \mathrm{wal}(n, k), \tag{2.77}$$

$$n = 0, 1, \ldots, N-1.$$

The inverse is given by $\underline{T}^{-1} = \underline{W}_N$, which implies that (see Figure 2.20)

$$\hat{s}_k = \sum_{n=0}^{N-1} \hat{c}_n \, \mathrm{wal}(k, n), \tag{2.78}$$

$$k = 0, 1, \ldots, N-1.$$

By analogy with frequency, the term " sequency" has been used[122] and defined as one-half the average number of zero-crossings per second. Also by analogy with the frequency domain representation, the Walsh functions can be denoted by sal(sine Walsh) and cal(cosine Walsh) functions, which are defined by

$$\mathrm{sal}(n, k) = \mathrm{wal}(2n-1, k) \tag{2.79}$$

$$n = 1, 2, \ldots, N/2,$$

$$\mathrm{cal}(n, k) = \mathrm{wal}(2n, k) \tag{2.80}$$

$$n = 1, 2, \ldots, N/2 - 1,$$

where n represents the sequency.

Consider now a new ordering of the Walsh functions which is given by[123]

$$\mathrm{had}(n, k) = \mathrm{wal}(m, k) \tag{2.81}$$

$$n = 0, 1, \ldots, N-1,$$

where n is obtained from m by bit-reversal of the binary-to-Gray code conversion of m, i.e.,

$$m = (m_{p-1}\ m_{p-2} \cdots m_1\ m_0)_{\text{binary}}$$

$$g = (g_{p-1}\ g_{p-2} \cdots g_1\ g_0)_{\text{binary}}$$

$$n = (g_0\ g_1 \cdots g_{p-2}\ g_{p-1})_{\text{binary}}$$

where $(g)_{\text{binary}}$ represents the Gray code of m, and is obtained by the following relations:

$$g_i = m_i \oplus m_{i+1}\ ,\ 0 \leq i \leq p-2 \tag{2.82a}$$

and

$$g_{p-1} = m_{p-1} \tag{2.82b}$$

where $\oplus$ denotes addition modulo-2 ($0 \oplus 0 = 1 \oplus 1 = 0$, and $1 \oplus 0 = 0 \oplus 1 = 1$). It can also be seen that m can be obtained from n by means of a Gray code-to-binary code conversion of the bit-reversal of n.

The ordering of the Walsh functions according to (2.81), i.e., had(n, k), n = 0, ..., N - 1, is known as natural, or Hadamard ordering.[111,123] Thus, by arranging the rows of the Walsh matrix in natural ordering, a Hadamard matrix $\underline{H}_N$ is obtained, which like the Walsh matrix, is orthogonal and symmetric. It has the advantage, however, of presenting a simple recursive structure, namely,

$$\underline{H}_{2^k+1} = \begin{bmatrix} \underline{H}_{2^k} & \underline{H}_{2^k} \\ \underline{H}_{2^k} & -\underline{H}_{2^k} \end{bmatrix} \tag{2.83}$$

$$k = 0, 1, 2, \ldots$$

where $\underline{H}_1 = [1]$.

The Hadamard transform is defined by $\underline{T} = (1/N)\underline{H}_N$, and its inverse is given by $\underline{T}^{-1} = \underline{H}_N$. Thus (see Figure 2.20), in this case

$$\underline{c} = \frac{1}{N} \underline{H}_N \cdot \underline{s} \tag{2.84a}$$

and

$$\hat{\underline{s}} = \underline{H}_N \cdot \hat{\underline{c}}. \tag{2.84b}$$

An important property of the Hadamard matrix has been shown by Turner.[128] It says that, apart from the ordering, the elements of the matrices given by

$$\underline{A} = \underline{H}\,\underline{X}\,\underline{H} \tag{2.85a}$$

and

$$\underline{B} = \tilde{\underline{X}}\,\tilde{\underline{H}} \tag{2.85b}$$

are identical, where $\underline{H}$ represents a $2^p \times 2^p$ Hadamard matrix, $\tilde{\underline{X}}$ is a row matrix comprised of the original 2^{2p} elements of $\underline{X}$, and $\tilde{\underline{H}}$ is the Hadamard matrix formed by taking the Kronecker product of $\underline{H}$ with itself. The importance of this property is that (2.85a) offers a computationally more efficient means of performing the transform.

Both Hadamard and Walsh matrices are comprised only of +1's and -1's elements, and this means that only additions and subtractions are required in order to compute the multiplication by a data matrix. Also, fast algorithms are available[123,129-131] in order to efficiently compute the respective transforms. These fast algorithms involve the use of $N \log_2 N$ additions and substractions, where $N = 2^p$ and p is the total number of iterations. A Fortran subroutine can be found in reference 123. It should also be noted that these transforms have the attractive property that the direct and inverse operations are essentially the same.

With regard to the characteristics of the Walsh (or Hadamard) transform, it should be mentioned that, unlike the Karhunen-Loève transform, the correlation is not fully removed in the transform domain. However, the non-vanishing elements not on the main diagonal of the covariance matrix in the transform domain are few,[114] and the advantages mentioned in the previous paragraph with respect to the simplicity of the transforms served to stimulate the interest of using them as a means of efficient digital transmission of speech.

The discrete Fourier transform (DFT)[123] is another orthogonal transformation. In this case, the sine and cosine functions are used and the transformation into the frequency domain is defined by

$$c_n = \frac{1}{N} \sum_{k=0}^{N-1} s_k \cdot e^{-j(2\pi nk)/N} \qquad (2.86)$$

$$n = 0, 1, \ldots, N-1$$

where $\{s_k;\ k = 0, 1, \ldots, N-1\}$ is the data sequence. Using the notation of Figure 2.20, the inverse DFT of $\{\hat{c}_n;\ n = 0, 1, \ldots, N-1\}$ is given by

$$\hat{s}_k = \sum_{n=0}^{N-1} \hat{c}_n \cdot e^{j(2\pi nk)/N} \qquad (2.87)$$

$$k = 0, 1, \ldots, N - 1.$$

If the data vector $\underline{s}$ is real-valued then the DFT components c_o and $c_{N/2}$ are real, and

$$c_{N-\ell} = c_\ell^* \quad , \qquad \ell = 0, 1, \ldots, N/2 - 1, \qquad (2.88)$$

where c_ℓ^* denotes the complex conjugate of c_ℓ. It can also be shown[123] that if s(t) is a real-valued signal of duration L, with no harmonic in the Fourier series expansion of its periodic extension above N'/2, then the Fourier transform $S(\omega)$ of the sampled version $\{s(kL/N);\ k = 0, 1, \ldots, N - 1;\ N \geq N'\}$ is related to the DFT coefficients $\{c_n;\ n = 0, 1, \ldots, N/2\}$ by

$$c_n = \frac{1}{L} S(n \cdot 2\pi/L). \qquad (2.89)$$

Now consider a sequence $\{s_k\}$ of length $N = 2^P$. It can be shown[134] that the N-point DFT of this sequence can be written as

$$c_n = a_n + e^{-j2\pi/N} b_n, \quad 0 \leq n \leq N/2 - 1, \qquad (2.90a)$$

and

$$c_n = a_{n - N/2} + e^{-j2\pi/N} b_{n - N/2}, \quad N/2 \leq n \leq N - 1, \qquad (2.90b)$$

where $\{a_n\}$ is the N/2-point DFT of the even components of $\{s_k\}$, and $\{b_n\}$ is the N/2-point DFT of the odd components of $\{s_k\}$. If this procedure is iterated up to the stage where only 2-point DFT's are left to be

calculated, then a total of approximately $N/2 \log_2 N$ complex multiplications will be required to compute the DFT as opposed to $(N - 1)^2$ for the case of direct evaluation of (2.86). This process is known as 'decimation-in-time' and illustrates the basic ideas of algorithms for the efficient computation of the DFT. These algorithms are known as fast Fourier transforms (FFT).[123,134]

More recently, another discrete transformation has been introduced which is signal-independent, has a fast algorithm for the computation of the transform coefficients, and performs almost as well as the Karhunen-Loève transformation. For these reasons, it has found application in some of the most recent developments of digital speech coding. It is known as the discrete cosine transform (DCT), and was introduced by Ahmed, Natarajan, and Rao[132] in 1974. The DCT of an N-length sequence $\{s_k\}$ is defined by

$$c_n = \frac{1}{N} \sum_{k=0}^{N-1} s_k \cdot g_n \cos[(2k + 1)\, n\pi/2N], \qquad (2.91)$$

$$n = 0, 1, \ldots, N - 1$$

where $g_0 = \sqrt{2}$, and $g_n = 2$ for $n = 1, \ldots, N - 1$. The inverse DCT $\{\hat{s}_k\}$ of the sequence $\{\hat{c}_n ; n = 0, 1, \ldots, N - 1\}$ is given by

$$\hat{s}_k = \sum_{n=0}^{N-1} \hat{c}_n \cdot \hat{g}_n \cos[(2k + 1)\, n\pi/2N], \qquad (2.92)$$

$$k = 0, 1, \ldots, N - 1$$

where $g_0 = \sqrt{2}/2$, and $g_n = 1$ for $n = 1, \ldots, N - 1$. The DCT matrix $\underline{C}_N$ satisfies

$$\underline{C}_N^t \, \underline{C}_N = (N/2)\underline{I}_N , \qquad (2.93)$$

where $\underline{I}_N$ is the N x N identity matrix. A fast computational algorithm has been suggested,[132] which uses a 2N-point FFT on a data sequence whose number of samples had been doubled by appending N zeros. This can be noted by writing (2.91) in the form

$$c_o = (\sqrt{2}/N) \sum_{k=0}^{N-1} s_k \tag{2.94a}$$

$$c_n = (2/N)\mathrm{Re}\ \{e^{-j(n\pi)/2N} \sum_{k=0}^{2N-1} s_k \cdot e^{-j(2\pi nk)/2N}\} \tag{2.94b}$$

$$n = 1, \ldots, N-1$$

where $s_k = 0$ for $k = N, N+1, \ldots 2N-1$. Storage and computational savings can, however, be obtained by an algorithm that was later developed,[133] which uses two FFT's of length N on the original data sequence.

If the transform coefficients $\{c_i\ ;\ i = 0, 1, \ldots, N-1\}$ are to be transmitted in digital form, then they should be quantised to $\{\hat{c}_i\ ;\ i = 0, 1, \ldots, N-1\}$ as shown in Figure 2.20. A number of bits b_i, $i = 0, 1, \ldots, N-1$, are allocated for each coefficient. The number of bits b_i used in this process can be determined by minimising the total average distortion measure given by

$$\bar{\varepsilon} = E[(\underline{c} - \hat{\underline{c}})^t . (\underline{c} - \hat{\underline{c}})]$$

$$= \frac{1}{N} \sum_{i=1}^{N} \bar{\varepsilon}_i \quad , \tag{2.95}$$

where $\bar{\varepsilon}_i$ is the average distortion measure associated with the i-th coefficient. A constraint on this minimisation is that

$$\bar{B} = (1/N) \sum_{i=0}^{N-1} b_i = \text{constant}, \tag{2.96}$$

and the resulting expression for b_i can be shown[10,124] to be

$$b_i = \bar{B} + (1/2)\log_2\sigma_i^2 - (1/2N) \sum_{j=0}^{N-1} \log_2\sigma_j^2 \quad ; \tag{2.97}$$

$$i = 0, 1, \ldots, N-1,$$

where σ_j^2 is the variance of the coefficient c_j.

Using N = 16, Campanella and Robinson[112] have shown that, without changing the signal-to-quantising noise ratio performance, the bit-rate of a 56 Kbits/sec μ-law log-PCM can be lowered to 42.5, 46.0, and 48.5 Kbits/sec, by means of the Karhunen-Loève, discrete Fourier, and Hadamard transform, respectively. In reference 113, for the Fourier and Hadamard techniques, the same authors achieve a further bit-rate reduction to 32.5 Kbits/sec, using a truncation of the transform coefficients at the high end. The resulting speech quality is found to be good.

Another investigation[114] has shown that the energy distribution of the Walsh coefficients is concentrated in the low-order coefficients for voiced sounds and in the high-order coefficients for unvoiced sounds. By means of a 64-point Walsh transform, and coefficient quantisation according to the distribution of voiced speech, bit-rates of 44 Kbits/sec and 22 Kbits/sec have been achieved, the first resulting in very good reproduction of speech. In this same work, the author also comments on the possibility of using the Walsh transform pitch-synchronously, i.e., sampling the signal in such a way that each pitch period is comprised of N samples.

In a comparison of the DFT with the Walsh-Hadamard transform (WHT) coding techniques, it has been observed[116] that the gain over log-PCM achieved by the latter is lower for all bit-rates ranging from 8 Kbits/sec to 64 Kbits/sec, and that it increases at low-bit-rates for both transforms. It has also been observed[10] that the DFT and the DCT show an assymptotically optimum performance with respect to this gain, when plotted as a function of the block size N. This means that both the DFT and the DCT approach the Karhunen-Loève transform (KLT) for increasing N. However, the latter is the nearest one to the KLT.

The results presented so far serve to illustrate some of the possibilities of using orthogonal transforms as a means of digital coding speech signals. As will be seen next in this section, transform coding can be employed in achieving quite significant savings in the transmission bit-rate. In fact, some of the most recent developments in the area of speech coding[48,119,120] belong to this class of techniques.

2.15.2. Dominant Coefficient Walsh-Hadamard Transform Coding.

It was reported[110] in 1970 that intelligible speech could be obtained by using only the two greatest valued coefficients resulting from the application of a 16-point Walsh transform. Three years later, Shum, Elliott, and Brown[11] published results of a further investigation on the subject, in which a 64-point transform (either Walsh or Hadamard), was applied to 8 msec speech segments, sampled at 8 Ksamples/sec, from which only 8 or 4 dominant (greatest absolute value) coefficients

were extracted for transmission. Using 6 bits to identify (label) each coefficient position, and 7 bits for amplitude, total bit-rates of 13 Kbits/sec and 6.5 Kbits/sec were obtained for the two cases, respectively, and fair to good quality of reproduced speech was claimed.

These results established the basis for a more extensive study of the technique,[12,117] which included properties of the Hadamard transform coefficients and system performance with change in number of dominant coefficients, sampling rate, and bit allocation. Some evidence of the existence of dominant coefficients is provided by the concentration of the average power in a small number of coefficients out of an N-point transform. For N = 64, the 6 to 8 dominant coefficients, for instance, contain between 60% and 90% of the total power when 8 KHz sampling is used. High intelligibility of reconstructed speech could be obtained at total bit-rates between 9.74 and 13 Kbits/sec, but the performance falls considerably if the sampling frequency is reduced to 6.1 KHz. Frangoulis and Turner[117] proposed a special technique for identifying 7 dominant coefficients out of 64, which made possible a reduction in transmission bit-rate from 11.375 Kbits/sec to 8 Kbits/sec.

The minimum number of dominant coefficients required to yield highly intelligible (though considerably noisy) reproduction of speech, has been indicated[12] as 4, 8, 16, and 32, for Hadamard transform sizes of 64, 128, 256 and 512, respectively. The system performance is, in general, lower for female than for male speech, a result which is in agreement with the smaller concentration of average power for the female speech (see reference 117).

At transmission bit-rates in excess of 9.6 Kbits/sec, the Hadamard transform coding technique can provide high speech intelligibility, as mentioned previously, and fairly good speech quality, though not sufficient for commercial application. At these bit-rates, however, the scheme produces better quality speech than the DCDM and APCM waveform coders.[12] At lower bit-rates, such as 4.8 Kbits/sec, the level of quality is considerably inferior to that obtained with vocoder-type techniques. The Hadamard coding scheme is, however, easily implemented and it appears worth investigating some new approaches to improve the performance of this system at low-bit-rates. This is, in fact, the subject of the research work reported on in Chapter 3 of this thesis.

2.15.3. Hadamard-Haar Transform Coding.

The Hadamard transformation does not completely remove the signal redundancy, since, as has been mentioned previously, the covariance matrix in the transform domain has a number of non-vanishing elements not on the main diagonal. In an attempt to exploit this property, Frangoulis and Turner[118] applied a second transform, namely, the Haar transform,[122,135] to the set of Hadamard coefficients. By using 64-point transforms, and 8 KHz sampled speech, they observed that from the total average power of the coefficients resulting from the successive application of the Hadamard and Haar transforms, between 44% and 65.6% is concentrated in the 5 to 8 dominant coefficients, with a difference of about 4% between male and female speech. This concentration is again lower for the latter.

At a bit-rate of about 10 Kbits/sec (6 dominant coefficients) the Hadamard-Haar technique has been reported[118] to provide intelligible and reasonably good reproduction of speech, whilst at lower bit-rates (5 and 3.5 Kbits/sec), corresponding to the transmission of 3 or 2 dominant coefficients, the speech quality is low, although intelligibility is not lost.

It should be observed that this coding scheme is computationally more complex than the straight dominant coefficient Hadamard transform technique, since now a three-valued second transformation is involved. In addition, for bit-rates below 8 Kbits/sec the speech quality obtained with the Hadamard-Haar scheme is more noisy than that produced by the straight Hadamard technique.[12] For bit-rates higher than 8 Kbits/sec the two systems give similar performance and hence, the straight Hadamard transform coding is more attractive.

2.15.4. Adaptive Transform Coding.

Since speech is a non-stationary signal, the bit assignment given by (2.97) would be more effective if it were based on the short-term statistics of speech sounds. This sort of approach was suggested by Zelinski and Noll[10] in 1977, and called adaptive transform coding (ATC). A block diagram of the scheme is shown in Figure 2.21.

In ATC, the input speech segment $\underline{s} = \{s_k;\ k = 0, 1, \ldots, N-1\}$ is gain-adjusted by a factor $V = (\hat{\sigma})^{-1}$, where $\hat{\sigma}$ is the estimated rms value of the input segment to be processed, i.e.,

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{j=0}^{N-1} s_j^2. \qquad (2.98)$$

A discrete cosine transformation of length N = 128 is used, resulting in a transform vector $V\underline{c}$, whre $\underline{c}$ is the transform of $\underline{s}$. Since $V\underline{c}$ is not known at the decoder, a coarse estimation of the variances of the transform coefficients is proposed, based on the assumption of a smooth distribution. The suggestion consists in averaging the squares of the transform coefficients over M neighbouring values, resulting in a set of L = N/M estimates $\hat{\sigma}_j^2$, which, after being quantised to $\hat{\hat{\sigma}}_j^2$, are transmitted as side information. Both transmitter and receiver use this information to calculate the quantiser step-sizes and bit assignment, based on the smooth approximation of the distribution, which is obtained by linear interpolation of logarithm values. The quantisers are adjusted by variable gain controls $\alpha_i = (\hat{\hat{\sigma}}_i)^{-1}$, i = 0, 1, ..., N - 1, where $\hat{\hat{\sigma}}_i$ is taken from the approximated distribution, and the bit assignment is given by

$$b_i = \bar{B}^* + \frac{1}{2}\log_2\hat{\hat{\sigma}}_i^2 - \frac{1}{2N}\sum_{j=0}^{N-1}\log_2\hat{\hat{\sigma}}_j^2 \qquad (2.99)$$

$$i = 0, 1, \ldots, N - 1,$$

where $\bar{B}^*$ is the average number of bits exclusive of side information. The number L of estimates $\hat{\hat{\sigma}}_i^2$ transmitted as side information is set to 16, and requires a bit-rate of approximately 2 Kbits/sec.

At bit-rates equal to or greater than 12 Kbits/sec this system provides a fairly good performance. A cosiderable amount of audible distortion is, however, increasingly observed at lower bit-rates, but the reproduced speech remains intelligible in the range of 4 to 8 Kbits/sec. Also, it has been shown by means of subjective tests[47] that ATC is better than sub-band coding and ADPCM with variable

predictor (ADPCM-V), at bit-rates of 24, 16, and 9.6 Kbits/sec. In terms of system complexity, however, SBC is simpler, and it has been argued[47] that it involves approximately the same degree of complexity as ADPCM-V.

In a recent work,[120] Zelinski and Noll have proposed two modifications to the basic structure of the ATC scheme. One had the purpose of obtaining a simpler implementation. The other aimed to improve the speech quality for lower bit-rates, which has been observed to be affected by the low-pass effect resulting from the non-transmission of the high end portion of the transform domain at lower bit-rates. The general complexity of the system is, however, higher than the dominant coefficient Hadamard transform coding scheme, and the performance is still not good enough at low transmission bit-rates (between 4.8 Kbits/sec and 8 Kbits/sec).

Another recent approach in ATC techniques is known as vocoder-driven ATC, and has been suggested in 1978 by Tribolet and Crochiere.[119] At bit-rates between 8 and 16 Kbits/sec, this system, by means of a much more complex adaptation strategy, performs considerably better than the original ATC scheme of Zelinski and Noll. However, it is again noticeably degraded when the region of 4.8 Kbits/sec is approached.

The vocoder-driven adaptive transform coder is based on a model of the speech production mechanism, as in the case of vocoder systems, in order to obtain a more efficient bit allocation strategy. In a sense, by using the DCT coefficients this scheme performs an LPC analysis and synthesis to provide a more adequate spectrum approximation including the pitch structure, thus enabling its preservation. In this method, side information is composed of LPC parameters and pitch information.

Another possibility in transform coding is to consider the set of orthogonalised exponentially damped sinusoids with fixed frequencies and fixed damping for each frequency.[121] These form an interesting set of functions on account of their similarities to waveforms of voiced speech. In general, however, transform coding techniques still do not provide good quality speech at bit-rates of about 4.8 Kbits/sec, and only by means of complex adaptation techniques these schemes can produce good results at 8 Kbits/sec.

2.16. Time Coding.

In this section an attempt will be made to bring together a number of time-domain developments on the subject of speech coding. These techniques have in common the basic characteristic of extracting information from the speech waveform only at certain time instants, or intervals. They thus do not transmit information at every sampling instant, as in the case with waveform coders, nor are they based on a model of speech production mechanism, as the vocoder-type devices, and neither do they use transform domain representation of speech as does the class of transform coding techniques. The methods considered in this section will be referred to as time coding techniques, a term which has been used by King and Gosling.[18]

It should be noted that since time coding techniques do not extract information at regular intervals, they generally produce variable-information-rate outputs which require buffer storage in order to obtain a constant data rate for transmission. This aspect has been investigated in more detail, and theoretical and experimental results relating to the problem are given in Chapter 5 of this thesis.

Some basic structures which have been developed in the area of time coding will now be considered.

2.16.1. Clipped Speech.

In 1946, Licklider[136] reported that intelligibility is little affected by clipping the speech waveform, and that even with infinite peak clipping, in which the waveform is reduced to rectangular waves with zero-crossings identical to those of the original signal, intelligible speech could be attained. It was found, however, that intelligibility was critically affected if centre rather than peak clipping was introduced. This can be explained by the fact that, depending on the amount of centre clippling, formant structure can be completely lost.

In another work, published by Licklider and Pollack[137] in 1948, it was verified that intelligibility is affected if speech is first differentiated or integrated before infinite clipping, the first improving and the latter reducing it. In fact, highly intelligible speech (over 90% word articulation) can be obtained if differentiation precedes the infinite clipping process. Licklider and Pollack also verified that differentiation or integration following infinite clipping has no significant influence on intelligibility, but affects quality, the latter improving and the former worsening it. In any case, however, the quality of clipped speech is poor. Clipped speech sounds unnatural and talker recognition is not preserved.

Experiments on time quantisation of infinitely clipped speech have also been carried out,[138] and two quantisation methods have been tested. In one of them, the quantised waveform switches at the end of a quantisation interval if the input signal has switched in that interval (method A); and in another (method B), if the input signal has switched an odd number of times during that interval. The arrangement differentiation-infinite clipper-time quantiser-integrator has been used. Intelligibility tests with a quantisation interval $\Delta T = 0.1$ msec have shown that method B gives about the same score as that of the system without quantisation, and that with $\Delta T = 0.2$ msec intelligible speech is attained, but with a reduced articulation score (about 65%). Lower scores are obtained using method A, but if ΔT is very small (< 0.1 msec) both methods are expected to give identical results, since the input wave is unlikely to switch more than once per interval.

Many years after the work of Licklider, Ainsworth[15] carried out some experiments in which zero-crossings were represented by pulses of short duration with positive or negative polarity. Best intelligibility scores were obtained if positive-going and negative-going zero-crossings could be distinguished, independently of polarity, and the worst results were achieved when no distinction could be made.

Using a quantisation interval equal to 0.1 msec, the bit-rate required for the transmission of clipped speech is seen to be 10 Kbits/sec. If advantage is taken of the fact that the average rate at which data (zero-crossings) is produced is less than the higher rates

at which source sometimes produces its information, then the necessary bit-rate for the transmission of clipped speech is estimated[14] to be between 4700 bits/sec and 6170 bits/sec. In this case, however, the delay that has to be introduced into the communication process may be such as to make two-way conversation difficult.

Since speech quality, and to a certain extent, intelligibility, are not important factors in space applications, clipped speech was the technique chosen for use in the Apollo moon exploration program.[139]

An analysis of speech waveforms has shown[17] that for small intervals τ between zero-crossings (τ < 1 msec), the half-waves with one extreme only are the most frequent one. For larger values of τ, two-absolute maxima half-waves are more likely to occur, and for even larger values of τ three-absolute maxima half-waves are found. Based on these observations, Sobolev and Telepnev[17] proposed a simple method to improve the quality of infinitely clipped speech, in which the half-waves between zero-crossing intervals ($\tau_i = t_{i+1} - t_i$) are approximated by a smooth shape with number of absolute maxima (1 or 2) controlled by τ_i. The shaping function is given by

$$g_i(t) = (-1)^i\left[\sin \frac{\pi}{\tau_i}(t - t_i) + \frac{\tau_i^2}{\epsilon} \sin \frac{3\pi}{\tau_i}(t - t_i)\right] \qquad (2.100).$$

where ϵ is set equal to 5.7×10^{-6} sec^2 in order to result in a shape with very small valley for τ_i = 0.8 msec and large valley for τ_i = 1.6 msec.

In the above method, amplitude information is also obtained every 10 msec by averaging the absolute amplitude values of speech over this period. In terms of quality of regenerated speech, the system has been judged to give better results than clipped speech. The improvement, however, was not considered to be sufficient for commercial application.

2.16.2. Time-Shape Pattern Matching Coder.

The idea of using smooth shapes between zero-crossings in order to achieve a better regeneration quality of speech has more recently been further developed.[18] The technique is referred to by King and Gosling[18] as time-encoded speech (TES). However, it is more appropriate in this review to refer to the technique as time-shape pattern matching coder.

In the system of King and Gosling, zero-crossing intervals τ_i are logarithmically quantised, and shapes between crossings are defined by their number of absolute maxima η_i . An array composed of elements defined by time-shape patterns (τ_i, η_i) is constructed, and a finite number N of patterns with high probability of occurrence is selected to compose the array which will be actually used in the coding process. Thus, the array is comprised of an N-symbol alphabet.

The coding process consists in transmitting for each speech segment between zero-crossings a b-bit word corresponding to a symbol (τ_i, η_i) of the alphabet. This symbol is selected in such a way that the difference between the log-quantised zero-crossing interval τ of the speech segment and τ_i is minimum, and the difference between

the number of absolute maxima η_i associated with τ_i and the number of absolute maxima of the speech segment is minimum. An n-bit word is also transmitted after every eight symbols, which signifies the average amplitude information for those symbols.

Experiments have been carried out[18] with: (a) N = 23, b = 5, n = 8; (b) N = 16, b = 4, n = 8; and (c) N = 8, b = 3, n = 8. In the case of the specification (a), good speech intelligibility can be obtained, but the level of quality is not sufficient for telephone communication applications. For the specifications (b) and (c) intelligible speech has been reported.[18] In these cases, however, the quality of reconstructed speech is substantially reduced and naturalness is lost. Based on a 900 sec^{-1} average zero-crossings rate, bit-rates of 5.4, 4.5 and 3.6 Kbits/sec, respectively, have been estimated[18] when specifications (a), (b), and (c) are used. It is important to note that the bit-rates being referred to are average bit-rates, and that delays of the order of 1 sec may be involved (see Chapter 5 of this thesis) in order to convert the variable-rate-source data into constant-transmission-rate data. However, from an implementation point of view the time-shape pattern matching coder is a simple system, and it appears worth investigating on the possibilities of some further development of this technique.

2.16.3. Extremal Coding.

Another important suggestion in the class of time coding techniques is known as extremal coding.[14,140] Whereas differentiated and infinitely-clipped speech uses only the time instants of extremes, the extremal coding uses as further information about the speech waveform, the amplitudes $\{A_1, A_2, \ldots\}$ at these instants. Thus, the transmitted

information consists of these amplitudes plus the intervals $t_{i+1} - t_i$ between extremes. At the receiver, a smooth function is used to interpolate between them. Additional operations are also introduced in order to remove the extremes that occur during silence intervals due to background noise, and also to remove the extremes that arise from small 'bumps' (i.e., extrema A_{j+1} and A_{j+2} are removed if $|A_{j+2} - A_{j+1}| < B\,|A_{j+1} - A_j|$, where $0 < B < 1$).

Subjective tests have shown that with an average bit-rate of 15 Kbits/sec, the extremal coder can provide speech quality which is much the same as with 30 Kbits/sec PCM. Thus, a fairly good quality of reconstructed speech can be obtained at an average bit-rate of 15 Kbits/sec. However, the performance is not much improved for higher bit-rates. In addition, fairly long delays (of the order of 1 sec) may be involved[14] in order to convert the variable-rate-source data into constant-transmission-rate data having the average bit-rate referred to above.

2.16.4. Pitch-Synchronous Time Coding.

In experiments conducted in 1956 by David and McDonald[13] an attempt was made to exploit, in a simple way, the quasi-periodicity of voiced speech by transmitting only one of every N pitch periods, and at the receiver, repeating the retained period N - 1 times. For N = 6 and a monotone pitch of 5 msec only a small degradation in intelligibility and quality of reconstructed speech was noted. However,

the use of vocoder-generated monotone speech precluded the observation of the effects of overlapping of reconstructed periods onto the retained ones, when a considerable variation of pitch exists over the N periods. In the same work, experiments revealed that the elimination of 50% of the pitch period implies a considerable degradation of speech intelligibility, and that when this proportion is increased the speech is almost unintelligible.

Another technique, considered many years later by Stover,[16] also used elimination of portions of the pitch period but in a somewhat different way. Natural speech was used as input. Since it has been observed[16] that the quality of voiced sounds is not much affected when only 3 msec is extracted from every fifth pitch period, with the condition that the data be recirculated at the speaker's pitch rate, Stover selected only the 3 msec of every fifth pitch period of infinitely clipped speech for transmission. For unvoiced sounds the recirculation occurs at a pseudorandom rather than a periodic rate. Intrapitch exponential damping is restored by an exponential function generator, and voiced to unvoiced intensity ratio is kept to 20 dB. This arrangement has been observed to restore some naturalness which is eliminated by the clipping process. Word articulation scores of between 60% and 96% have been obtained[16] for bit-rates in the range of 1.2 to 9.6 Kbits/sec, but the indications are that the speech quality obtained is still poor at these low-bit-rates.

These experiments, however, provide some information on the possibilities of exploiting the quasi-periodicity of the voiced speech waveform, in a simple way, using the time coding approach.

2.16.5 . Adaptive Aperture Coding.

A recently developed technique[19] of time coding is based on the concept of aperture coding in which sample amplitudes are only transmitted at sampling instants where the amplitude crosses an aperture centred on the last transmitted sample. In the reconstruction process the non-transmitted samples are interpolated (according to some rule) between each pair of transmitted samples.

A method of aperture coding without explicit transmission of the time instants of non-transmitted samples has been described by Jayant and Christensen,[19] and is illustrated in Figure 2.22. The aperture length L is finite and its shape is convergent, defined by a number L of P-points $\{P_1, P_2, \ldots P_L\}$, a number L of N-points $\{N_1, N_2, \ldots, N_L\}$ and a "zero" level Z, according to the equation

$$A(t) = A_0.2^{-Jt}, \qquad (2.101)$$

which has been found appropriate for speech with values of J between 0.5 and 1.

As shown in Figure 2.22, the previously transmitted sample is s(o) by convention. If a sample s(j) crosses the aperture, its nearest point P_j (or N_j) is transmitted, indicating that its amplitude has been quantised to P_j (or N_j), and that samples at t = 1, ..., j, have not been transmitted. In the case that no crossing has occurred up to (and including) t = L then the sample at t = L is encoded by the " zero" level Z.

In Figure 2.22, the dashed horizontal lines indicate that in the reconstruction process, the amplitude of each non-transmitted sample is set equal to the last transmitted sample amplitude. This is equivalent to a 1st-order prediction with coefficient $a_1 = 1$. First-order prediction with coefficient $a_1 < 1$, higher-order prediction, and adaptation of the prediction coefficients have also been considered.

Adaptive apertures have been suggested which follow the rule

$$A_o^{(r+1)} = G_1 A_o^{(r)} + G_2 (ADAPT)^{(r)} \tag{2.102a}$$

where

$$(ADAPT)^{(r)} = 1 \text{ if } \sum_{s=1}^{4} R_{r-s} < K \tag{2.102b}$$

$$= 0 \quad \text{otherwise.}$$

In these expressions, R_j is the run length of the aperture j, defined as the number of samples which did not cross the aperture plus 1, $0.95 \leq G_1 \leq 0.99$, and K = 6, 6, and 8, for L = 1, 2, and 3, respectively.

For 8 KHz sampling, $G_2 = 45$, and L = 3 with Huffman coding, the delay involved has been reported as 80 msec. This can be reduced to 35 msec if $G_2 = 35$, and L = 2 without Huffman coding.[19]

The use of 1st-order adaptive prediction and sampling rates of 8 and 12 Ksamples/sec resulted in highly intelligible speech at average bit-rates of 9.6 and 14.4 Kbits/sec, and the reconstructed speech has been judged to be better than adaptive delta-modulated speech at similar bit-rates. However, a compromise exists since the latter is a simpler coding scheme.

At the bit-rates mentioned above, some improvement has been obtained using other variations of adaptive aperture coding.[20]

It should be mentioned that the results presented in this section on time coding techniques served to indicate that the time coding schemes can produce better results than waveform coders at low-transmission bit-rates, without there being any major sacrifice in terms of complexity (although they are generally not as simple as waveform coders). However, there is the disadvantage that the delay which must be introduced into the communication process may sometimes be too long to permit normal two-way conversation.

2.17. Discussion.

From the large number of speech coding techniques presented in this review, it can be seen that at very low-bit-rates (such as 4.8 Kbits/sec and less), the quality of the speech obtained is still not sufficiently good to be acceptable for normal telephone communication. Vocoder-type techniques are the only ones which, in most cases, are capable of producing fairly good results at low-bit-rates - they do, however, suffer from being talker and environment dependent. Furthermore, from an implementation point of view, vocoder-type schemes are very complex. At bit-rates of roughly 9.6 Kbits/sec, those techniques which avoid the pitch extraction problem, such as the phase vocoder and the voice-excited vocoder can provide, in general, better quality (more natural) speech than can pitch-excited vocoders. Those methods are, however, still fairly complex and the speech quality obtained is not sufficient for commercial application. Systems which are very easy

to implement, such as waveform coders, can give reasonably good quality speech at bit-rates above 16 Kbits/sec, but are not useful for low transmission bit-rates. The signal quality becomes poor when the bit-rate approaches the 8 Kbits/sec range. Transform coding techniques, on the other hand, can usually provide reasonably good quality speech at bit-rates in the range of 9.6 Kbits/sec, but commercial quality is still not achieved at these rates. At 4.8 Kbits/sec, even the most complex forms of transform coding cannot produce sufficiently good quality speech. However, the simplicity of the Hadamard transform coding technique, in particular, is an attractive aspect of the system. Further investigation, forming part of the research work reported on in this thesis, has been carried out. Time coding techniques constitute another class of systems, which is further investigated in Chapter 4 of this thesis. These techniques exploit the characteristics of the speech waveform to a larger extent than waveform coders, and can potentially produce more significant results at bit-rates between 4.8 Kbits/sec and 9.6 Kbits/sec, without involving any major sacrifice in terms of system complexity. It is, however, important to note that time coding methods suffer from delay problems due to the variable-rate-data which is in general produced by these systems. These problems are treated in more detail in Chapter 5 of this thesis.

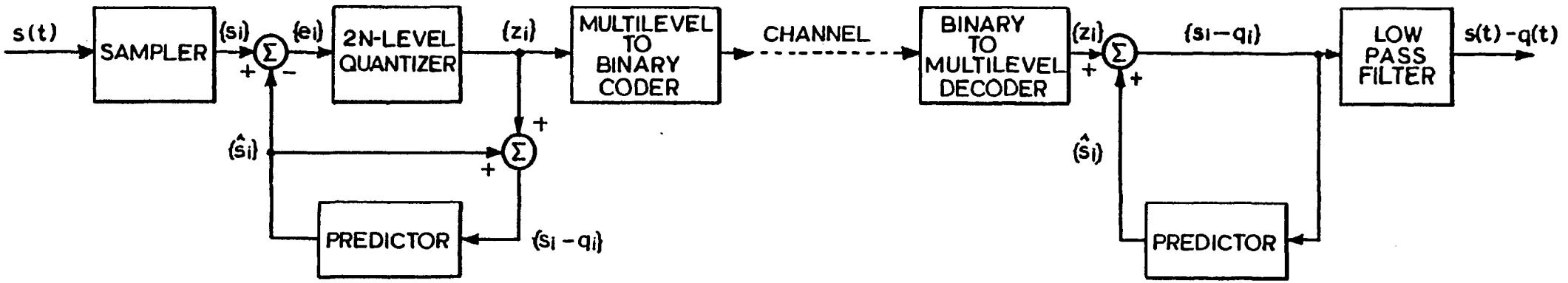


Figure 2.1. DPCM system.

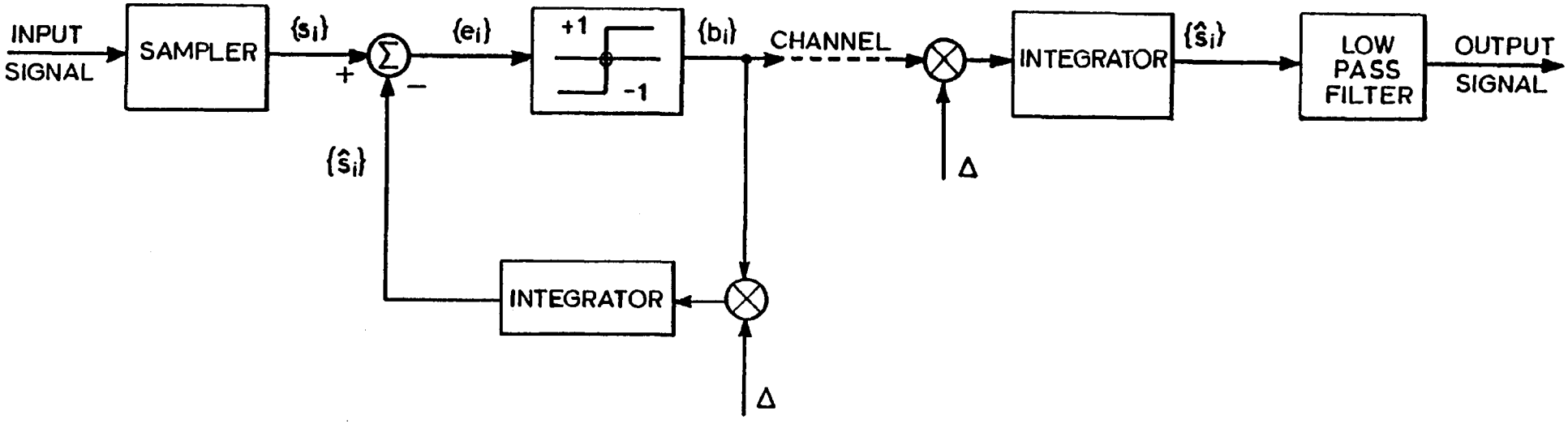


Figure 2.2. DM system.

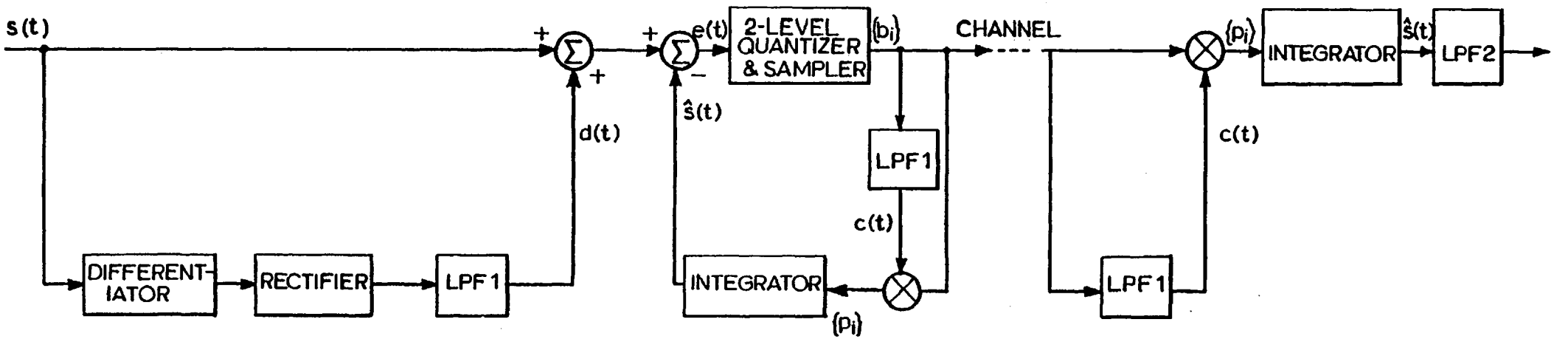


Figure 2.3. Continuous DM.

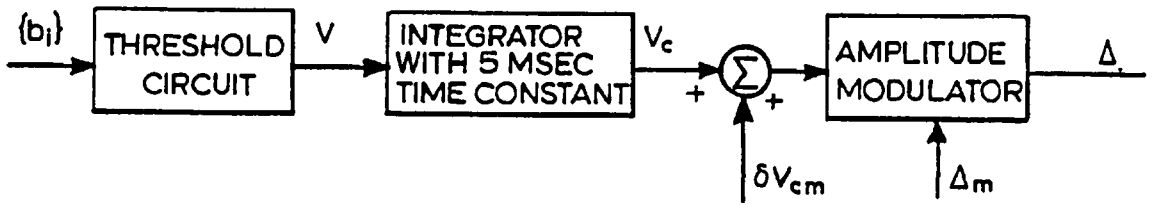


Figure 2.4. Step-size adaptation logic in DCDM.

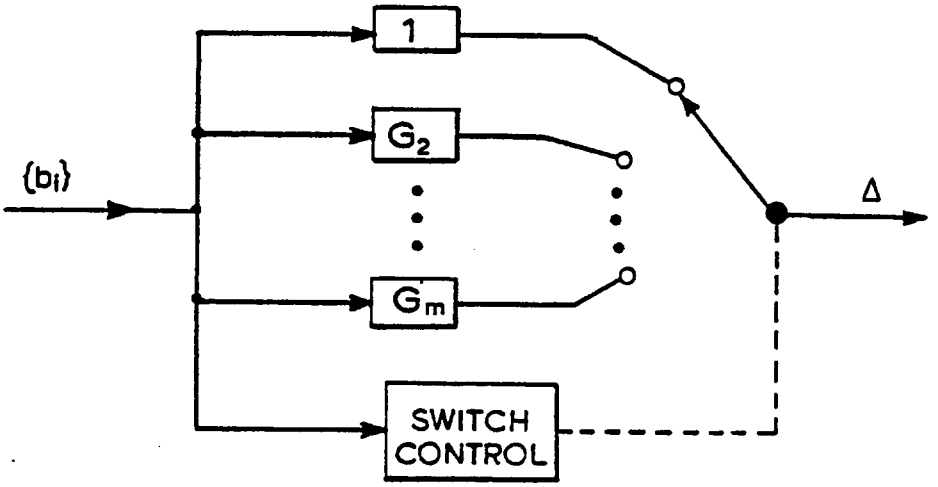


Figure 2.5. Step-size adaptation logic in discrete adaptive DM.

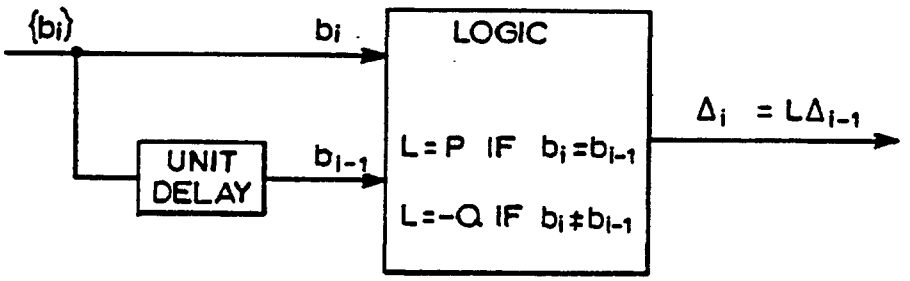


Figure 2.6. Step-size adaptation logic in one-bit memory DM.

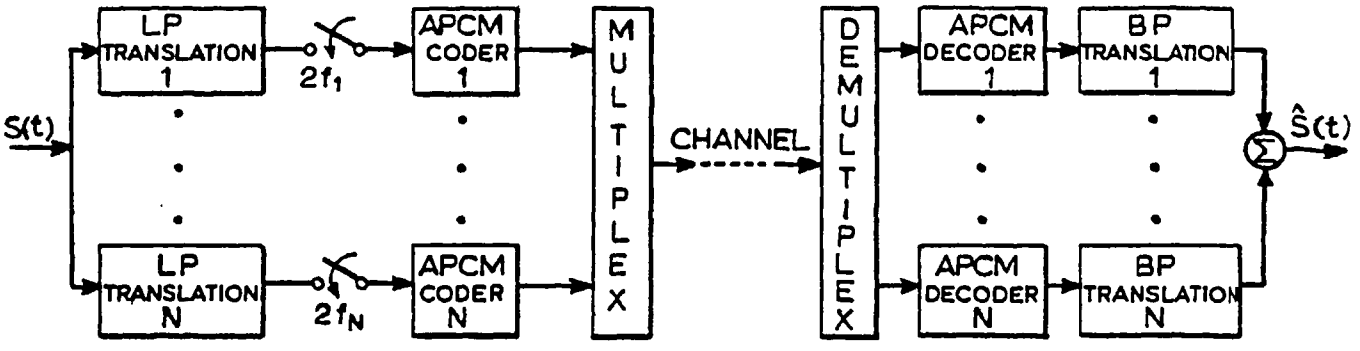


Figure 2.7. Sub-band coder using low-pass translation technique.

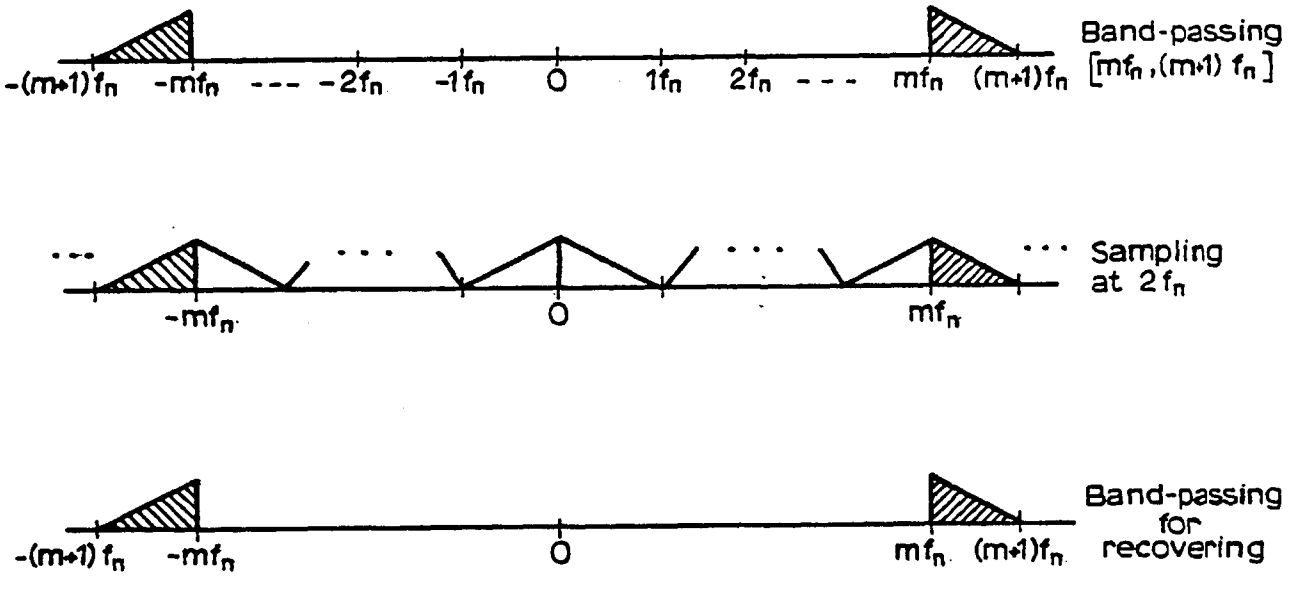


Figure 2.8. Technique of integer-band sampling.

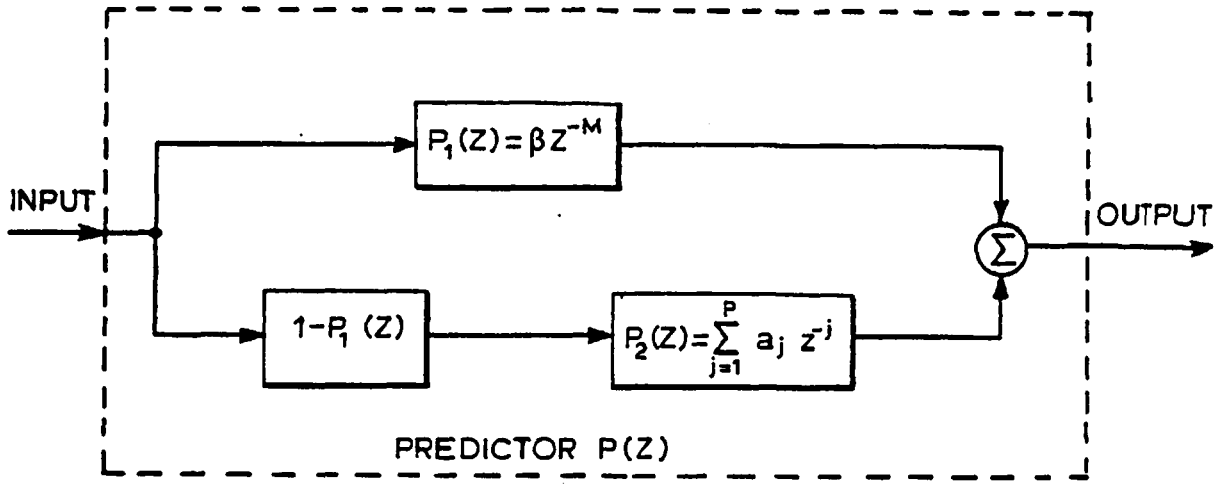


Figure 2.9. Predictor used in adaptive predictive coding.

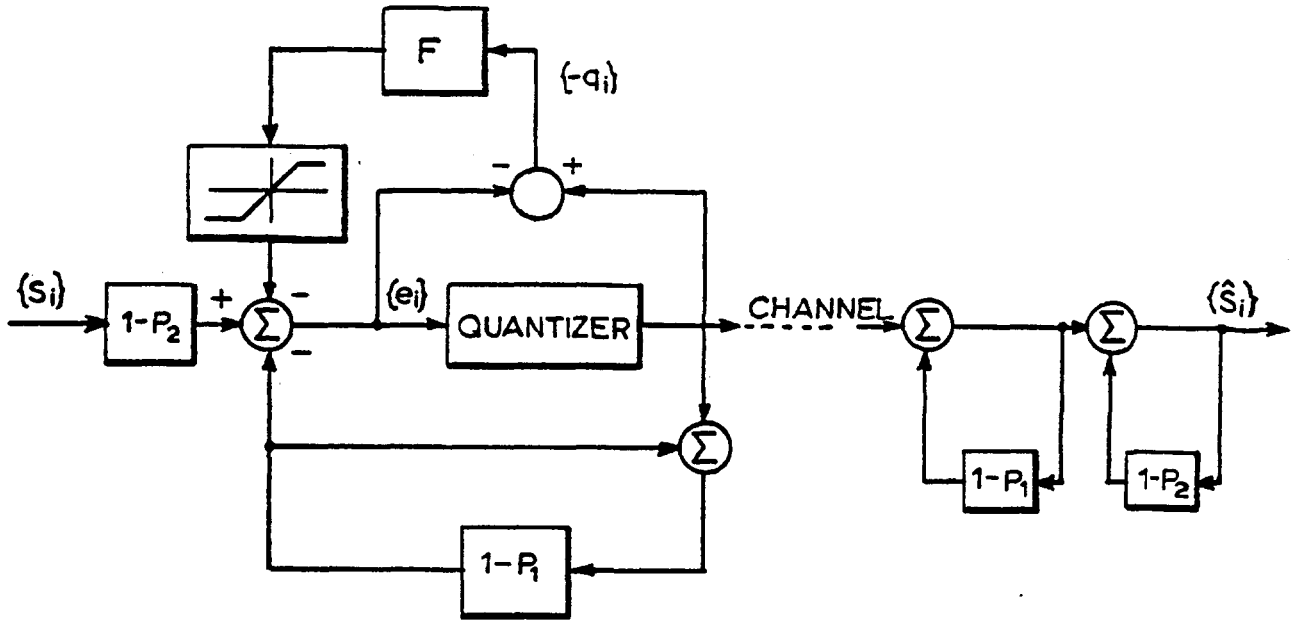


Figure 2.10. Adaptive predictive coder with adjustable noise spectrum.

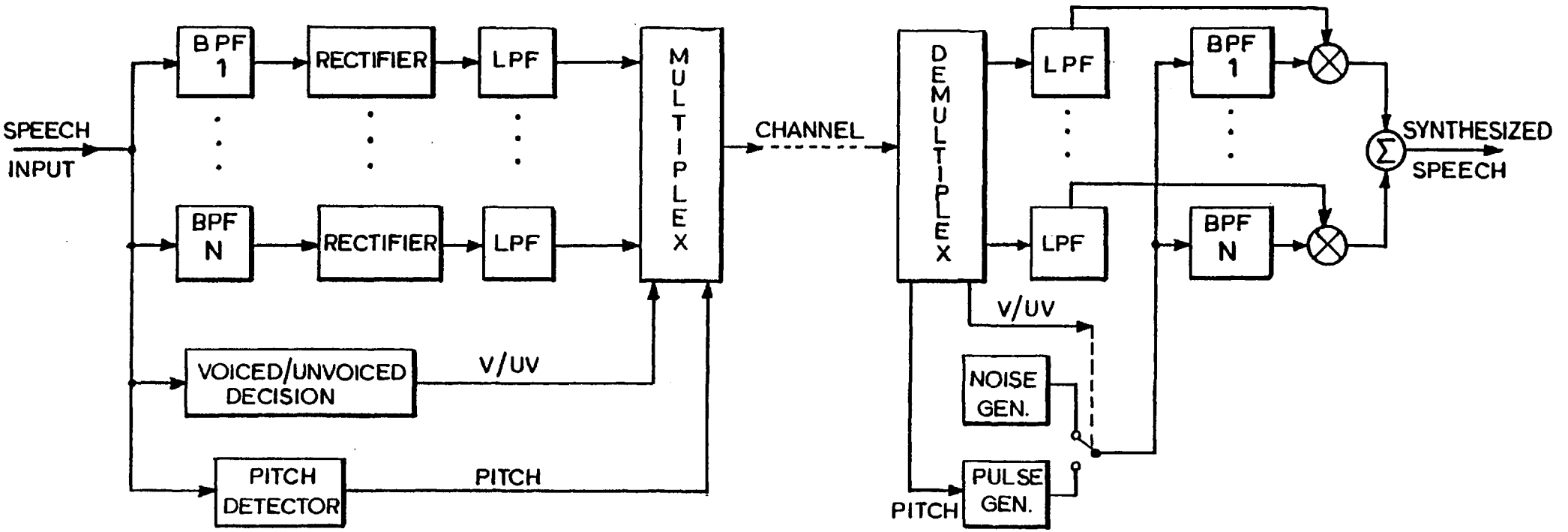


Figure 2.11. Channel vocoder.

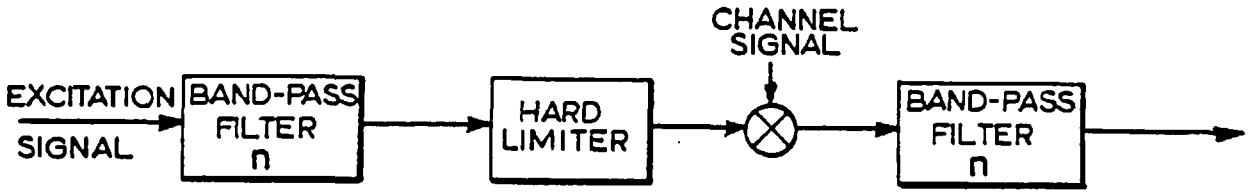


Figure 2.12. Channel synthesis using spectral flattening.

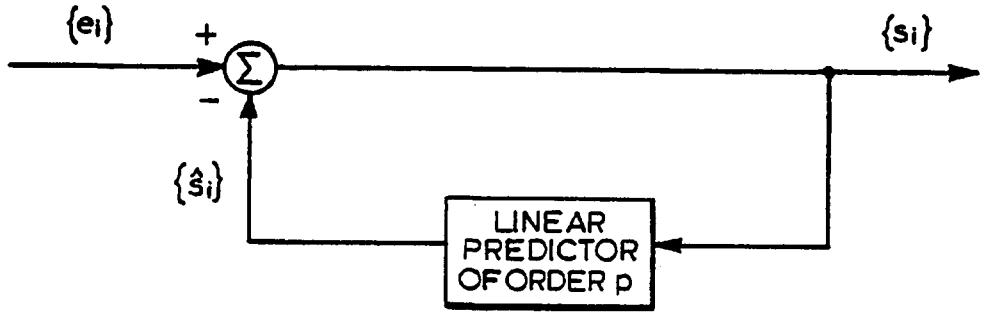


Figure 2.13. Discrete all-pole model in the time domain.

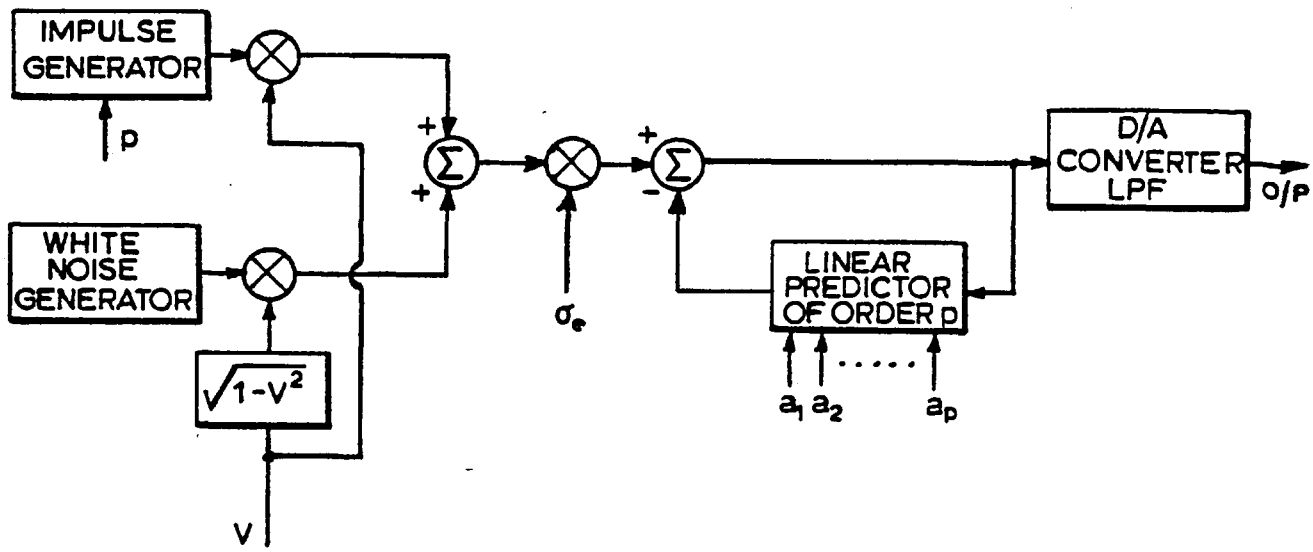


Figure 2.14. Synthesis procedure in the maximum likelihood vocoder.

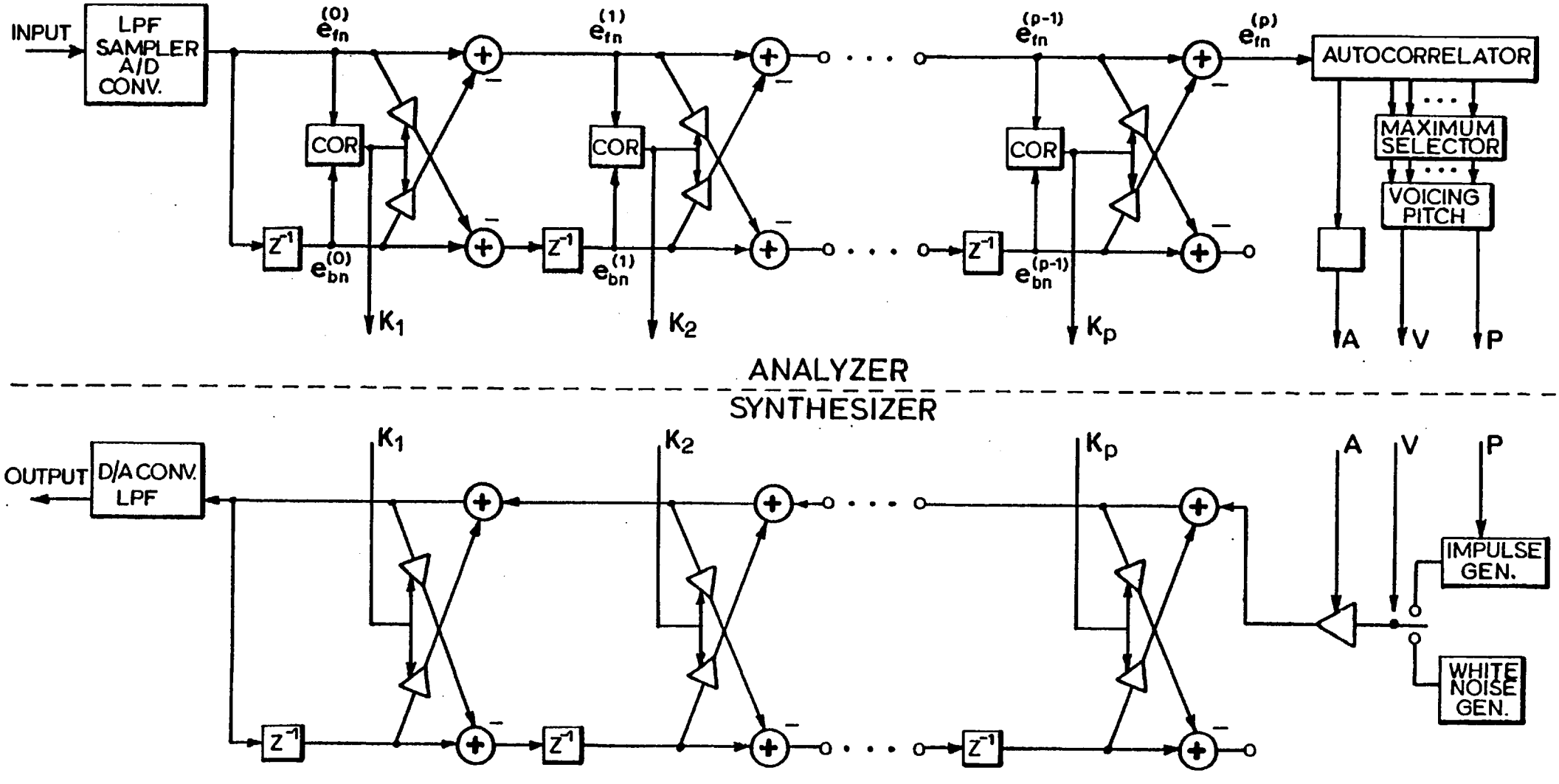


Figure 2.15. PARCOR vocoder.

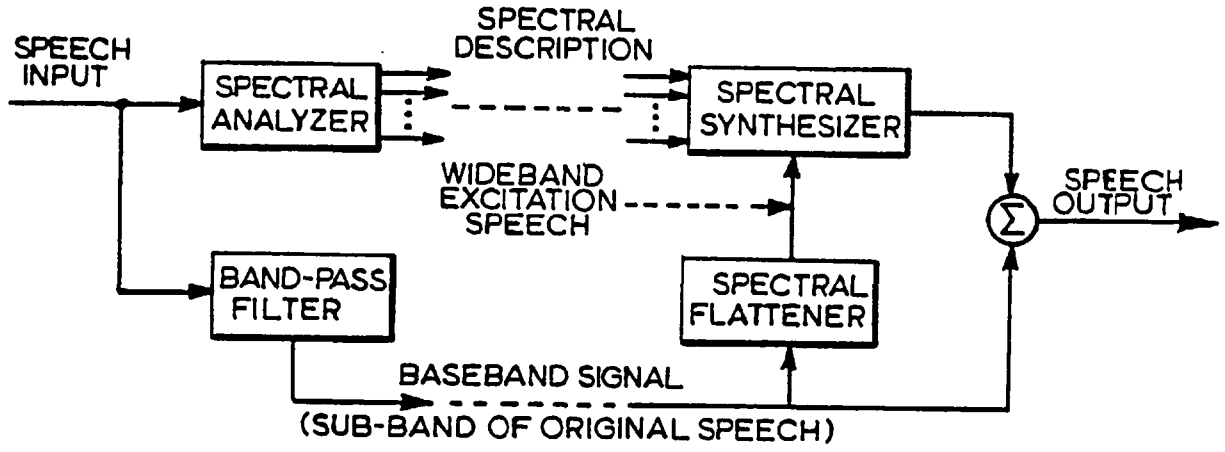


Figure 2.16. Voice-excited vocoder.

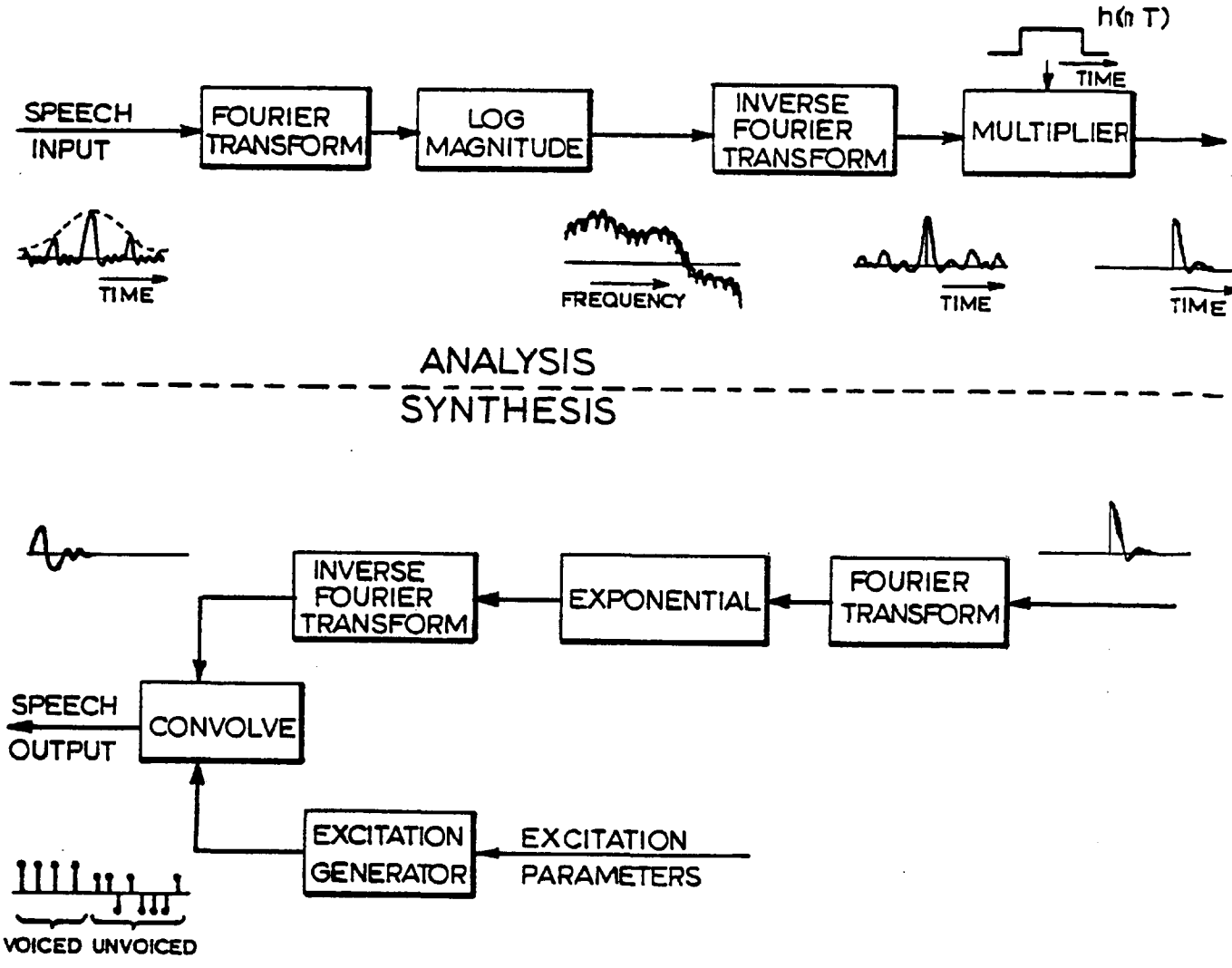


Figure 2.17. Homomorphic vocoder.

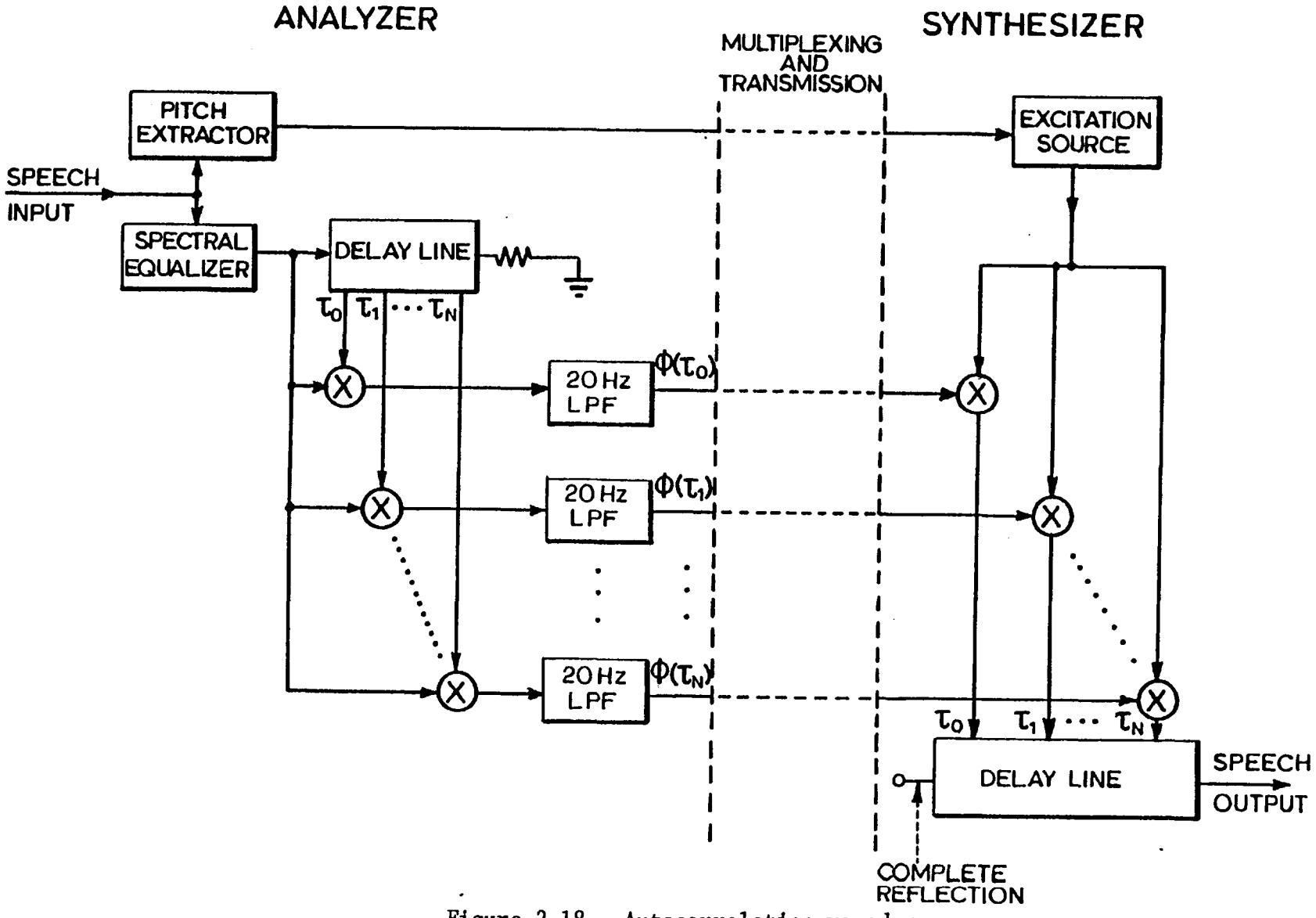


Figure 2.18. Autocorrelation vocoder.

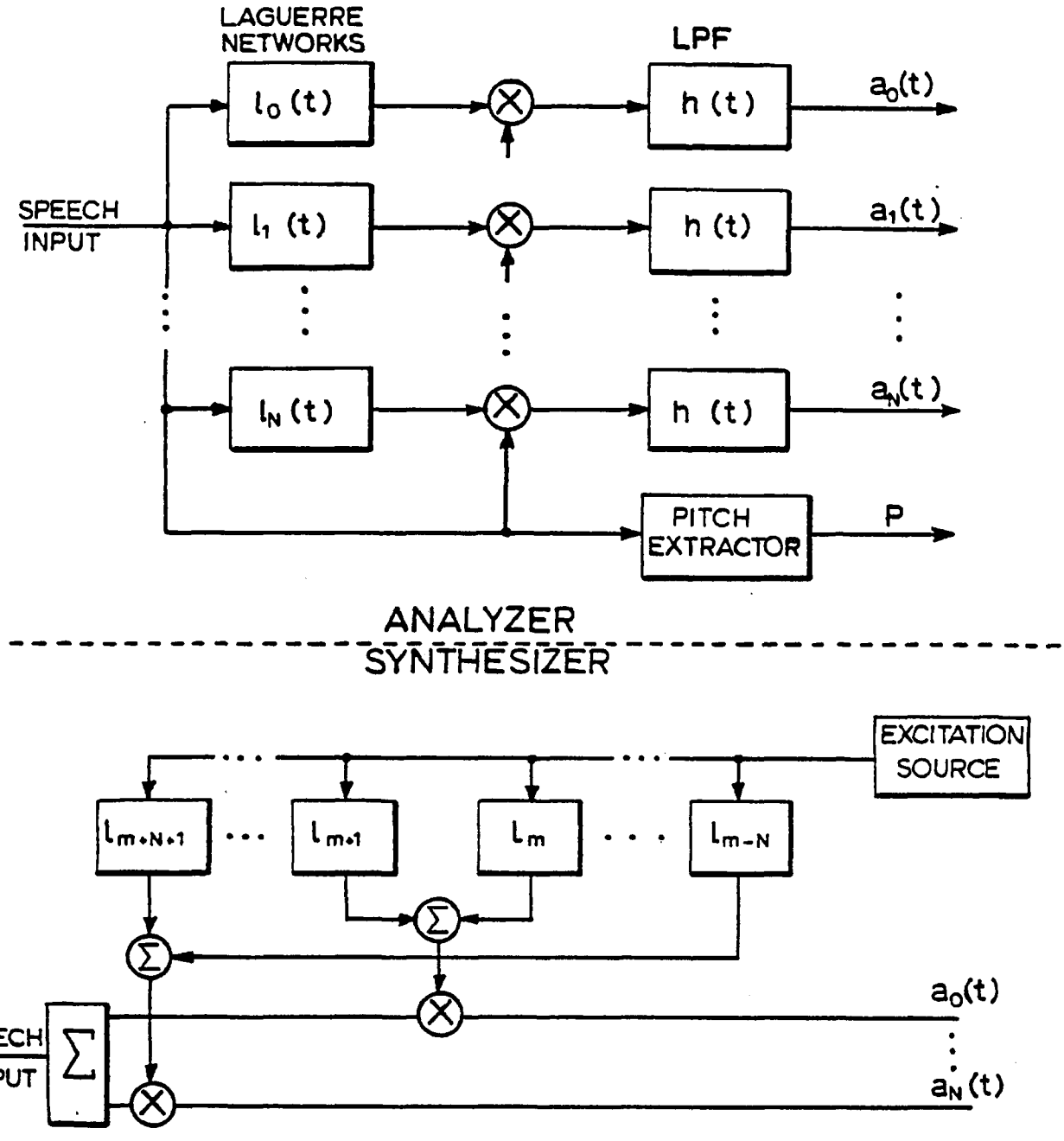


Figure 2.19. Laguerre vocoder.

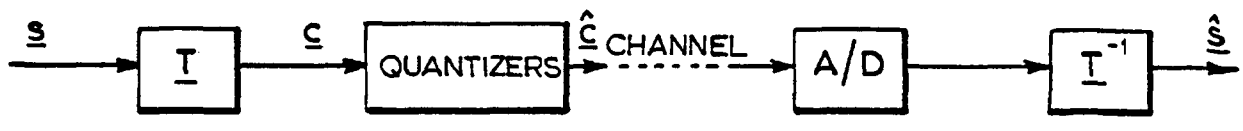


Figure 2.20. Transform coding.

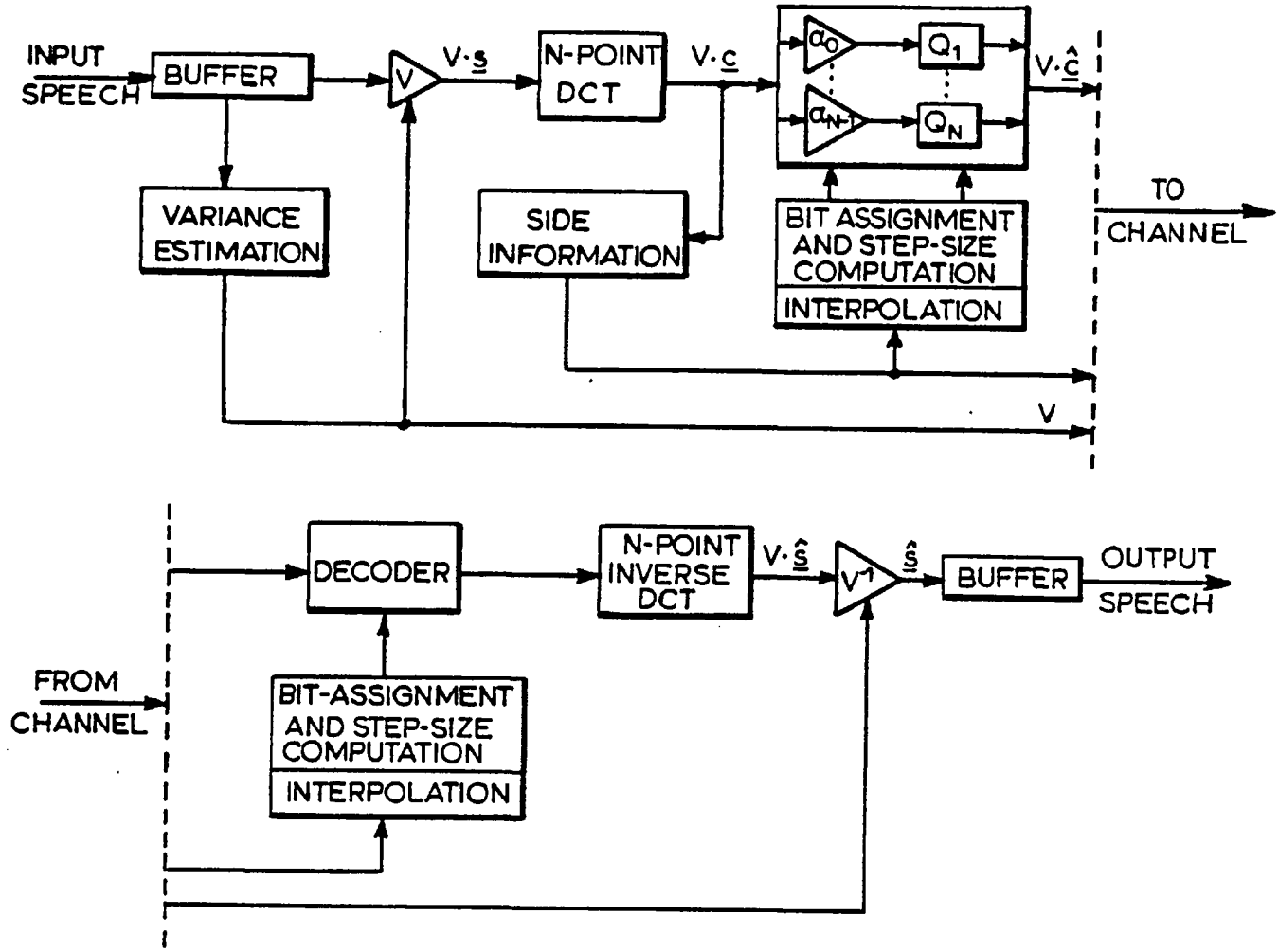


Figure 2.21. Adaptive transform coder.

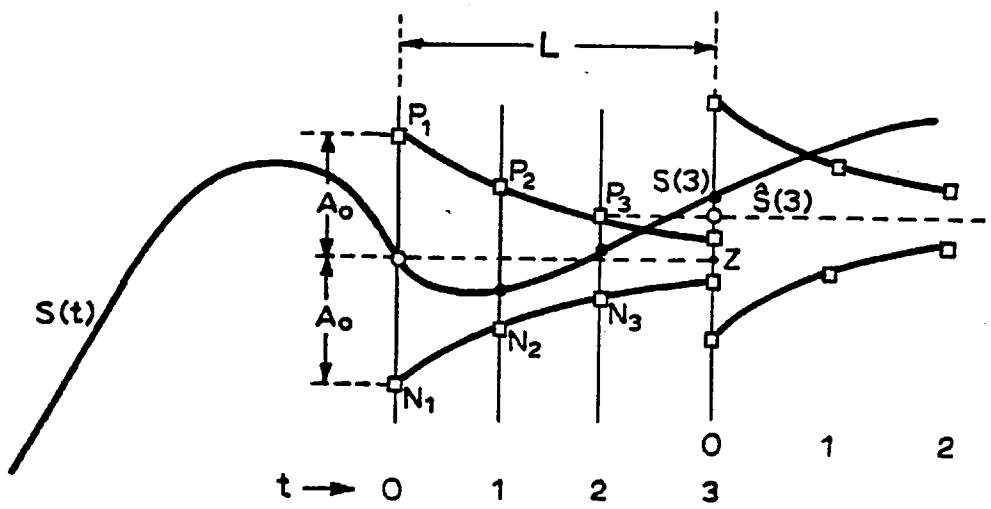


Figure 2.22. Illustration of aperture coding of length L = 3.

CHAPTER 3

NEW APPROACHES TO THE HADAMARD TRANSFORM CODING OF SPEECH

CHAPTER 3

NEW APPROACHES TO THE HADAMARD TRANSFORM CODING OF SPEECH

As mentioned in Chapter 2, the selection and use of dominant (high-value) Hadamard coefficients as a means of obtaining low-bit-rate digitally coded speech has received much attention in recent years. It has been found, however, that at bit-rates below 6 Kbits/sec the decoded speech obtained with the simple Hadamard transform coding technique is not of commercial quality.[12] In fact, at these low-bit-rates, the performance is substantially lower than the results obtained at bit-rates of 9.6 Kbits/sec and above.

In this chapter, two methods aimed at improving the performance of the system at these low-bit-rates are examined. In the first part of the chapter, the possibility of further exploiting the residual correlation between the dominant coefficients of the speech segments in the Hadamard domain is investigated. In the second part, consideration is given to the effects of spectrum aliasing and discontinuities in the Hadamard transform coded speech. Certain modifications for reducing the influence of these effects without changing the bit-rate are then discussed. Illustrative examples of the results obtained with these methods are included in the tape recording attached to this thesis.

3.1. Double Hadamard Transform Coding Technique.

3.1.1. Motivation

Unlike the Karhunen-Loève transform, the signal correlation is not completely eliminated by the Hadamard transformation. In fact, one finds a number of non-vanishing elements which are not located on the main diagonal of the covariance matrix in the transform domain.[114] Consequently, more efficient methods of transmission can be obtained by exploiting this residual redundancy which exists in the Hadamard transform representation of the information signal. One such possibility has been investigated by Frangoulis and Turner.[118] The method used consisted of the successive application of the Hadamard and Haar transforms, such that the remaining correlation between the Hádamard coefficients could be further exploited by the Haar transform.

As noted in Chapter 2, a reasonably good quality of speech can be attained if, out of a total of 64, 7 or 8 dominant coefficients from the Hadamard transform representation are used.[117] If some residual redundancy between these dominant coefficients still remains then it is feasible to reduce their representation resulting in a more efficient method of transmission. One possibility involves the application of a second Hadamard transformation to the set of dominant coefficients.

In the double Hadamard transform method an N-point transform is first applied to a block consisting of N speech samples $\{s_0, s_1, \ldots, s_{N-1}\} = \underline{s}^t$, resulting in a set of coefficients $\{c_0, c_1, \ldots, c_{N-1}\} = \underline{c}^t$ given by

$$\underline{c} = \frac{1}{N} \underline{H_N} \cdot \underline{s}, \tag{3.1}$$

where $\underline{H_N}$ is the Hadamard matrix of order N. From the set of coefficients $\{c_0, c_1, \ldots, c_{N-1}\}$ given by (3.1), a sub-set is selected, which consists of n (< N) dominant coefficients $\{d_0, d_1, \ldots, d_{n-1}\} = \underline{d}^t$, where each d_j, j = 0, 1, ..., n-1, denotes one of the dominant coefficients selected from $\underline{c}$. This sub-set of dominant coefficients is then n-point Hadamard transformed to $\underline{h}$, which is the double Hadamard transform representation of the input signal, i.e.,

$$\underline{h} = \frac{1}{n} \underline{H_n} \cdot \underline{d}, \tag{3.2}$$

where $\underline{h}^t = \{h_0, h_1, \ldots, h_{n-1}\}$. It should be noted that the successive application of the transformation, where the second application is performed on a sub-set of the complete set of N Hadamard coefficients differs fundamentally from a successive application in which the second one is performed on the complete set of N Hadamard coefficients. The latter procedure would simply reconvert the N Hadamard coefficients to the original speech signal.

3.1.2. The Average Power Density Function in the Double Hadamard Transform Domain.

In the investigation of the potential offered by the double Hadamard transform method, specific values of N = 64 and n = 8 were considered. The main factor governing this choice was the fact that the 8 dominant coefficient representation of 64-point Hadamard transformed speech provides a reasonably good reproduction of the signal.

As observed in Chapter 2, the possibilities offered by orthogonal transform methods lie mainly in the degree to which the distribution of the coefficients in the transform domain can be efficiently exploited. In the case of straight Hadamard transform coding, for example, the average power density function in the transform domain shows that, depending on the spoken material and the speaker, between 55% and 88% of the total power is concentrated[117] in the 8 dominant coefficients out of the total 64.

The aim of introducing the double Hadamard transform technique was to determine whether or not the use of only a sub-set of the coefficients obtained after the application of the second transformation would effectively represent the original set of dominant terms. In order to obtain quantitative data on this question, speech signals were examined in terms of their average power density function in the double Hadamard transform domain. The speech material used in this investigation was low-pass filtered to 3.04 KHz and then sampled at a frequency of 8 KHz.

Figures 3.1 to 3.4 show the average power distribution of the 8 coefficients (ordered from the largest to the smallest values) obtained after the application of the second Hadamard transformation. The figures relate to four different sentences taken from the Harvard list number 1 (see appendix). From the figures it can be seen that most of the power is concentrated in only a small sub-set of the coefficients, independent of the speech material. In addition, the results indicate that this concentration occurs irrespective of the sex of the speaker, although the percentage of the total power contained in a few dominant coefficients can be seen to be slightly lower in the case of the female speaker. The specific values of the percentage power contained in 1, 2, and 3 dominant coefficients are presented in Table 3.1 for both male and female speakers, and for the same four sentences used in Figures 3.1 to 3.4. These results show that for the male speaker between 40.2% and 79.8% of the total average power is contained in the 1 to 3 dominant coefficients, respectively, while for the female speaker, these figures are 37.4% and 78.2%, respectively. If should also be noted that the percentage power contained in 1 dominant coefficient is between 38.4% and 39.7% less than the concentration of power in 3 dominant coefficients.

Table 3.1. Percentage Power Contained in Hadamard Coefficients After Second Transformation.

Sentence Number	Number of Coefficients	Percentage Power Male Speaker	Percentage Power Female Speaker
1	1	40.7	38.2
	2	64.6	62.3
	3	79.7	77.4
2	1	41.0	37.4
	2	65.4	61.8
	3	79.8	76.9
3	1	40.2	39.4
	2	63.3	63.3
	3	78.6	78.2
4	1	40.2	37.8
	2	63.5	62.0
	3	79.2	77.5

3.1.3. Bit Allocation for Primary and Secondary Labelling.

A simplified block diagram of the double Hadamard transform coding scheme is shown in Figure 3.5, where N = 64 and n = 8. The input shown in this figure consists of blocks of 64 samples of a speech waveform which has been low-pass filtered and sampled at a rate of 8 Ksamples/sec. These blocks ($\underline{s}$) of samples are 64-point Hadamard transformed, according to equation (3.1), to $\underline{c}^t = \{c_0, c_1, \ldots, c_{63}\}$ and a set of 8 dominant coefficients $\underline{d}^t = \{d_0, d_1, \ldots, d_7\}$ is selected. The corresponding position of each d_j in $\underline{c}$ is extracted for transmission and is denoted as primary labelling. After the selection and extraction of 8 dominant Hadamard coefficients, the next stage in the encoding procedure is to apply an 8-point Hadamard transform to the 8 previously selected coefficients ($\underline{d}$), according to equation (3.2), and to select from the new resulting Hadamard coefficients $\underline{h}^t = \{h_0, h_1, \ldots, h_7\}$ a reduced number (m) of dominant coefficients. The amplitude of these coefficients are quantised and encoded for transmission, and their corresponding location in $\underline{h}$, which is denoted as secondary labelling in Figure 3.5, is also transmitted.

At the receiver all coefficients in $\underline{h}$ are set equal to zero, except for the set of m retained coefficients. The 8-point inverse Hadamard transform (IHT) is then taken. This results in a set of coefficients which represents an approximation to $\underline{d}$. Using the primary labelling technique, these coefficients are placed in their

appropriate locations in $\underline{c}$, with all the remaining components being set equal to zero. In the final stage of decoding, a 64-point inverse Hadamard transform (IHT) is applied, and the resulting sequence of samples is then low-pass filtered to produce the output speech.

As has already been mentioned, the primary labelling is obtained by selecting the 8 high absolute value Hadamard coefficients from the total of 64. In order to obtain a reduced number of labelling bits this selection was carried out bearing in mind the following two observations from reference 117, concerning dominant coefficients associated with 64-point Hadamard transformed speech:

A. Each one of 7 dominant coefficients is associated with each of the Walsh-Fourier spectral components G_0, G_1, ..., G_6, which comprise the following sets of Hadamard coefficients:

G_0 : {Hadamard coefficient : 0}

G_1 : {Hadamard coefficients: 32 to 63}

G_2 : {Hadamard coefficients: 16 to 31}

G_3 : {Hadamard coefficients: 8 to 15}

G_4 : {Hadamard coefficients: 4 to 7}

G_5 : {Hadamard coefficients: 2 and 3}

G_6 : {Hadamard coefficient : 1} .

B. There is a high probability that the 8 high value Hadamard coefficients are contained within a sub-set S of 32 coefficients, where S is given by

$S = \{0, 3, 8, 10, 12, 14, 16, 19, 20, 24, 26, 28,$
$30, 32, 34, 36, 38, 39, 40, 42, 44, 46, 48,$
$50, 52, 54, 56, 58, 59, 60, 62, 63\}.$

On the basis of these observations, the set of 8 dominant coefficients were selected by choosing as a member of the set the largest absolute value coefficient in each of the following groups:

$S_0 = \{0\}$

$S_1 = \{\text{from 32 to 63}\}$

$S_2 = \{16, 19, 20, 24, 26, 28, 30\}$

$S_3 = \{8, 10, 12, 14\}$

$S_4 = \{4, 5, 6, 7\}$

$S_5 = \{3\}$

$S_6 = \{1\}$

$S_7 = S - \{0, 3\}.$

With coefficients selected in this way, the number of bits necessary to identify the coefficient associated with each of the sets are 0, 5, 3, 2, 2, 0, 0, and 5, respectively. This constitutes the bit allocation for primary labelling. Thus, a total of 17 bits are used in this procedure.

The secondary labelling, which consists of the identification of the m dominant coefficients out of 8 after the application of the second Hadamard transformation, is obtained by a straightforward selection from the 8 coefficients. Thus, a further 3 bits per coefficient are used for the purpose of labelling the 8 coefficients, and thereby signifying to the receiver which have been selected to represent the speech.

3.1.4. Discussion of Experimental Results.

In order to obtain some indication of the subjective quality of the reconstructed speech, the double Hadamard transform coding technique was simulated on a digital computer. The speech signals used in the system input were sentences taken from the Harvard list number 1. The sentences were first low-pass filtered to 3.04 KHz and then sampled at a rate of 8 Ksamples/sec. The samples were A/D converted using a 10-bit convertor, and then fed into a PDP-15 computer which was used to simulate the system. After processing, the digitised samples were again converted to analogue form by means of a 10-bit D/A convertor, and then low-pass filtered to recover the continuous signals.

In the double Hadamard transform coding approach the resulting bit-rate (bits/sec) is given by

$$B = (b_p + b_s + \sum_{i=1}^{m} b_i).f_s/N \qquad (3.3)$$

where b_p is the total number of bits per block used in the primary

labelling; b_s is the total number of bits per block used in the secondary labelling; f_s is the sampling frequency; and the b_i's are the number of bits used for coding the amplitude of each transmitted coefficient. In the present case N = 64, b_p = 17, f_s = 8 KHz, and b_s = 3.m.

The number (m) of dominant coefficients selected from the second Hadamard transformation was varied from 1 to 3. In the case of m = 1, an eight-bit linear quantisation of the coefficient amplitude was used; for m = 2 , seven bits per coefficient were used; and for m = 3 seven bits per coefficient were used for each of the two largest coefficients and six bits for the third coefficient. Under these conditions, the resulting bit-rates associated with the encoding were 3.5, 4.625, and 5.75 Kbits/sec, for m = 1, 2, and 3, respectively.

Informal listening tests indicated that even though the quality of the reproduced speech was lower than would be expected from the results given by the average power density function, fairly good speech intelligibility could be obtained with the 4.625 and 5.75 Kbits/sec schemes, especially for male voices. The system's increased ability to handle the male rather than the female voice can be accounted for by the higher percentage power contained in the dominant coefficients for the male speaker.

When only one dominant coefficient was used (m = 1), with its associated bit-rate of 3.5 Kbits/sec, the resulting speech quality was poor but intelligibility was not lost. However, the quality was considerably worse than that obtained when 3 dominant coefficients were

used. It should be noted that the concentration of the total average power in the latter (approximately 80%) was about twice that obtained when only 1 dominant term was used.

The first three sections of the tape recording include some of the results obtained from the simulations at the three previously mentioned bit-rates, in addition to the results obtained with the original Hadamard and Hadamard-Haar schemes, using the same speech material, at 4.875 Kbits/sec and 3.25 Kbits/sec.

The following two important results are worth noting with regard to the performance and ease of implementation of the 4.625 and 3.5 Kbits/sec double Hadamard technique as compared with the corresponding Hadamard and Hadamard-Haar schemes operating at similar bit-rates (4.875 and 3.25 Kbits/sec):

A. The speech reproduced by the double Hadamard transform coding method sounds more natural (less 'metallic') than that produced by the Hadamard and Hadamard-Haar techniques; and
B. the double Hadamard scheme is more easily implemented than the Hadamard-Haar technique, since the latter involves a three-valued second transformation and is of much higher order (N = 64) than the second Hadamard transform (n = 8) used in the double Hadamard technique. The system is only slightly more complicated to implement than the straight Hadamard encoding scheme.

The conclusions can be summarised as follows: the technique of reducing the residual redundancy in dominant Hadamard coefficients by applying another Hadamard transformation can provide intelligible,

though somewhat noisy, speech reproduction at low-bit-rates (5.75, 4.625 and 3.5 Kbits/sec). It has the advantage of sounding more natural than that obtained when using the Hadamard and Hadamard-Haar techniques, and furthermore the system presents a simpler structure than the Hadamard-Haar scheme.

3.2. The Combined Use of Oversampling and Smoothing in Decoding.

As has already been mentioned, the quality of the decoded speech obtained with the simple Hadamard coding is substantially degraded for bit-rates below 6 Kbits/sec. The work presented in this section aims to determine some of the factors which may contribute to this degradation, and to introduce some modifications for reducing the influence of these factors.

As the main interest was in low-bit-rate Hadamard transform coding schemes, two particular cases, both using a 64-point transform, were considered. In one case, 3 dominant coefficients were selected, with 7 bits being used for coefficient-amplitude quantisation. In another case, 4 dominant coefficients were selected, and 4 bits were used for amplitude quantisation. Using a sampling frequency of 8 KHz, the corresponding bit-rates are 4.875 Kbits/sec and 5 Kbits/sec for the 3 and 4-dominant coefficient systems respectively.

3.2.1. Motivation.

In the decoding process associated with the dominant coefficient Hadamard technique, an N-dimensional sequence consisting of n (< N)

transmitted coefficients in their correct position, and N-n zero elements in the remaining positions, is inverse Hadamard transformed. The result is a sequence of N samples at a rate equal to the sampling rate, which is usually 8 Ksamples/sec. This sequence forms a staircase-shaped waveform since the decoded waveform is the result of a linear combination of a small number (as compared to N) of square-wave functions.

In order to illustrate the type of regenerated signal which is obtained from a Hadamard transform coder at low-bit-rates, Figures 3.6 to 3.9 show 16 msec segments of original and processed speech waveforms. Figures 3.6 and 3.7 illustrate the reconstruction by the 3 dominant Hadamard coefficients out of a total of 64. Figures 3.8 and 3.9 present, for the same speech segments respectively, the regenerated signal obtained when the 4 dominant coefficient system was used. As can be seen from these figures, the reconstructed waveform is of the staircase type.

Spectral characteristics of both the original and Hadamard processed speech are presented in Figures 3.10 to 3.15. The short-time spectrum magnitudes of two different 32 msec segments of voiced speech are shown in Figures 3.10 and 3.13. The corresponding spectrum magnitudes for Hadamard processed speech using 3 dominant coefficients are presented in Figures 3.11 and 3.14. Figures 3.12 and 3.15 show the spectra resulting from the use of a 4 dominant coefficient system. The spectral characteristics of the Hadamard processed speech shown in these figures are typical examples of the magnitude spectra which were obtained for a wide variety of speech sounds.

Some interesting observations can be noted from various comparisons among Figures 3.6 to 3.15. Although the pitch and the low-frequency spectrum of the original signal are preserved relatively well by the encoding process, it is clear that high-frequency distortions occur. The spectrum of the Hadamard coded speech exhibits stronger high-frequency components with respect to the low ones than is the case with the original spectrum. In addition, the spectral envelope at these high frequencies is completely lost. This can be explained, in part, by an aliasing effect resulting from the wide-band spectrum of the regenerated signal and the low output sampling rate (8 Ksamples/sec).

Another aspect of the Hadamard reconstruction of speech signals is the discontinuities in the processed waveform due to the staircase shaped characteristic resulting from the Hadamard transform coding process. This has been considered as a factor affecting speech quality when only a small number of Hadamard coefficients (in the present study, 3 or 4 out of 64) are used to transmit speech information.

In the final part of this section, a number of modifications are introduced in the Hadamard coded speech in order to determine the extent to which the above two factors contribute to the quality degradation observed in the regenerated signal.

3.2.2. Proposed Modifications.

In an attempt to reduce the influence of the aliasing effect, the number of output samples per second was increased. The resulting waveform, which is the low-bit-rate Hadamard representation of the original speech, is thus described in the time-domain by a number of samples per second at a much higher rate than the original sampling rate. The direct consequence of this modification is that when the discrete samples are passed through the output low-pass filter the extent of the aliasing is reduced so that the spectral characteristics represent the processed speech spectrum more accurately. Output sampling rates of 2 and 4 times the original sampling rate (i.e., 16 and 32 Ksamples/sec respectively), were examined.

A further modification of the coding scheme concerned the smoothing of the processed waveform in order to avoid the discontinuities introduced by the Hadamard transform processing. The smoothing of the Hadamard waveform representation of speech can be thought of as adaptive low-pass filtering. This produces a smooth waveform following the characteristics of the Hadamard representation of the original speech, and as such differs significantly from the usual output low-pass filter.

Various types of smoothing were considered, with the functions $F(X) = X^2(3-2X)$ and $F(X) = \sin^2(\pi X/2)$ being used as interpolating functions. In some methods of smoothing, the zero-crossings of the

Hadamard processed speech were preserved and the interpolating functions $\cos(2\pi X /4)$ and $\sin(2\pi X /4)$ were used for the preservation of the zero-crossings. The smoothing operation can be described as follows: if A and B are the sample amplitudes at the time instants t_A and t_B, then the amplitude of the sample interpolated at the instant t between t_A and t_B is given by

$$s(t) = A + (B - A).\ F\left(\frac{t - t_A}{t_B - t_A}\right). \qquad (3.4)$$

The amplitudes A and/or B were taken either as the extremes of the Hadamard processed speech, or as waveform levels of constant amplitude in the Hadamard reconstructed speech. In these cases, the instant (t_A or t_B) associated with each of these amplitudes (A or B) was taken as the midpoint, in time, of the constant amplitude interval. Obviously, in situations where the zero-crossings were preserved, A (or B) was equal to zero, and t_A (or t_B) was a zero-crossing position. Figure 3.16 illustrates the smoothed waveform of the Hadamard processed speech segment shown in Figure 3.7.

Figure 3.17 shows the effects on the short-time spectrum magnitude of the Hadamard processed speech, when output oversampling of 16 Ksamples/sec and waveform smoothing were employed. The same speech segment used to obtain the spectrum magnitude shown in Figure 3.13 was also used in Figure 3.17. The spectrum shown in Figure 3.17 is a typical example illustrating the elimination of the

extraneous high-frequency components appearing in the Hadamard processed speech (compare with Figures 3.13 and 3.15) and a better shaping of the spectral envelope. These figures do, however, illustrate the inability of the low-bit-rate Hadamard transform coding to preserve the high-frequency characteristics of the original speech information.

3.2.3. Speech Quality Obtained with Modified Hadamard Coding.

Original and modified Hadamard coding systems were simulated on a PDP-15 computer, using sentences (spoken by both male and female speakers) taken from the Harvard list number 1. Two Hadamard transform coding systems were simulated - one used 3 dominant coefficients and 7 bits per coefficient amplitude, and the other used 4 dominant coefficients and 4 bits to quantise each coefficient amplitude. In both cases the order of the transformation was 64.

Section 4 and 5 of the enclosed tape recording illustrate the effect of output oversampling at 16 K samples/sec for the 3 and 4 dominant coefficient systems respectively. Sections 6 and 7 illustrate the effect of the combined use of oversampling and smoothing in decoding.

The quality of the original Hadamard systems, at the low-bit-rates considered in this study, is very poor as compared with existing complex schemes operating at similar bit-rates. Although spectral improvements were obtained with the modified Hadamard coding, simulation of the systems revealed that the resulting speech quality is still not good enough. Increasing the output sampling rate was found to produce only a marginal reduction in the noise characteristics associated with the

processed speech. This can be explained by the fact that although the aliasing effect is important, the other degrading factors associated with the Hadamard transform method of coding are in comparison very much more serious. The introduction of waveform smoothing, however, tends to reduce the 'harshness' of the reconstructed signal, particularly when combined with the increased output sampling rate. Although significantly better than that obtained with straight Hadamard coding, the resulting speech is still not of commercial quality.

Although not included in the attached recorded material, the effect of using only waveform smoothing is to produce a 'muffled' speech which is more degraded than that obtained from the combined use of oversampling and smoothing in decoding. This indicates that if the aliasing effect is not reduced, the low-pass characteristic resulting from the smoothing process is not sufficient to produce results which are noticeably superior to those obtained with the original Hadamard technique.

The speech quality obtained when using the increased output sampling rate of 16 Ksamples/sec differed very little from that obtained with 32 Ksamples/sec. In addition, the difference between the speech quality obtained with the various methods of smoothing was not significant.

The main conclusions can be summarised as follows: the investigation has revealed that aliasing results in a small contribution to the distortions, and that the inherent degradation due to Hadamard coding

is much greater. When the elimination of the aliasing effect is coupled with the smoothing of the waveform, a noticeable improvement in the quality of the regenerated speech results. Although this improvement does not result in speech of commercial quality, it indicates that the method may be useful when applied to other coding techniques which involve similar signal regeneration schemes.

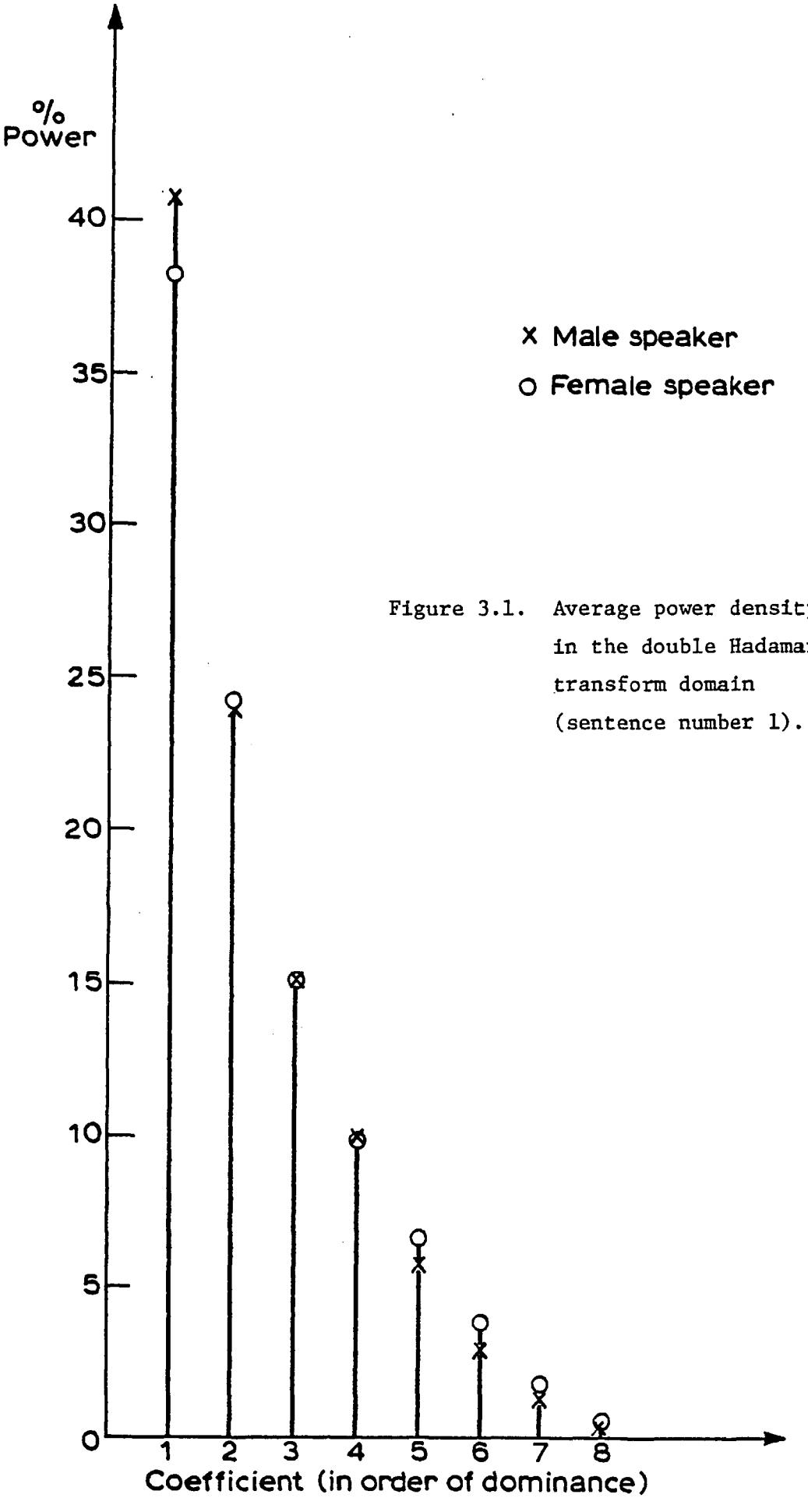


Figure 3.1. Average power density in the double Hadamard transform domain (sentence number 1).

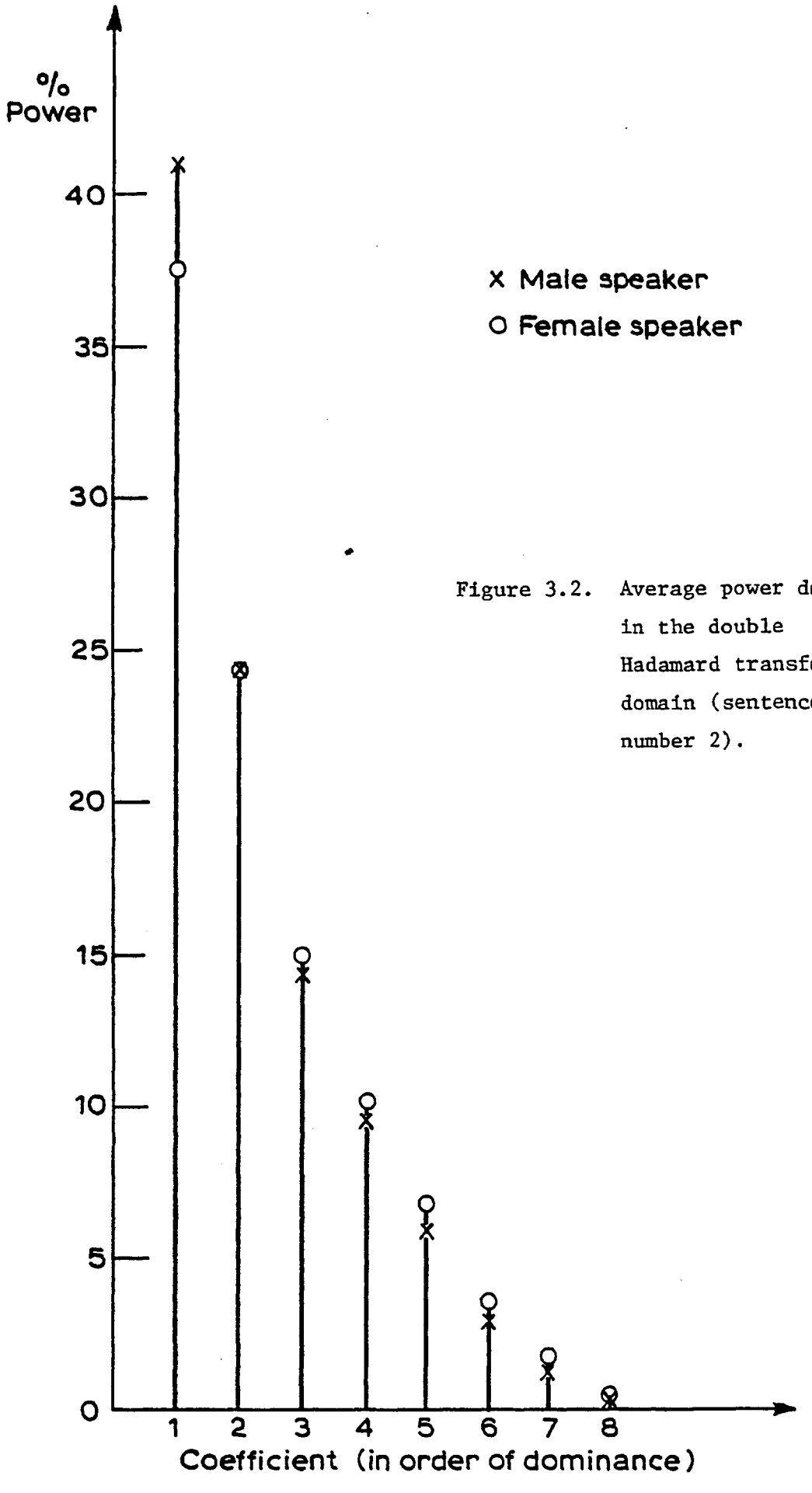


Figure 3.2. Average power density in the double Hadamard transform domain (sentence number 2).

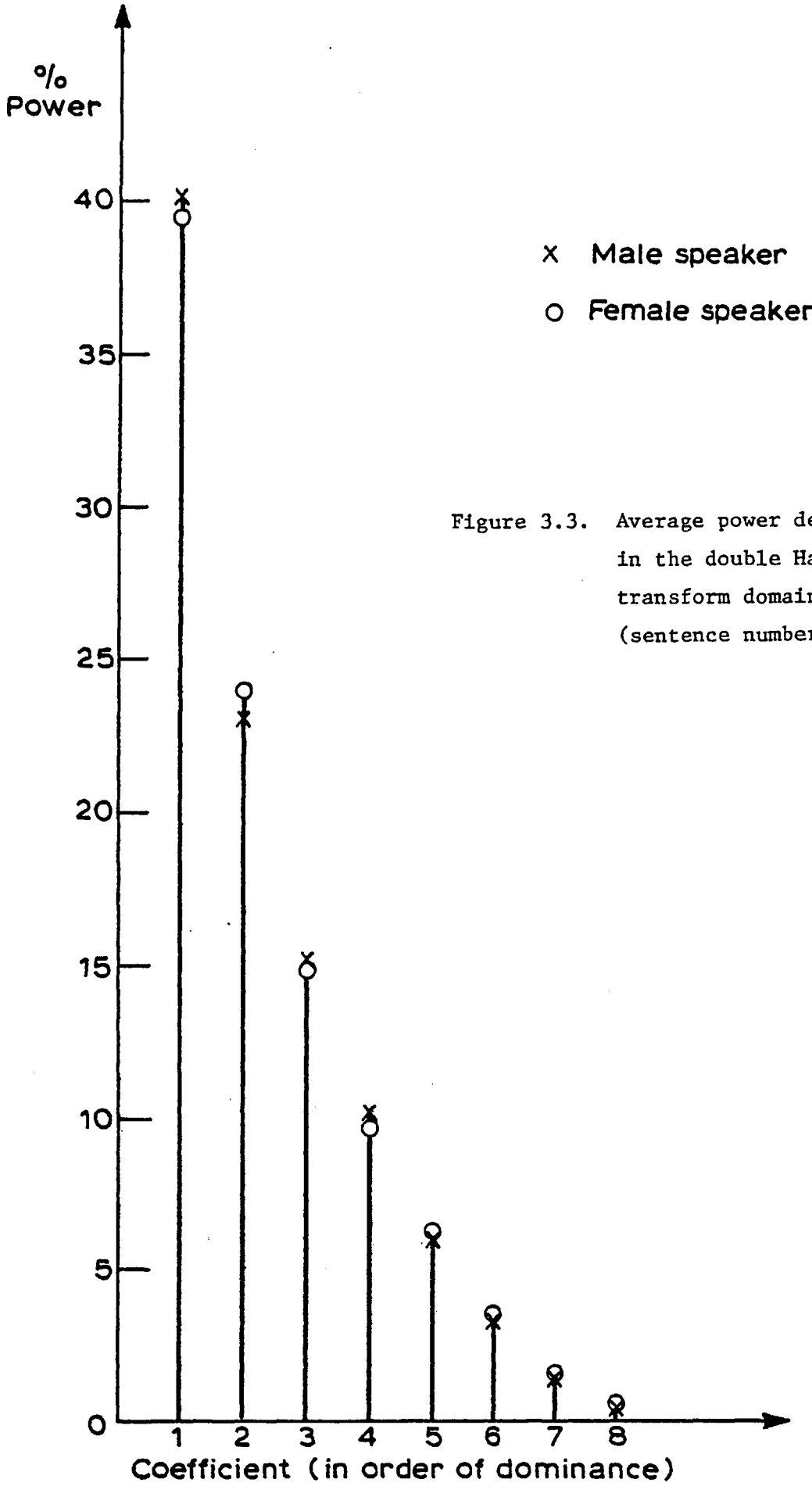


Figure 3.3. Average power density in the double Hadamard transform domain (sentence number 3).

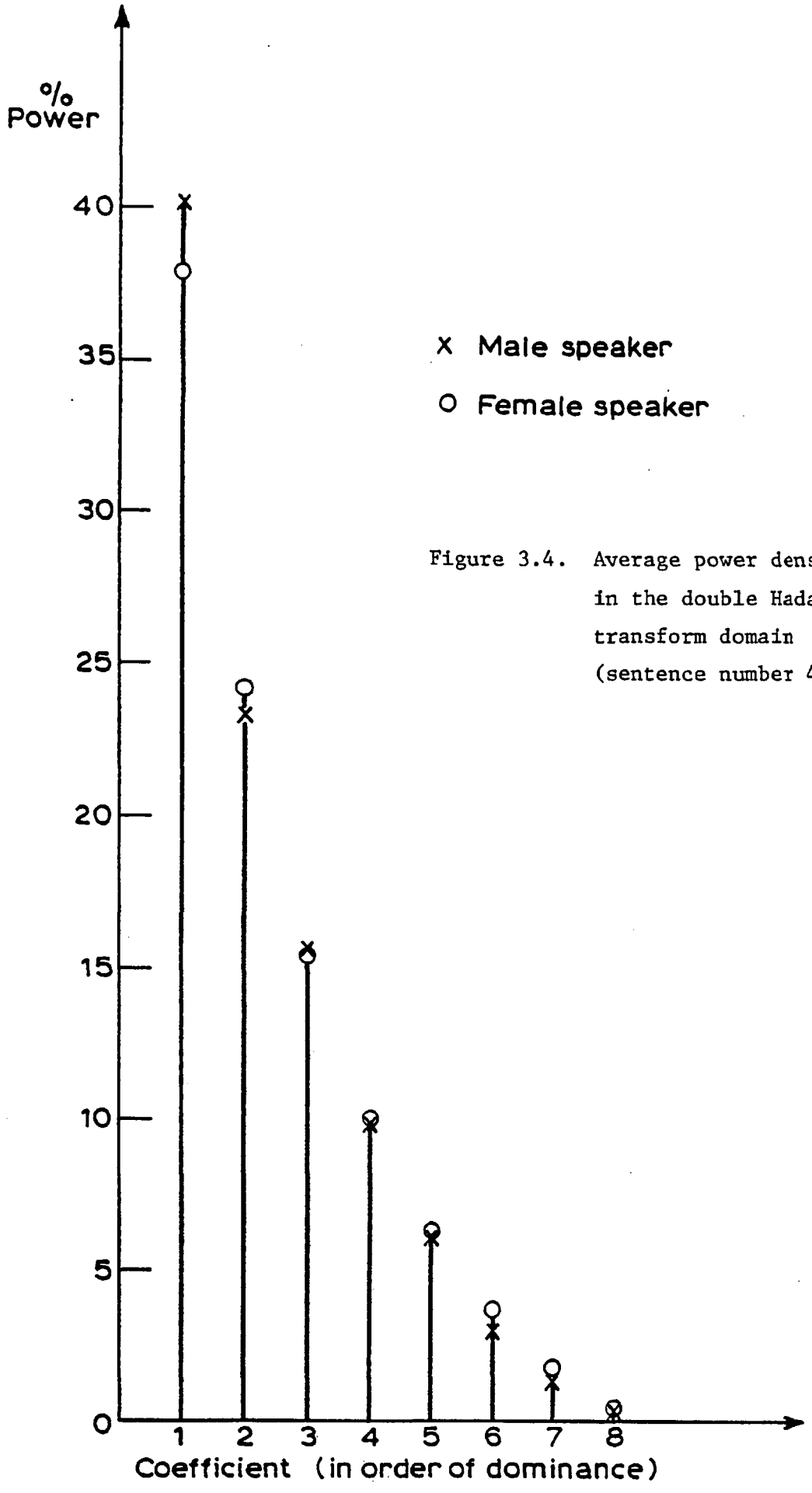


Figure 3.4. Average power density in the double Hadamard transform domain (sentence number 4).

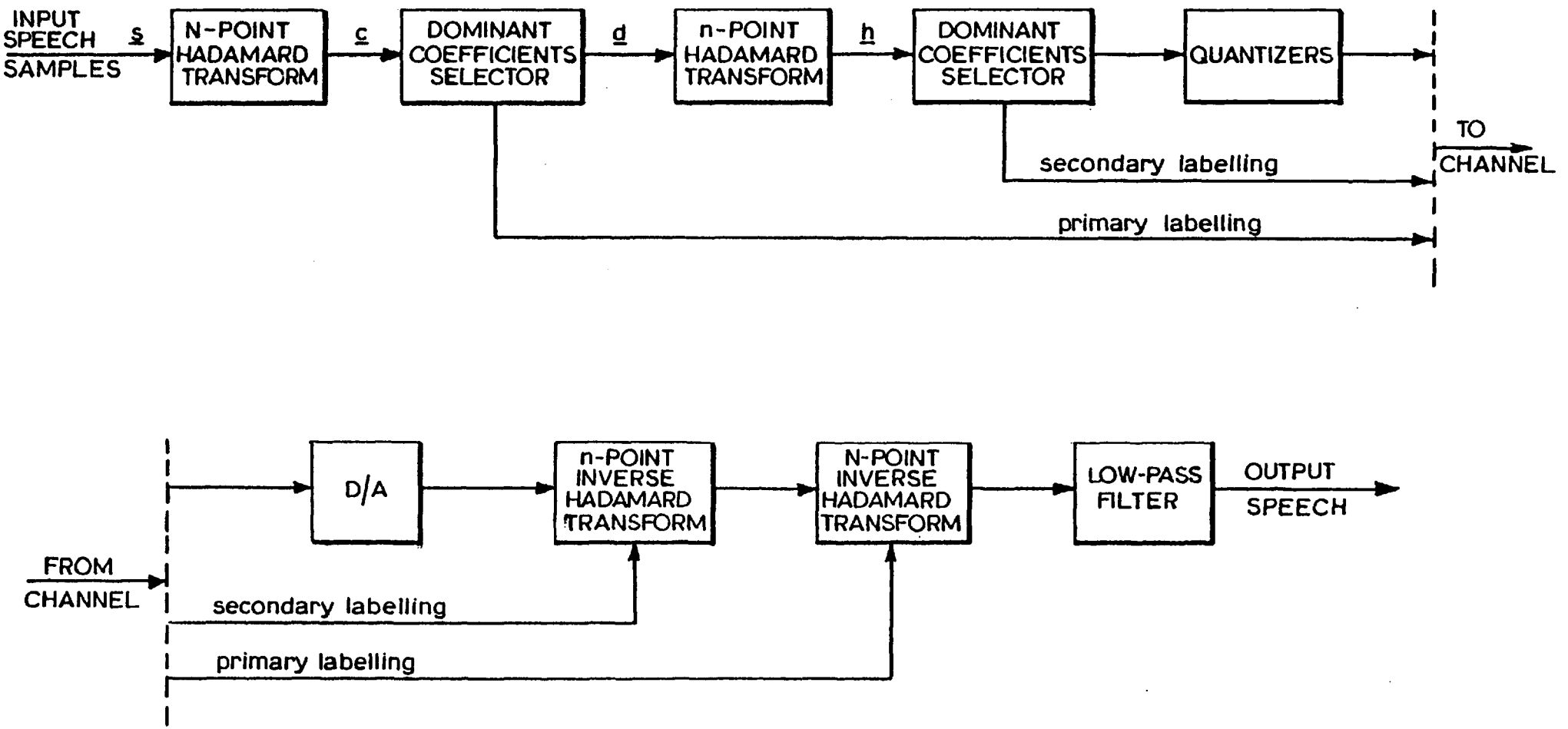


Figure 3.5. Double Hadamard transform coding.

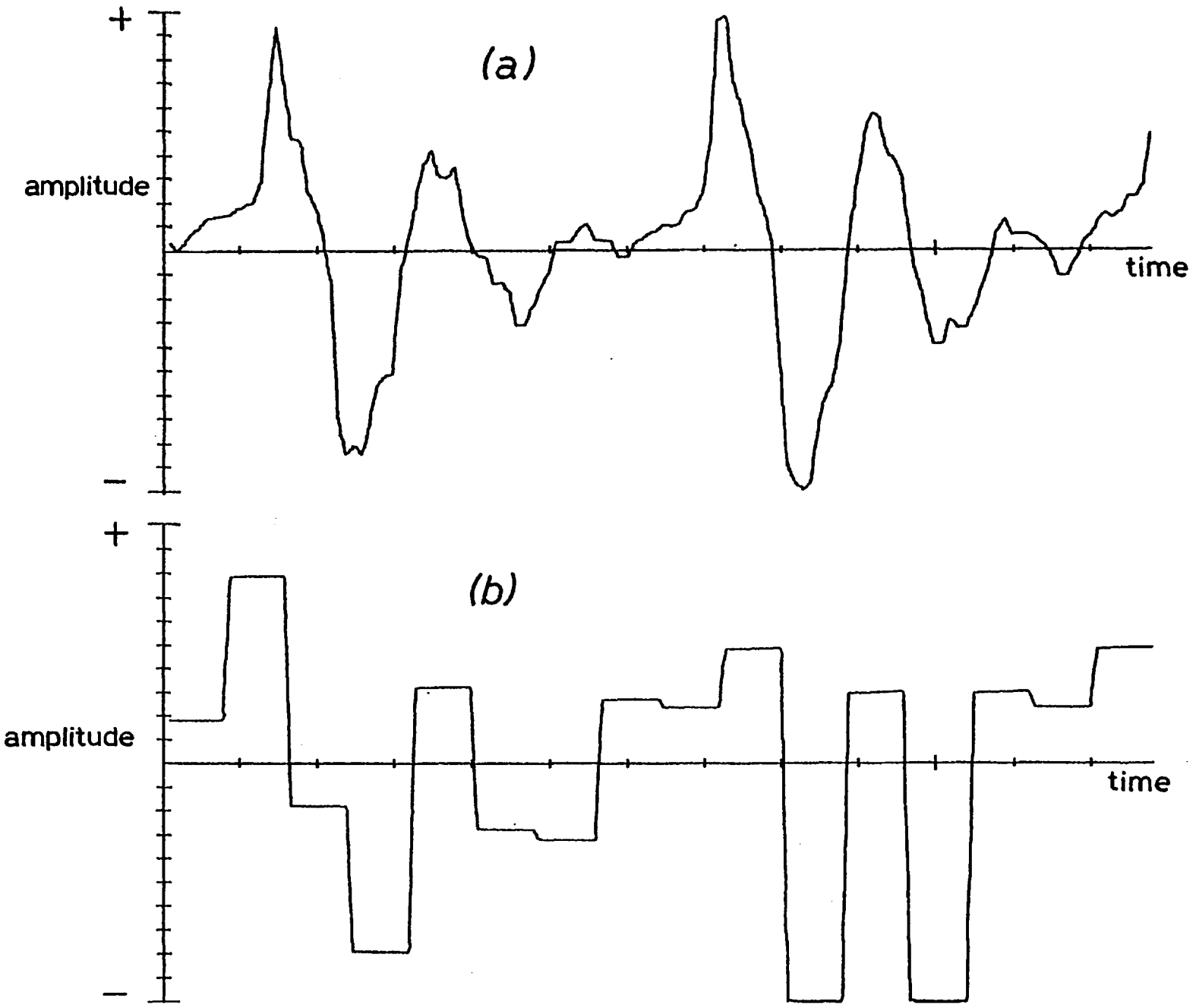


Figure 3.6. Original (a) and Hadamard processed (b) speech waveform (16 msec), using 3 dominant coefficients out of 64.

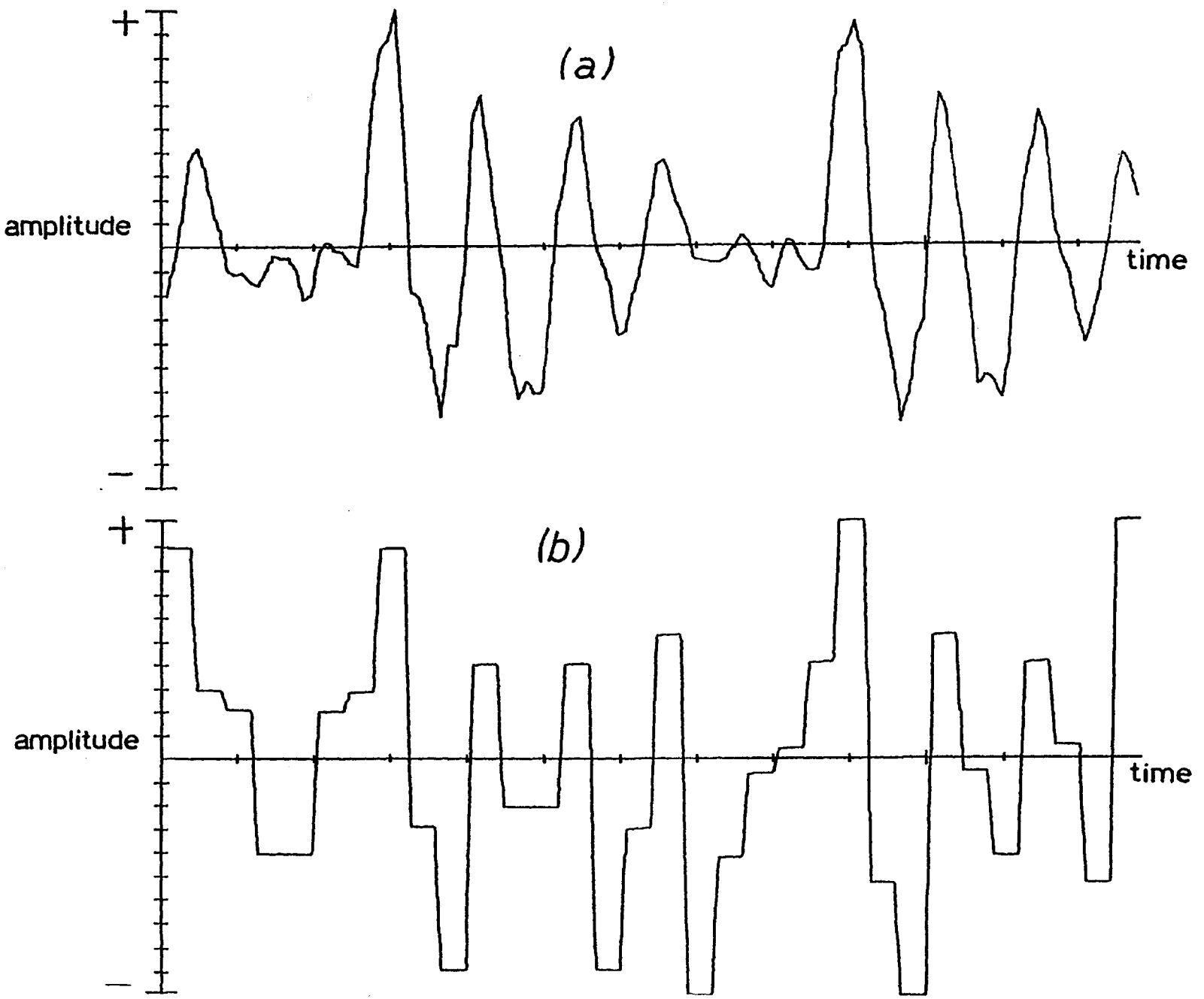


Figure 3.7. Original (a) and Hadamard processed (b) speech waveform (16 msec), using 3 dominant coefficients out of 64.

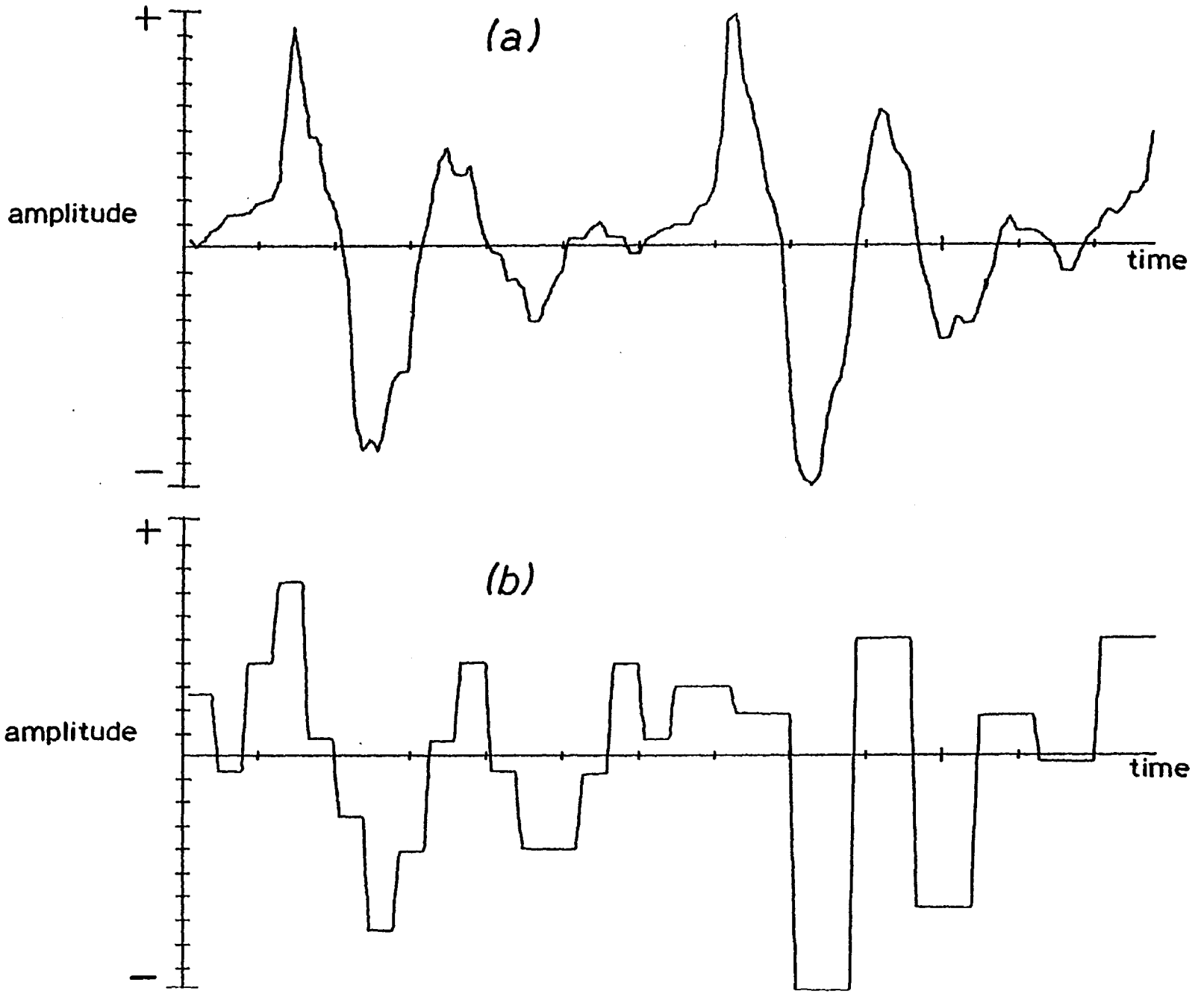


Figure 3.8. Original (a) and Hadamard processed (b) speech waveform (16 msec), using 4 dominant coefficients out of 64.

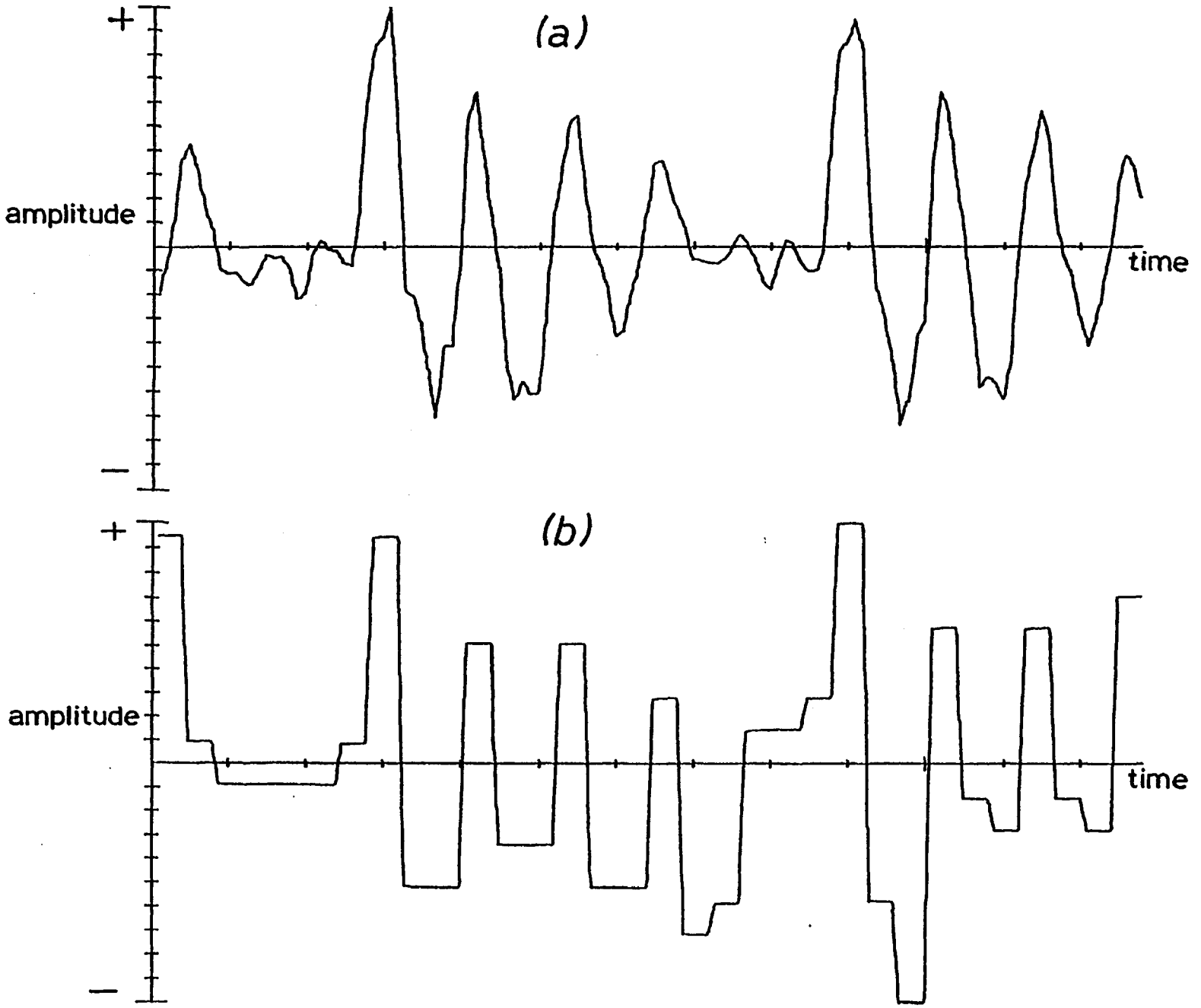


Figure 3.9. Original (a) and Hadamard processed (b) speech waveform (16 msec), using 4 dominant coefficients out of 64.

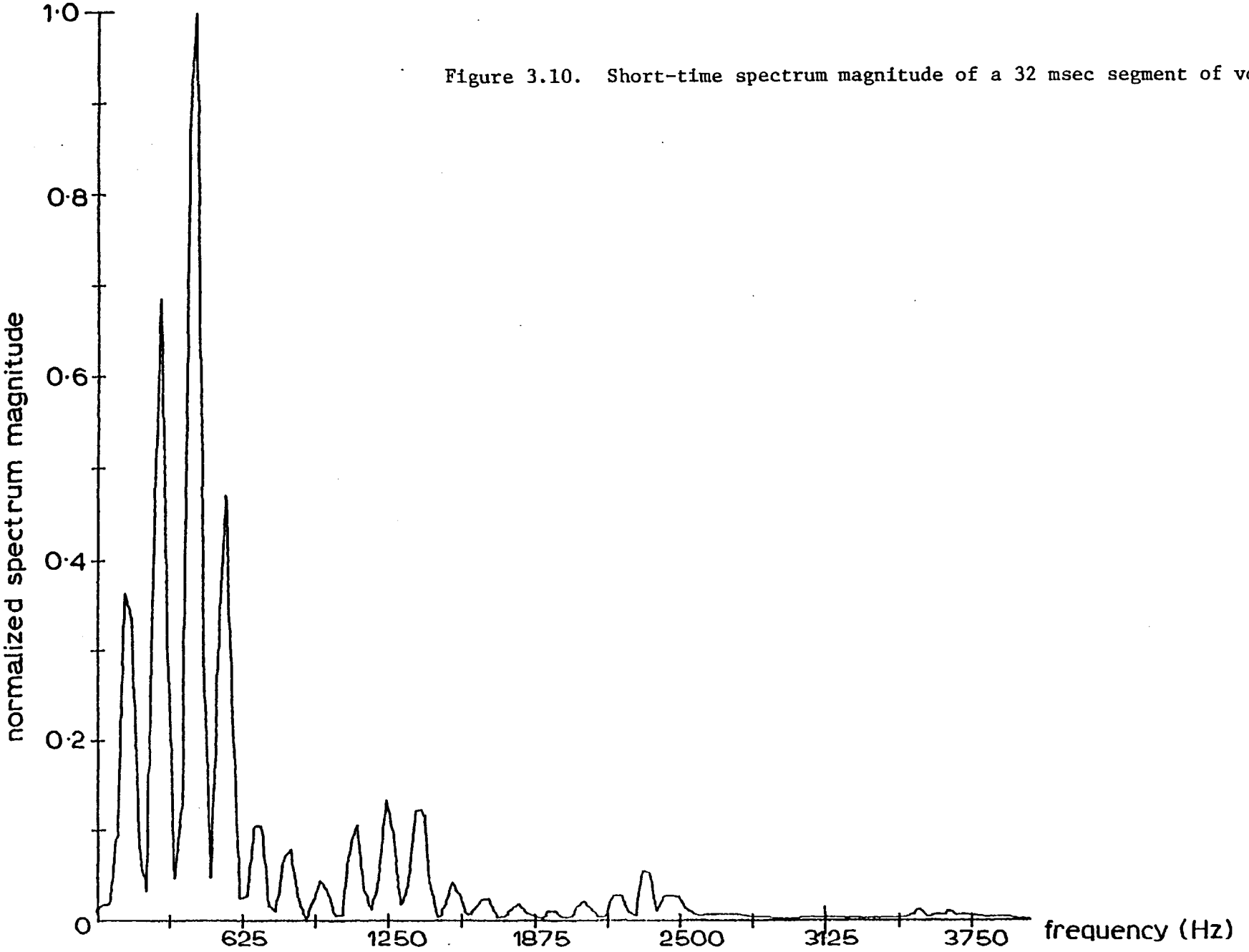


Figure 3.10. Short-time spectrum magnitude of a 32 msec segment of voiced speech.

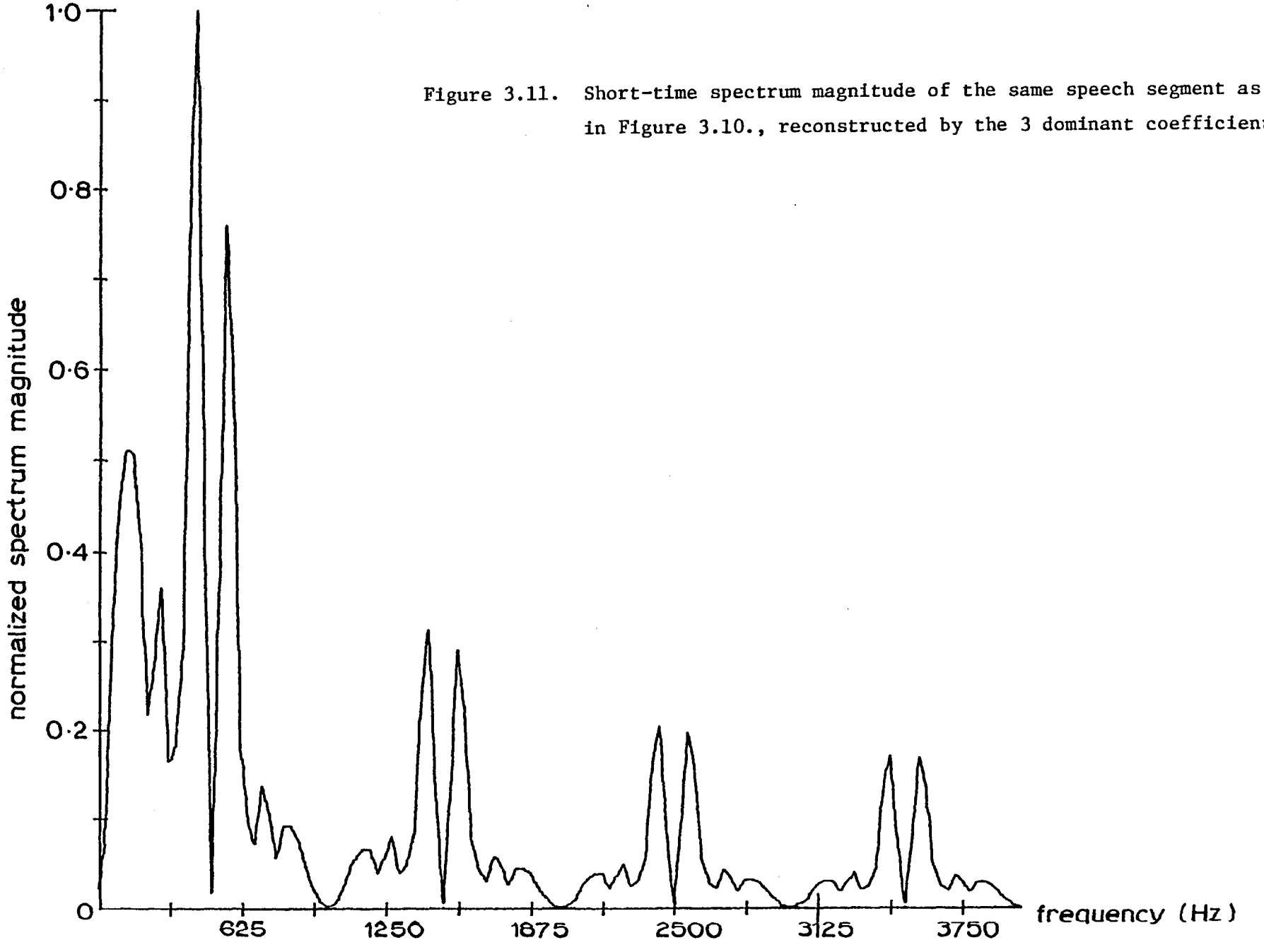


Figure 3.11. Short-time spectrum magnitude of the same speech segment as used in Figure 3.10., reconstructed by the 3 dominant coefficient system.

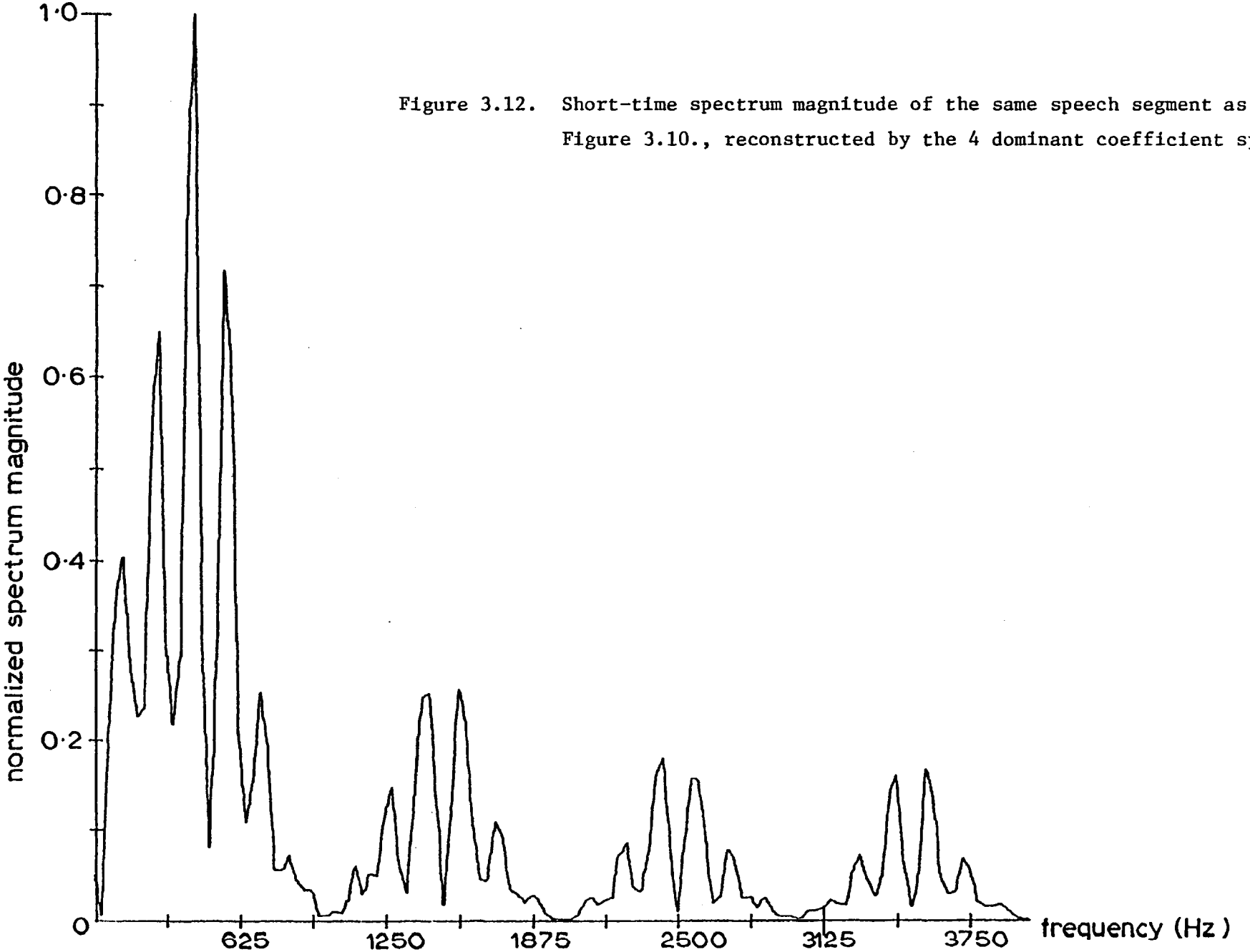


Figure 3.12. Short-time spectrum magnitude of the same speech segment as used in Figure 3.10., reconstructed by the 4 dominant coefficient system.

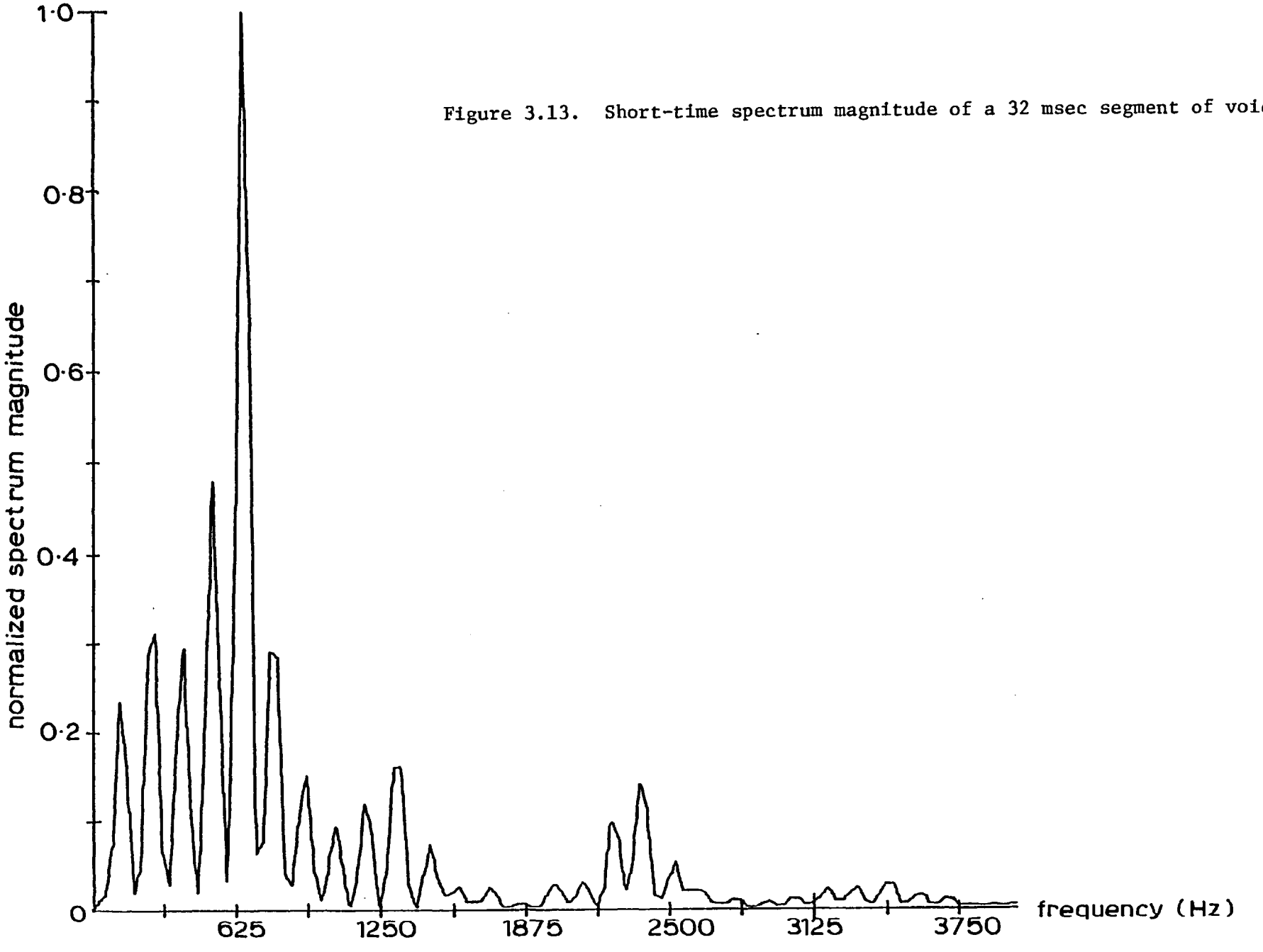


Figure 3.13. Short-time spectrum magnitude of a 32 msec segment of voiced speech.

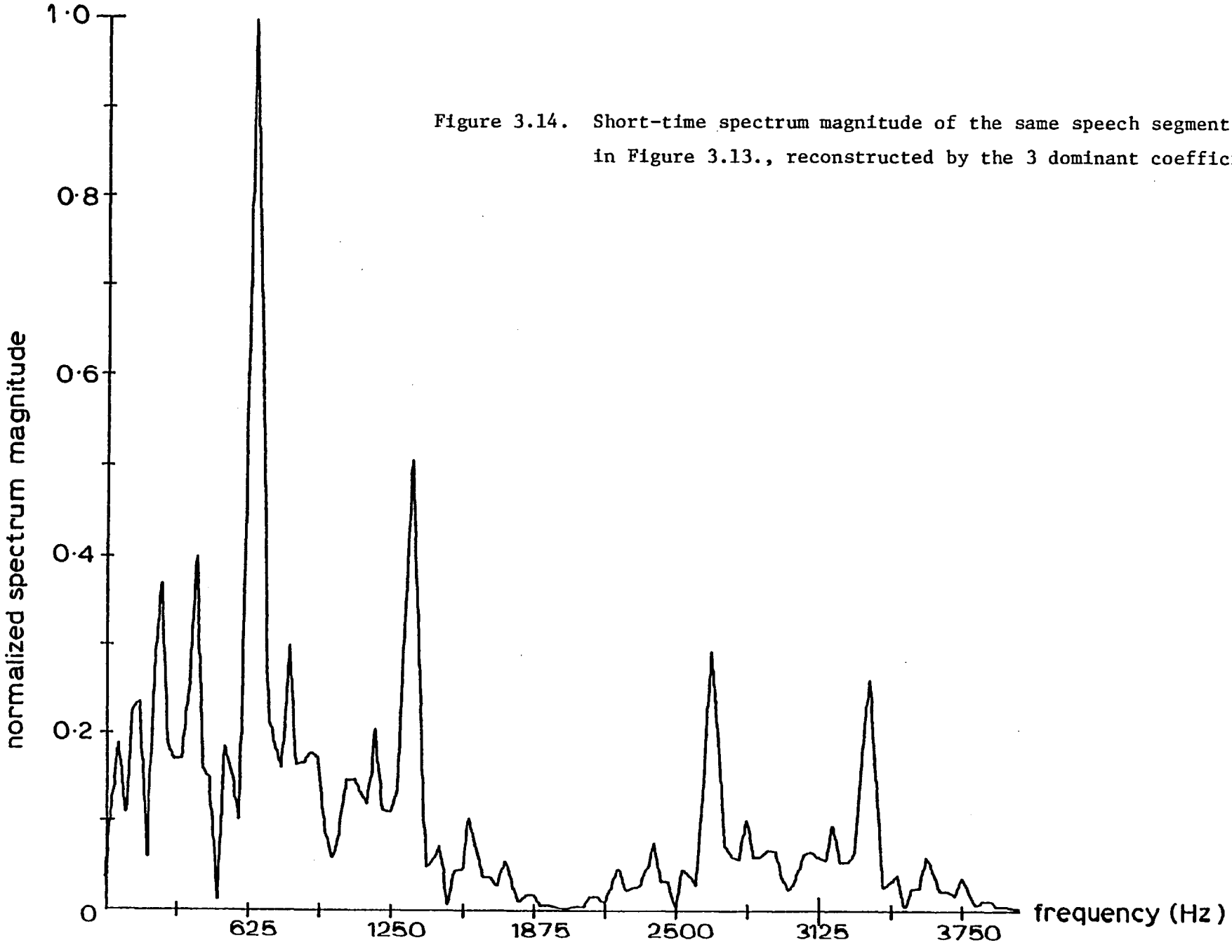


Figure 3.14. Short-time spectrum magnitude of the same speech segment as used in Figure 3.13., reconstructed by the 3 dominant coefficient system.

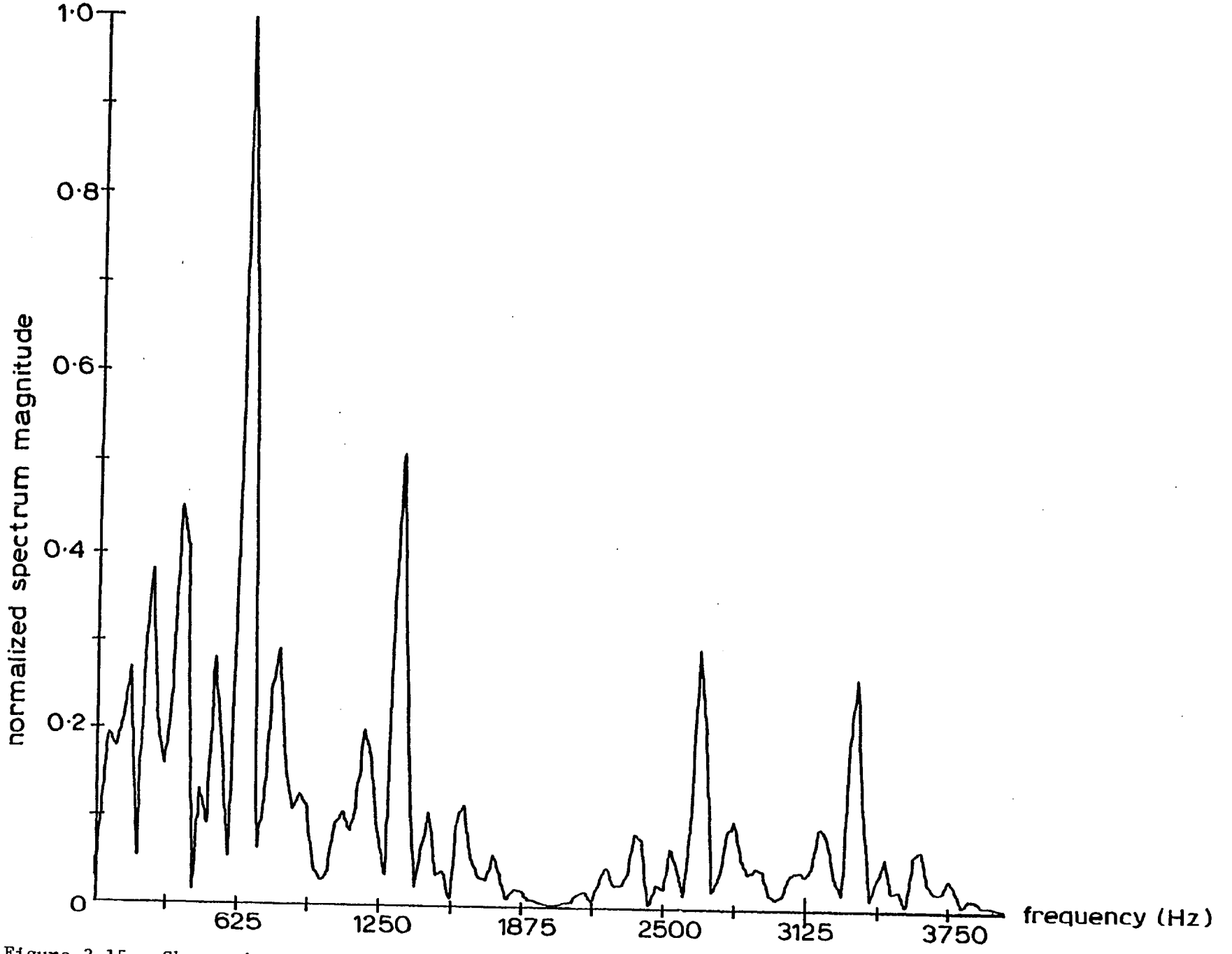


Figure 3.15. Short-time spectrum magnitude of the same speech segment as used in Figure 3.13 reconstructed by the *4 dominant coefficient system.*

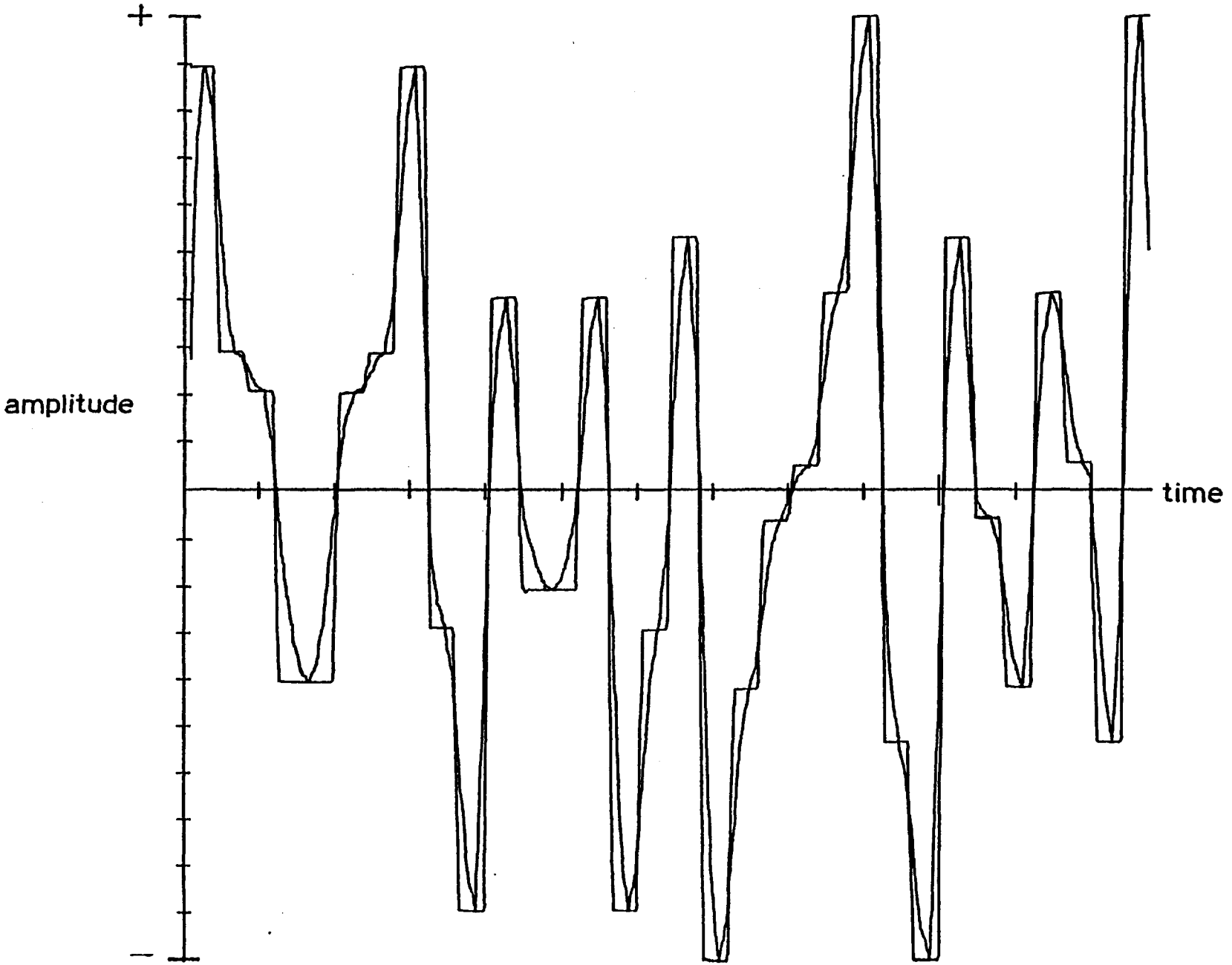


Figure 3.16. Hadamard processed waveform and smoothed equivalent of the speech segment used in Figure 3.7.

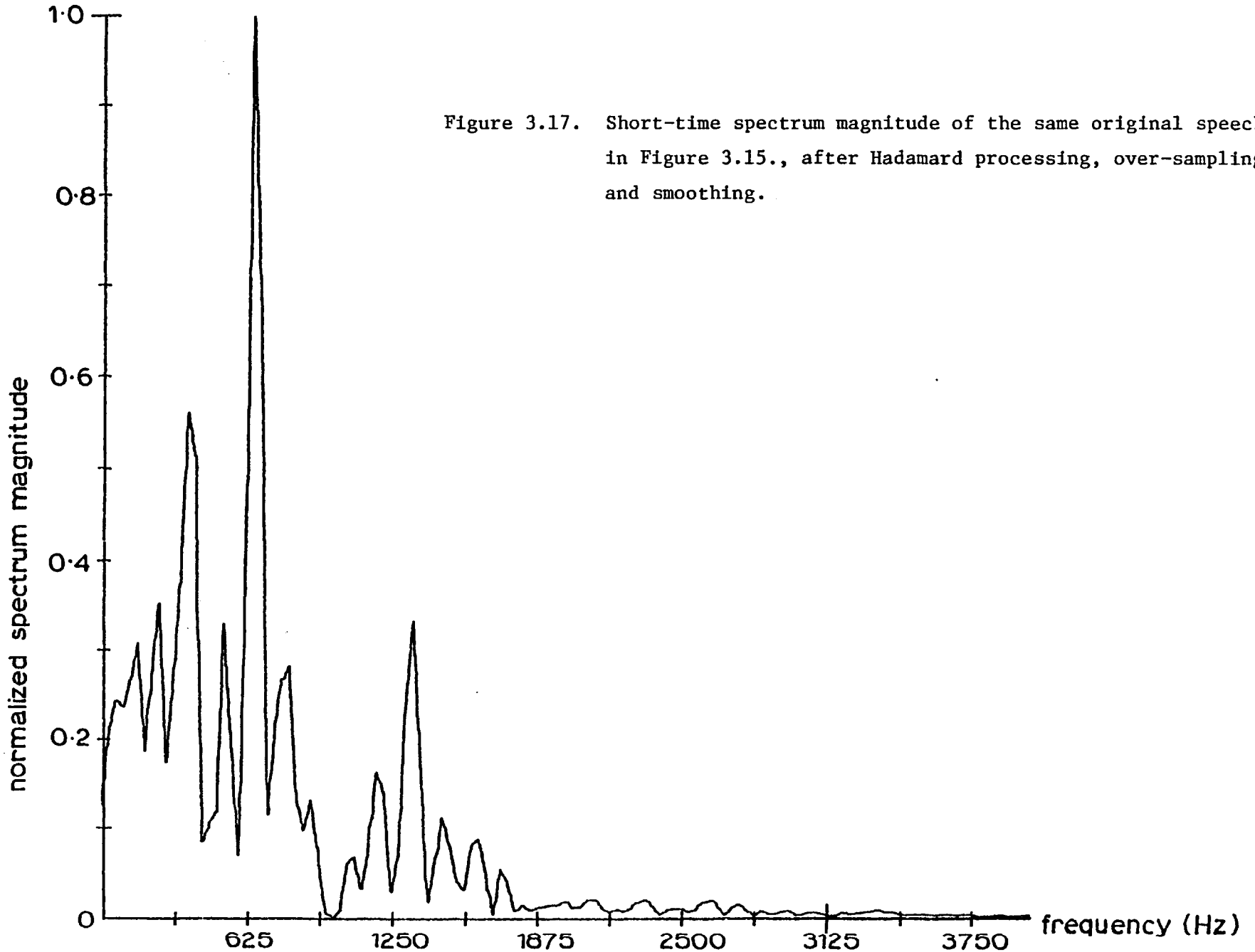


Figure 3.17. Short-time spectrum magnitude of the same original speech used in Figure 3.15., after Hadamard processing, over-sampling (16 KHz) and smoothing.

CHAPTER 4

TIME CODING OF SPEECH SIGNALS

CHAPTER 4

TIME CODING OF SPEECH SIGNALS

4.1. Introduction.

It was seen in Chapter 2 that, of the available methods of speech coding, schemes of intermediate complexity do not produce good quality speech at bit-rates below 8 Kbits/sec, even though intelligibility is usually preserved. From the point of view of implementation, waveform coding schemes are simple, but the speech quality obtained at bit-rates of about 8 Kbits/sec is poor. Also, by their very nature, these methods cannot operate at rates below 1 bit per sample. Delta-modulation, for instance, operates at 1 bit/sample, but requires a sampling rate considerably above the Nyquist rate. On the other hand, vocoder-type schemes can operate at bit-rates below 4.8 Kbits/sec, give reasonable though not commercial quality speech, but are complex systems to implement. In addition, performance is dependent on both the speaker and the environment. In the case of transform coding methods, relatively simple techniques of the Hadamard type do not appear to be capable of giving fairly good speech quality at low-bit-rates. As was seen in Chapter 3, significant improvements are not obtained even if certain special strategies are employed in the Hadamard domain. Hence, some schemes of intermediate complexity which may be able to produce good quality speech at bit-rates less than 8 Kbits/sec must be considered.

Of the schemes of intermediate complexity which have the potential for operating at bit-rates below 8 Kbits/sec, time coding appears to be promising. On account of this, a major part of the research reported on in this thesis was devoted to the study of systems of this kind, and the question of developing a time coding scheme capable of giving the required quality at rates of 8 Kbits/sec or lower. The time coding approach is promising since it can be arranged to take account of the quasi-periodic nature of voiced speech waveforms, and to exploit the characteristics of the speech waveform to a larger extent than waveform coders. Time coding techniques are potentially capable of producing more significant results at low-bit-rates.

In the development of a new time coding technique it was necessary to formulate (1) a method of classifying voiced and unvoiced sections of speech; (2) a pitch detection algorithm which gives pitch period markers directly on the pitch peaks of voiced speech waveforms; (3) a simple technique for encoding the voiced segments of the speech waveform; and (4) a simple technique for encoding the unvoiced segments of the speech waveform.

In the following section a new time-domain pitch period detector is described. Then, in section 4.3 a series of five experiments on pitch-synchronous time coding of voiced speech are presented. The new pitch period detection method forms the basis of the approach. In section 4.4 a method of clipped unvoiced speech regeneration is suggested. Then, the ideas and methods set out in sections 4.2, 4.3,

and 4.4 are employed in the development of a new scheme for the time coding of speech signals, which is described in section 4.5. The new scheme operates at an average bit-rate of 7.6 Kbits/sec. The results obtained from a computer simulation of the system are included on the attached tape recording, and a discussion of these results is given in section 4.6.

4.2. A Time-Domain Pitch Period Detector.

4.2.1. General Comments.

The determination of the quasi-periodicity existing in voiced sound waveforms is known as pitch detection. Pitch detection is an important problem in most areas of speech research. In particular, a reliable pitch period detection is an essential element of low-bit-rate speech coding systems (e.g., vocoder-type systems). Such detectors strongly influence the quality of the regenerated speech since the human ear is sensitive to small variations in this parameter. The major improvements of speech quality obtained with the voice-excited vocoder are in fact indicative of the important role played by pitch in the subjective quality of speech. In addition, the quality of vocoder-synthesized speech is considerably improved if the voiced source of the model consists of a series of pulses with separation obtained by eye detection from the original waveform.

In most cases the identification by eye (on a period-by-period basis) of the quasi-periodicity of the voiced speech waveforms constitutes an accurate and often easy method of pitch detection. However, the development of a machine to perform this task presents serious

difficulties. Firstly, the classification of speech signals into purely voiced or unvoiced regions is an oversimplification. There are, for instance, sounds, such as the voiced fricatives, whose generation involves both types of excitation. Consequently the excitation source is not well defined by a periodic sequence of pulses. Secondly, the excitation for voiced sounds is not exactly periodic but exhibits, in general, small (but sometimes quite significant) variations from period to period. This is true not only with regard to the time intervals but also refers to amplitudes and shapes. In fact, the large peaks of the speech waveform (the pitch peaks), which often appear at the beginning of each pitch period, usually have a wide dynamic range. This makes it difficult to automatically detect the peaks unless special techniques are employed. Another factor which contributes to the complexity of the problem of pitch detection is related to the effects of the formant structure on the periodicity of the waveforms. Furthermore, the lack of precise information with respect to what points should be chosen as the beginning and end of each pitch period leads in some instances to different measurements being taken from the same speech material.

In the time coding approach considered in this chapter, the pitch detection algorithm is expected to satisfy the following requirements: (1) involves only simple operations; (2) is compatible with the operations employed in time coding, which involve direct time-domain measurements of the speech waveform; (3) provides pitch period markers directly on the pitch peaks of the speech waveform; and (4) yields pitch period markers which are in good agreement with those obtained by visual inspection of the original waveform.

Despite the difficulties involved in the estimation of the periodicity of voiced speech waveforms, the need for a reliable and simple algorithm has stimulated the development of a large number of pitch detection schemes.[141-155] However, most of these schemes involve complex signal processing operations (e.g., cepstral analysis and autocorrelation analysis) and do not satisfy requirements (2) and (3) above. Pitch detection algorithms which satisfy condition (2) use time-domain techniques, and attempt to determine the signal periodicity from the waveform information only.[141-145] However, these time-domain algorithms either require a large number of tests to obtain an estimate of the pitch period, or do not give reasonably accurate estimates. For these reasons it was found more appropriate to develop a new algorithm which would involve both a reduced number of tests for obtaining an estimate of the pitch period, and also satisfy the four requirements mentioned above. The main application intended for pitch detection was in the new time coding technique described in section 4.5.

It should be mentioned at this point that the performance of a pitch detector can be considerably improved if a non-linear smoothing algorithm is used.[156-158] The effect of non-linear smoothing on a pitch period contour is illustrated in Figure 4.1. This is a typical result obtained by Rabiner et al.[156] when applying the algorithm to smooth the pitch data. However, the application of non-linear smoothing to the new pitch detection algorithm will not be considered in the present work.

4.2.2. Description of the New Pitch Detection Algorithm.

The pitch detecion algorithm to be described uses only the time-domain information relating to the voiced speech signal. The algorithm is partially based on the 'peakedness' of the waveform, but does not require threshold detection of peaks thus avoiding the difficulties arising from the wide dynamic range of speech.

The original speech was initially low-pass filtered at 3.04 KHz and then sampled at a rate equal to 8 Ksamples/sec, thus providing a time scale resolution of 125 μsec on the measurements. The algorithm was designed to span a range of pitch periods varying from 2.5 msec to 16 msec.

For a given segment of voiced speech waveform the method aims to provide two pitch markers at sample positions NP1 and NP2, which correspond to two successive pitch peaks of the waveform.

The time-domain pitch period detection algorithm (TDPD) comprises the following steps:

Step 1.

Because of the upper limit of 16 msec (128 samples) considered in this algorithm, a speech segment of 32 msec duration was used in the search for the pitch markers. The first marker NP1 was selected in this first step as the location of the maximum absolute value sample amplitude in the initial 16 msec portion of the segment. The reason for taking into consideration both positive and negative amplitudes concerns the fact that the pitch peaks are, in some instances, in greater evidence in either one or the other polarity. This, in fact,

is one of the main causes of errors in the algorithm developed by Gold[142] which considers only the positive polarity. This property of the speech signal is illustrated in Figure 4.2 for two segments of 32 msec-long speech waveforms.

The sign of the sample in NP1 gives the polarity for pitch period search in the segment being analysed. In this initial step the aim was to select a few samples of the segment which are most likely to contain the second pitch marker (NP2). The purpose of this selection was to reduce the problems created by the formant structure on the measurements related to the waveform periodicity. One way of reducing the effects of the formant structure, considered by Sondhi,[148] consists of applying a centre clipping process. In the case of the TDPD algorithm, however, a number (N) of samples were chosen from the range R = [NP1 + 20, NP1 + 128] such that: a) the interval between two successive zero-crossings was allowed to contain, at most, one of these samples; and b) the samples selected were of the largest absolute amplitudes in the range R. It should be noted that only a single polarity (as given by the sample at NP1) is considered.

Since the minimum and maximum values of the pitch period spanned by this algorithm (TDPD) are 2.5 msec and 16 msec, respectively, it is expected that the range R contains between one and six pitch periods. Thus, for the case of high-pitch speakers, N should not be set to a value less than 6, since in this situation the second pitch peak could be missed. For this reason N was set to 6. However, it should be noted that this is the maximum value of N because it it possible

that, for low-pitch speakers and sounds with low first formant, for instance, the number of samples N found in the interval R (with the properties a) and b) previously mentioned) is less than 6.

The notation used in the TDPD algorithm was {s(1), s(2),..., s(256)} for the segment (s) of 32 msec length, {L(1), L(2), ..., L(N)} for the positions of the samples selected in this initial phase of the algorithm, and NP1 and NP2 for the location of the two pitch period markers. Figure 4.3 illustrates the selection of NP1 and the N (in this case 6) peaks. Figure 4.4 shows the output of step 1 for the speech segment shown in Figure 4.3. The set of positions {L(j); j = 1, ..., N} was ordered such that L(m) > L(n) if m > n.

Step 2.

This step involves a simple first estimate of NP2 (the second pitch period marker), and is based on the regularity of voiced speech waveforms. The following test was applied to the sample locations obtained from step 1.

Consider the set {L(j); j = 1, ..., N} starting with j = 1, and do NP2 = L(j). If

$$|(NP2 - NP1) - [L(j + 1) - NP2]| < \alpha_1 |(NP2 - NP1)|, \qquad (4.1)$$

then perform step 3. In this expression, the most desirable value of α was found, experimentally, to be $\alpha = 0.17$. If condition (4.1) was not satisfied, then the largest absolute value amplitude between s(NP2) and s[L(j + 1)] should be chosen, i.e.,

$$\max_{n} \{|s(n)|; n = NP2, L(j + 1)\}.$$

Then, NP2 is set equal to the location of this maximum, j is set to j + 1, and the regularity test (4.1) is repeated. It should be noted that the selection of the maximum was included in order to eliminate the peaks which are unlikely to be good candidates for the second pitch peak. This selection was based on the fact that the pitch period is generally characterised by larger absolute value amplitudes in its initial than in its final portion.

In cases when the last peak [position: L(N)] was used without (4.1) being satisfied, perform step 5.

Step 3.

The following test aims to avoid the situation which arises when a strong peak occurs approximately half way between two adjacent pitch peaks, as illustrated in Figure 4.5. The test is required in order to avoid erroneously halving the estimate of the pitch period.

The test consisted of determining the minimum absolute value amplitude ($\hat{s}$) between s(NP1) and s[L(j + 1)] (as given by step 2), i.e.,

$$\hat{s} = \min_{n} \{|s(n)|;\ n = NP1,\ L(j+1)\} \tag{4.2}$$

and, if

$$|s(NP2)| < \beta \cdot \hat{s} \tag{4.3}$$

then set NP2 to L(j + 1) and j to j + 1. If j = N, perform step 5; otherwise perform step 2 [repeat test (4.1)]. In (4.3), $\beta = 7/8$ was found to be a good choice for a wide range of pitch periods and speech material. If condition (4.3) was not satisfied, then perform step 4.

Step 4.

The purpose of the test applied in this step was to prevent errors likely to occur when the formant structure was not sufficiently removed. This is usually the case for low-pitch speakers and high first formant sounds, which exhibit a high number of zero-crossings. Figure 4.6 shows a typical segment of speech which illustrates this situation.

The test first takes the minimum absolute value between s(NP1) and s(NP2) (as given by step 3), i.e.,

$$\tilde{s} = \min_{n} \{|s(n)|;\ n = NP1,\ NP2\} \tag{4.4}$$

and, if

$$|s[L(j+1)]| < \beta . \tilde{s} \tag{4.5}$$

where again, $\beta = 7/8$ was found, experimentally, to be a good choice for this parameter, then there are two possibilities. When $j + 1 = N$, perform step 5. When $j + 1 \neq N$, j is set to $j + 2$ and NP2 is set to L(j) - if $j = N$, perform step 5, otherwise perform step 2. If (4.5) does not hold, then NP2 is the detected second marker of the pitch period (NP2 - NP1).

Step 5.

This step is used only when the last peak [location: L(N)] has been used and the second pitch marker requires further testing. This may occur for one of the two following reasons; (1) two pitch periods, starting from position NP1, cannot be fitted into the analysis segment;

or (2) the polarity of the pitch peaks was wrongly detected. As shown in Figure 4.7, the polarity may be wrongly detected in spite of the high amplitude of the sample located at the first pitch marker (NP1). However, this situation was found to occur in only a few cases.

In this step the algorithm was repeated from step 1 with samples of opposite polarity. The values of NP1 and NP2 resulting from the first search were stored in KP1 and KP2, respectively, and the markers obtained from the second search involved new values, NP1 and NP2.

The greatest sample amplitude in absolute value located in the first half of the analysis segment (32 msec) was at the position KP1 (NP1 of the first search). Thus, the pitch markers obtained in the first search are more likely to constitute a better estimate of the pitch period. However, cases such as that illustrated in Figure 4.7 may occur. In these cases, on account of the high amplitude characteristics of the pitch peaks, the amplitudes obtained in the second search [s(NP1) and s(NP2)] are not expected to be significantly smaller than those resulting from the first search [s(KP1) and s(KP2)]. The following test was found to be efficient in dealing with such situations. If

$$|s(NP1) + s(NP2)| \geq \gamma \cdot |s(KP1) + s(KP2)| \qquad (4.6)$$

then the values of NP1 and NP2 of the second search are the detected pitch markers. Otherwise, the values of the first search are the estimated ones. In (4.6), the most desirable value of γ was found, experimentally, to be $\gamma = 0.8$.

Results of performance of the TDPD algorithm will be given in sub-section 4.2.4.

4.2.3. Voiced-Unvoiced Classification.

The voiced-unvoiced classification method to be used in connection with the TDPD algorithm is based on a simplified version of a pattern recognition technique proposed by Atal and Rabiner[159] for classifying speech segments into 'voiced', 'unvoiced' or 'silence'.

The method suggested by Atal and Rabiner[159] is based on a classical minimum probability of error decision rule applied to five measurements taken from the speech segment to be classified. The measurements relate to the signal energy, zero-crossings, the auto-correlation coefficient, the first predictor coefficient, and the energy of the prediction error. The probability density function associated with the measurements is assumed, for each class i ($i = 1, 2, 3$), to be a multi-variable Gaussian distribution with mean $\underline{m}_i$ and covariance matrix $\underline{R}_i$, where $i = 1$ for the class 'voiced', $i = 2$ for the class 'unvoiced', and $i = 3$ for the class 'silence'. If $\underline{x}$ denotes a column vector representing the measurements, then the decision rule consists of choosing class i if

$$\hat{d}_i = (\underline{x} - \underline{m}_i)^t \, \underline{R}_i^{-1} (\underline{x} - \underline{m}_i) \tag{4.7}$$

is minimum, i.e., if $\hat{d}_i < \hat{d}_j$ for all $j \neq i$.

The algorithm of Atal and Rabiner was tested[159] using only one of the five measurements at a time. It was found that no isolated measurement was capable of classifying the speech sound accurately. However, it was observed that certain measurements worked well for

classifying the speech segments into two classes only. In particular, the zero-crossings measurement is capable of accurately distinguishing between voiced and unvoiced segments, but is inefficient in discriminating between voiced sounds and silence.

In the classification method to be used in connection with the TDPD algorithm it was decided to use the zero-crossings measurement (N_Z) associated with a simple silence detector. The silence detector used in the present case differs slightly from that used in connection with the modified autocorrelation (MAUT) algorithm.[149, 157] In both cases the silence detector is based on the peak signal level. In the present case, the test for silence detection was initially applied to a set of three frames, each of 10 msec duration. If this set is detected as a silence segment, then the test is applied to the successive frames one at a time, rather than to a set of three. This helps to prevent long segments of either background noise or unvoiced speech being classified as voiced speech. The application of the test to three frames initially is a convenient method of avoiding detection of low-amplitude sections of voiced speech as non-voiced. A threshold value of 1/30 of the peak absolute signal amplitude, which is approximately the same for most tested utterances, was found to be adequate. Clearly, this level may vary in accordance with the background noise, and the threshold value should be chosen so as to cover the range of situations in which the system will be used in practice. An alternative could involve the use of some form of automatic adaptation based on the signal level.

The low-pass filtered speech at the system input was first high-pass filtered at 200 Hz so as to remove d.c. or low-frequency noise which could affect the measurements. The training data for distinguishing between the classes 'voiced' (i = 1) and 'unvoiced' (i = 2) consist of the mean m_i, i = 1, 2, and the standard deviation σ_i, i = 1, 2, of zero-crossings per frame. As observed by Atal and Rabiner,[159] the training data obtained from one speaker can be used for different speakers. This was also verified experimentally in the present study.

Frames of 10 msec duration were considered. Using T as the silence threshold, $m_1 = 11.224$, $\sigma_1 = 4.838$, $m_2 = 48.614$, and $\sigma_2 = 14.911$, the method of classifying voiced and non-voiced (silence, unvoiced) frames can be summarised as follows:

Step 1.

A segment consisting of three successive frames (j, j + 1, j + 2) is initially considered. Then, the sample of maximum absolute amplitude M3, is selected from this segment. If M3 < T then the three frames are classified as non-voiced (silence), j is set to j + 3, and perform step 3. Otherwise, perform step 2.

Step 2.

The number of zero-crossings, N_Z, in frame j is computed. If

$$(N_Z - m_1)^2/\sigma_1^2 < (N_Z - m_2)^2/\sigma_2^2 \qquad (4.8)$$

then the frame j is classified as voiced; otherwise as unvoiced. Then j is set to j + 1, i.e., the next frame is taken, and perform step 1.

Step 3.

The sample of maximum absolute amplitude, M1, is selected from frame j. If M1 < T, then frame j is classified as non-voiced (silence) j is set to j + 1, and this test is repeated. Otherwise, perform step 2.

4.2.4. Experimental Results.

An initial evaluation of the performance of the TDPD algorithm (with respect to pitch estimation) was obtained from the analysis of 160 segments of voiced speech. The duration of each of these segments was 32 msec, and both male and female speech was used. The TDPD algorithm was applied to these segments, and the voiced-unvoiced classification was not used in this preliminary test. It was found that, apart from 7 errors due mainly to transitional segments, the results obtained were as good as those obtained by human visual detection. Typical results obtained are shown in Figures 4.8 to 4.15 for different voiced speech segments of 32 msec duration. In these figures, two vertical lines superimposed on the waveform indicate the pitch markers (positions NP1 and NP2) as given by the TDPD algorithm.

A more detailed evaluation of the TDPD algorithm, when used in connection with the voiced-unvoiced classification method described in sub-section 4.2.3., was carried out. The speech signal used in this investigation was segmented into frames of 10 msec duration. Different utterances, spoken by male and female speakers, were analysed by the TDPD algorithm and the results compared with those obtained by the modified autocorrelation (MAUT) technique.[149]

The MAUT algorithm uses 900 Hz low-pass filtered speech. For each 30 msec segment a clipping level is fixed at 64% of the minimum peak absolute value sample between the initial and final 10 msec sections of the segment. This segment is then centre-clipped and infinitely-clipped, taking the values -1, 0, and + 1. The autocorrelation function is computed and normalised to its value at 0-delay. Its maximum is then determined, and if this exceeds 0.3 the section is considered to be voiced and the position of the maximum is regarded as the estimated pitch period. Otherwise, the section is considered to be unvoiced. A silence detector similar to the one previously described is also used.

The error parameters used in the comparison were:

1) Percentage of gross pitch errors, i.e., errors larger than, or equal to , 10 samples (the error being relative to detection by visual inspection of the waveform).
2) Mean ($\bar{e}$) and standard deviation (σ_e) of the difference of pitch estimates given by the MAUT (P_{MAUT}) and TDPD (P_{TDPD}) algorithms when neither involves gross errors. This gives a measurement of the degree of agreement between the two algorithms when no gross errors occur.

3) Percentage of voiced frames classified incorrectly as non-voiced (unvoiced, silence).
4) Percentage of non-voiced frames classified incorrectly as voiced.

Sentences of approximately 2.5 sec duration were used in the comparison. The speech signal was first low-pass filtered to 3.04 KHz, sampled at a rate of 8 Ksamples/sec, and then A/D converted using a 10 bit-convertor. Pitch period detection was performed every 10 msec, and the analysis segment was 30 msec long for the MAUT algorithm and 32 msec long for the TDPD algorithm. A total of 533 voiced frames and 330 non-voiced frames were analysed by the two techniques. The error parameters obtained are shown in Table 4.1.

From Table 4.1 it can be seen that, with respect to gross pitch errors, the performance of the TDPD algorithm is better than the MAUT algorithm for male speech. The performance (with respect to gross pitch errors) of the MAUT algorithm is, however, better for female speech. When no gross errors occur, the pitch period estimates provided by the schemes are in close agreement for all the utterances analysed. TDPD, however, gives superior performance in terms of the misclassification of voiced frames as non-voiced frames, and vice-versa.

Figures 4.16 and 4.17 show, respectively, the pitch period contours for utterance number 1 (" the box was thrown beside the parked truck") and utterance number 5 (" rice is often served in round bowls") analysed by the TDPD and the MAUT pitch detection algorithms.

Table 4.1. Error Parameters for MAUT (M) and TDPD (T) Pitch Period Detection Algorithms.

Error Parameter	Utterance Number			
	1 (male)	2 (male)	3 (female)	4 (female)
% Gross Errors	15.2(M) 14.4(T)	19.8(M) 14.1(T)	5.1(M) 18.1(T)	5.6(M) 17.0(T)
$\bar{e}$	-0.12	-0.26	0.04	-0.20
σ_e	0.98	1.77	1.33	1.13
% Voiced Errors	11.5(M) 2.1(T)	5.1(M) 2.5(T)	0.8(M) 0.0(T)	0.8(M) 0.0(T)
% Non-voiced Errors	11.6(M) 5.3(T)	15.6(M) 9.3(T)	20.0(M) 7.1(T)	21.1(M) 11.5(T)

These figures illustrate the typical results obtained with each method. It should be noted that an improved performance could be expected for both methods if a non-linear smoothing technique, such as the one suggested by Rabiner et al.[156] (see illustration in Figure 4.1), were applied to the pitch contours.

An interval of 10.34 seconds of speech was analysed using a PDP-15 computer by both the TDPD and the MAUT algorithms. The average speed of execution was τ_T = 139 seconds per second of processed speech for the TDPD algorithm and τ_M = 181 seconds per second of processed speech for the MAUT algorithm. Thus, for the relatively long segment of speech analysed in this investigation the TDPD algorithm runs faster than the MAUT algorithm ($\tau_M \cong 1.3\ \tau_T$). This result suggests that the TDPD algorithm requires in general a shorter execution time than the MAUT algorithm. This can be seen more clearly by observing that both algorithms require, basically, the same kind of operations (comparisons and additions).

From the experimental results obtained in this study it can be concluded that the TDPD algorithm meets most requirements in that:

1) the algorithm involves, in most cases, only three simple tests to obtain the pitch estimate;
2) the operations involved are simple in nature (additions and comparisons);
3) the algorithm is compatible with the operations employed in time coding (e.g., zero-crossing measurements, peak selections);

4) the algorithm provides pitch period markers directly on the speech waveform; and
5) the pitch estimates are generally in good agreement with those obtained by visual inspection of the waveform, and the accuracy compares favourably with the MAUT algorithm.

As mentioned earlier, the available pitch detection algorithms which use time-domain techniques, such as those reported in references 141 to 145, do not satisfy all the conditions listed above. The time-domain algorithms involve, in general, simple operations. However, a large number of tests is usally required to obtain an accurate estimate of the pitch period.[143-145] Thus, the indications are that the complexity of the TDPD algorithm, which basically involves only comparisons, additions, and three simple tests to identify the pitch markers, is significantly lower than for the other methods.

For all these reasons, the TDPD algorithm was considered an attractive pitch detection algorithm for use in the time coding technique to be described in section 4.5.

4.3. Experiments on Pitch-Synchronous Time Coding of Voiced Speech.

This chapter is concerned with the development of a new time coding scheme which treats voiced and unvoiced portions of the speech waveform independently, and which exploits the quasi-periodic nature of voiced speech. It is, therefore, necessary to develop simple techniques to encode the voiced and unvoiced portions of the speech

waveform. In this section the encoding of voiced segments of speech is considered and a method for encoding the unvoiced segments of speech is presented in section 4.4.

With respect to the encoding of voiced speech, a number of characteristics of the waveform (such as the periodicity, the zero-crossings, the amplitude of extremes, etc) are manipulated with the aim of reducing the amount of information to be transmitted. Several experiments were performed on voiced speech waveforms in order to determine which features are most important in terms of subjectively acceptable speech.

In all the experiments described in this section one speech segment of 32 msec duration was initially used for pitch period determination. The resulting period is given by the TDPD algorithm described in the pervious section, with two markers being positioned at NP1 and NP2, corresponding to two adjacent pitch peaks in the 32 msec-long segment. Unvoiced sounds were not processed, and voiced-unvoiced classification was obtained by visual inspection of the original speech waveforms.

The original speech signal used in the experiments was low-pass filtered at 3.04 KHz, sampled at a rate of 8 Ksamples/sec, and 10-bit A/D converted. Sentences spoken by both male and female speakers were used, and typical results of the experiments are included on the tape recording attached to this thesis.

4.3.1. Experiment I.

The main aim of this initial experiment was to determine the extent to which the quasi-periodicity of the voiced speech waveform can be exploited, in a simple way, if normal pitch variable speech is used. The method consisted of retaining only one pitch period of a given segment of speech waveform, and then repeating the retained period for the entire duration of each segment.

In Experiment I, the duration of each segment was set equal to the time interval used in the determination of the pitch period. This time interval was 32 msec. A simple fixed segmentation of the waveform was considered in order to avoid the use of a complex pitch dependent algorithm to obtain a variable segmentation. In the experiment, the waveform intrapitch (i.e., the waveform of the retained pitch period) was normalised with respect to the pitch peak amplitude before the pitch period was recirculated within the segment. The largest positive amplitude and the largest negative amplitude were selected from each half interval (16 msec) of every segment. After the pitch period was recirculated within the segment, these four amplitudes were then used to restore some amplitude variation to the regenerated speech segment. It should be noted that this experiment is fundamentally different from that reported by David and McDonald[13] in that natural speech, rather than vocoder-generated monotone speech, is used.

Figures 4.18 and 4.19 show 128 msec segments of original speech and the reconstructed waveform obtained by Experiment I. Figure 4.20 shows the original and reconstructed waveforms for a shorter segment

(32 msec) . As can be seen, the waveforms of reconstructed speech closely approximate the original waveforms. However, since speech is not monotone, it should be noted that there are pitch and waveform discontinuities at the junctions between segments. Also, on account of the simple fixed segmentation used, it can be seen that some voiced portions were regenerated as unvoiced, and vice-versa.

Original and processed utterances are included in section 8 of the attached recorded material and illustrate the effects obtained in this experiment. It is clear that distortions do occur due to waveform discontinuities at the junctions between segments and also due to the use of fixed segmentation. However, the effects of these distortions are not critical and the regenerated speech is highly intelligible and suffers only a limited degradation in quality. The amount of information in the voiced portions of the waveform was reduced by a factor of approximately P/T, where P is the pitch period and T = 32 msec is the the duration of each segment. The factor P/T can vary over an approximate range of 1/12 to 1/2.

4.3.2. Experiment II.

In Experiment I the quasi-periodic nature of the voiced portions of the speech waveform was exploited, while no attempt was made to reduce the amount of information of the waveform intrapitch. If, for instance, this waveform is to be coded by means of a PCM technique in which every sample of the waveform intrapitch is transmitted, little is gained in terms of information rate reduction. In Experiments II,

III, and IV, the basic approach of Experiment I was employed and the information contained in the waveform intrapitch was manipulated in various ways. The aim was to examine whether it is possible, by means of simple techniques, to obtain a reduced representation of the waveform intrapitch and a subjectively good regeneration of voiced speech.

Experiment II is concerned with the effect on the quality of the regenerated speech if only the following features are extracted and used in regenerating the waveform intrapitch: (1) the average distance between successive zero-crossings as measured over the 32 msec segment; and (2) the position of the largest amplitude sample of the waveform intrapitch, selected from the samples with opposite polarity to that of the pitch peaks. The use of feature (1) was based on the fact that the zero-crossings of the waveforms of vowel sounds are observed to occur regularly over time.[160] Feature (1) was used in order to restore the zero-crossings pattern of the waveform intrapitch. The purpose of using feature (2) was to preserve, for the polarity opposite to that of the pitch peaks, the periodicity of the waveform of the 32 msec segment.

The regeneration of the waveform intrapitch is illustrated in Figure 4.21. The pitch peaks are positioned at NP1 and NP2, and the peak of opposite polarity to that of the pitch peaks is positioned at NP3, where NP3 is given by feature (2) above. Zero-crossings are separated by an interval Δ, where Δ is the average distance between successive zero-crossings as measured over the 32 msec segment. As

shown in Figure 4.21, exponential damping and smooth interpolation by a function of the type

$$F(X) = \sin^2(\pi X/2) \qquad (4.9)$$

were used in the intrapitch reconstruction. After the waveform intrapitch was reconstructed, the approach used in Experiment I was then employed.

Section 9 of the attached recorded material illustrates the resulting processed speech. The speech quality was noticeably degraded as was its intelligibility. The inaccuracy of the zero-crossing positions and the artificial regularity introduced on the reconstructed waveform seem to be important factors contributing to this degradation. The degradation is reflected in artificial and 'metallic' sounding speech. It can be concluded, therefore, that the use of features (1) and (2) is not sufficient to produce reasonable quality speech.

4.3.3. Experiment III.

The above results indicate that additional information is required from the waveform intrapitch in order to obtain a good regeneration of voiced speech. Experiment III was conducted with this purpose in mind.

In Experiment III, the zero-crossing locations in a short time interval (I) around the pitch peaks (I/2 for each one) were extracted. Additionally, the positions of the zero derivative of the waveform intrapitch (i.e., peaks and valleys) were also extracted in the interval I.

This information - location of zero-crossings, peaks and valleys - was used to obtain a more accurate description of the waveform intrapitch in the short time interval I in the region of the pitch peaks. The short time interval I around the pitch peaks was chosen since this interval has been found to be most important with respect to the correct perception of sounds.[16] For the entire duration of the waveform intrapitch, excluding the interval I, the separation between zero-crossings was approximated by a fixed distance D. This procedure was similar to that used in Experiment II. In the present case the distance D was taken either as the maximum interval between zero-crossings (D = max) found in the waveform intrapitch, or as the average distance between zero-crossings (D = ave) measured over the duration of the pitch period excluding the interval I. The waveform intrapitch was reconstructed in a manner similar to that described for Experiment II.

Two values of I were considered: I = 2.5 msec and I = 3 msec. Figure 4.22 shows, for the speech segment used in Figure 4.20, the original and reconstructed waveform for the case of D = ave and I = 2.5 msec. Simulations were run for various combinations of D and I. The quality of the regenerated speech was found to be poor for all cases, with the same artificial sounds observed in Experiment II. Section 10 of the attached tape recording illustrates the quality of the regenerated speech obtained for three different combinations of D and I.

The results obtained indicate that the information extracted from the reduced interval around the pitch peaks is still not sufficient to give a clearly intelligible reproduction of voiced speech.

4.3.4. Experiment IV.

Experiment IV was conducted on the basis of the following two observations: (1) the previous experiment indicated that the information extracted from the waveform intrapitch, for a periodic recirculation in segments of 32 msec duration, was still not sufficient to provide good regeneration of voiced speech; and (2) a considerable amount of information is carried by points of zero derivative of the waveform.[137]

In the first part of Experiment IV extremal coding[14] was applied to the waveform intrapitch. This means that the location and amplitude of every extreme of the waveform intrapitch was extracted and used to reconstruct this waveform. The regeneration of the waveform intrapitch was obtained by means of a smooth interpolation between the extremes, using an interpolating function of the type given by (4.9).

In the second part of the experiment the position and amplitude of every local maximum (in absolute value) was used to regenerate the waveform intrapitch. The local minima, or valleys, were not extracted from the waveform but were artificially created in the reconstruction of the waveform intrapitch, with amplitude equal to a fixed proportion (80%) of the smallest absolute amplitude between the two adjacent maxima. A smooth interpolation with a function of the type given by (4.9) was again used.

Finally, in the third part of the experiment, a further possibility was considered in which only the position and amplitude of the absolute extreme between each pair of adjacent zero-crossings was used in the reconstruction process. Again, as in the first and second parts of the experiment, a smooth interpolation was utilised to reconstruct the waveform intrapitch.

The features extracted from the waveform intrapitch in these three methods are indicated by crosses on Figure 4.23. Original and reconstructed waveforms of the speech segment used in Figure 4.22 are shown in Figures 4.24 and 4.25, for the situations when only the absolute extreme between zero-crossings is used and when all peaks and valleys are considered, respectively.

The resulting speech quality obtained in this experiment was improved as compared with the results of Experiments II and III. This is particularly true of the first two methods of feature extraction since the quality of the speech obtained when only the absolute extreme between each pair of zero-crossings was used was more degraded than that obtained with the first two methods. For all cases, however, degradation in speech quality can still be perceived. In section II of the attached recorded material, processed utterances are included to illustrate the effects produced by the three methods considered.

4.3.5. Experiment V.

It should be noted that in all experiments, whether those described in the literature or carried out in this study, one pitch period is detected and recirculated periodically in the extension of a segment.

This means that small variations in the pitch period have not been taken into account. In Experiment V this feature was exploited with the aim of further improving on the reasonably good quality speech obtained in Experiment IV.

In the first step of the experiment a 32 msec segment was taken and the positions NP1 and NP2, corresponding to two adjacent pitch peaks, were determined. The position of the first sample of the 32 msec segment is denoted by NS1. In the next step the amplitudes and positions of local maxima (in absolute value) in the interval [NP1, NP2] were extracted. Thus, the following parameters, illustrated in Figure 4.26, were obtained:

Positions: NP1, K1, K2, ..., KN, NP2

Amplitudes: A(NP1), A(K1), A(K2), ..., A(KN), A(NP2).

The other parameters used in the experiment were the positions and amplitudes of other pitch peaks on a period by period basis. To find the other pitch peaks (and consequently the pitch periods) in the backward direction, the following procedure was followed. If the present pitch period is JP and the position of the present pitch peak is NP, then the maximum amplitude sample, of same sign as that located at NP1, is sought in the interval [I1, I2], where

$$I1 = (NP - JP) - \alpha.JP \tag{4.10a}$$

$$I2 = (NP - JP) + \alpha.JP \tag{4.10b}$$

and $\alpha = 0.17$ is the same parameter appearing in equation (4.1). It should be observed that initially JP = NP2 - NP1 and NP = NP1. Figure 4.27 illustrates the relative positions of NP1, NP2, NP, I1 and I2.

The parameters resulting from the search constitute, period-by-period, a series of amplitudes [AP(L1), AP(L2), ...] and positions (L1, L2, ...) of the pitch peaks.

In the reconstruction process, the amplitudes and positions of the local maxima for the other pitch periods were obtained by scaling those in [NP1, NP2]. For instance, if A(K) is the amplitude and K the position (measured from NP1) of a local maximum in [NP1, NP2], then the corresponding local maximum in a pitch period JP with pitch peak amplitude AP(L) is given by the location K' (measured from L) and the amplitude A'(K'), where

$$K' = \frac{K}{NP2 - NP1} \cdot JP \tag{4.11}$$

and

$$A'(K') = \frac{A(K)}{A(NP1)} .AP(L) \tag{4.12}$$

As in Experiment IV, in the reconstruction process a valley was created at the mid-point between each pair of adjacent maxima of same sign. The amplitude of the valley was taken as a fixed proportion (80%) of the smallest amplitude value between the two maxima. A smooth interpolation of the type given by (4.9) was used between two peaks or a peak and a valley.

When analysis (determination of other pitch peaks based on the [NP1, NP2] period) has been completed in the backward direction, the duration of the interval analysed is equal to (NP2 - NS1). If the duration of this interval is greater than, or equal to, NUP, where NUP is a pre-fixed time updating parameter, then NS1 is set equal to

NP2 and the analysis is restarted from position NS1. If that interval is less than NUP the analysis is continued, but in the forward direction. In this case, if at any time a new pitch peak is found, the above mentioned interval is increased by the new pitch period. Consequently, the updating test is applied in order to determine whether the analysis should be restarted (interval > NUP) or continued (interval < NUP). In terms of the quality of the regenerated speech, a good choice of the parameter NUP was found to be NUP = 128 samples.

To illustrate the type of speech reconstruction obtained in Experiment V, a segment of 128 msec (1024 sampling intervals) of original and reconstructed voiced speech, using NUP = 128 samples, is shown in Figure 4.28. From the figure it can be seen that the processed speech follows the original waveform very closely. This is a criterion that is not necessary, but is sufficent to indicate that fairly good subjective quality may be expected. In fact, the resulting signal did sound natural and was clearly intelligible, despite the occurrence of some degradation of quality. Section 12 of the attached tape recording includes some original utterances and the same utterances processed according to Experiment V. Overall, the experiment revealed that the small pitch period variations do, in fact, play an important role in relation to the perceived quality of speech, and that the use of the local maxima for reconstruction of the waveform intrapitch is sufficient for fairly good speech regeneration to be achieved.

4.3.6. Discussion.

Experiments have been described relating to the time coding of voiced speech. The experiments were conducted in order to examine the extent to which simple features of the speech time waveform can be used to obtain good quality regeneration of voiced speech.

In the first four experiments a fixed segmentation of the speech waveform into intervals of 32 msec was considered. The results of the first experiment indicate that if the waveform contained in one pitch period is recirculated in the extension of each segment, the regenerated speech which results is highly intelligible with only a small degradation in quality. However, if the information concerning the waveform intrapitch description is substantially reduced, as in Experiments II and III, then a significant degradation of speech quality can be perceived. In Experiment IV, improvement was obtained by using more information, such as extremal points or local maxima, from the waveform intrapitch. In Experiment V good quality regenerated speech was obtained by using the local maxima of the waveform intrapitch, and by taking the small variations of pitch period (extraction of amplitudes and locations of pitch peaks, period by period) into account.

4.4. A Method of Clipped Unvoiced Speech Regeneration.

As mentioned previously, in order to develop a complete time coding scheme it is necessary to encode unvoiced as well as voiced speech. This section is concerned with the development of a simple technique for encoding the unvoiced portions of the speech waveform.

It has been fairly well established[6] that unvoiced speech can be represented in terms of parameters much more crudely than can voiced speech. For instance, to encode the voiced segments of speech by means of the LPC vocoder technique, 10 or more linear prediction coefficients are usually required.[79-82] However, the quality of the synthesised speech is not affected if only 4, rather than 10, linear prediction coefficients are used to encode the unvoiced segments of speech.[86] This suggests that a cruder representation of unvoiced speech (as compared with voiced speech) can be used without affecting speech quality.

Although intelligibility is not completely lost by peak clipping the speech waveform, the quality of the resulting speech is poor. Clipped speech soundsunnatural and speaker recognition is not preserved. However, on account of the comments made in the previous paragraph, the crude representation of the speech waveform by means of its zero-crossing positions is almost certainly more critical for voiced than for unvoiced speech. The present section examines the feasibility of obtaining good quality regenerated unvoiced speech using the zero-crossing positions of the waveform and a simple method of unvoiced clipped speech regeneration.

Consider a segment of unvoiced speech waveform and let $\{t_1, t_2, \ldots\}$ be the time instants associated with the points of zero derivative (extremes) of the waveform. Consider the instants associated with alternate extremes (i.e., either $\{t_1, t_3, t_5, \ldots\}$ or $\{t_2, t_4, t_6, \ldots\}$).

In the case of short segments of unvoiced speech waveforms observation shows that the distance between two successive instants $(t_{j+2} - t_j)$ associated with alternate extremes is nearly constant for the entire duration of the segment. Figure 4.29 shows a 16 msec segment of unvoiced speech waveform and the time instants at which alternate extremes occur.

In order to obtain a quantitative indication of the extent to which this observation applies more generally, the distances $\{D_i \; ; \; i = 1, \ldots, N\}$ between each pair of successive instants associated with the alternate extremes of the waveform were measured. The measurements were taken for 32 msec sections of unvoiced speech. The average distance

$$\bar{D} = \frac{1}{N} \sum_{i=1}^{N} D_i \tag{4.13}$$

was computed for each segment. The results obtained revealed that, in general, the distances D_i differ from their short-term average $\bar{D}$ by less than one sampling interval (0.125 msec). Histograms relating the difference $D_i - \bar{D}$ are presented in Figures 4.30 and 4.31 for male and female speech, respectively. These results indicate that $\bar{D}$ is a useful parameter for estimating the number of peaks of the waveform between two adjacent zero-crossings for each segment. For a zero-crossings interval T, the number of peaks, P, can be estimated from the ratio of T to $\bar{D}$.

A method of unvoiced clipped speech regeneration, involving the use of a smooth shape between adjacent zero-crossings of fixed segments of the waveform, was developed. For each zero-crossings interval of duration T, the smooth shape was defined by a number of peaks P given

by the integer part of the ratio $T/\bar{D}$. $\bar{D}$ was computed for each segment of 30 msec duration. It should be noted that although the histograms shown in Figures 4.30 and 4.31 were obtained for segments of 32 msec duration, the use of 30 msec segments was found to be more convenient in view of the time coding technique to be described in section 4.5. The smooth shapes were obtained by means of sinusoidal functions, and the amplitudes of the valleys were taken as 0.6 times the amplitudes of the peaks. Thus the regenerated signal consists of piecewise approximation of the speech waveform intervals between successive zero-crossings. This is accomplished by the use of smooth shapes which are defined by a number of peaks estimated from the speech waveform for each interval. The rms value of the speech amplitudes was also obtained for each 10 msec interval. This rms value was used in the reconstruction of the waveform in order to restore the loudness envelope of the original speech.

Different types of unvoiced sounds were processed according to this method. In section 13 of the attached tape recording, original and processed unvoiced sounds for five different words containing the fricatives /ʃ/, /f/, and /s/, and the plosives /p/ and /t/ are presented. From the recordings it is clear that the method provides very good regeneration of unvoiced speech, with the processed samples being almost indistinguishable from the originals. These results indicate that it is possible to obtain good quality of unvoiced speech using the zero-crossing positions of the waveform and a simple method

of unvoiced clipped speech regeneration. The method suggested in this section involved: (1) the short-term (30 msec) average distance $\bar{D}$ between two successive instants associated with the alternate extremes of the waveform; and (2) the rms value of the speech amplitudes for every 10 msec section of the waveform.

4.5. Low-Bit-Rate Time Coding of Speech Signals.: A New Scheme.

The procedure adopted in the development of a new time coding scheme comprised the experimental investigation of a number of simple features of the speech waveform. Voiced and unvoiced speech waveforms were treated independently and the results of the experimental investigation were presented in sections 4.3 and 4.4, respectively. The basic features exploited in the experiments were: (1) the zero-crossings of the waveform; (2) the amplitudes and positions of extremes; and (3) the quasi-periodicity of voiced speech waveforms. Using the ideas and results presented in sections 4.3 and 4.4, and the time-domain pitch period detector (TDPD) developed in section 4.2, a complete time coding technique has been devised. A detailed description of this technique and the main results obtained follow.

4.5.1. System Description.

The basic elements of the system are: (1) an algorithm which classifies the speech signal into 'voiced' or 'non-voiced' segments; (2) a pitch period detection technique; (3) a method which analyses and reconstructs the voiced portions of the waveform; and (4) a method which analyses and reconstructs the non-voiced portions of the waveform.

The input speech signal was initially low-pass filtered at 3.04 KHz, sampled at a rate of 8 Ksamples/sec, and then digitised using a 10-bit A/D convertor. The resulting signal was then segmented into frames of 10 msec (80 samples) duration, which were classified as 'voiced' or 'non-voiced' by the algorithm described in section 4.2. Each 30 msec interval consisting of three successive frames (j, j + 1, j + 2) was classified as voiced if at least two frames were voiced. Otherwise, the interval was classified as non-voiced.

The procedure adopted for the coding of voiced speech is identical to that used in Experiment V of section 4.3. This procedure was chosen due to the fact that good quality speech can be obtained and the operations involved in the analysis and reconstruction of the voiced speech waveform are relatively simple.

If the whole interval consisting of three successive frames is classified as voiced, then the first sample position of the first frame is denoted by NS1, and a 32 msec analysis segment (30 msec plus an additional 2 msec interval) is taken. The application of the TDPD pitch detection algorithm to the 32 msec segment gives the pitch period markers at locations NP1 and NP2. These locations and the corresponding amplitudes A(NP1) and A(NP2) are extracted and transmitted. In addition, the locations {K1, K2, ..., KN} and amplitudes {A(K1), A(K2), ..., A(KN)} of local maxima (in absolute value) in the interval [NP1, NP2] are also extracted and transmitted. The amplitudes and positions of the other pitch peaks are obtained, on a period by period basis, according to the following search procedure: if NP and JP are, respectively, the present pitch period and the position of the present pitch peak, then

the location of the maximum amplitude sample is sought in the interval [I1, I2] given by (4.10a) and (4.10b). This location and amplitude correspond to the location and amplitude of the pitch peak adjacent to JP. This procedure is then repeated in order to determine the next pitch peak. After completion of the analysis (determination of other pitch peaks) in the backward direction, an updating test is applied. This test involves determining whether or not the interval analysed, i.e., (NP2 - NS1), exceeds a pre-fixed time updating parameter NUP. In this case, NS1 is set equal to NP2 and the analysis is restarted from this instant (NS1). If (NP2 - NS1) $\leq$ NUP, the analysis of pitch peaks is continued in the forward direction and, as in Experiment V, if at any time a new pitch peak is found, the updating test is again applied. In this case the interval analysed is given by the difference between the location of the new pitch peak and NS1.

For voiced speech intervals, the following parameters are transmitted:

1) A(NP1), A(NP2): amplitudes of pitch peaks as determined by the TDPD algorithm;
2) NP1, NP2: locations of pitch peaks as determined by the TDPD algorithm;
3) A(K1), A(K2), ..., A(KN): amplitudes of local maxima of the waveform intrapitch;
4) K1, K2, ..., KN: locations of local maxima of the waveform intrapitch;
5) AP(L1), AP(L2), ... : amplitudes of the other pitch peaks on a period by period basis; and
6) L1, L2, ... : locations of the other pitch peaks on a period by period basis.

The reconstruction of the voiced speech waveform is obtained by: (1) scaling the positions and amplitudes of local maxima in the period [NP1, NP2] to the other pitch periods [see expressions (4.11) and (4.12)]; (2) creating one valley between each pair of maxima of same sign (the amplitude of the valley is set equal to 80% of the smallest amplitude between the two maxima); and (3) using a smooth interpolation [of the type given by (4.9)] between two peaks or one peak and one valley.

The analysis and reconstruction of unvoiced (and silence) segments of the waveform use the method described in section 4.4. The encoding of silence as unvoiced intervals was chosen for two reasons. Firstly, since the amplitude variation is followed by the rms value taken from short segments of 10 msec, the reconstructed waveform will not introduce any degradation in the quality of the regenerated speech. Secondly, the silence detector may assign as silence segments some low-level portions of unvoiced speech.

If the 30 msec interval consisting of three successive frames was classified as unvoiced, then the following parameters were extracted and transmitted to the receiver for use in the reconstruction process:

1) Z(I): zero-crossing locations in the 30 msec segment of unvoiced speech;
2) RMS(J), J = 1, 2, 3 : rms value of the speech amplitudes in each sub-interval of 10 msec duration; and
3) $\bar{D}$: short-term (30 msec) average distance between two successive instants associated with the alternate extremes of the waveform.

The method for reconstructing the unvoiced segments of the speech waveform was based on the use of a smooth shape (with P peaks) between each pair of adjacent zero-crossings. The number of peaks, P, was estimated from the ratio of T to $\bar{D}$, where T is the time interval between adjacent zero-crossings and $\bar{D}$ is the parameter mentioned above. After reconstruction, the loudness envelope of the original speech was restored by means of the rms value of the speech amplitudes in each 10 msec interval.

The transmission parameters extracted by the analysis of voiced and unvoiced speech were quantised according to the nature of each parameter. The difference between adjacent pitch periods, rather than the position of each pitch peak, was used to transmit information relating to the positions of the pitch peaks (L1, L2, ...). This information was quantised using a 6-bit quantiser, the accuracy of which was within 1 sampling interval. Since the zero-crossing intervals for unvoiced sounds are generally very small, it was found that the most suitable strategy for quantising the zero-crossing locations Z(I) involved a level of accuracy equal to 1 sampling interval. Thus, 1 bit per sample is used for this purpose. The rest of the parameters are quantised as shown in Table 4.2. As can be seen from this table, 6-bit log-quantisation is used to quantise amplitude parameters including both polarities. The polarity of A(NP2), and AP(L1), AP(L2), ..., need not be transmitted since it is implied by the polarity of A(NP1).

4.5.2. Experimental Results.

The waveforms of the original and the reconstructed speech are compared in Figures 4.32 to 4.37. The 128 msec reconstructed waveform segments shown in these figures are typical examples of the results obtained for a wide variety of speech sounds spoken by both male and female speakers. As can be seen from these figures, the waveform information is in general well preserved, and the transition from one type of sound to another closely follows that of the original speech (see, for instance, Figure 4.36).

Table 4.2. Quantisation of some Parameters used in the Time Coding Technique.

Parameter	Polarity	Type of Quantiser	No. of Bits
$\bar{D}$	+	log (μ = 100)	3
RMS(J)	+	log (μ = 100)	5
A(NP1)	+, -	log (μ = 100)	6
A(NP2)	+	log (μ = 100)	5
NP1, NP2	+	linear	7 each
A(K1), A(K2), ...	+, -	log (μ = 100)	6 each
K1, K2, ... (differences)	+	linear	4 each
AP(L1), AP(L2), ...	+	log (μ = 100)	5 each

Comparisons of original and reconstructed speech are presented in the form of spectrograms in Figures 4.38 to 4.43. The spectrogams of original speech for the utterances " The box was thrown beside the parked tru(ck)" and " Glue the sheet to the dark blue backgroun(d)" are presented in Figures 4.38 and 4.40, respectively, and the spectrograms of unquantised and quantised versions of processed speech for each of these utterances are shown in Figures 4.39 and 4.41, respectively. The spectrograms of the original and the reconstructed (quantised) signal for the utterance " It's easy to tell the depth of the well" spoken by a male and a female speaker respectively, are shown in Figures 4.42 and 4.43. As can be seen from these figures, for the low-frequency range, which includes the first and second formants, the reconstructed speech closely resembles the original. In particular, the first formant region is well preserved in the processed speech, but the third formant is either missing or attenuated in most portions of the regenerated speech. This effect can be explained by the smooth interpolation employed in the reconstruction process of the voiced sounds, which introduces a low-pass type characteristic to the processed speech. Comparison of the spectrograms of original and reconstructed speech also reveals that the important features of voiced-unvoiced discrimination and pitch period are well preserved. With regard to the quantisation effects, no significant differences can be observed when comparing the spectrograms of the unquantised and quantised versions of the processed speech. The subjective quality of the reconstructed speech will be discussed in section 4.6.

Table 4.3 shows the average bit-rate obtained for each of twelve sentences, when the parameters transmitted by the time coding scheme were encoded according to the bit allocation given previously. An average bit-rate of 7.6 Kbits/sec resulted when the entire interval of 29.696 seconds of speech (the twelve sentences) was considered. As can be seen from Table 4.3, the variation in bit-rate between the various sentences is small. This indicates that one can expect only reasonably short delays to be required for the conversion of the variable-rate-source data into constant-transmission-rate data having the average bit-rate referred to above. Estimates of these delays and buffer storage for the time coding technique at a constant-transmission-rate of 7.6 Kbits/sec will be presented in Chapter 5. Table 4.3 also shows that the bit-rates obtained for female speech are lower than those obtained for male speech. This is primarily linked to the pitch period length. The number of information bits required to encode the shorter pitch periods of the female speech is correspondingly smaller than the number required for the male speech.

4.6. Considerations Relating to Speech Quality and System Complexity.

In informal listening tests male and female speech coded with the Hadamard technique were compared with the same speech material processed by the time coding scheme. The Hadamard coding technique used a 64-point Hadamard transformation, 8 dominant coefficients, and a special technique for labelling 7 of these coefficients (see reference 117). The Hadamard coder operated at a bit-rate of 9.6 Kbits/sec

Table 4.3. Time Coding Technique : Average Bit-Rate Per Sentence.

Sentence Number	Sentence Duration (msec)	Bit-Rate (Kbits/sec)
1 (male speaker)	2560	7.7
2 (male speaker)	2752	8.0
3 (male speaker)	2368	8.9
4 (male speaker)	2112	8.3
5 (male speaker)	2560	7.2
6 (male speaker)	2560	7.8
7 (female speaker)	2560	7.6
8 (female speaker)	2496	7.7
9 (female speaker)	2496	6.8
10 (female speaker)	2048	6.9
11 (female speaker)	2624	7.6
12 (female speaker)	2560	6.8

and the time coder at 7.6 Kbits/sec. With regard to speech intelligibility it was observed that both schemes produced good results since the processed speech is effortlessly intelligible in both cases. However, the speech processed with the time coding technique was found to be better in that the small degradation observed in the speech quality was less annoying than that obtained with Hadamard coding. In addition, more natural speech regeneration was obtained with the time coding approach. It was also clear that the increased degradation for female speech, characteristic of the Hadamard coding technique, did not occur with the time coder.

At bit-rates below 12 Kbits/sec the Hadamard transform coder has been found[12] to give better results, in terms of performance, than simple waveform coding techniques such as digitally controlled delta-modulation (DCDM). It is, therefore, reasonable to argue that the 7.6 Kbits/sec time coding scheme also provides better quality speech than the simple methods of waveform coding such as DCDM. This argurment is well founded since the time coder exploits the characteristics of the speech waveform to a larger extent than waveform coders and, thus, is expected to give better results at low-bit-rates.

Section 14 of the attached tape recording includes processed utterances obtained with the new time coding method at 7.6 Kbits/sec. In section 15 of the recordings each processed utterance is followed by the same utterance resulting from the Hadamard transform coding scheme operating at 9.6 Kbits/sec.

The new time coding technique described in this chapter is of moderate complexity, the most complex part being that concerned with the analysis of the voiced portions of the waveform. The analysis of unvoiced segments of speech is much less complex. The operations involved (zero-crossings detection, counting of number of extremes, and averaging) are very simple in that little arithmetic is required for their implementation. Furthermore, the reconstruction of voiced and unvoiced speech segments can be implemented relatively easily. The smooth shapes used to reconstruct the speech waveform can be obtained from a read-only memory by means of look-up table techniques.

Table 4.4 shows the lower and upper limits of the number of operations required for the classification of speech into 'voiced' or

Table 4.4. Lower and Upper Limits of the Number of Operations Required for Voiced-Unvoiced Classification and Pitch Detection.

Operation	Voiced-Unvoiced Classification				Pitch Detection	
	Unvoiced		Voiced			
	Lower	Upper	Lower	Upper	Lower	Upper
Addition	0	24	16	24	3	32
Comparison	240	640	480	640	348	1581
Multiplication by a constant	0	6	4	6	3	21

'unvoiced' segments and for pitch period determination. These limits were obtained by counting the number of operations necessary to perform V/UV classification and pitch detection, according to the minimum and maximum possible routes of the algorithms. As can be seen from Table 4.4, no multiplications are required since the small number of multiplications by a constant can be performed employing a few additions and arithmetic shift instructions. In addition to the number of operations needed to perform voiced-unvoiced classification and pitch detection, the coding of voiced speech segments also requires the following operations:

1) In order to obtain the local maxima of the waveform intrapitch (i.e., between pitch peak locations NP1 and NP2), it is necessary to perform

$$\sum_{i=1}^{M} [(NP2 - NP1) - i] = M(NP2 - NP1) - \frac{M(M + 1)}{2}$$

comparisons, where M is the number of local maxima found in the interval (NP1, NP2). Typical values of M are between 3 and 14.

2) If N_P is the number of pitch peaks (excluding NP1 and NP2) occurring in the analysis segment, then the number of operations needed to determine these pitch peaks can be estimated as approximately $3N_P$ additions, $Np[(1/3)(NP2 - NP1) - 1]$ comparisons, and N_p multiplications by a constant.

3) If N_F pitch peaks are found from the analysis in the forward direction, then $(N_F + 1)$ additions and $(N_F + 1)$ comparisons are also needed in order to perform the updating test.

The above estimates were applied to the particular situation in which the number of operations involved is maximal for a wide range of speech signals. This situation corresponds to NP2 - NP1 = 56 samples, M = 14, $N_P = 2$, $N_F = 1$, and $L_V = 21.7$ msec, where L_V is the duration of the analysis segment. In this case it was found that a total of 64 additions, 2938 comparisons and 29 multiplications by a constant were required. The upper limit of the number of operations needed for voiced-unvoiced classification and pitch detection was included in this counting. If the number of operations is then divided by the duration (L_V) of the analysis segment, it results that approximately 2950 additions, 135390 comparisons and 1336 multiplications by a constant are required per second. It is also clear that most of the operations are used to perform pitch detection and to obtain the local maxima of the waveform intrapitch.

The above estimates indicate that V/UV classification and analysis of voiced speech, including pitch detection, can be implemented in real-time using currently available microprocessors (such as the Intel 8086, the Zilog Z8000, and the Motorola MC 68000).[161] For instance, the execution time for 16-bit additions and comparisons taken by the Motorola MC 68000 microprocessor[162] is 1.5 μ sec for 16-bit addressing. If multiplication by a constant is taken as three

additions then the overall execution time for V/UV classification and analysis of voiced speech (including pitch detection) can be estimated at approximately 213 msec per second of speech. This corresponds to the least favourable situation considered above. The accuracy of the 16-bit microprocessors is believed to be sufficient, since the scheme employs numerical processing which does not cause any kind of accumulative round-off error.

In order to analyse 1 second of speech with the 9.6 Kbits/sec Hadamard coding scheme mentioned at the beginning of this section, 48000 additions and 14122 comparisons are required.[12] Thus, in the least favourable situation for the time coder, the most complex part of the system (analysis of voiced speech) requires about 2.3 times as many operations as the Hadamard coder. However, it should be noted that this factor is usually lower. Furthermore, the time coding scheme shows promise at low-bit-rates whilst the speech quality obtained with the Hadamard coding falls substantially for bit-rates below 6 Kbits/sec. Furthermore, the complexity involved in the time coding scheme is less than that required in other speech coding techniques, such as adaptive predictive coding and adaptive transform coding, which also show promise at low-bit-rates.

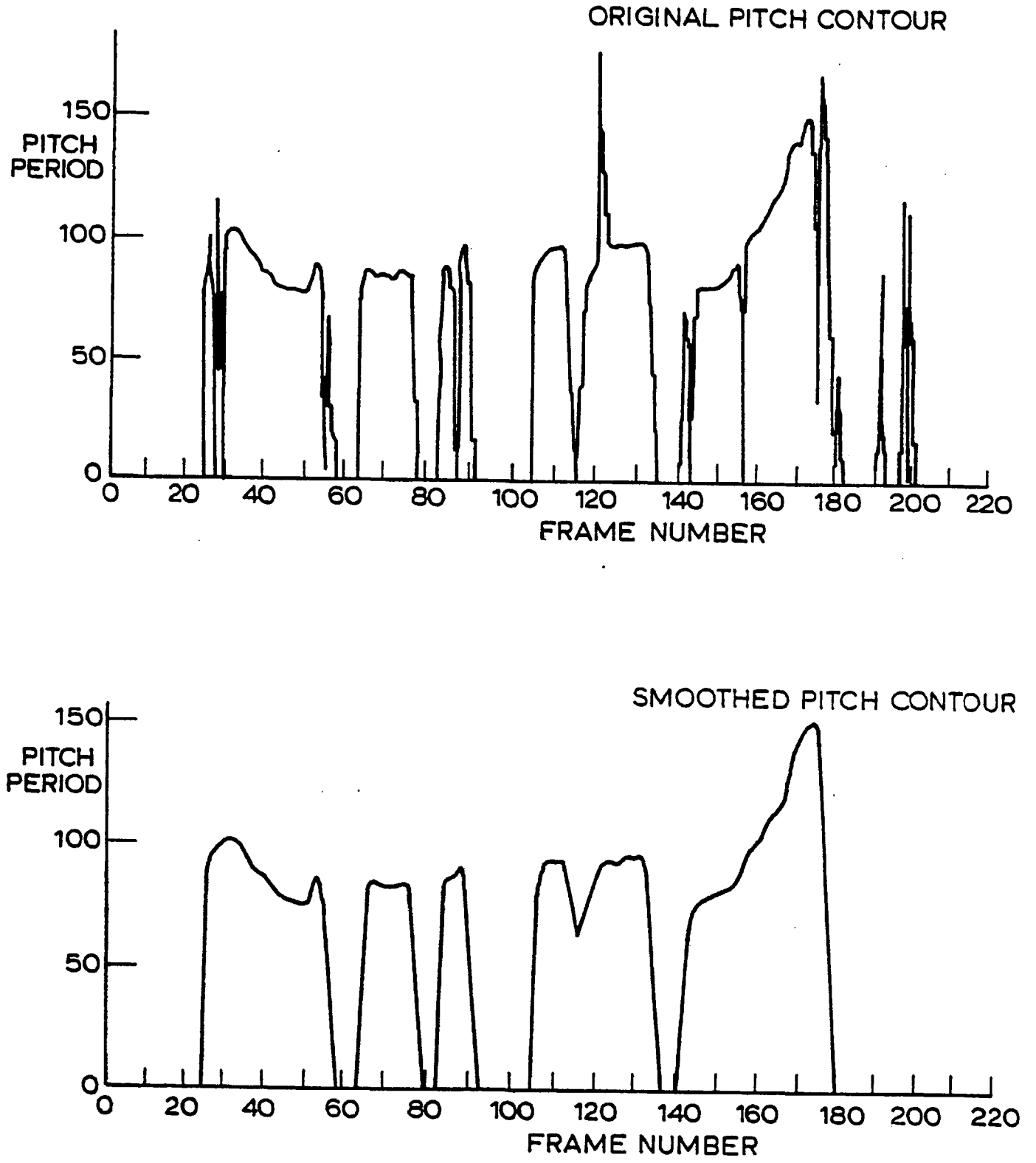


Figure 4.1. Non-linear smoothing of pitch period contours (after Rabiner, Sambur, and Schmidt[156]).

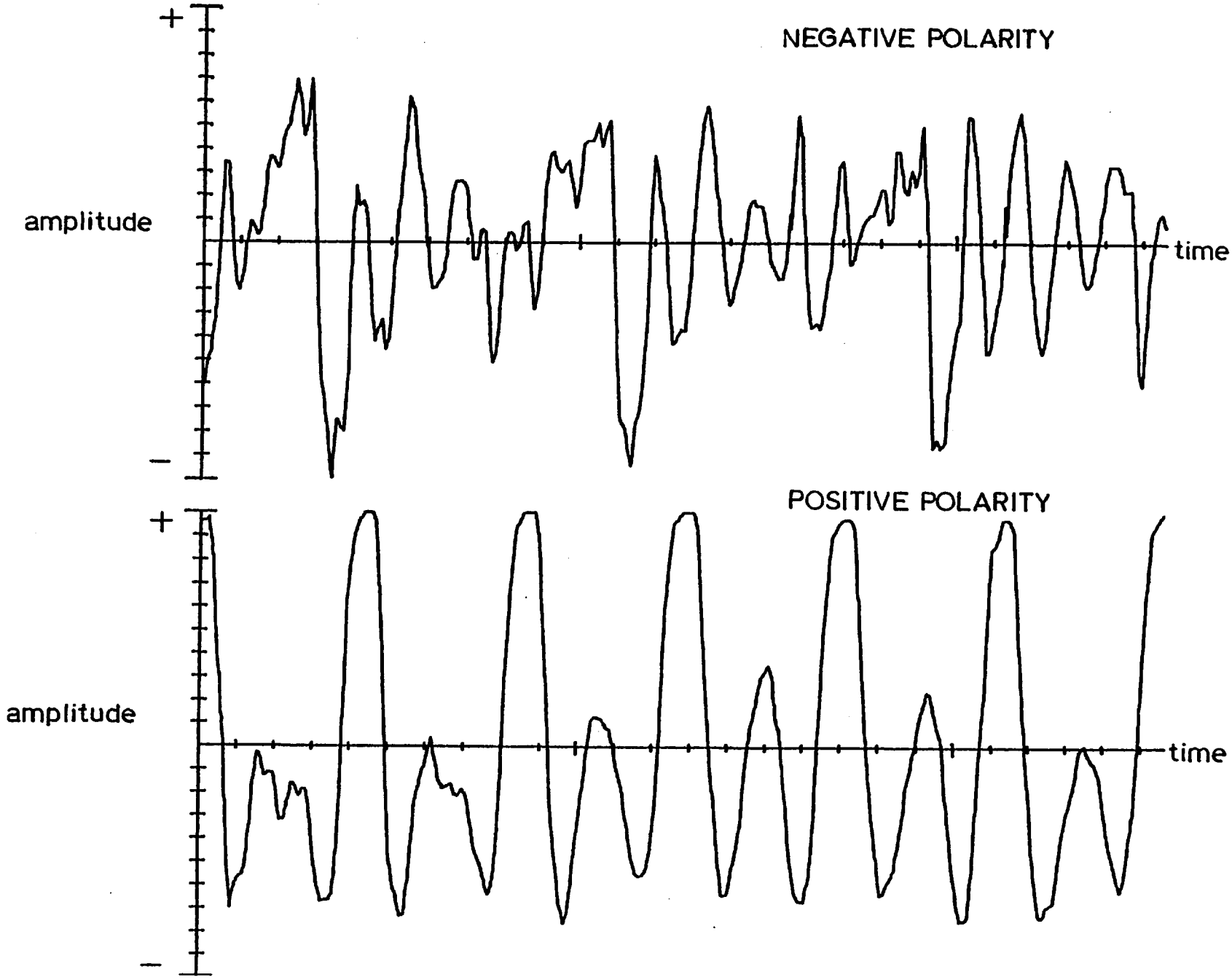


Figure 4.2. Illustration of pitch peaks polarity (time interval: 32 msec).

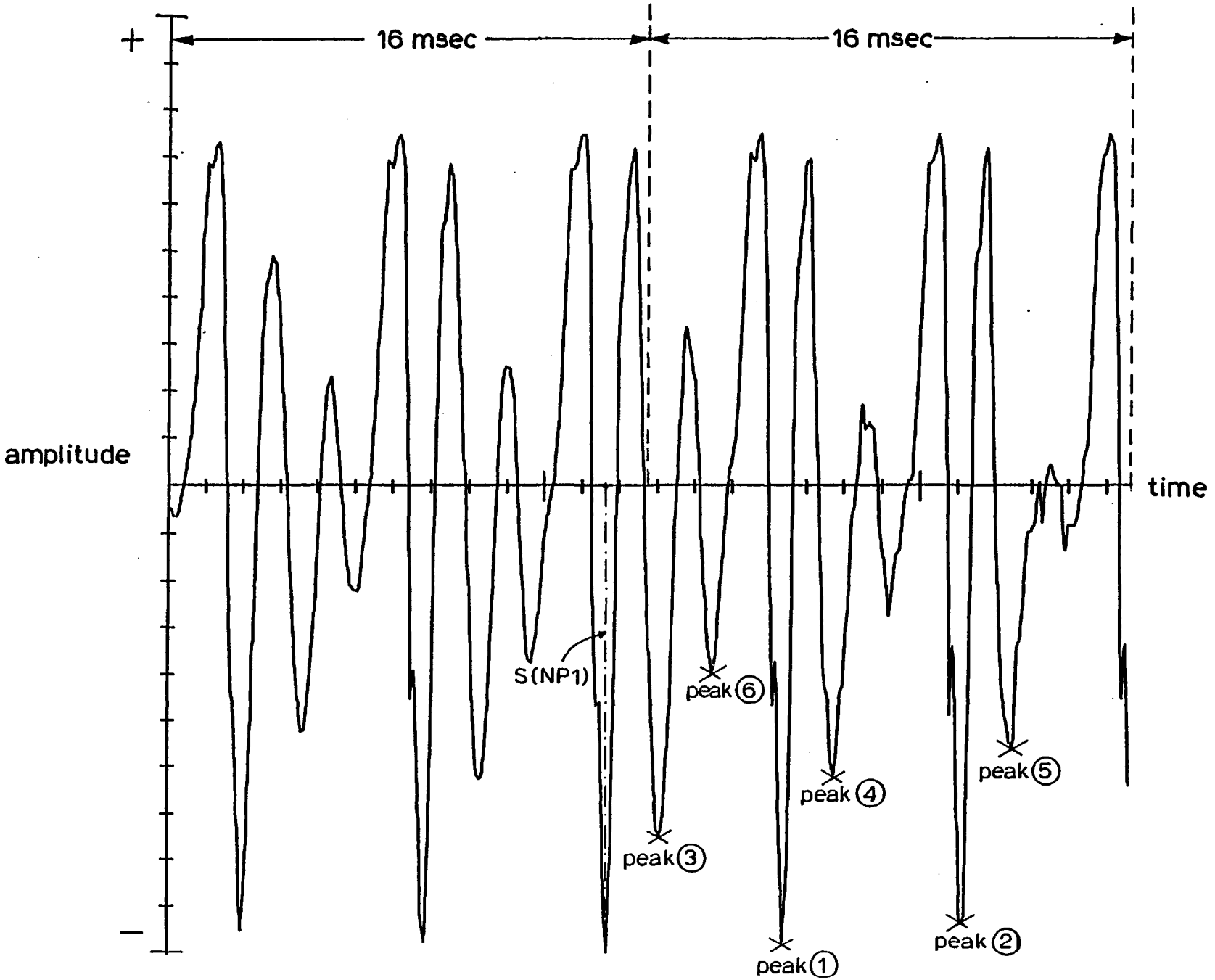


Figure 4.3. TDPD algorithm: selection of NP1 and the N peaks (N = 6) according to step 1.

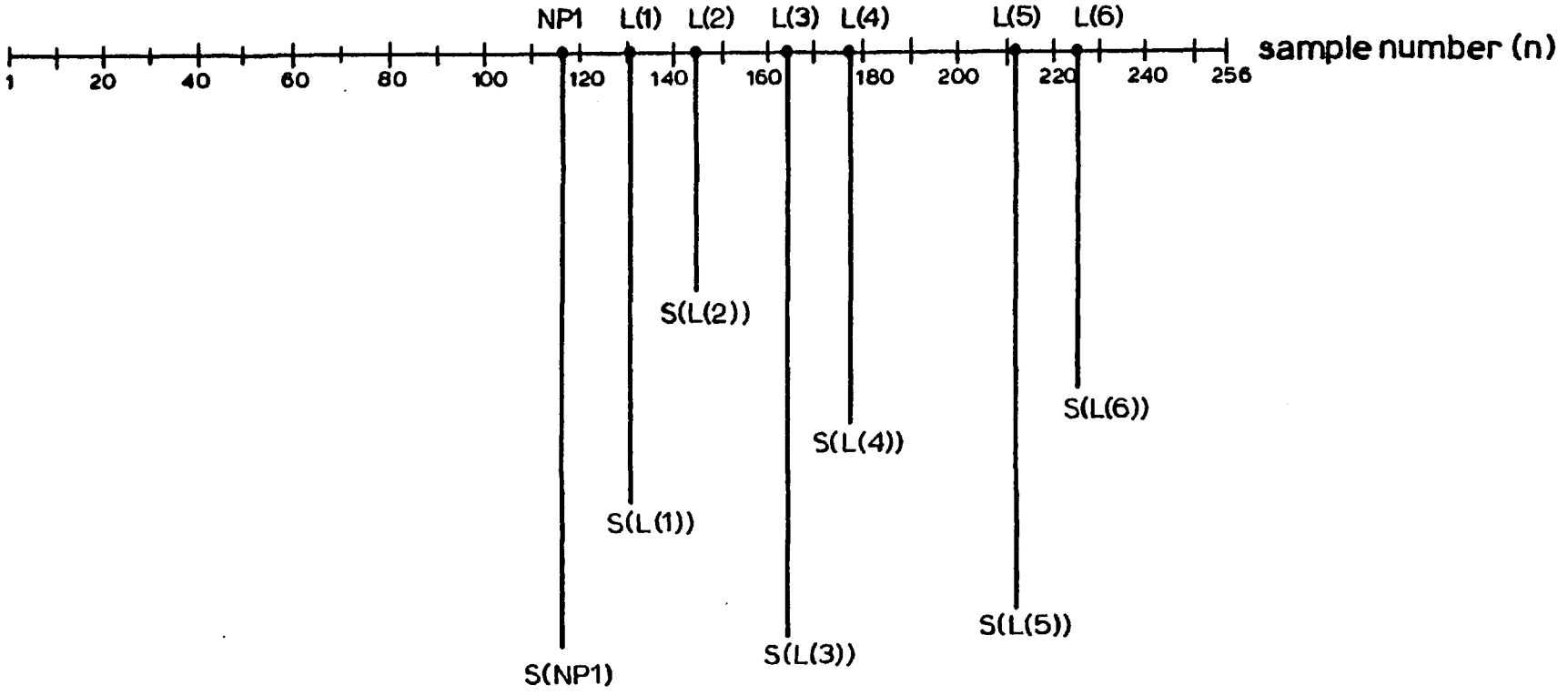


Figure 4.4. TDPD algorithm: result of step 1 for the speech segment used in Figure 4.3.

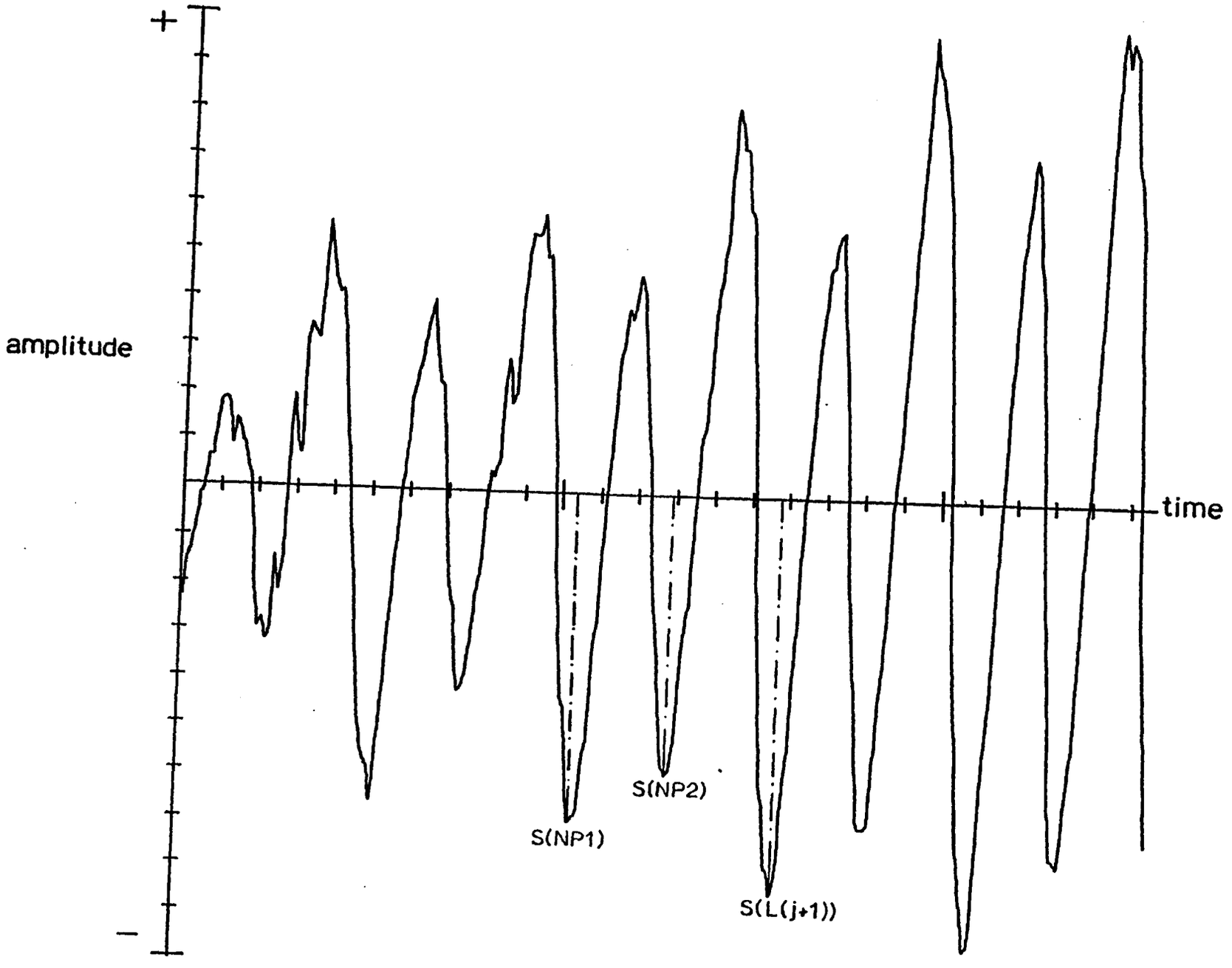


Figure 4.5. TDPD algorithm; illustration of an incorrect pitch estimation to be corrected by step 3.

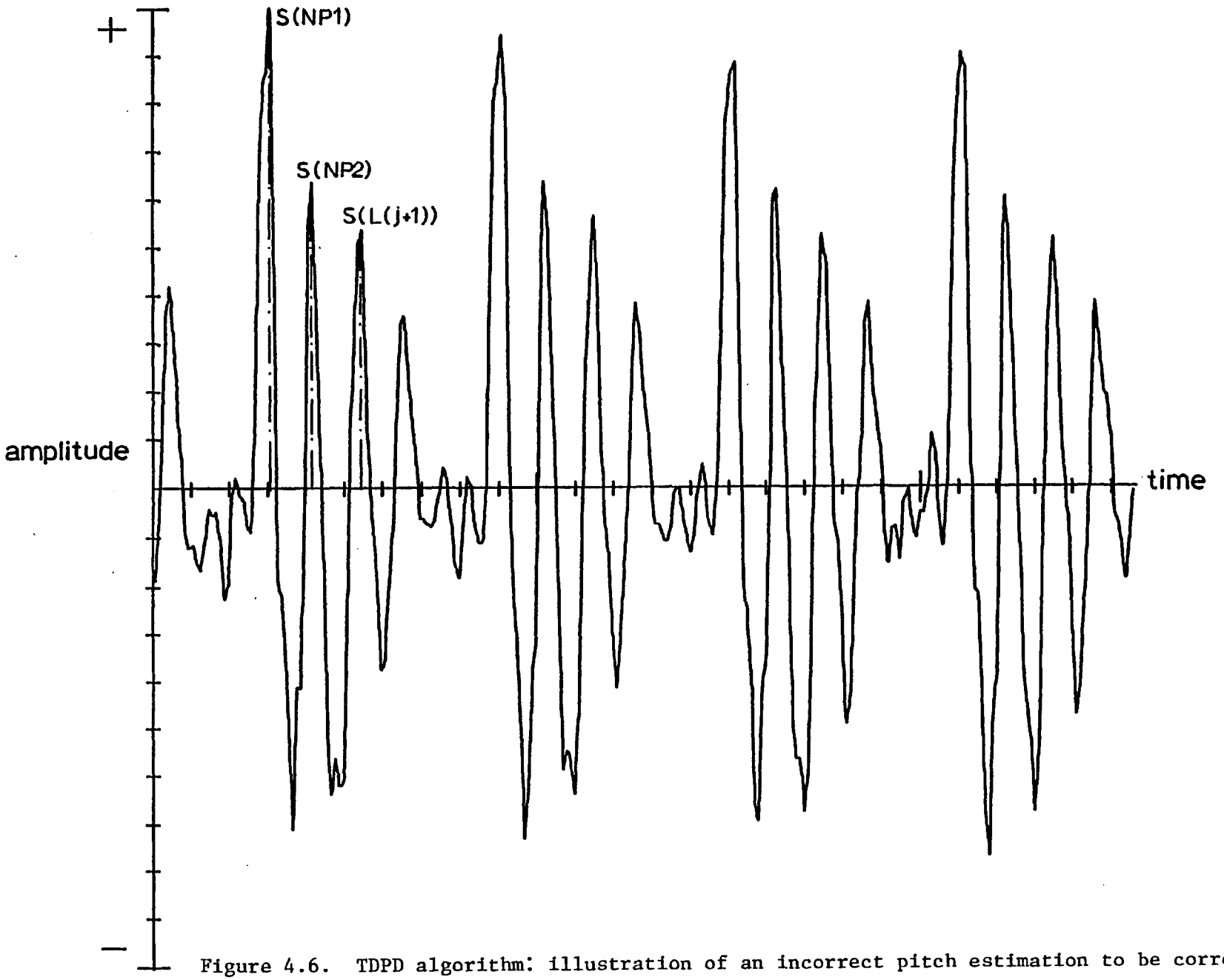


Figure 4.6. TDPD algorithm: illustration of an incorrect pitch estimation to be corrected by step 4.

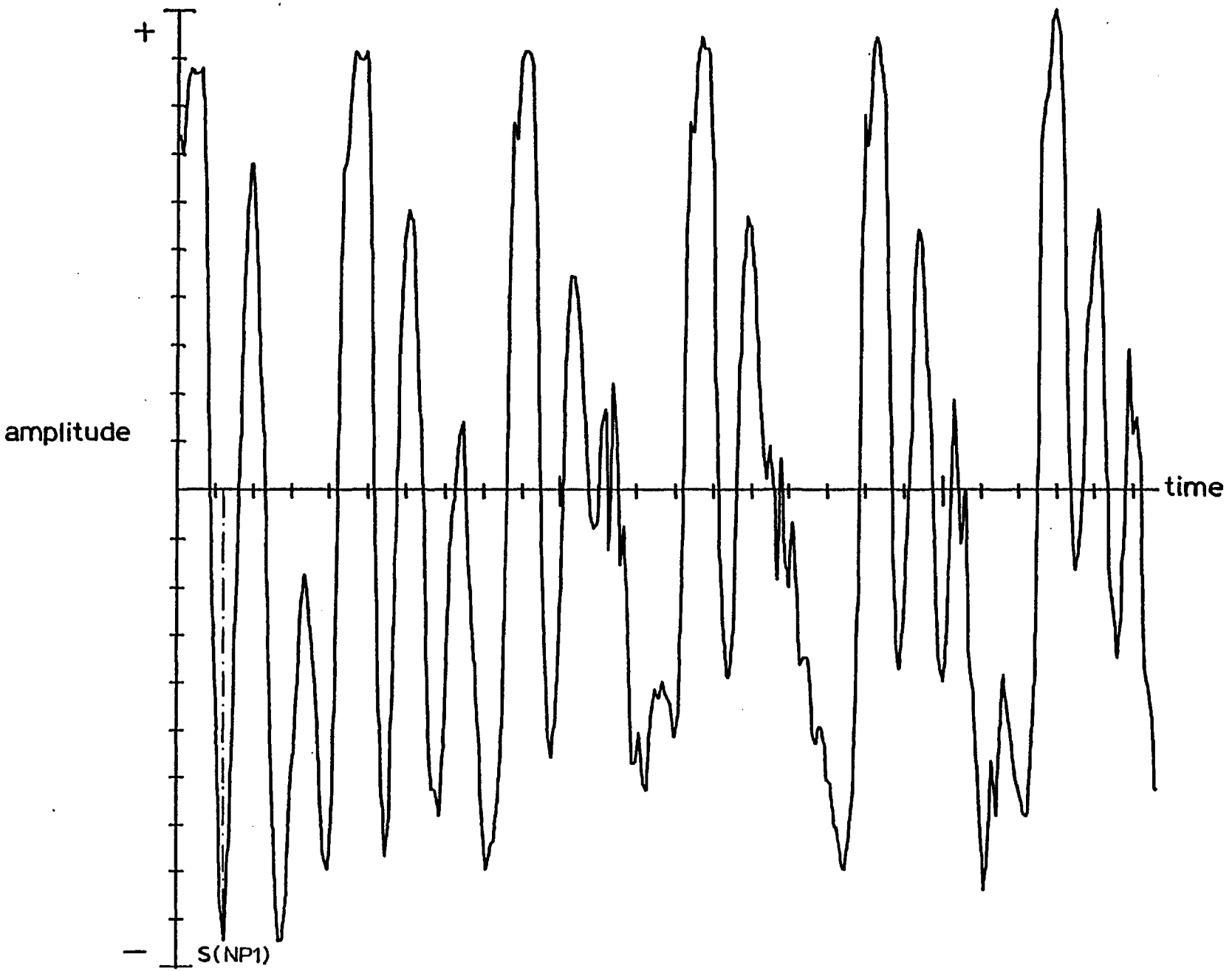


Figure 4.7. TDPD algorithm: illustration of a typical speech segment to which step 5 is to be applied.

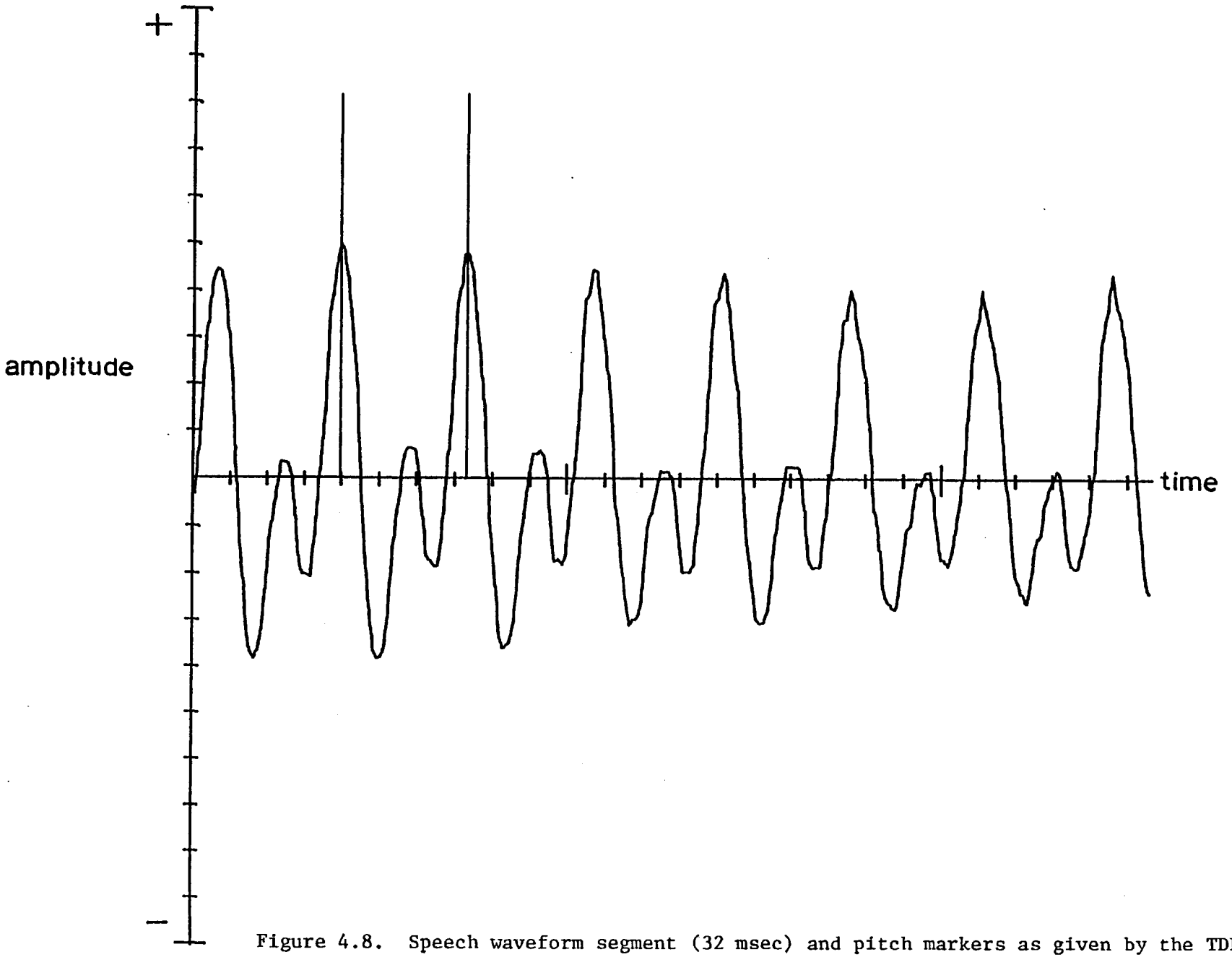


Figure 4.8. Speech waveform segment (32 msec) and pitch markers as given by the TDPD algorithm.

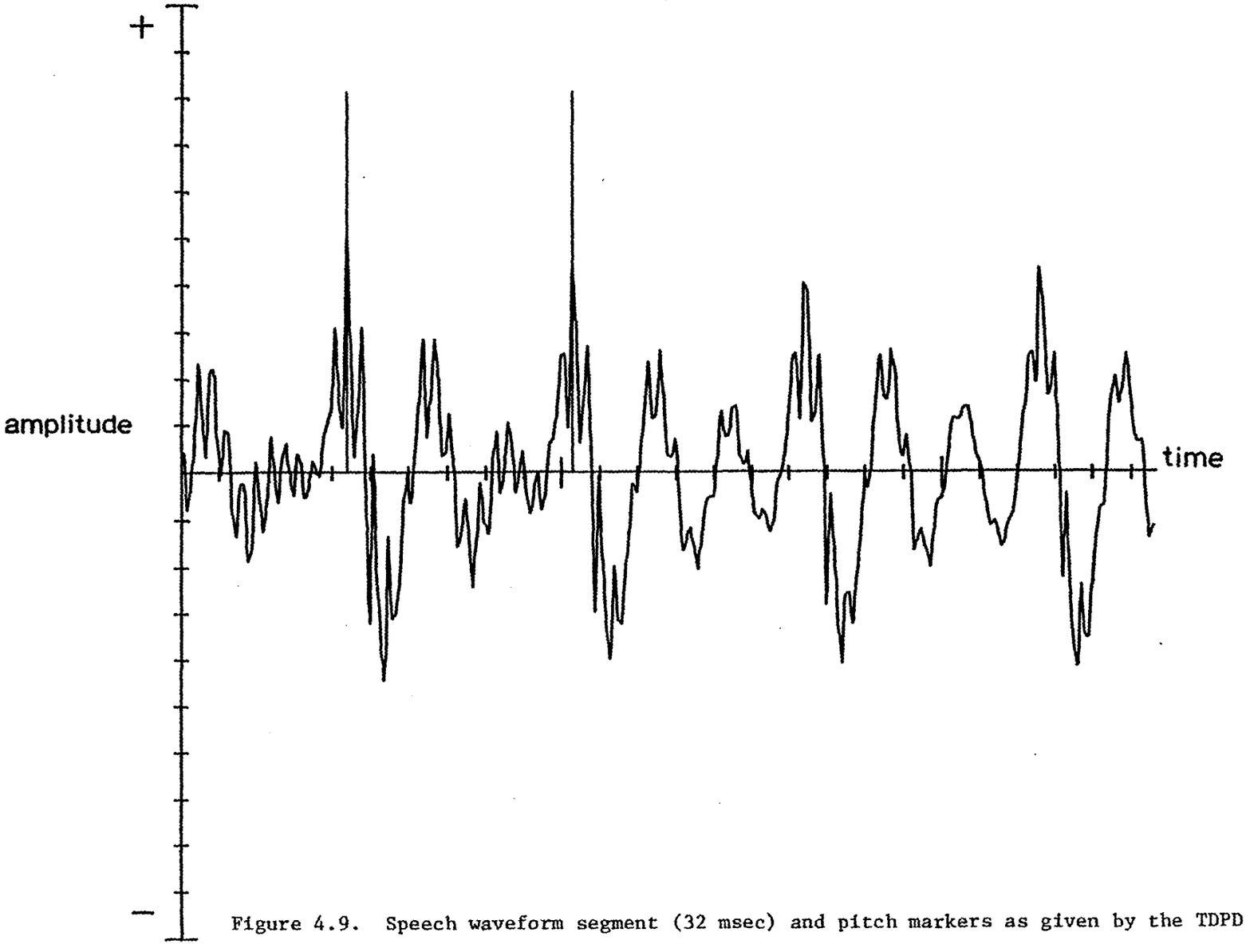


Figure 4.9. Speech waveform segment (32 msec) and pitch markers as given by the TDPD algorithm.

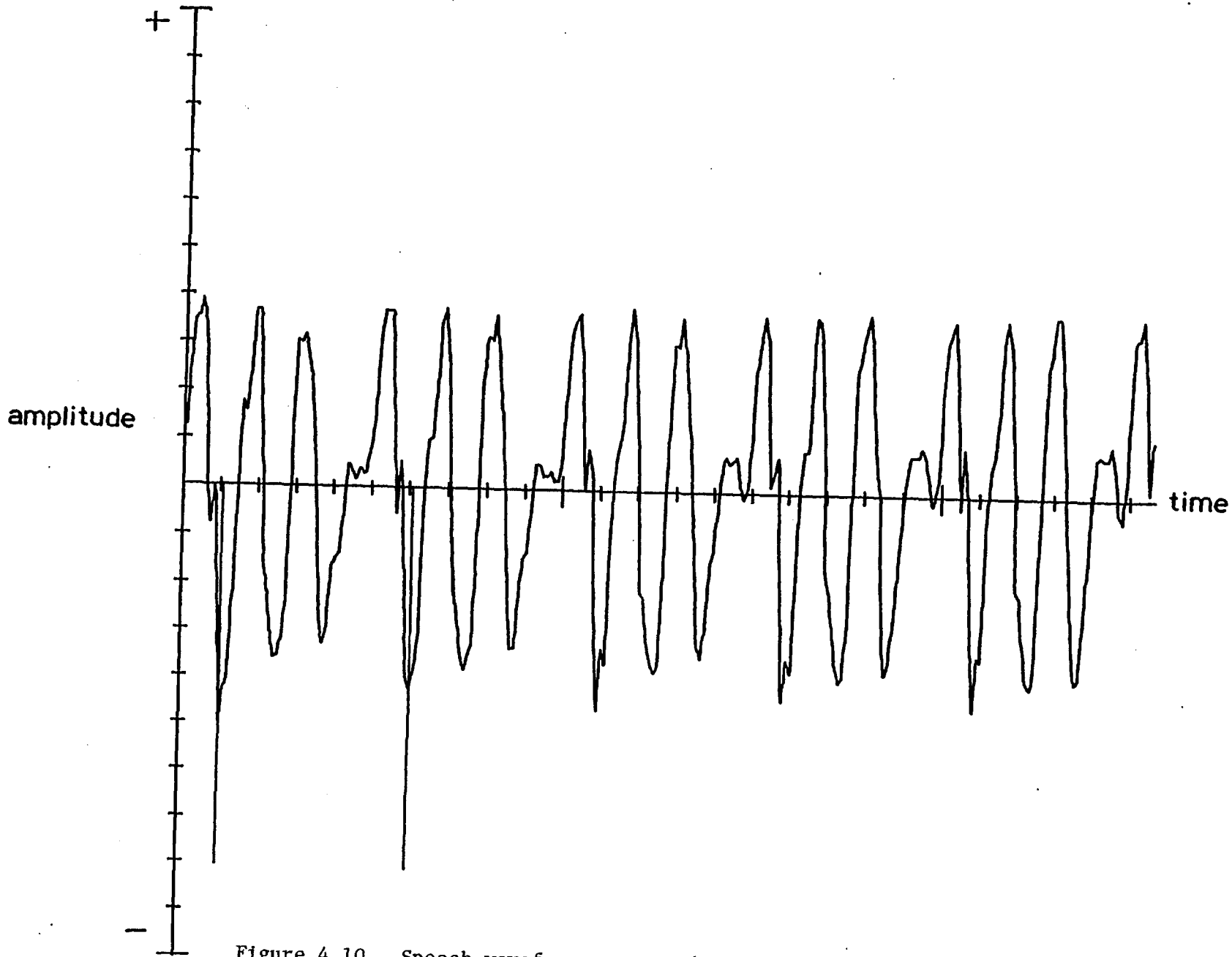


Figure 4.10. Speech waveform segment (32 msec) and pitch markers as given by the TDPD algorithm.

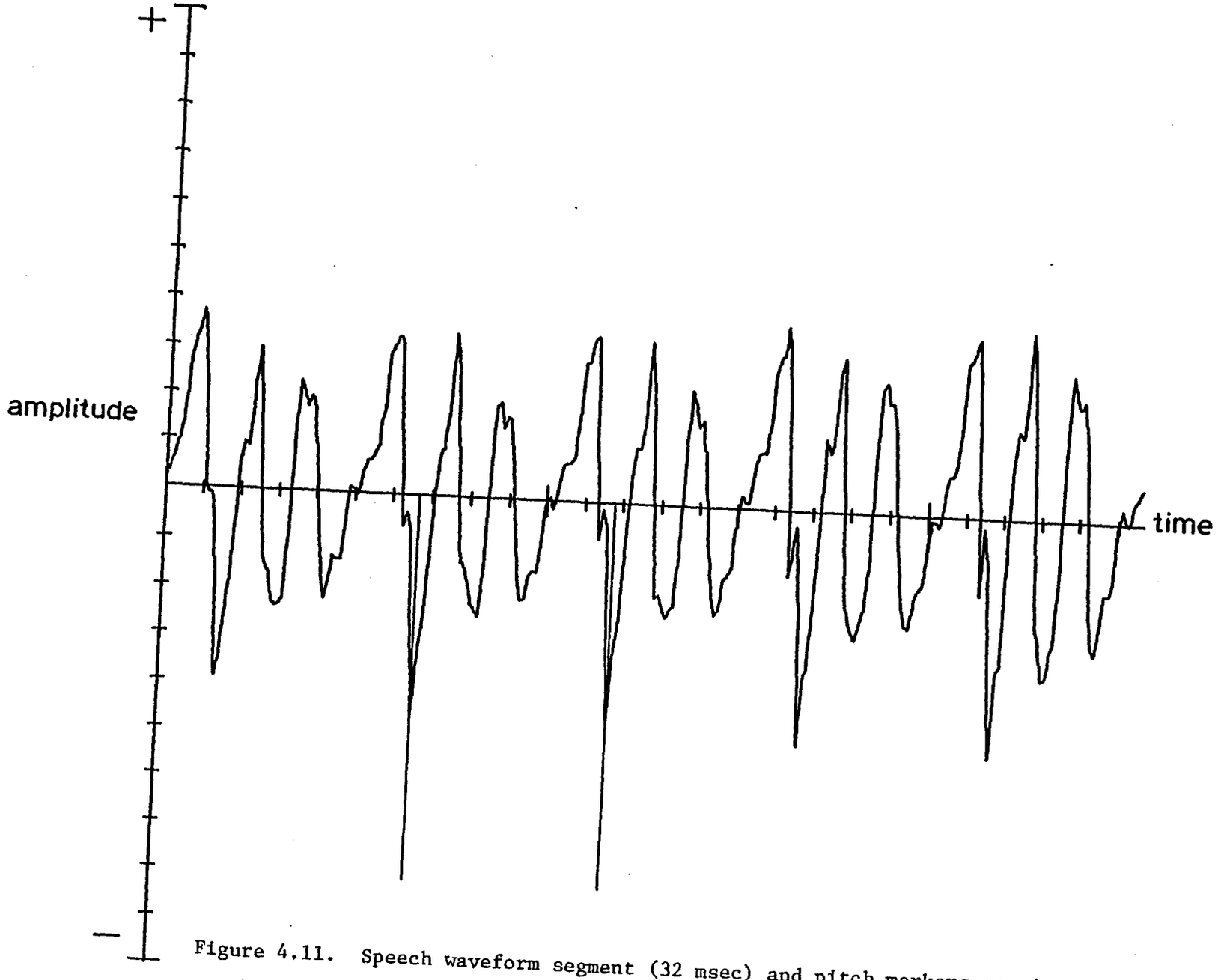


Figure 4.11. Speech waveform segment (32 msec) and pitch markers as given by the TDPD algorithm.

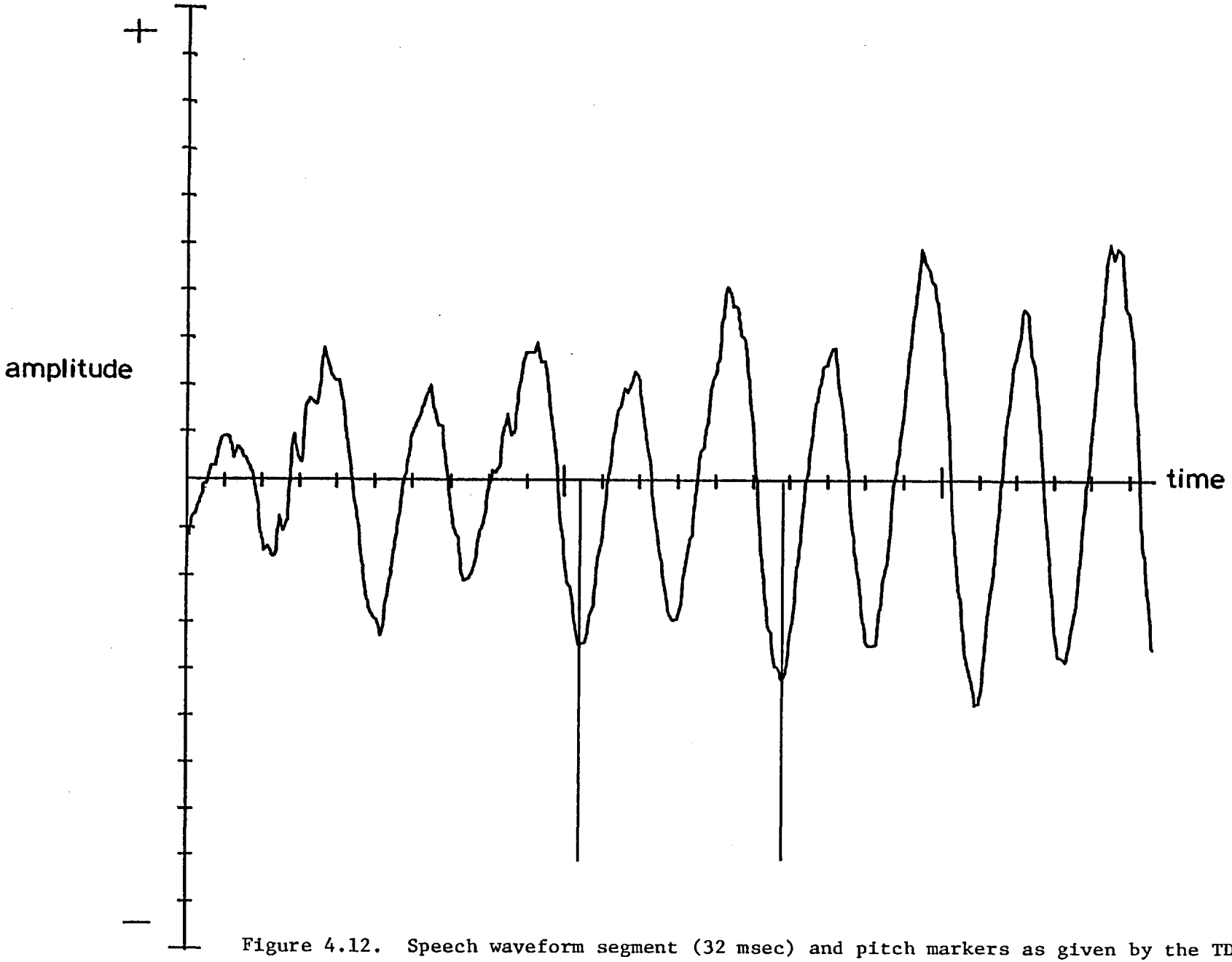


Figure 4.12. Speech waveform segment (32 msec) and pitch markers as given by the TDPD algorithm.

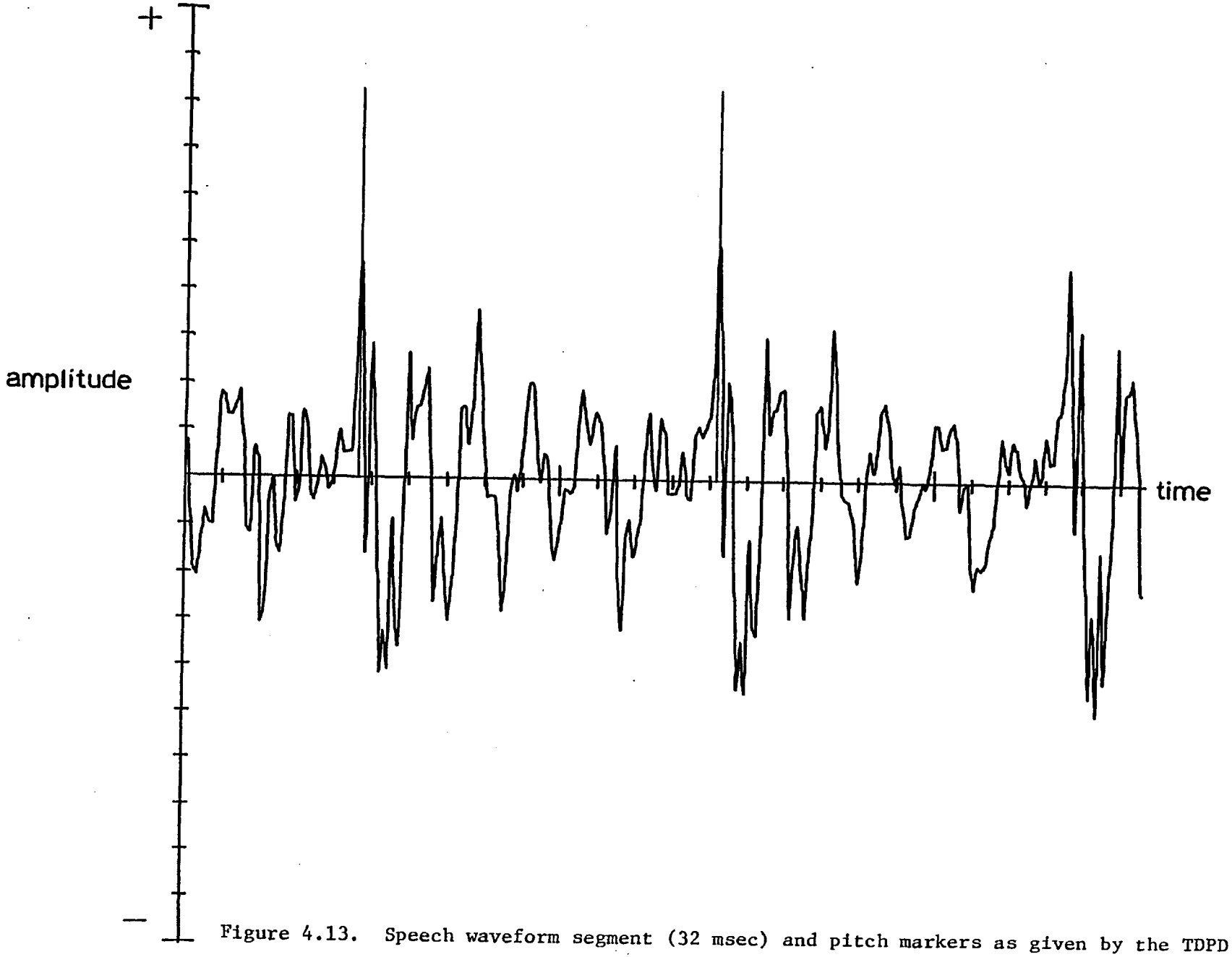


Figure 4.13. Speech waveform segment (32 msec) and pitch markers as given by the TDPD algorithm.

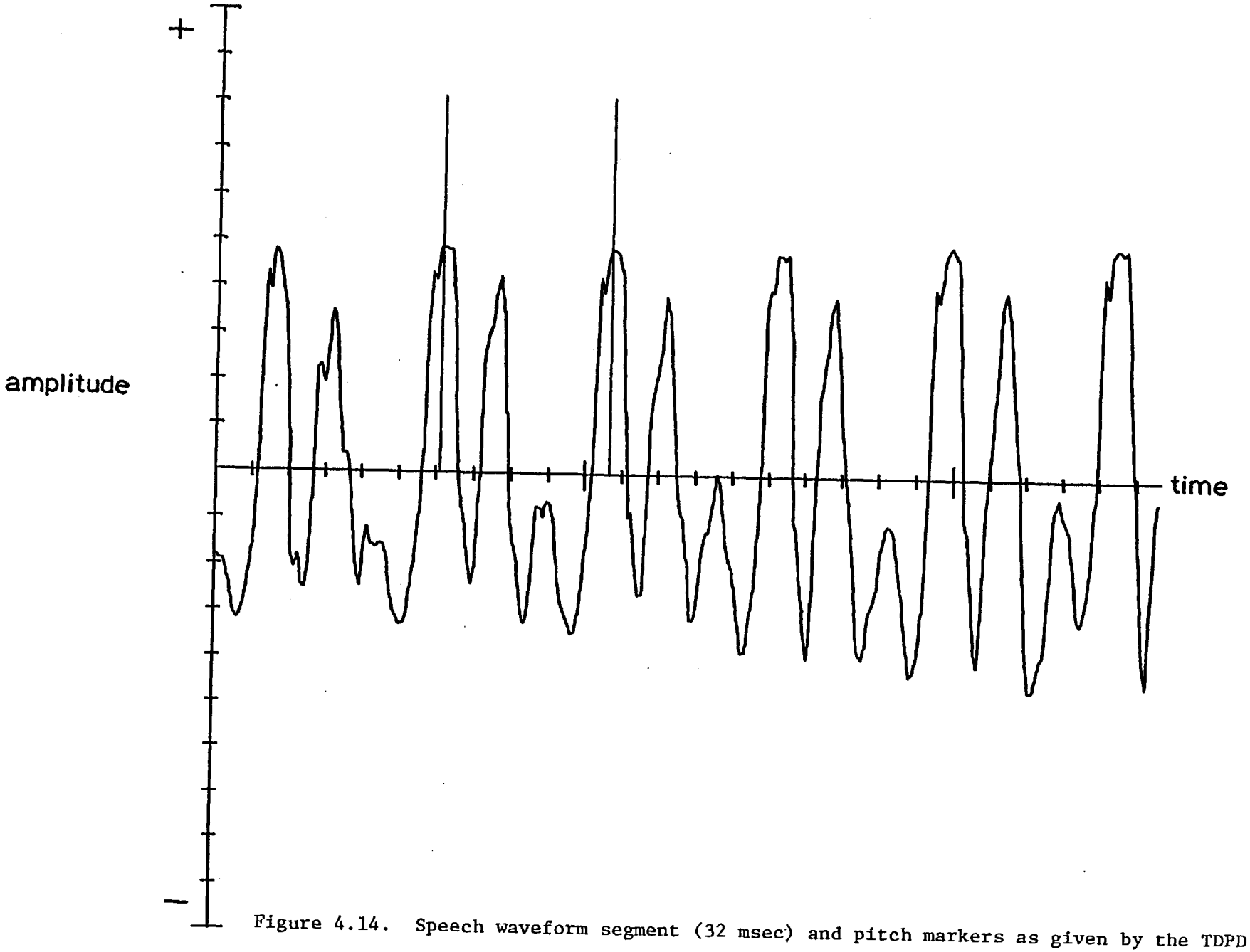


Figure 4.14. Speech waveform segment (32 msec) and pitch markers as given by the TDPD algorithm.

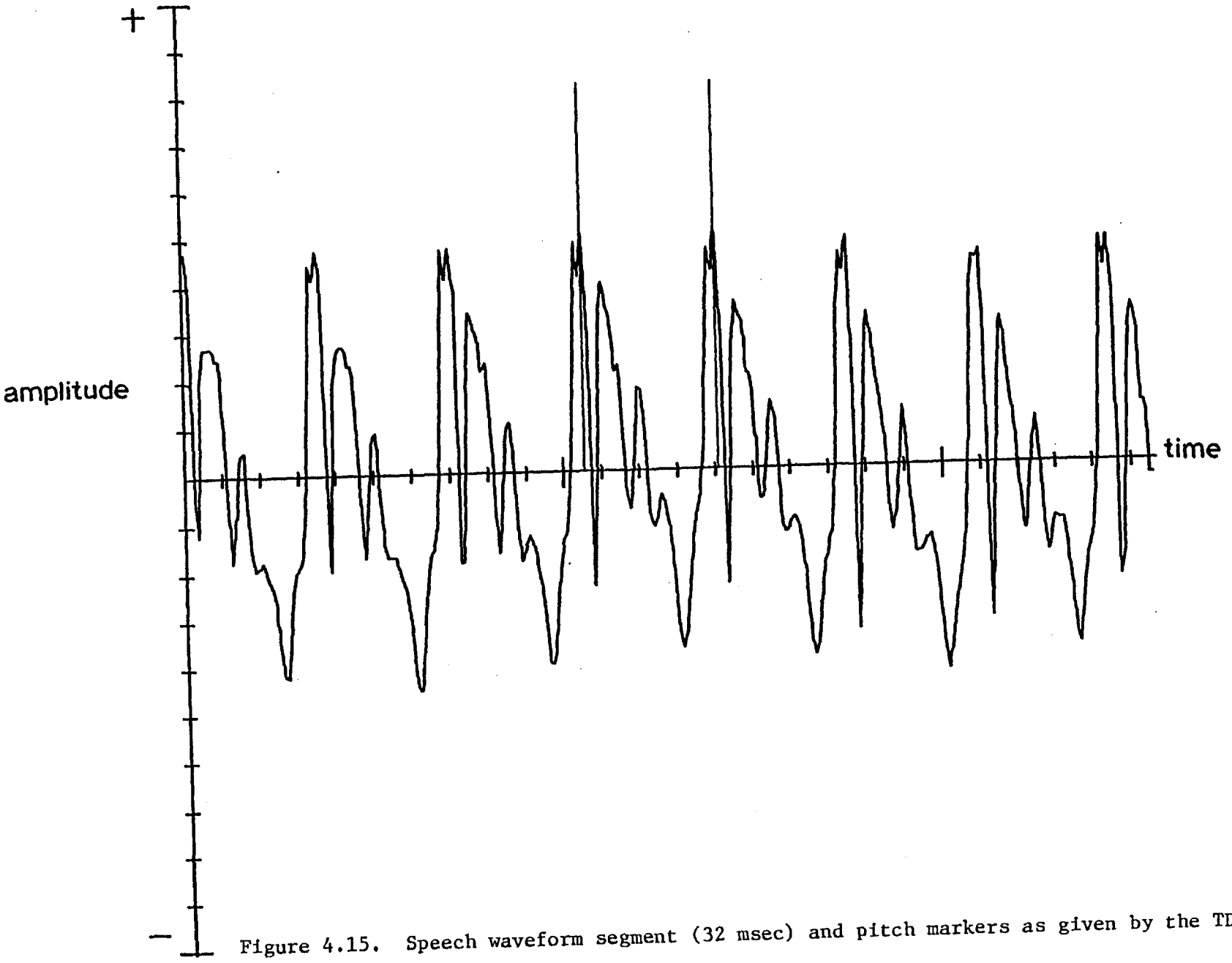


Figure 4.15. Speech waveform segment (32 msec) and pitch markers as given by the TDPD algorithm.

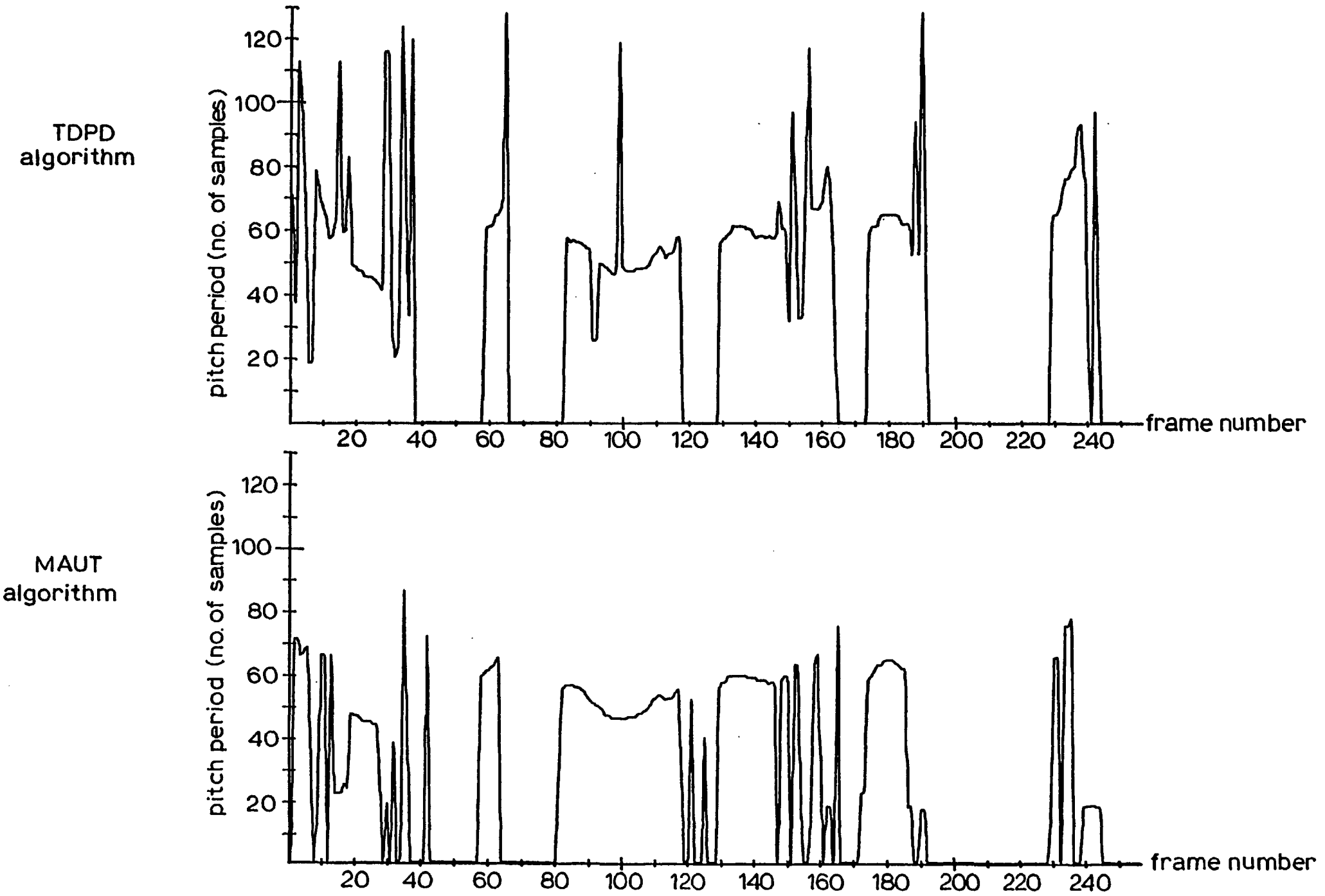


Figure 4.16. Pitch period contours for utterance number 1.

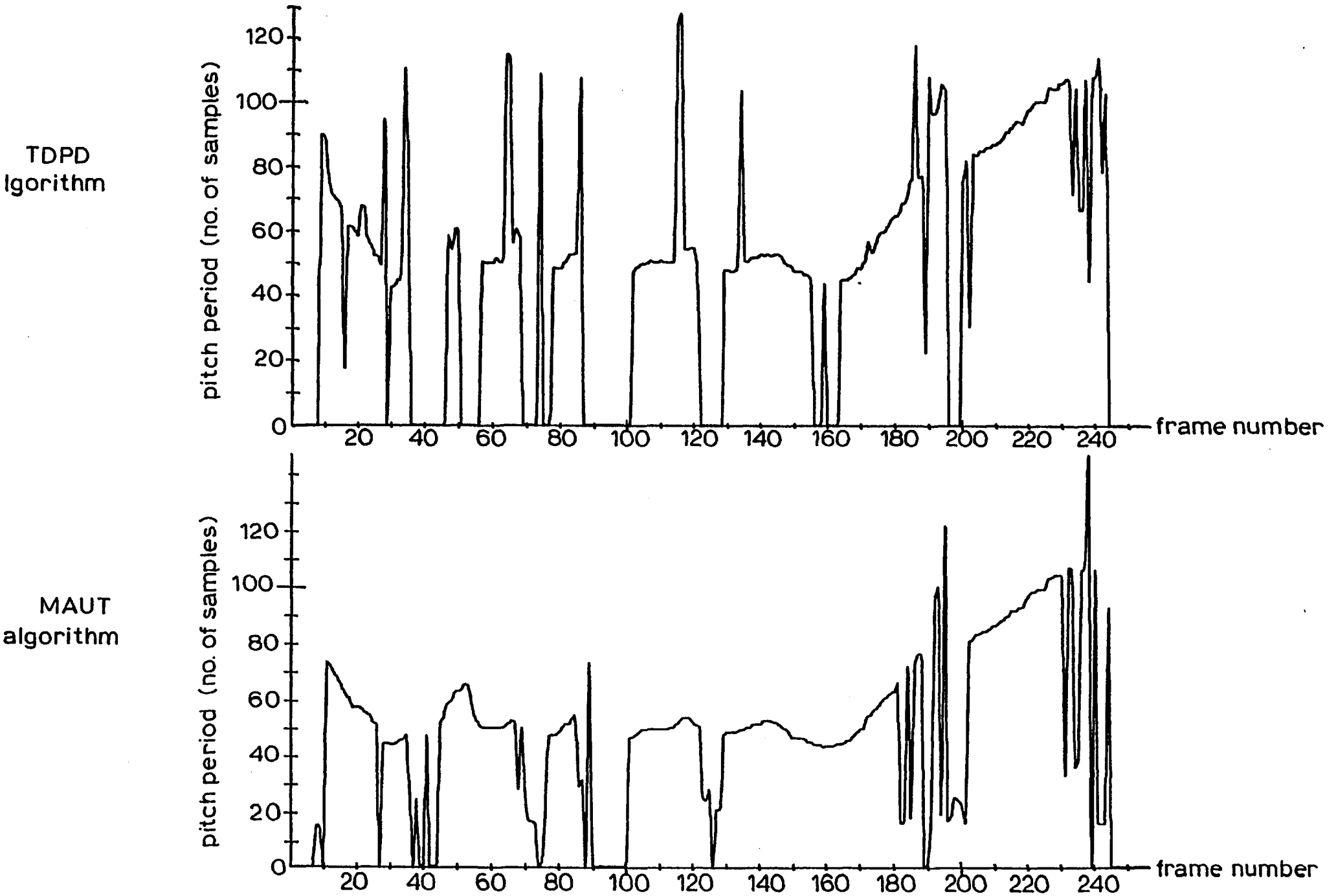


Figure 4.17. Pitch period contours for utterance number 5.

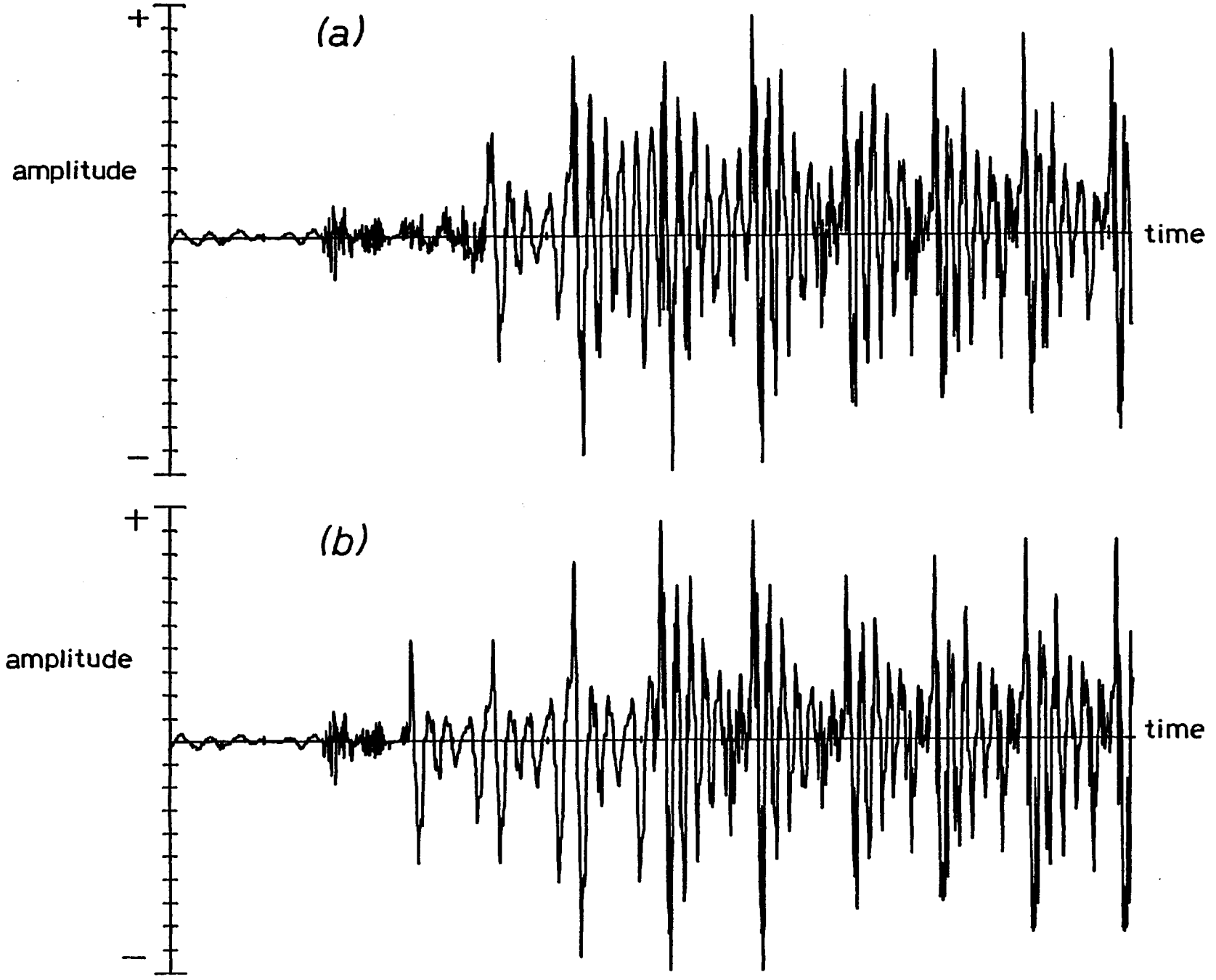


Figure 4.18. Experiment I: (a) original waveform; (b) reconstructed waveform. Time interval = 128 msec.

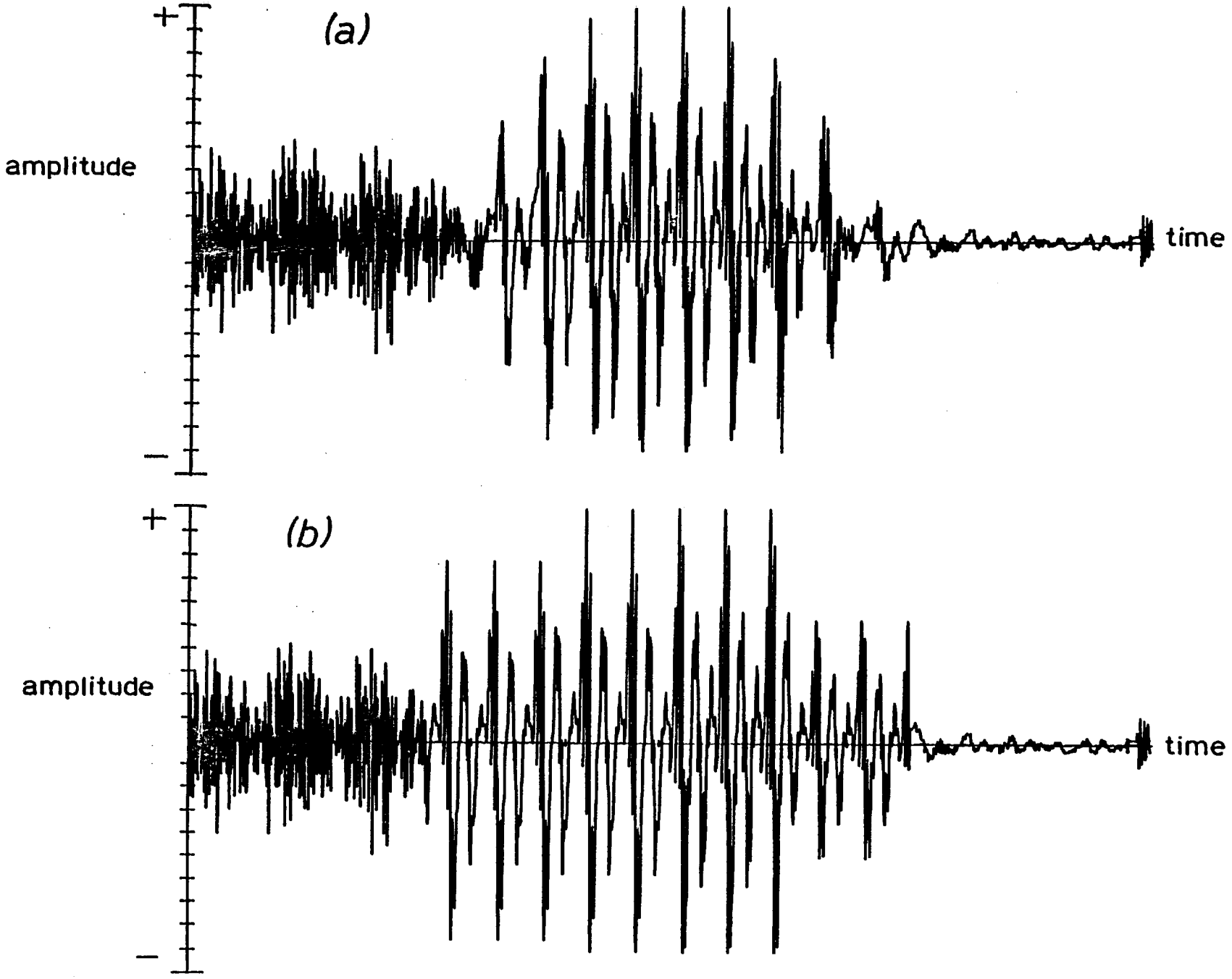


Figure 4.19. Experiment I: (a) original waveform; (b) reconstructed waveform. Time interval = 128 msec.

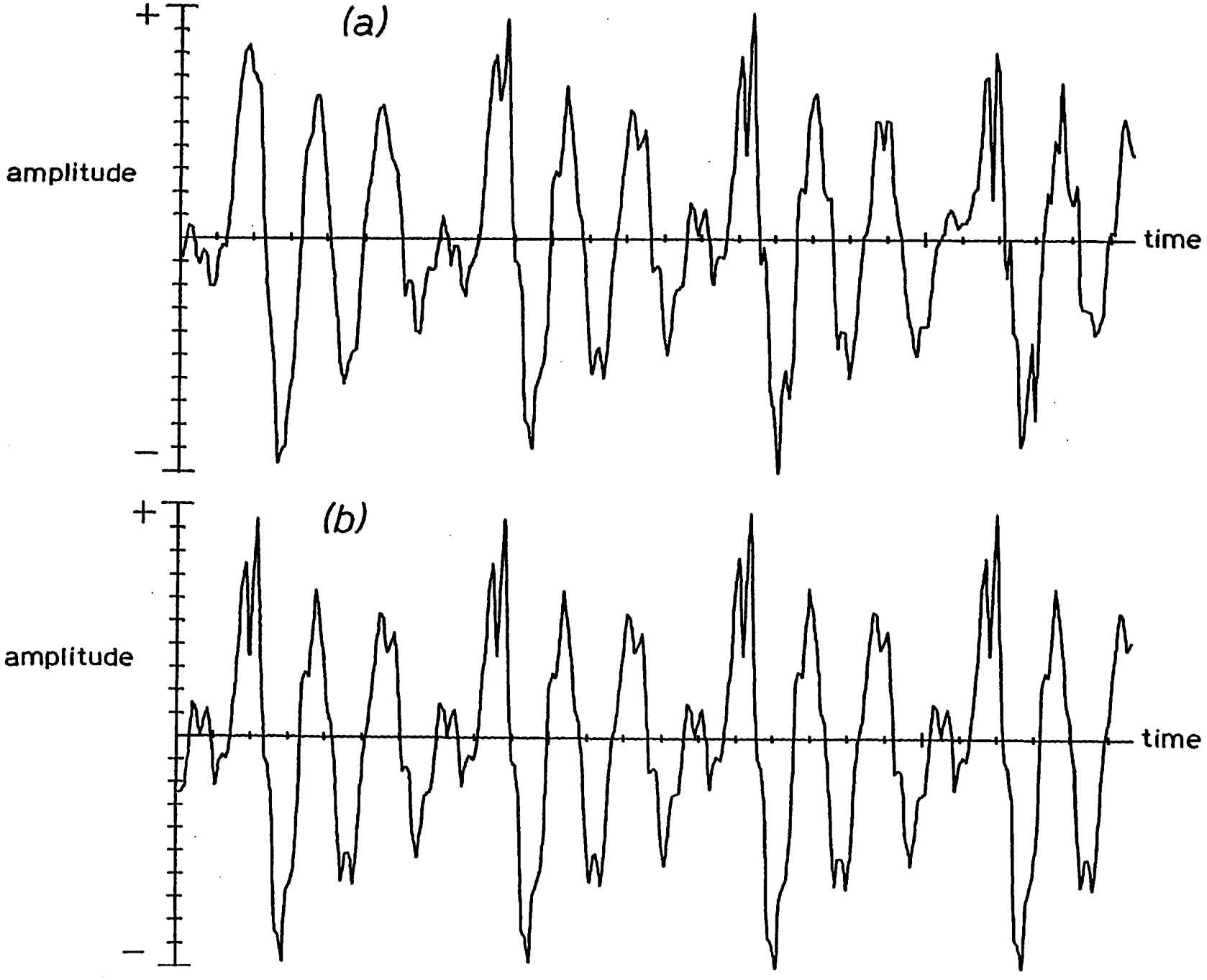


Figure 4.20. Experiment I: (a) original waveform; (b) reconstructed waveform. Time interval = 32 msec.

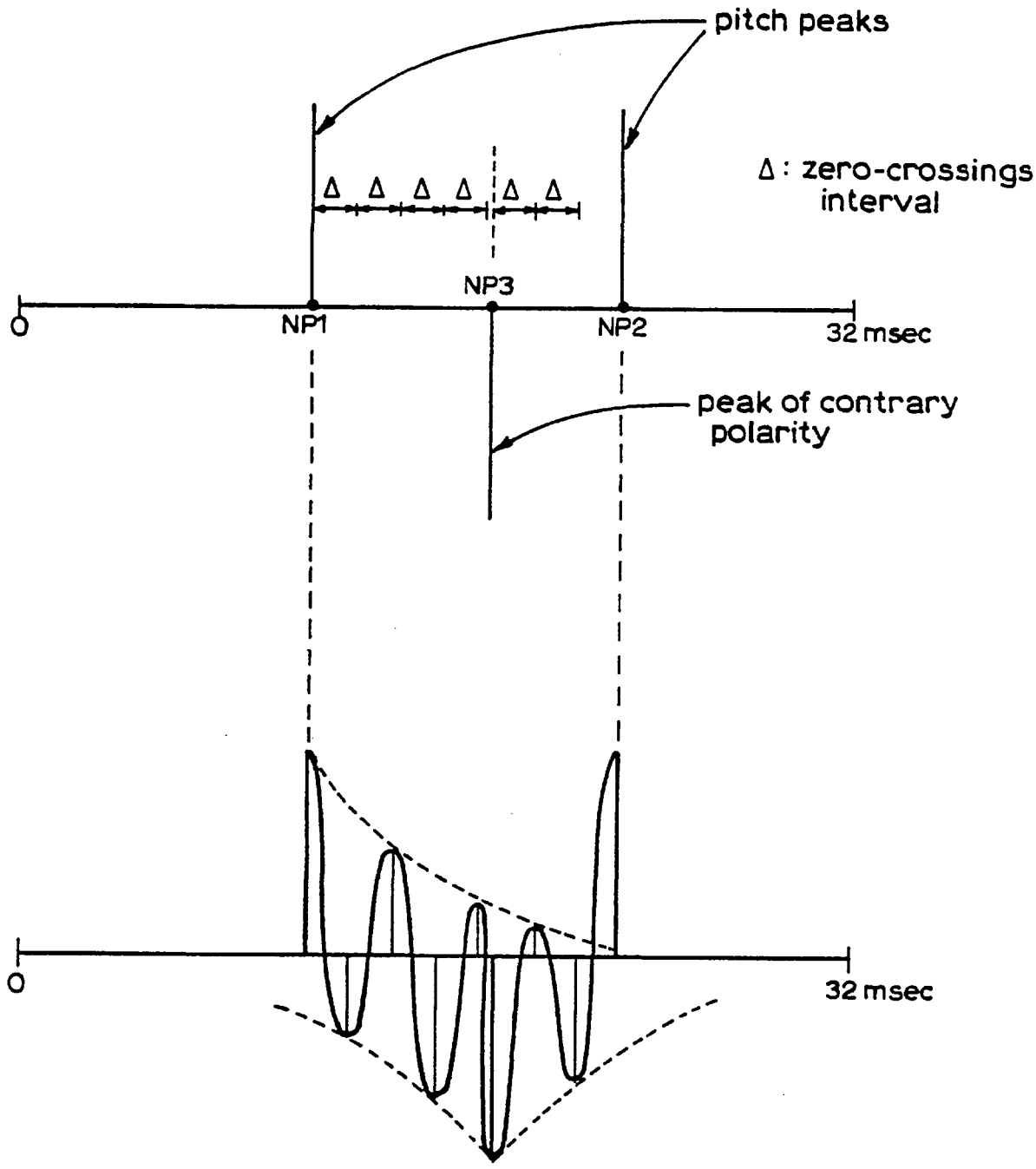


Figure 4.21. Experiment II: illustration of intrapitch waveform reconstruction.

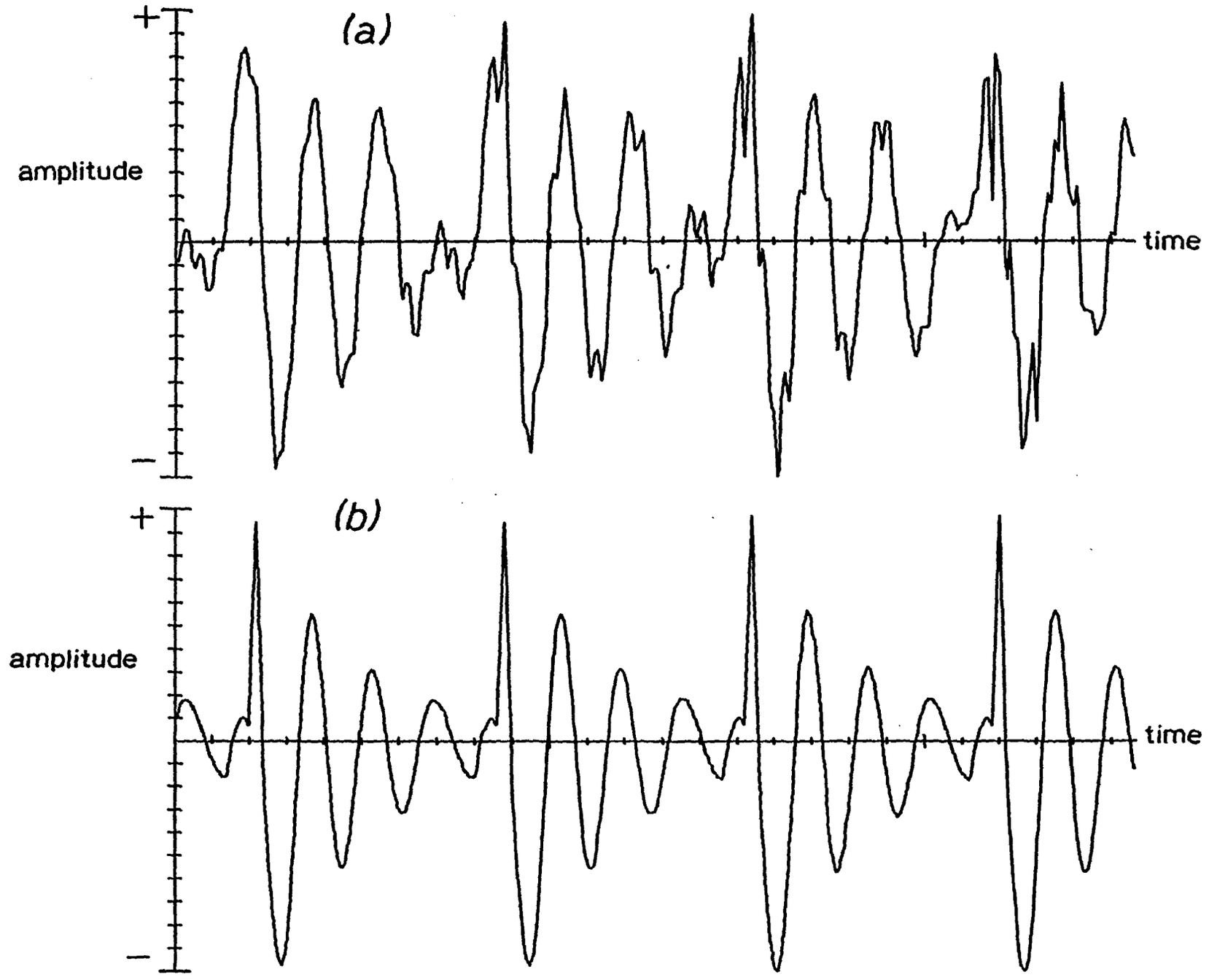


Figure 4.22. Experiment III: (a) original waveform; (b) reconstructed waveform. Time interval = 32 msec.

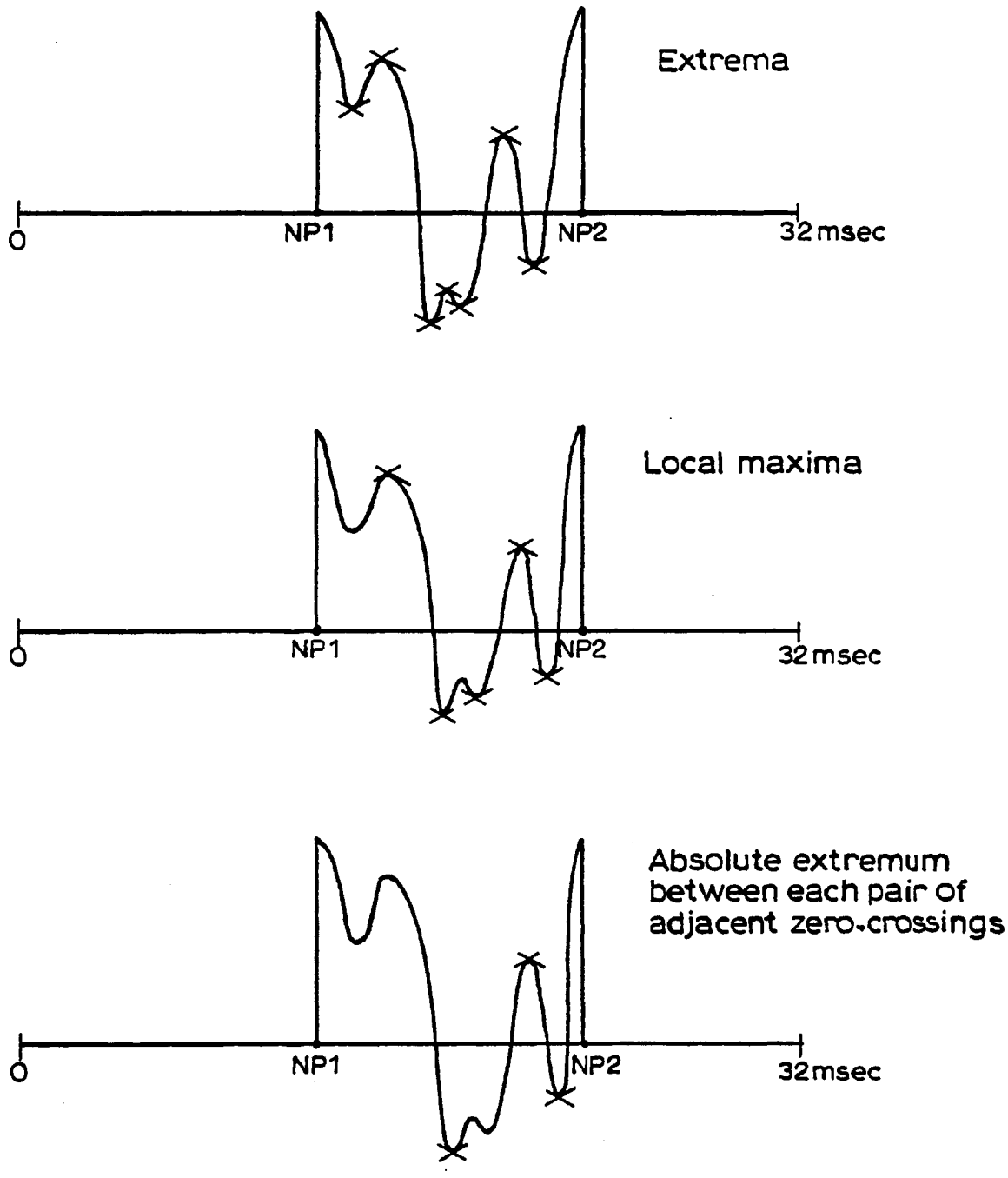


Figure 4.23. Experiment IV: illustration of features extracted from the waveform intrapitch.

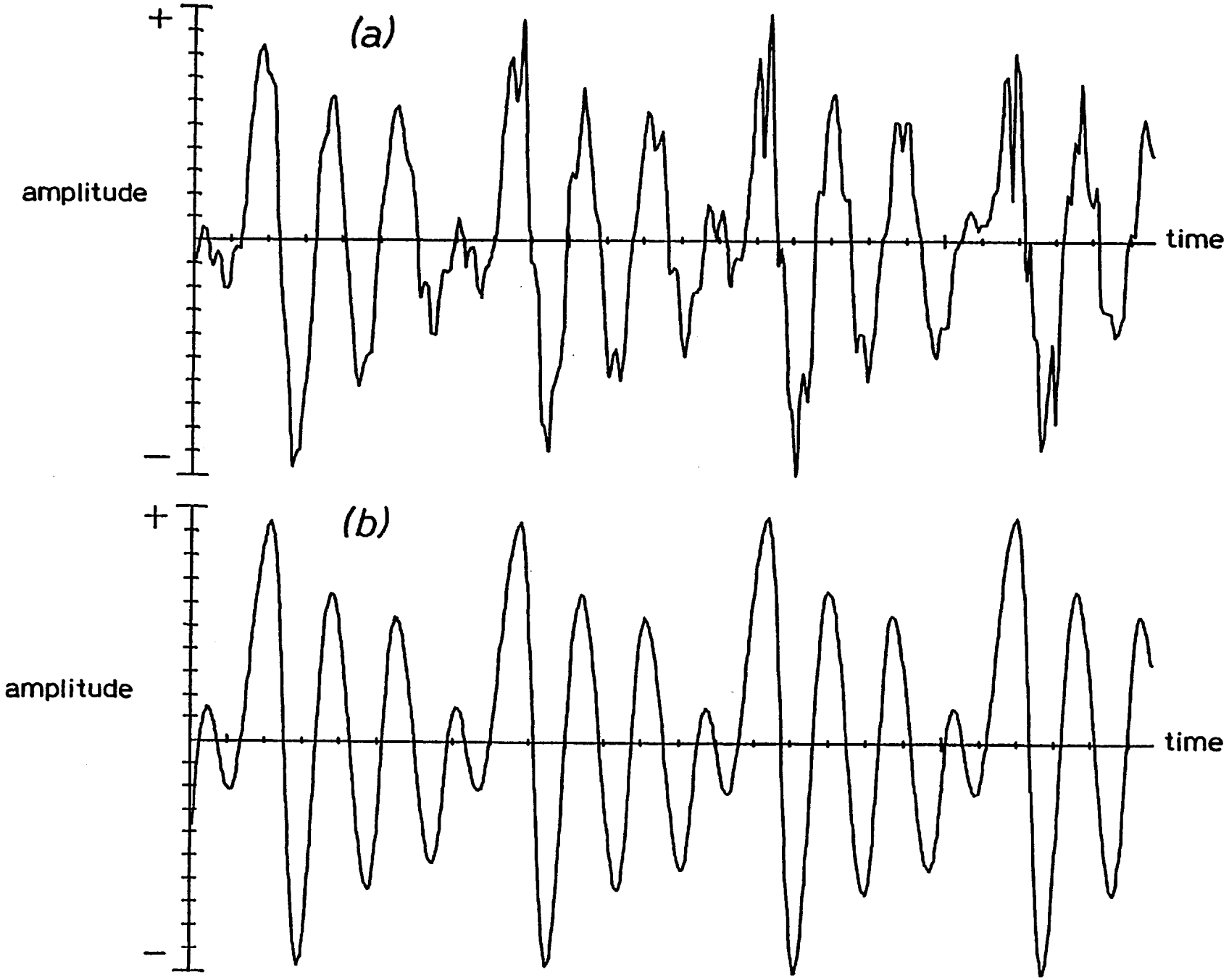


Figure 4.24. Experiment IV: (a) original, and (b) processed waveform using only the absolute extreme between zero-crossings.

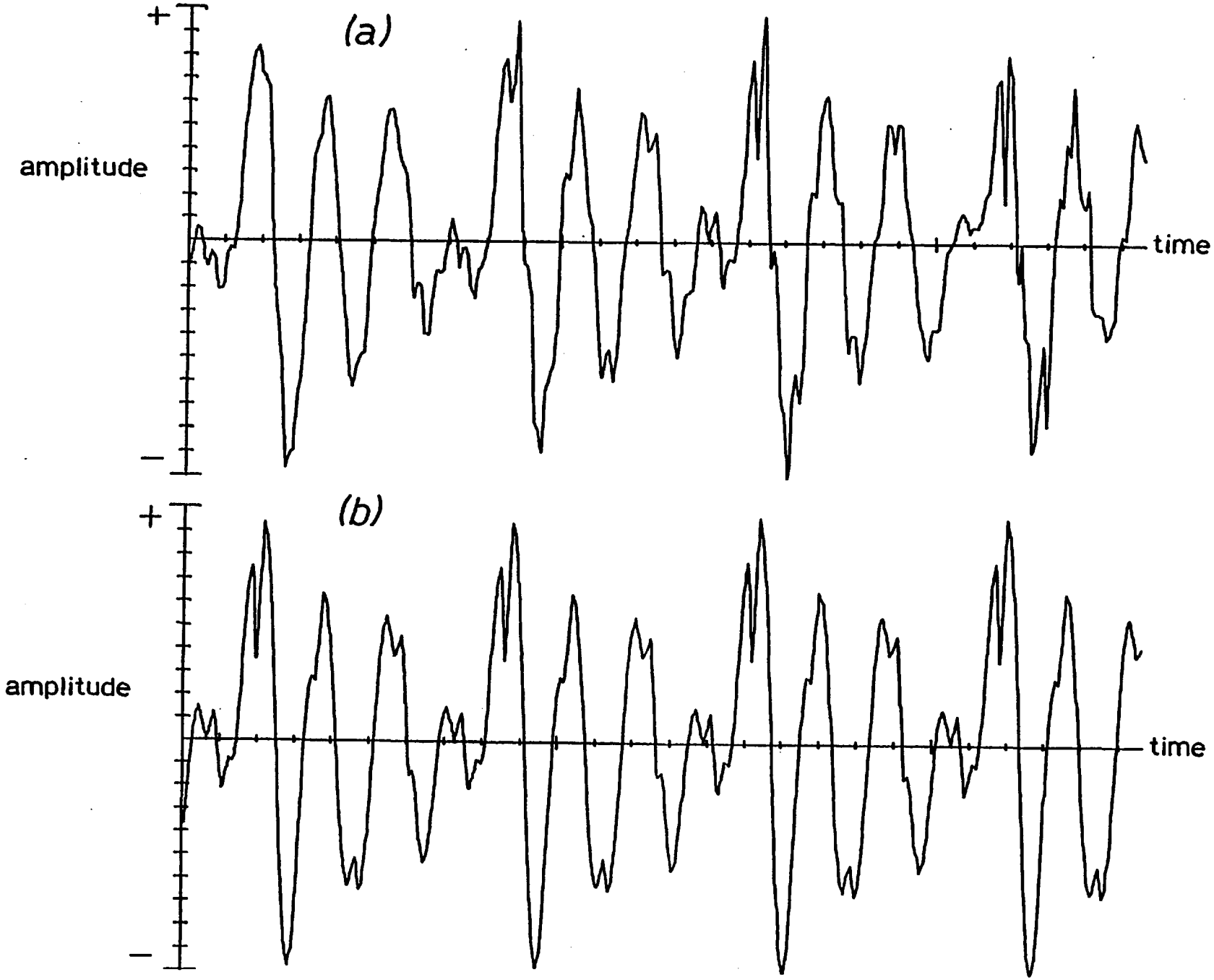


Figure 4.25. Experiment IV: (a) original, and (b) processed waveform using all peaks and valleys.

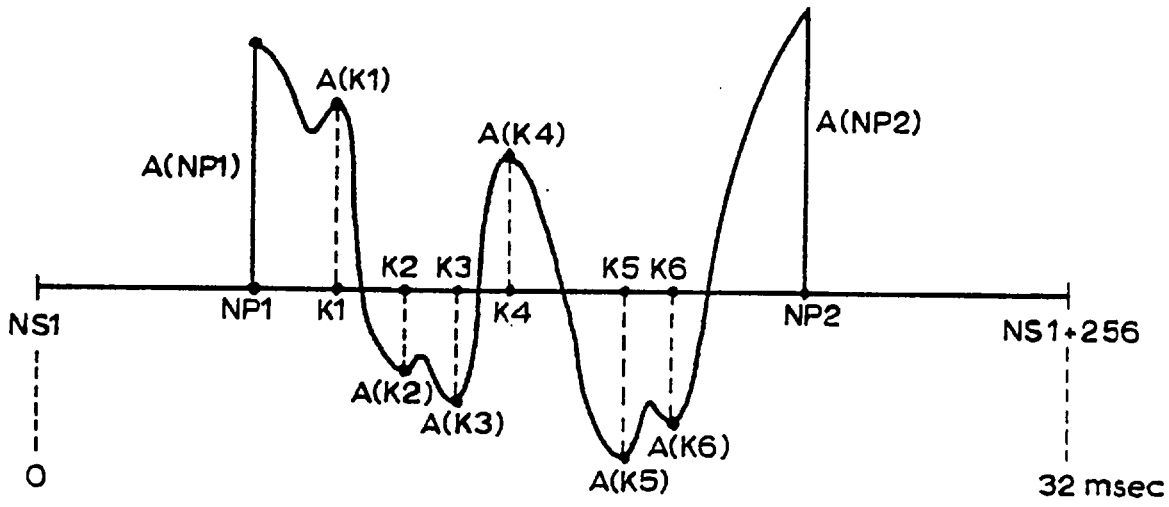


Figure 4.26. Experiment V: extraction of intrapitch parameters.

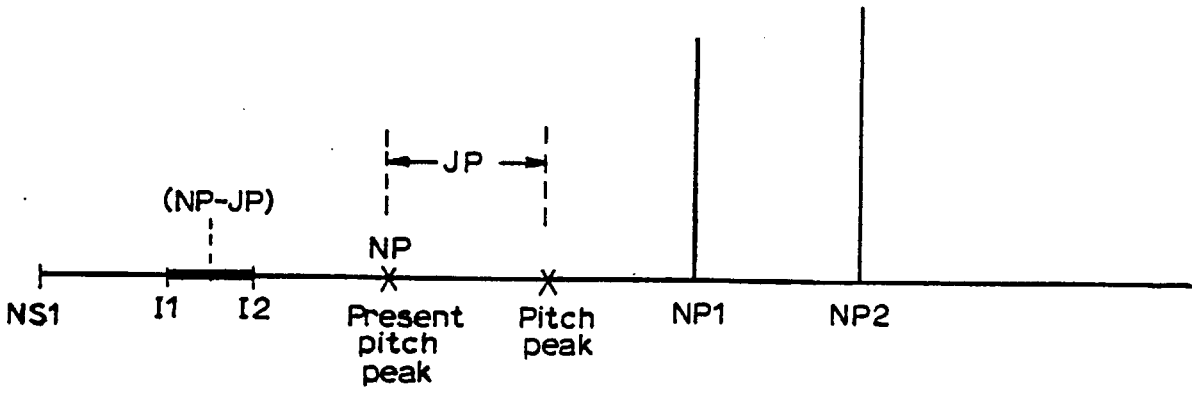


Figure 4.27. Experiment V: extraction of other pitch peaks.

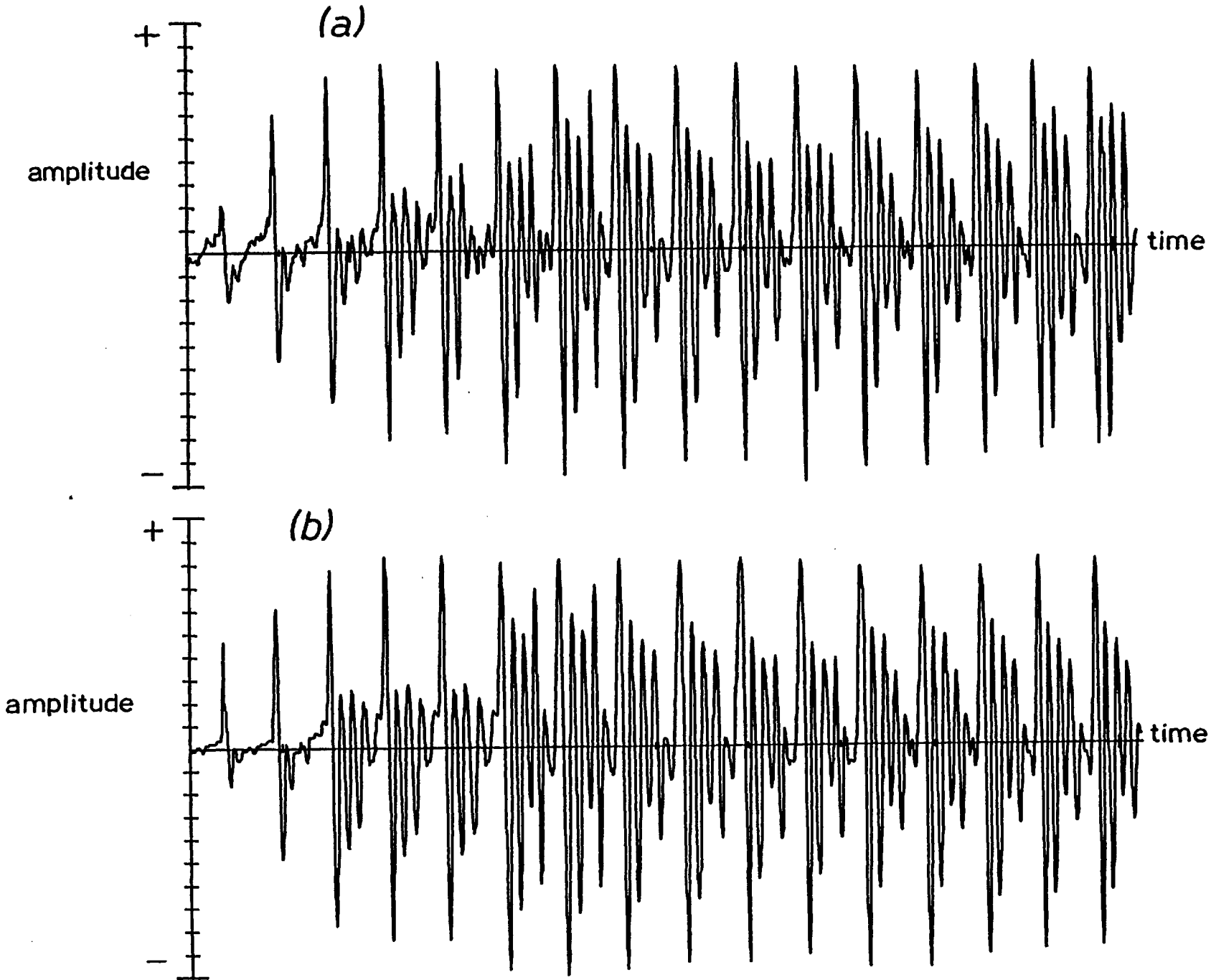


Figure 4.28. Experiment V: (a) original, and (b) processed waveform (NUP = 128 samples). Time interval = 128 msec.

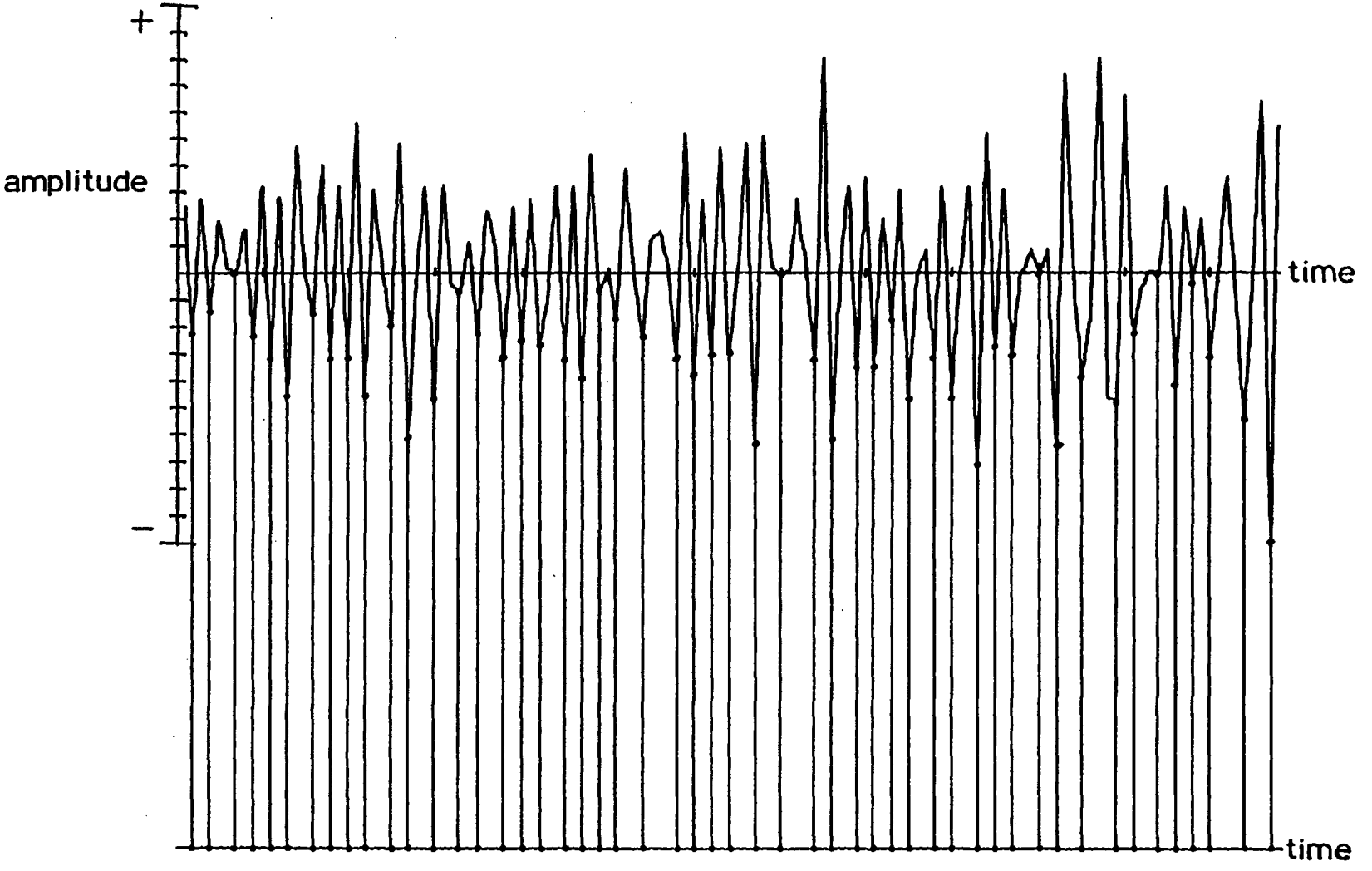


Figure 4.29. Time occurrence of alternate extremes of a 16 msec segment of unvoiced speech waveform.

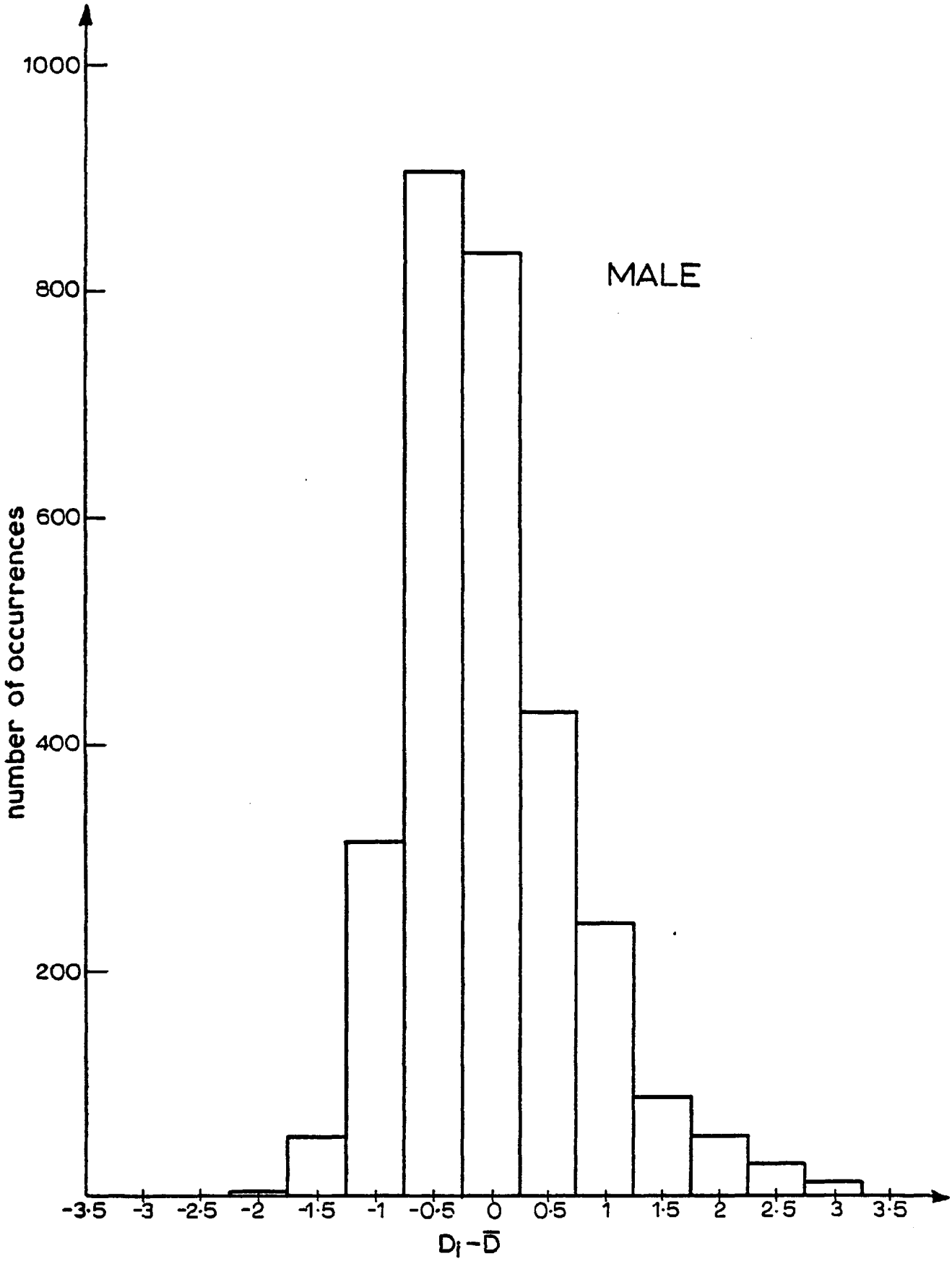


Figure 4.30. Histogram relating $D_i - \bar{D}$, where D_i is the time interval between alternate unvoiced waveform extremes and $\bar{D}$ is the average over each 32 msec segment (male speech).

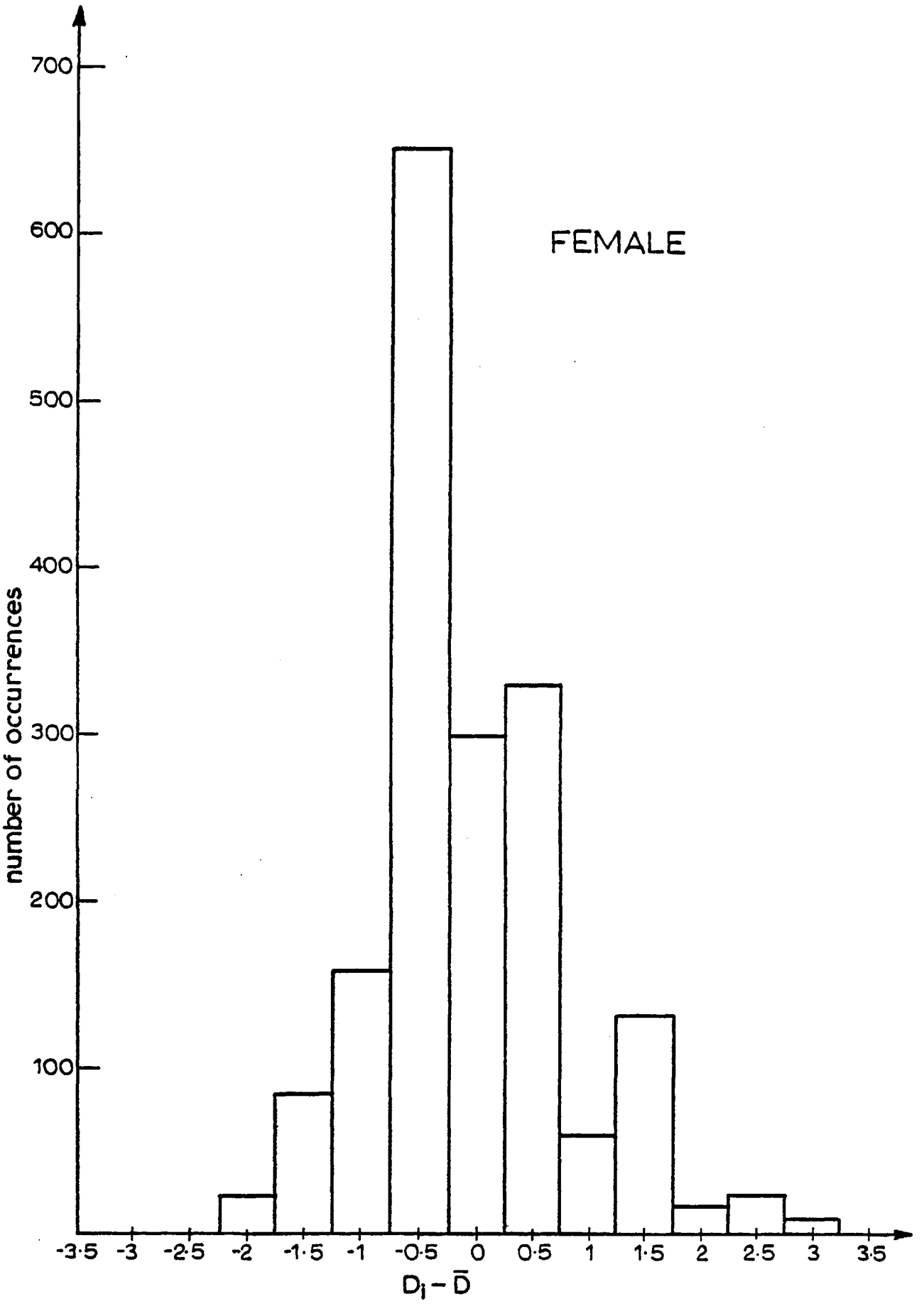


Figure 4.31. Histogram relating $D_i - \bar{D}$, where D_i is the time interval between alternate unvoiced waveform extremes and $\bar{D}$ is the average over each 32 msec segment (female speech).

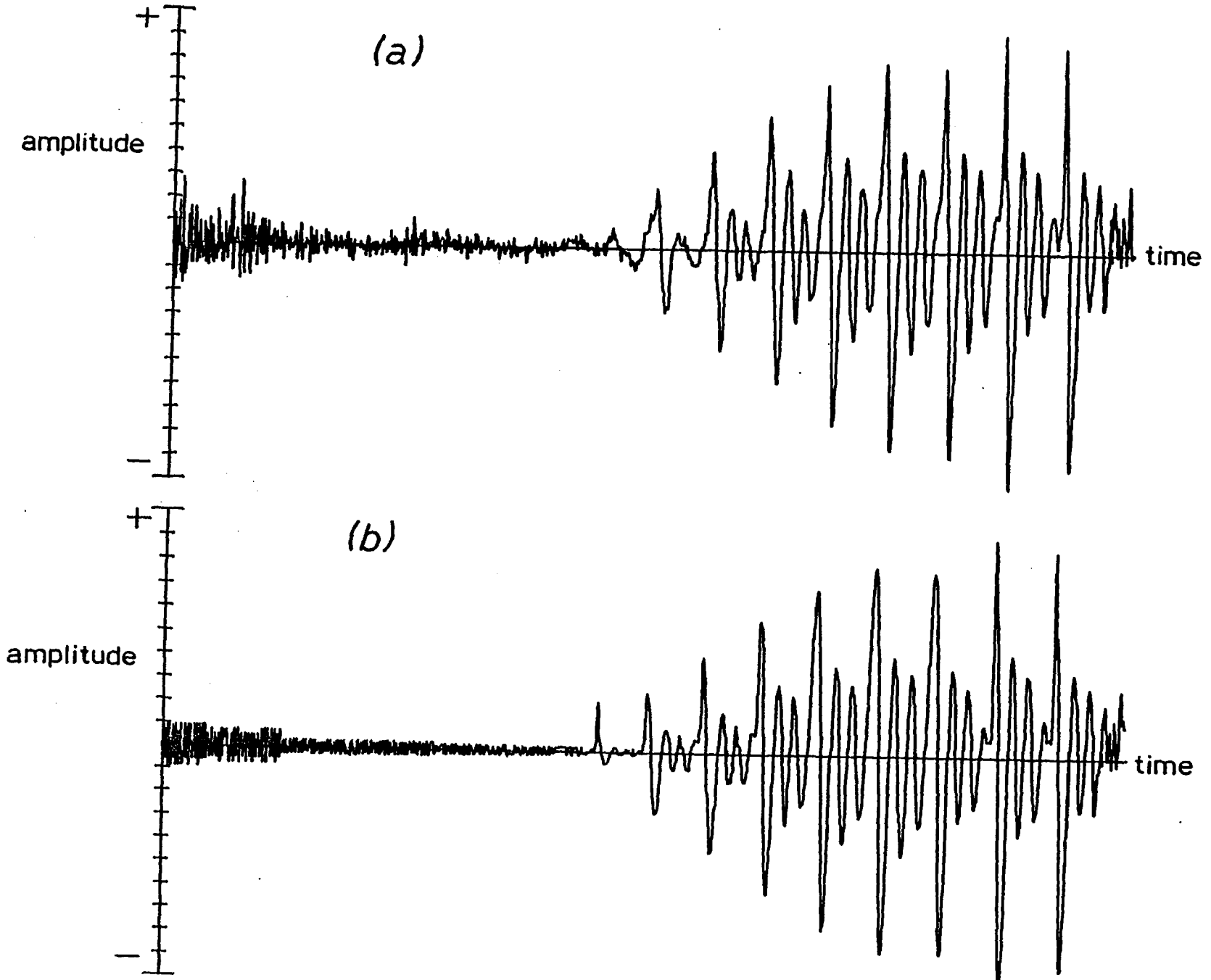


Figure 4.32. Time coding technique: (a) original, and (b) processed waveform. Time interval = 128 msec.

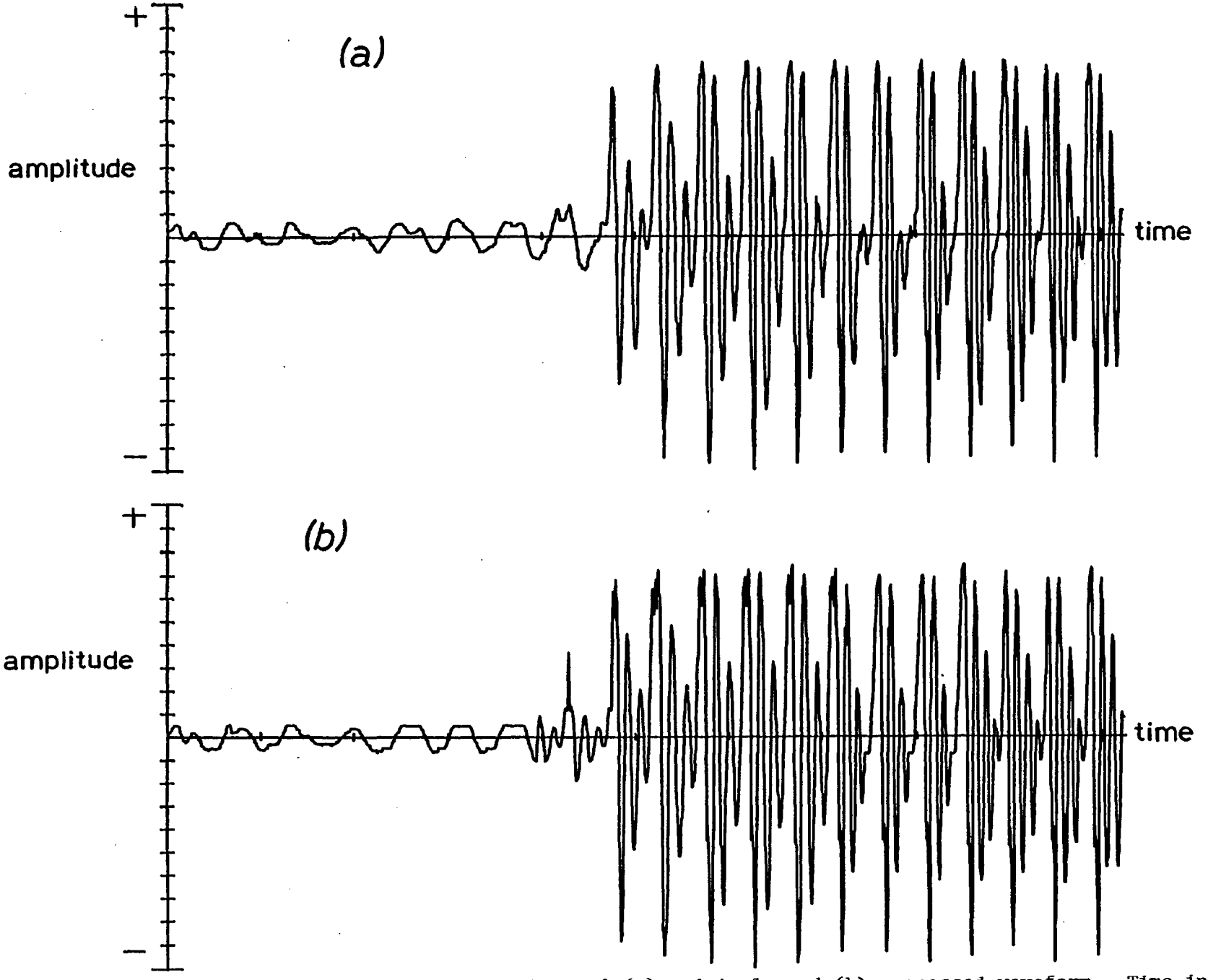


Figure 4.33. Time coding technique; (a) original, and (b) processed waveform. Time interval = 128 msec.

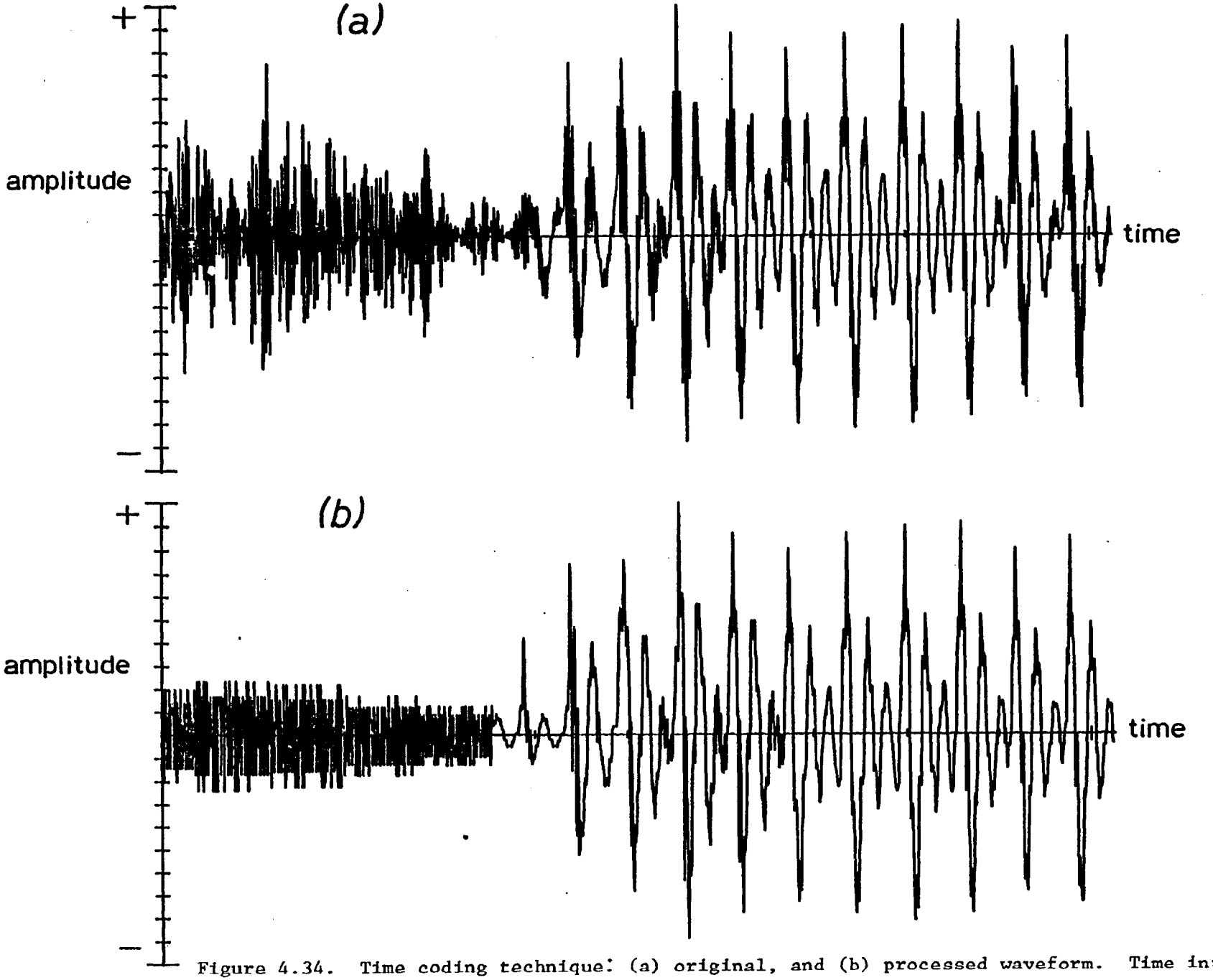


Figure 4.34. Time coding technique: (a) original, and (b) processed waveform. Time interval = 128 msec.

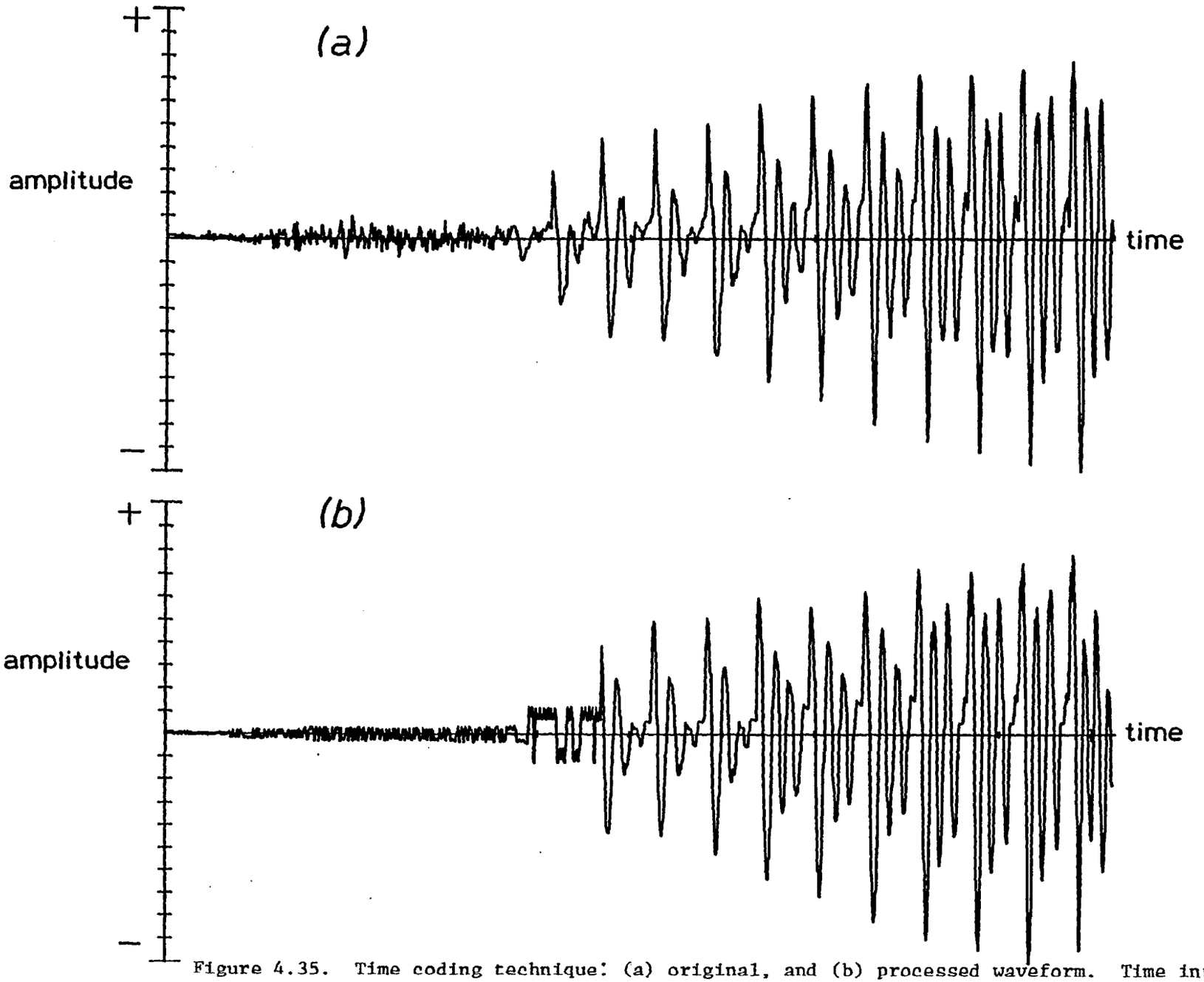


Figure 4.35. Time coding technique: (a) original, and (b) processed waveform. Time interval = 128 msec.

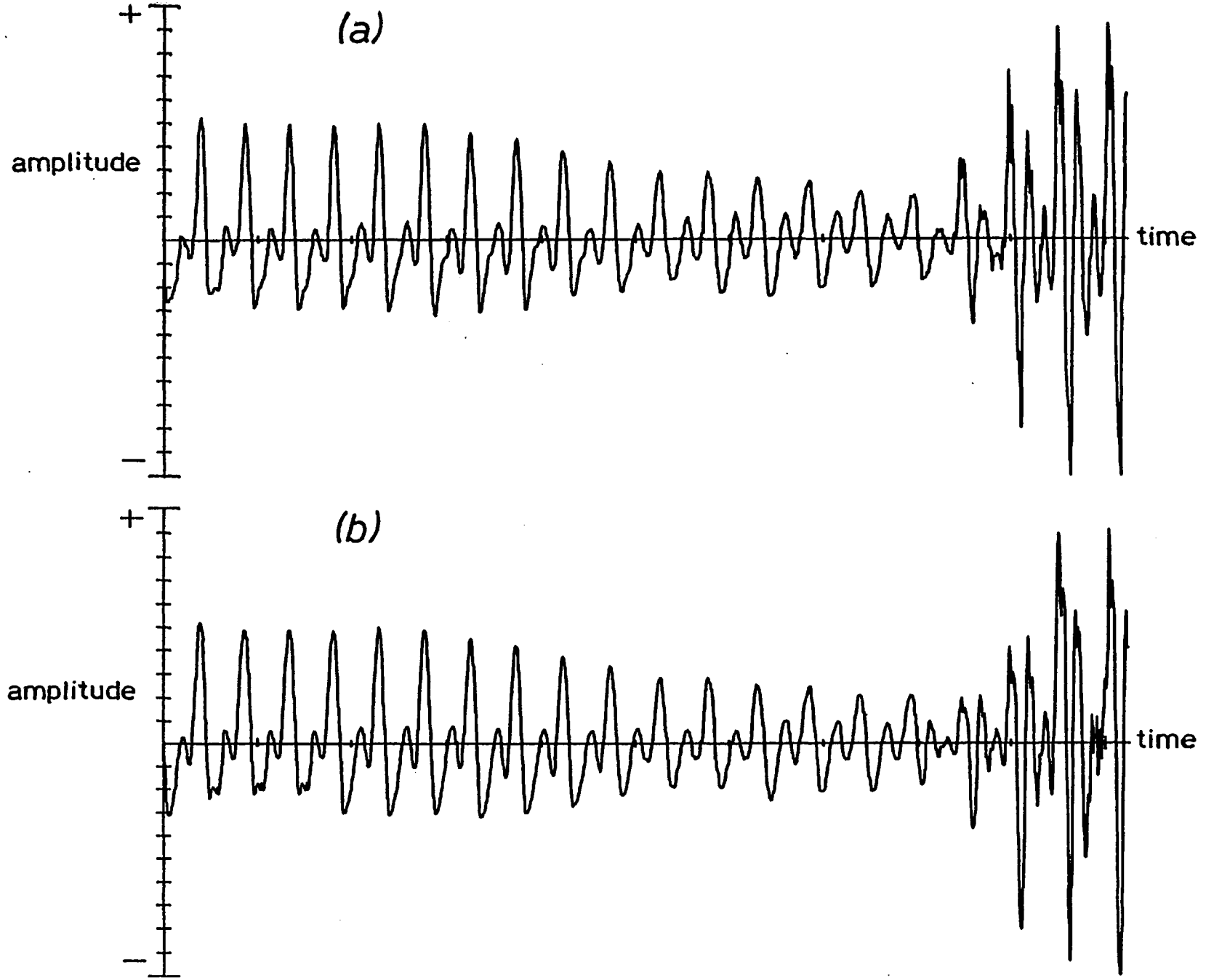


Figure 4.36. Time coding technique: (a) original, and (b) processed waveform. Time interval = 128 msec.

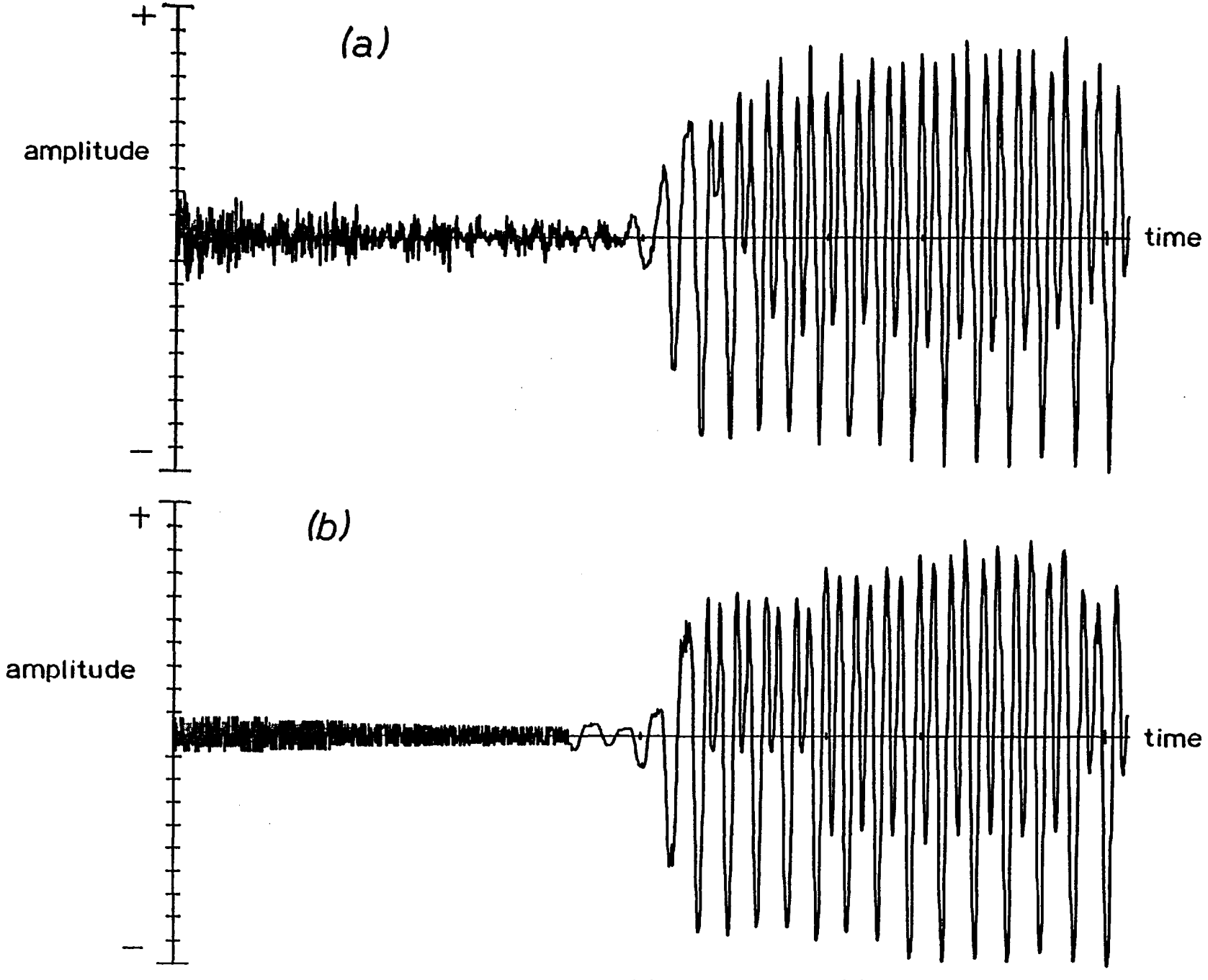


Figure 4.37. Time coding technique: (a) original, and (b) processed waveform. Time interval = 128 msec.

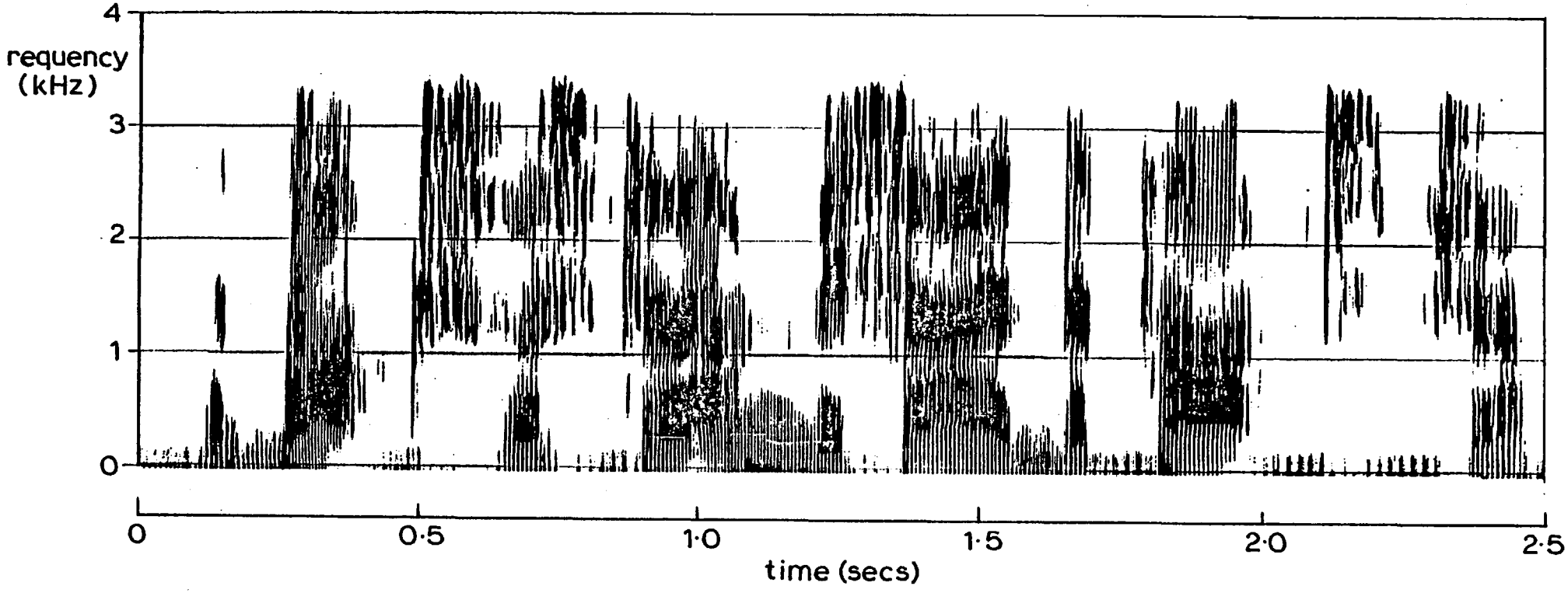


Figure 4.38. Spectrogram of original speech for the utterance " The box was thrown beside the parked tru(ck)" spoken by a male speaker.

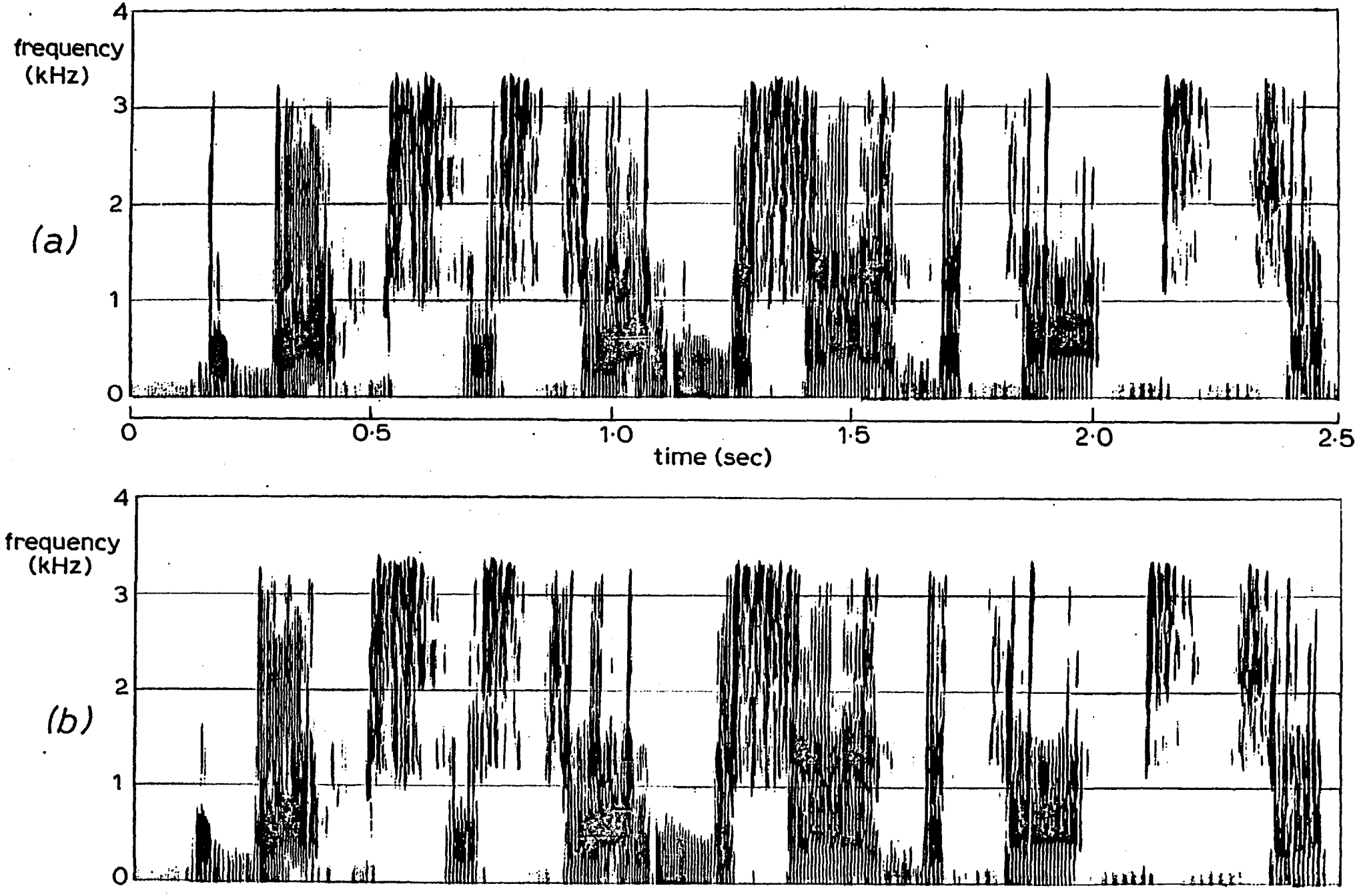


Figure 4.39. Spectrograms of processed speech, (a) unquantised and (b) quantised, for the utterance " The box was thrown beside the parked tru(ck)" spoken by a male speaker.

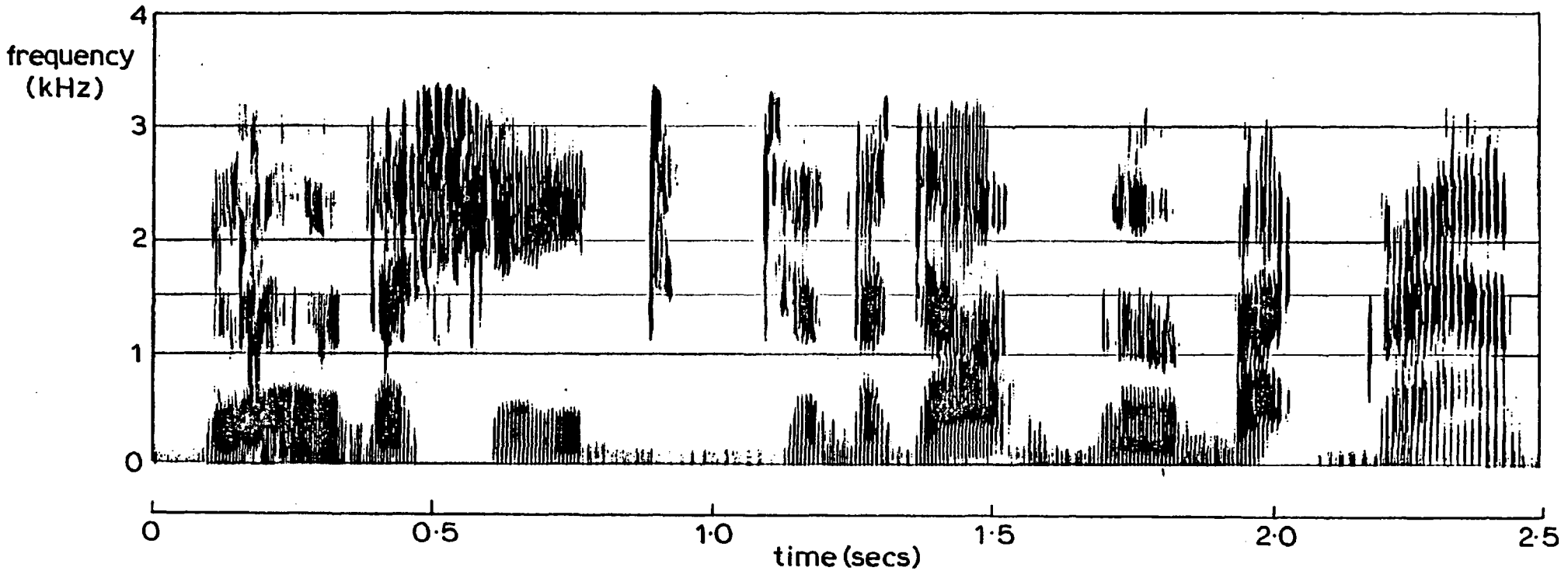


Figure 4.40. Spectrogram of original speech for the utterance " Glue the sheet to the dark blue backgroun(d)" spoken by a male speaker.

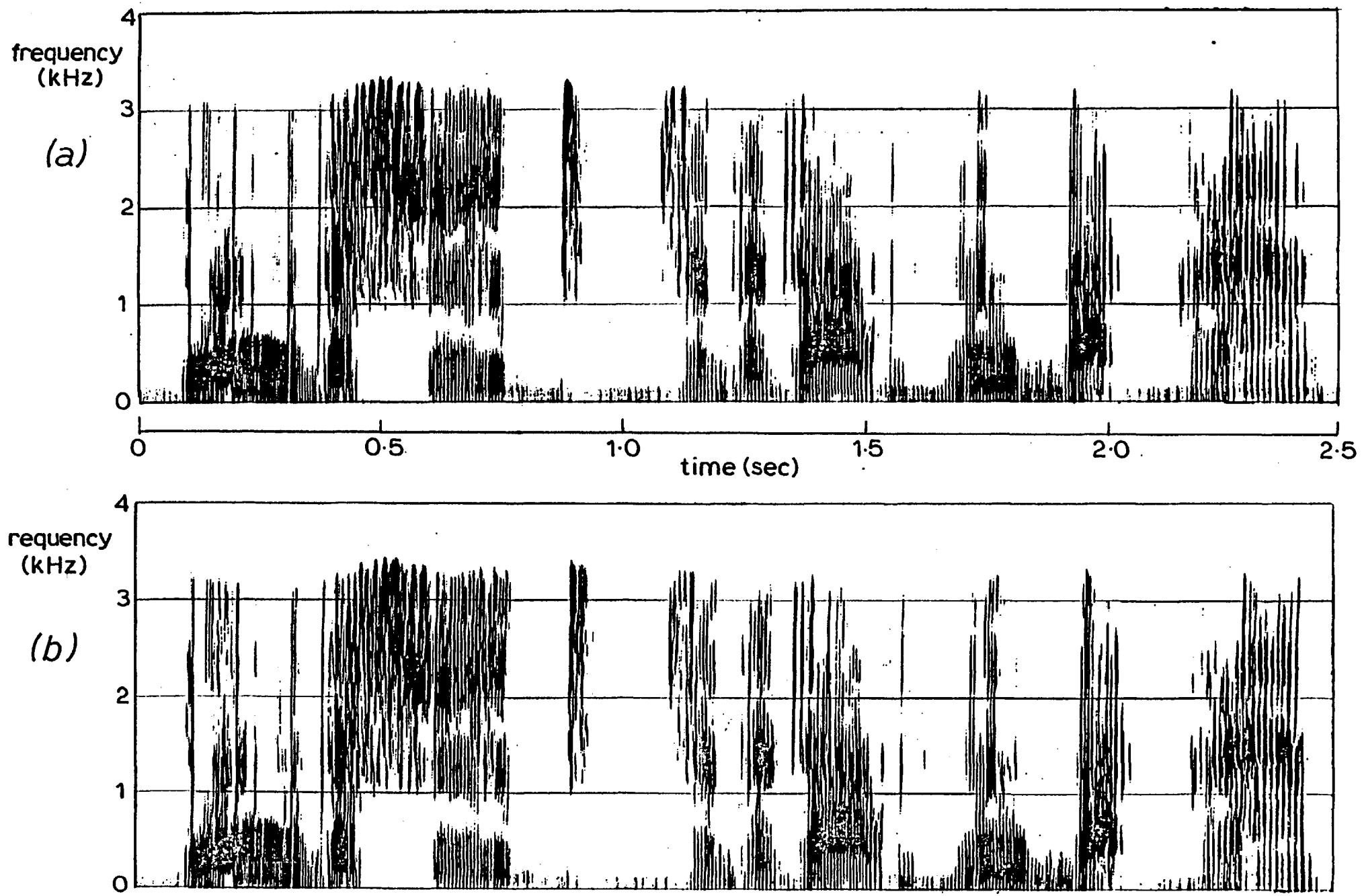


Figure 4.41. Spectrograms of processed speech, (a) unquantised and (b) quantised, for the utterance " Glue the sheet to the dark blue backgroun(d)" spoken by a male speaker.

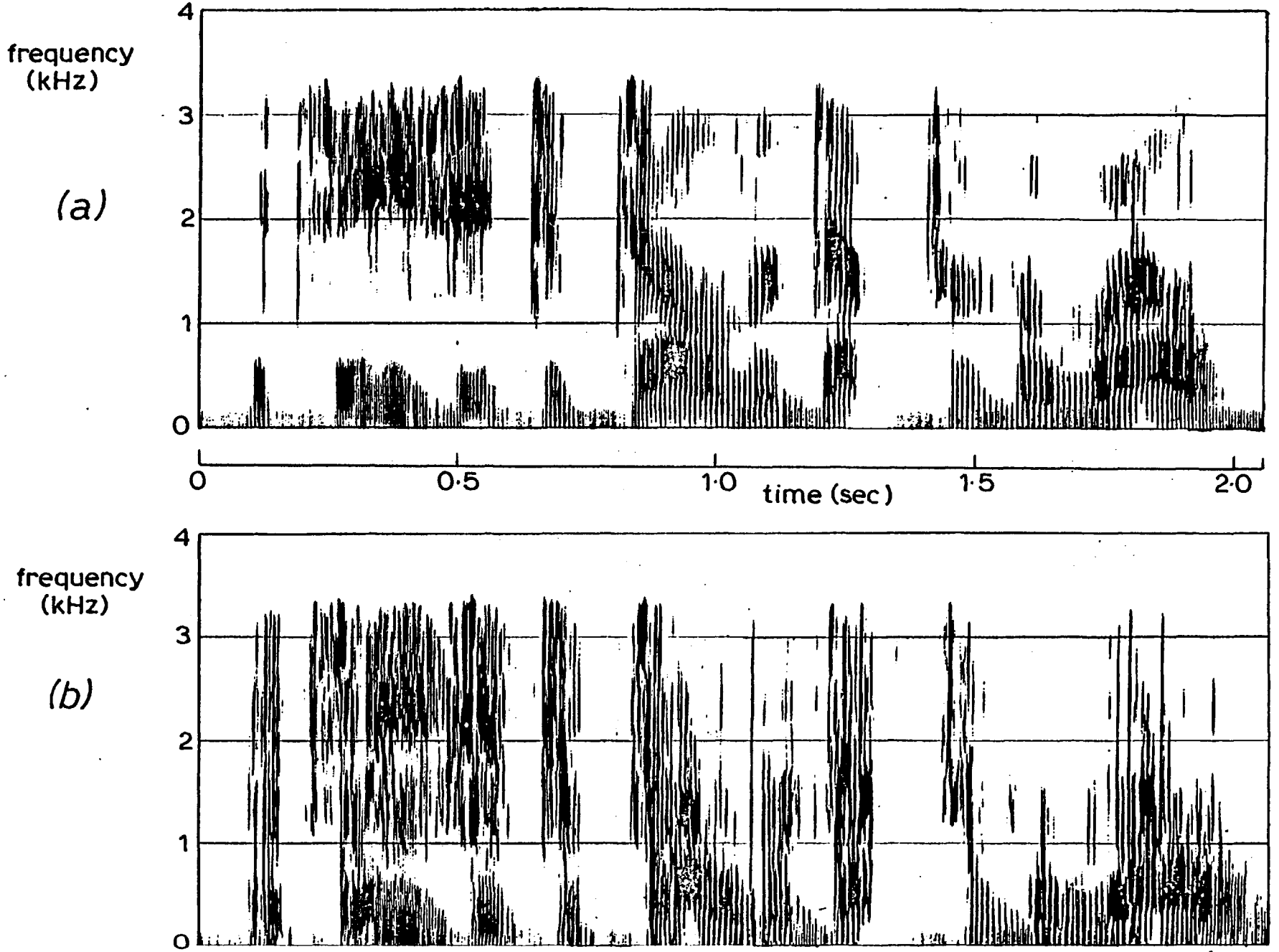


Figure 4.42. Spectrogram of (a) original, and (b) processed (quantised) speech for the utterance " It's easy to tell the depth of the well" spoken by a male speaker.

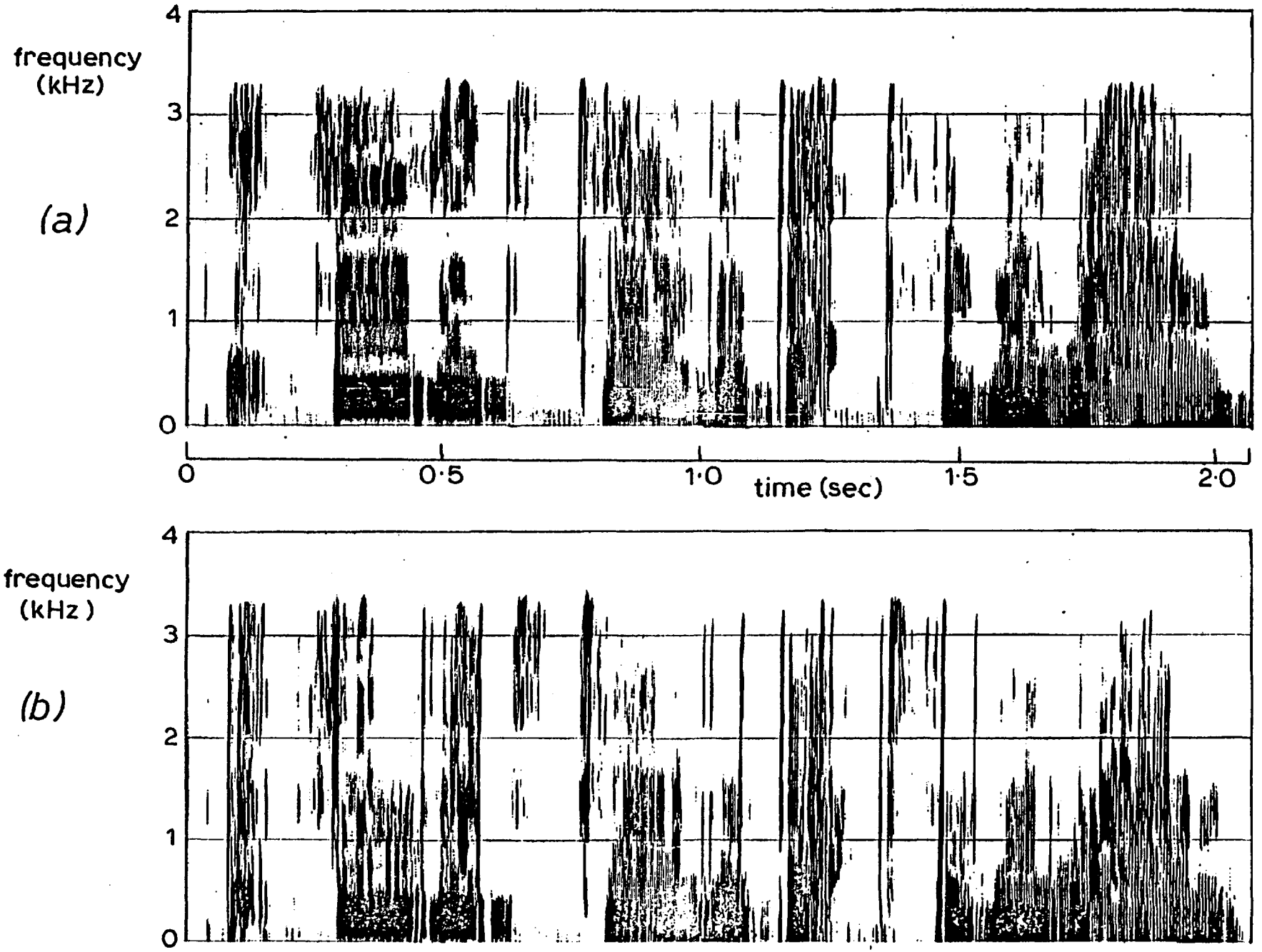


Figure 4.43. Spectrogram of (a) original, and (b) processed (quantised) speech for the utterance " It's easy to tell the depth of the well" spoken by a female speaker.

CHAPTER 5

DELAYS IN VARIABLE-INFORMATION-RATE TO CONSTANT-TRANSMISSION-RATE SYSTEMS

CHAPTER 5

DELAYS IN VARIABLE-INFORMATION-RATE TO CONSTANT-TRANSMISSION-RATE SYSTEMS

5.1. Introduction.

In recent years a number of schemes of data compression have been suggested in which advantage is taken of the fact that the average rate at which data is produced is often less than the higher rates at which the source sometimes produces its information.[14, 18-20]

In order to smooth the variable data rate so that it can be transmitted using a constant-rate transmission facility, it is necessary to provide buffer storage at the transmitter and/or the receiver. An unavoidable and obvious consequence of this buffering is that storage has to be provided and delays are introduced into the communication process. A detailed discussion of this problem is the subject of the present chapter, which has been stimulated mainly by the recently developed suggestion on time coding of speech[18] and the low-bit-rate speech coding technique proposed in Chapter 4 of this thesis.

The organisation of this chapter is as follows. In section 5.2 some theoretical aspects relating to the buffering of variable-rate sources are considered, and expressions are derived for the minimum delay required at the transmitter and the receiver in order to avoid distortion. The delay involved when a symbol-injection procedure is

used is also derived, and the relationship between this procedure and the delay-operated approach is discussed. Also, the time intervals during which the output will contain distortion if the overall delay used in a system is less than the necessary minimum value to avoid distortion is another aspect considered. In section 5.3, the theoretical results are used in order to obtain estimates of communication delay and buffer storage necessary when using zero-crossings-based speech coders. In section 5.4, similar estimates are obtained for the case in which the proposed time coding technique of Chapter 4 is used.

5.2. Theoretical Results.

5.2.1. Delay—Operated Systems.

Let it be assumed that the instantaneous rate at which a variable-rate source produces information is R(t) symbols per unit time, and that this information is fed into a buffer. Suppose that the transmission rate, the rate at which information is taken from the buffer is r, where r is a constant.

If the situation of transmit buffer underflow, that is, the situation in which a request is made to the buffer for information when in fact the buffer is empty, is to be avoided, then it is necessary that

$$\int_0^t R(t)\, dt \geq rt, \text{ for all } t > 0 . \qquad (5.1)$$

Now, if maximum use is to be made of the transmission data rate, it follows that r should be equal to the long-term average of R(t), i.e.,

$$r = \lim_{T \to \infty} \frac{1}{T} \int_0^T R(t)\, dt \qquad (5.2)$$

and it therefore also follows, since R(t) is variable, that, in general, there will be periods of time during which $\int_0^t R(t)\, dt$ will be less than rt. It is during these periods, unless certain steps are taken to avoid it, that buffer underflow will occur. Underflow can be avoided by the introduction of a delay before information is taken from the buffer. If this delay is denoted by δ, then the necessary and sufficient condition for avoiding underflow is that

$$\int_0^t R(t)\, dt \geq r(t - \delta), \text{ for all } t > \delta, \qquad (5.3)$$

i.e.,

$$R_{cum}(t) \geq r(t - \delta) \qquad (5.4)$$

where $R_{cum}(t)$ is the cumulative total number of information samples entering the buffer up to time t. From condition (5.4), it is clear that $\delta > [rt - R_{cum}(t)]_{max}/r$, where the term $[rt - R_{cum}(t)]_{max}$ is the maximum difference between rt and $R_{cum}(t)$. Equation (5.4) is depicted in Figure 5.1, from which it can be seen that the minimum necessary value of δ is determined by the maximum extent by which the cumulative total $R_{cum}(t)$ falls below the cumulative total rt. This factor is of fundamental importance since it shows at once that it is the relatively long-term persistence of lower than average rate data which determines the necessary delay at the transmitter.

It is also necessary that buffering be provided at the receiver to convert the received constant-rate data into the required variable-rate output from the system. Just as at the transmit buffer, a delay has to be introduced before information can be taken from the receiver buffer. This delay is necessary to prevent the situation arising in which a request is made to the buffer for information before it actually arrives at the buffer. The effect of having too small a delay is illustrated schematically in Figure 5.2.

The necessary minimum delay Δ at the receiver buffer can be determined from the following argument.

When the k-th data sample is required for decoding, it is necessary that it be present in the receiver buffer. The k-th data sample arrives in the receiver buffer $(k-1)\tau$ sec after the arrival of the first data sample, where τ is the transmission period ($\tau = 1/r$). If the first sample is taken from the receiver buffer Δ sec after it arrives, then the k-th sample is required from the buffer at time

$$\Delta + \sum_{i=1}^{k-1} T_i$$

where T_i is the time interval associated with the i-th data symbol.

If distortion of the type illustrated in Figure 5.2 is to be avoided, then it is necessary and sufficient for Δ to have a value such that

$$\Delta + \sum_{i=1}^{k-1} T_i \geq (k-1)\,\tau + \tau, \text{ for all } k > 1, \qquad (5.5)$$

where the term 'τ' may be assumed (for large k) negligible in comparison with $(k - 1)\tau$. Now, in condition (5.5), the term $\sum_{i=1}^{k-1} T_i$ corresponds to the time t taken at the output of the receiver buffer to decode the k - 1 source symbols, with time being measured from the instant Δ at which decoding commences. Also, k - 1 is the cumulative sum, $R_{cum}(t)$, of symbols taken from the receiver buffer in the t sec interval. Clearly, $R_{cum}(t)$ is the same cumulative function (time delayed) as was discussed in relation to the transmit buffer. It thus follows that equation (5.5) can be rewritten, neglecting the term τ, as

$$\Delta + t \geq \frac{R_{cum}(t)}{r} ,$$

i.e.,

$$R_{cum}(t) \leq (\Delta + t)r \tag{5.6}$$

and, on substituting $t' = \Delta + t$, this becomes

$$R_{cum}(t' - \Delta) \leq rt'. \tag{5.7}$$

From condition (5.6), Δ can be seen to be equal to, or greater than, $[R_{cum}(t) - rt]_{max}/r$, where $[R_{cum}(t) - rt]_{max}$ is the maximum value of the difference between $R_{cum}(t)$ and rt.

Condition (5.7) is illustrated in Figure 5.3, from which it is clear that the delay Δ is determined by the maximum extent to which $R_{cum}(t')$ exceeds rt', the cumulative total of information symbols received up to that instant.

Since t' is delayed relative to t by δ (since no information arrives at the receiver until δ sec have elapsed) it follows that the overall delay γ introduced by the system is

$$\gamma = \delta + \Delta \tag{5.8}$$

5.2.2. The Symbol-Injection Procedure.

In the argument previously presented relating to the performance of the transmitter buffer it was assumed that the system would operate in such a way as to avoid buffer underflow, and condition (5.4) was derived, which determines the minimum delay necessary to achieve this. An alternative to avoiding transmit buffer underflow by the provision of delay is to adopt a symbol-injection procedure at those instants at which underflow occurs. This symbol-injection mode of operation can be performed without introducing any serious misuse of the system's transmission rate. For example, if the source symbols are each represented, for the purpose of transmission over a binary system, by n binary digits, then the use of one of the 2^n n-digit patterns to denote symbol-injection simply reduces the effective transmission rate to $(2^n - 1)/2^n$ of its original value.

It is important that symbol-injection systems be considered in some detail since, as will now be shown, their performance is not the same as non-injection (delay-operated) systems. In addition, any system operating on the delay principle, but in which the delay is less than the minimum necessary to avoid buffer underflow, will have to invoke symbol injection at some time during the course of its operation.

The effect of symbol injection at the transmitter is illustrated in Figure 5.4. From the figure, it is clear that symbol injection removes the need to provide delay to prevent buffer underflow.

As regards symbol injection, it is important to note exactly under what conditions this occurs. Injection begins when both the transmitter buffer is empty and the rate at which information is required from the buffer is greater than the rate at which it enters the buffer from the source. Once started, injection continues until the rate at which information enters the buffer is equal to the rate at which it is required from the buffer. If the number of symbols injected up to time t is $S_I(t)$, the total number of symbols generated by the source up to time t is $R_{cum}(t)$ and the transmission rate is r, then it follows that injection begins when both

$$R_{cum}(t) + S_I(t) = rt \tag{5.9}$$

and

$$\frac{d[R_{cum}(t)]}{dt} < r \tag{5.10}$$

Injection ceases when $dR_{cum}(t)/dt = r$. It should be noted that during periods of injection the rate of symbols generated by the source plus the injected symbols is equal to the transmission rate r, i.e.,

$$\frac{d}{dt}\left[R_{cum}(t) + S_I(t)\right] = r \tag{5.11}$$

In decoding at the receiver of an injection system it is necessary to discard the injected symbols and also to introduce a delay to avoid the type of output distortion illustrated in Fgiure 5.2. The minimum necessary delay which is the total system delay, can be determined from the following argument, which is similar to that presented in connection with the derivation of condition (5.7).

To avoid distortion in the output it is necessary that, at any time t, the information required for presentation at the output of the system be available in the receiver buffer. To achieve this, it is necessary to introduce a delay β such that

$$\beta + \sum_{i=1}^{k-1} T_i \geq (k-1)\tau + S_I(t + T_k)\tau + \tau, \qquad (5.12)$$

where, as before,

$$\sum_{i=1}^{k-1} T_i = t$$

$$k - 1 = R_{cum}(t) \qquad (t \text{ measured from instant } \beta)$$

and

$$\tau = 1/r.$$

$S_I(t + T_k)$ corresponds to the number of symbols injected at the transmitter in the period up to the instant $t + T_k$, where T_k is the time duration between the instants at which the k-th and (k + 1)-th data symbols are required by the decoder.

It thus follows that condition (5.12) can be written as

$$\beta + t \geq \frac{1}{r} \{R_{cum}(t) + S_I(t) + \lambda_k + 1 \} \tag{5.13}$$

which means that the minimum necessary delay is

$$\beta \geq \frac{1}{r} \{R_{cum}(t) + S_I(t) + \lambda_k + 1 - rt\}_{max} \tag{5.14}$$

where λ_k is the number of symbols injected during the time interval T_k. Expression (5.14) is a fully accurate statement of the minimum decoding delay, β, necessary in an injection system.

If it is assumed that the number of symbols, λ_k, injected in the time interval T_k is negligible in comparison with $S_I(t)$, the total number of symbols injected up to time t, and if the unity factor in equation (5.13) is neglected then the following simplified expression results:

$$R_{cum}(t) + S_I(t) \leq (t + \beta)r \tag{5.15}$$

and, on substituting $t' = t + \beta$ this becomes

$$R_{cum}(t' - \beta) + S_I(t' - \beta) \leq rt' \tag{5.16}$$

which must hold for all $t' > \beta$.

The aspect of performance relating to condition (5.16) is illustrated in Figure 5.5. This condition differs importantly from condition (5.7) in that the extra term $S_I(t' - \beta)$ appears in condition (5.16). This term

is due to the fact that in certain cases the arrival at the receiver of the k-th information symbol may be delayed by an amount $S_I(t).\tau$ caused by the injection of dummy symbols. From conditions (5.7) and (5.16) it is clear that for $S_I(t' - \beta)$ greater than zero, β is greater than Δ.

To determine the minimum delay β in an injection scheme it is necessary that the function $S_I(t)$ be determined. This can be done quite simply using the previously mentioned fact that digit injection begins when both $R_{cum}(t) + S_I(t) = rt$ and $R(t) < r$. Figure 5.6a shows the $R_{cum}(t)$, the $R_{cum}(t) + S_I(t)$, and the rt curves, and Figure 5.6b shows the $S_I(t)$ curve where the instants when symbol injection occurs are also indicated. It should be noted that, as a result of injection, the cumulative total number of symbols curve is effectively coincident with the rt curve during periods of injection and then climbs above it, following the shape of $R_{cum}(t)$ shifted in amplitude by a constant value. From Figure 5.6b, it is clear that the total number of injected symbols up to an instant in time is equal to the sum of injections during all previously completed injection periods plus the number injected up to the instant being considered in the current injection period.

A comparison of overall system delays involved in the delay and symbol-injection systems can be obtained by considering conditions (5.6) and (5.15), and Figure 5.6. From condition (5.15), it is clear that if distortion is to be avoided in a symbol-injection system, β must satisfy the condition

$$\beta \geq \frac{[R_{cum}(t) - rt + S_I(t)]_{max}}{r} \quad . \tag{5.17}$$

That is to say, the necessary minimum value of β is determined by the maximum value of $R_{cum}(t) - rt + S_I(t)$.

Now, from condition (5.6) it is clear that $[R_{cum}(t) - rt]_{max}/r$ is equal to Δ, the necessary minimum receiver delay in a delay-operated system and, from Figure 5.6, that the maximum value $[S_I(t)]_{max}/r$ of $S_I(t)/r$ is equal to δ , the necessary minimum transmitter delay in a delay-operated system. In the delay-operated system, the minimum overall system delay is given by

$$\dot{\gamma} = \delta + \Delta = \frac{[R_{cum}(t) - rt]_{max}}{r} + \frac{[S_I(t)]_{max}}{r} \tag{5.18}$$

From conditions (5.17) and (5.18), it is clear that the minimum value of β is only equal to the minimum value of $\dot{\gamma}$ in the special case in which the maximum values of $[R_{cum}(t) - rt]$ and $S_I(t)$ occur at the same instant. It should be noted that this condition is satisfied if, and only if, the maximum value of $[rt - R_{cum}(t)]$ occurs before the maximum value of $[R_{cum}(t) - rt]$. Observe that the maximum value of $S_I(t)$ occurs when $[rt - R_{cum}(t)]$ achieves its maximum and this value is then maintained. Thus, at this instant, and onwards, the cumulative total number of symbols $[R_{cum}(t) + S_I(t)]$ curve is equal to the $R_{cum}(t)$ curve shifted in amplitude by $[rt - R_{cum}(t)]_{max}$, and for this reason $[R_{cum}(t) + S_I(t) - rt]_{max}$ will be equal to $[R_{cum}(t) - rt]_{max} + [S_I(t)]_{max}$ if and only if $[R_{cum}(t) - rt]_{max}$ occurs after $[rt - R_{cum}(t)]_{max}$. This is illustrated in Figure 5.7.

At this point it is important to note that when operating at a transmission rate equal to the average rate at which data is produced by the source, the effect of the simplifying approximations previously made is negligible, since the errors in delay estimates are small in comparison with the actual delays involved. When dealing with higher transmission rates, however, the necessary minimum decoding delays are significantly reduced and it is then necessary to remove the approximations in order to obtain sufficiently accurate estimates of the delays.

Although in the case of general transmission rates, it does not appear possible to obtain a simpler result that (5.14), it is possible in the case in which the transmission rate r is such that

$$r > \{d[R_{cum}(t)]/dt\}_{max}$$

to simplify the result in an important way. In this case

$$S_I(t + T_k) = r(t + T_k) - R_{cum}(t + T_k) \tag{5.19}$$

and since

$$R_{cum}(t + T_k) = R_{cum}(t) + 1$$

it follows from (5.19) that

$$S_I(t + T_k) = r(t + T_k) - R_{cum}(t) - 1$$

and hence on substituting for $S_I(t + T_k)$ in (5.14), the minimum necessary delay, β, is found to be

$$\beta \geq \frac{1}{r}\{r.T_k\}_{max}$$

i.e.,

$$\beta \geq \{T_k\}_{max}. \qquad (5.20).$$

5.2.3. Distortion Intervals in Systems with Insufficient Delays.

One important aspect of system performance remains to be considered, and that relates to the time intervals during which the output will contain distortion if the overall delay used in a system is less than the necessary minimum value to avoid distortion.

Suppose a system works on a delay-operated basis, with buffering being provided at both transmitter and receiver. Suppose also that the $R_{cum}(t)$ curve is as shown in Figure 5.8a, in which case the minimum delay δ necessary to avoid transmit buffer underflow is as indicated on the diagram. Let it be assumed that the actual delay used at the transmitter buffer is σ, where σ is less than δ. Under these circumstances, symbol injection will occur in the manner shown in Figure 5.8b. Symbol injection effectively forces the $R_{cum}(t)$ curve to be shifted to the dotted curve shown in Figure 5.8a.

At the receiver let it be assumed that instead of using the delay value of β shown in Figure 5.8c, which would be sufficient to avoid distortion in the output, a value of ε is used, where ε is less than β. If this is done, then it is clear from Figure 5.8c that distortion will occur in the intervals A and B indicated. In general terms, distortion will occur during those periods for which

$$R_{cum}(t - \varepsilon) + S_I(t - \varepsilon) > rt, \; t > \varepsilon. \qquad (5.21)$$

5.3. Applications to Zero-Crossings-Based Speech Coders.

5.3.1. Statistical Properties of the Zero-Crossings of Low-Pass Filtered Speech Signals.

To obtain estimates of the necessary buffer storage, and the extent of the delays involved when information about the speech segments linking successive zero-crossings is used in speech coding, measurements were made relating to the occurrence of zero-crossings in the low-pass filtered speech. The measurements were made on a variety of speech samples ranging from isolated sentences to measurements on a long (approximately 15 min) passage of continuous speech recorded from the radio. The recording was made using a high-quality tuner and associated high-quality tape equipment. The recorded speech was subsequently low-pass filtered to 3.04 KHz before making the zero-crossings measurements. Throughout the test, the peak signal level was set at 1 V.

In the first test, five isolated sentences taken from the Harvard list number 1 were examined with each sentence being spoken by both a male and a female speaker. The number of zero-crossings in 0.5 sec, 1.0 sec, and 2.0 sec sections of speech were measured. The speech sections examined were each contained entirely within the duration of the associated sentence, thereby eliminating from consideration any long periods of silence between sentences.

The measurements were made using zero-level offsets ranging from zero to 30 mV. This was done so that results could be obtained that were not seriously affected by the presence of any low-level system

noise, which is known[27] to have a considerable influence on the number of zero-crossings. The results of the test are given in Tables 5.1 and 5.2.

In the second test, 2 sec sections of the long (15 min) passage of speech were examined. The 2 sec samples were selected one after another in time as the speech passage was played into the zero-crossings detector. The sections did not overlap, and each section was selected so that long inter-sentence silences were avoided. Zero, 10 mV and 20 mV offsets were again used, and the results of the tests are shown in Figures 5.9 to 5.14 in the form of histograms. With an offset voltage of 10 mV it can be seen that the average number of zero-crossings per second, measured over the 2 sec intervals, is very close to 900 sec^{-1}. The figures also show that, for male speakers, and a 10 mV offset, approximately 13% of the 2 sec samples contain a number of crossings less than 1400 (i.e., 700 sec^{-1} averaged over 2 sec) and that approximately 18% have more than 2200 crossings in the 2 sec period). In the case of female speakers the corresponding figures are seen to be 8.5% and 12%, respectively.

In the final test, the cumulative function $R_{cum}(t)$ was measured for a number of arbitrarily selected 7 sec intervals of the material recorded from the radio. Typical examples of the results obtained are shown in Figures 5.15 to 5.17. In the test, the functions were measured at 20 msec intervals for the first 1.4 sec, at 40 msec intervals for the next 0.6 sec, and then at 250 msec intervals for the period from 2 sec to 7 sec.

Table 5.1. Zero-Crossings Measurements of Low-Pass Filtered Speech Signals (Zero Offset Voltage).

Sentence Number	0.5 sec section of speech	1.0 sec section of speech	2.0 sec section of speech
	Number of Zero-crossings	Number of Zero-crossings	Number of Zero-crossings
1 (male speaker)	955	1680	2653
2 (male speaker)	785	1634	3163
3 (male speaker)	1316	2043	3804
4 (male speaker)	555	1147	2089
5 (male speaker)	1023	1988	2848
6 (female speaker)	666	1392	2452
7 (female speaker)	361	864	2027
8 (female speaker)	683	1125	2438
9 (female speaker)	322	667	1647
10 (female speaker)	503	876	2847

Table 5.2. Zero-crossings Measurements of Low-Pass Filtered Speech Signals (Offset Voltages of 10 mV, 20 mV, and 30 mV; 2 sec sections of speech).

Sentence Number	Offset 10 mV	Offset 20 mV	Offset 30 mV
	Number of Zero-crossings	Number of Zero-crossings	Number of Zero-crossings
1 (male speaker)	2457	1818	1696
2 (male speaker)	2941	2992	2873
3 (male speaker)	3593	2595	1506
4 (male speaker)	2184	1458	1297
5 (male speaker)	2944	2142	1917
6 (female speaker)	2099	1326	1522
7 (female speaker)	2294	1142	1015
8 (female speaker)	2537	1238	1149
9 (female speaker)	1585	938	821
10 (female speaker)	2069	1109	1140

5.3.2. Estimates of Communication Delay and Buffer Storage.

In the case of delay-operated systems, a simple first estimate of the necessary overall system delay can be obtained from the results of the second test. These results indicate that it is not at all uncommon for the number of zero-crossings in 2 sec of speech to differ from the average value of 1800 by $\pm$ 400, and that even larger differences can and do occur. On the basis of the $\pm$ 400 difference, the necessary minimum delay is seen to be 8/9 sec, and an even larger delay would be necessary if it were desired to eliminate distortion that can arise when greater differences occur. From the results of test 3, presented in Figures 5.15 to 5.17, a more reliable estimate can be made of the necessary delay. If the theory developed in section 5.2 is applied to these figures then it is seen that in one case (Figure 5.15) a delay Δ of approximately 2 sec is required at the receiver buffer with virtually no delay being required at the transmitter ($\delta = 0$). This thus means that a minimum overall system delay of approximately $\gamma = 2$ sec is required if distortion is to be avoided. In another case (Figure 5.16) the necessary minimum overall delay is again approximately $\gamma = 2$ sec, with virtually all of this being at the transmitter (i.e., $\Delta = 0$, $\delta = 2$ sec). In the case shown in Figure 5.17, the overall required delay, γ , is approximately 1.38 sec, with approximately 1.2 sec of this being taken up at the receiver. In the case of systems operating on a symbol-injection principle the overall system delays would be $\beta = 2$ sec, 0.3 sec, and 1.38 sec, respectively. As can be observed, $\gamma = \beta$ only in the cases when the maximum value of $[rt - R_{cum}(t)]$ occurs before the maximum value of $[R_{cum}(t) - rt]$.

Estimates can also be made of the necessary buffer size, at both transmitter and receiver. In the case of delay-operated systems, the transmitter and receiver buffers have to be large enough to hold a number of symbols equal to the maximum value of the difference between $R_{cum}(t)$ and $r(t - \delta)$, and $r(t + \Delta)$ and $R_{cum}(t)$, respectively. That is, the transmitter buffer size is given by $[R_{cum}(t) - r(t - \delta)]_{max}$, and the receiver buffer size by $[r(t + \Delta) - R_{cum}(t)]_{max}$. It is easy to see that these two values are the same. Using the results of Figures 5.15 to 5.17 it is estimated that the transmitter and receiver buffers will each have to be capable of holding approximately 2000 symbols, or 2000 n binary digits, where n is the average number of binary digits used to represent each symbol.

Using the arguments presented at the end of section 5.2, it is possible to determine the time periods during which distortion will occur if the allowed system delay is less than the minimum necessary to avoid distortion. If a 0.25 sec delay is allowed at both transmitter and receiver, then it is estimated using the material of Figures 5.15 to 5.17 that distortion will occur for periods of 6.48 sec, 0.0 sec, and 5.50 sec for the 7 sec samples of speech shown. It should be noted that in the case of Figure 5.16 the zero interval of distortion resulted because the system had to invoke symbol injection at some time during the course of its operation as discussed in section 5.2. With the symbol-injection system having a total overall delay of 0.5 sec, distortion will occur for periods of 6.16 sec, 0.0 sec, and 5.43 sec, respectively.

5.4. Applications to the Proposed Time Coding Technique for Speech Signals.

In the previous section estimates were obtained of the necessary buffer storage, and the extent of the delays involved when coding speech in terms of information relating to speech segments linking successive zero-crossings. By applying the basic theoretical results which were presented in section 5.2, the time coding method for speech signals developed in chapter 4 can also be evaluated in terms of the necessary buffer size and overall system delay. Both the delay-operated approach and the symbol-injection procedure will be considered.

In the case of the proposed time coding technique described in chapter 4 the average bit-rate is 7.6 Kbits/sec. This value will be used to obtain the estimates relating to delay and buffer size. For purpose of obtaining these estimates, the cumulative total number of information binary digits $R_{cum}(t)$ produced by the coding process up to time t was measured for 7 sec of speech spoken by both male and female speakers. The results obtained are given in Figures 5.18 and 5.19.

If the theory developed in section 5.2 is applied to these figures then it is seen that, when a delay-operated principle is used, the overall required delay, γ, is approximately 0.58 sec in one case (Figure 5.18), with approximately 0.49 sec of this being taken up at the receiver (i.e., $\Delta = 0.49$ sec), and the rest at the transmitter ($\delta = 0.09$ sec). In the case shown in Figure 5.19, the minimum necessary value of the delay δ at the transmitter in order to avoid buffer underflow is $\delta = 0.25$ sec,

and a delay of approximately $\Delta = 0.05$ sec is required at the receiver buffer. This thus means that a minimum overall system delay of approximately $\gamma = 0.3$ sec is required if distortion is to be avoided.

Now, if a symbol-injection procedure is adopted, then the overall system delay for the case shown in Figure 5.19 will be approximately $\beta = 0.17$ sec, that is, a 0.13 sec reduction as compared to the delay mode of operation. The application of the symbol-injection principle to the case presented in Figure 5.18, however, will not provide any change in the overall system delay since, in this case, the maximum value of $[rt - R_{cum}(t)]$ occurs before the maximum value of $[R_{cum}(t) - rt]$. In this case, $\beta = \gamma = 0.58$ sec.

As regards transmitter and receiver buffer sizes, from the results of Figures 5.18 and 5.19 it is estimated that each will have to be capable of holding approximately 4500 binary digits, if the coding method is to operate at a constant transmission rate of 7.6 Kbits/sec.

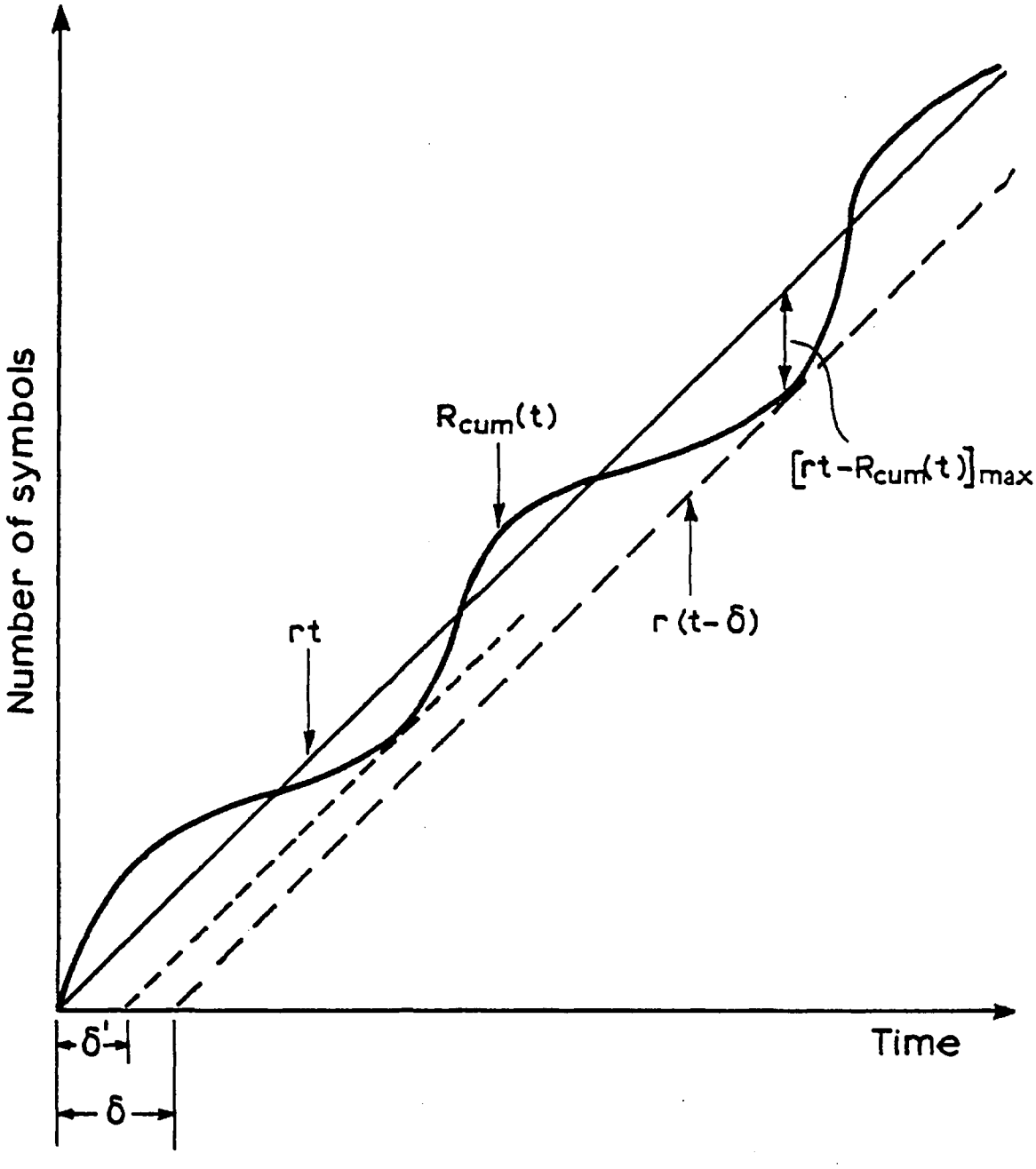
5.5. Discussion.

Some new general theory has been developed relating to the delays involved in variable-information-rate-source to constant-transmission-rate schemes of data compression. The results of a statistical analysis of the zero-crossings of low-pass filtered speech have been presented and the use of this information suggests that fairly long delays (of the order of 1 sec) may be involved in the distortionless transmission of speech encoded using information about the waveform segments linking successive real zeros of the speech signal. The delays involved are

such as to make two-way conversation difficult and it will, if speech is to be communicated on the basis of an average rate of zero-crossings of approximately 900 sec^{-1}, be necessary that some further redundancy removal techniques be developed for effectively removing the large variations that can occur in basic zero-crossings rates.

Estimates of overall system delay involved in the proposed time coding technique for speech signals have also been presented for illustrative sections of both male and female speech material. The results obtained for the proposed system, when operating at a transmission bit-rate of 7.6 Kbits/sec, indicate that a significant reduction of the necessary minimum overall delay may be involved. The maximum value obtained in this case was a delay of 0.58 sec, whereas for the zero-crossings-based speech coding methods this figure was about 3.5 times higher. It should also be noted that in the case of extremal coding of speech[14] the delay involved is of the order of 1 sec, if the system is to be operated at a transmission bit-rate of 15 Kbits/sec, which rate is about twice as high as that used in the proposed time coding scheme.

It is important to remember that all the above estimates of delay and buffer size are based on a limited set of measurements and can only be considered as indicative. In a system required to handle general speech input signals the delay and buffer storage requirements may be in excess of those quoted above, but this can only be determined by practical tests and the statistical analysis of a wide range of speech signals.



δ' = delay necessary to avoid first underflow condition (not sufficient to avoid second underflow condition)

δ = delay necessary to avoid worst case of underflow condition

Figure 5.1. Delay at transmit buffer to avoid underflow.

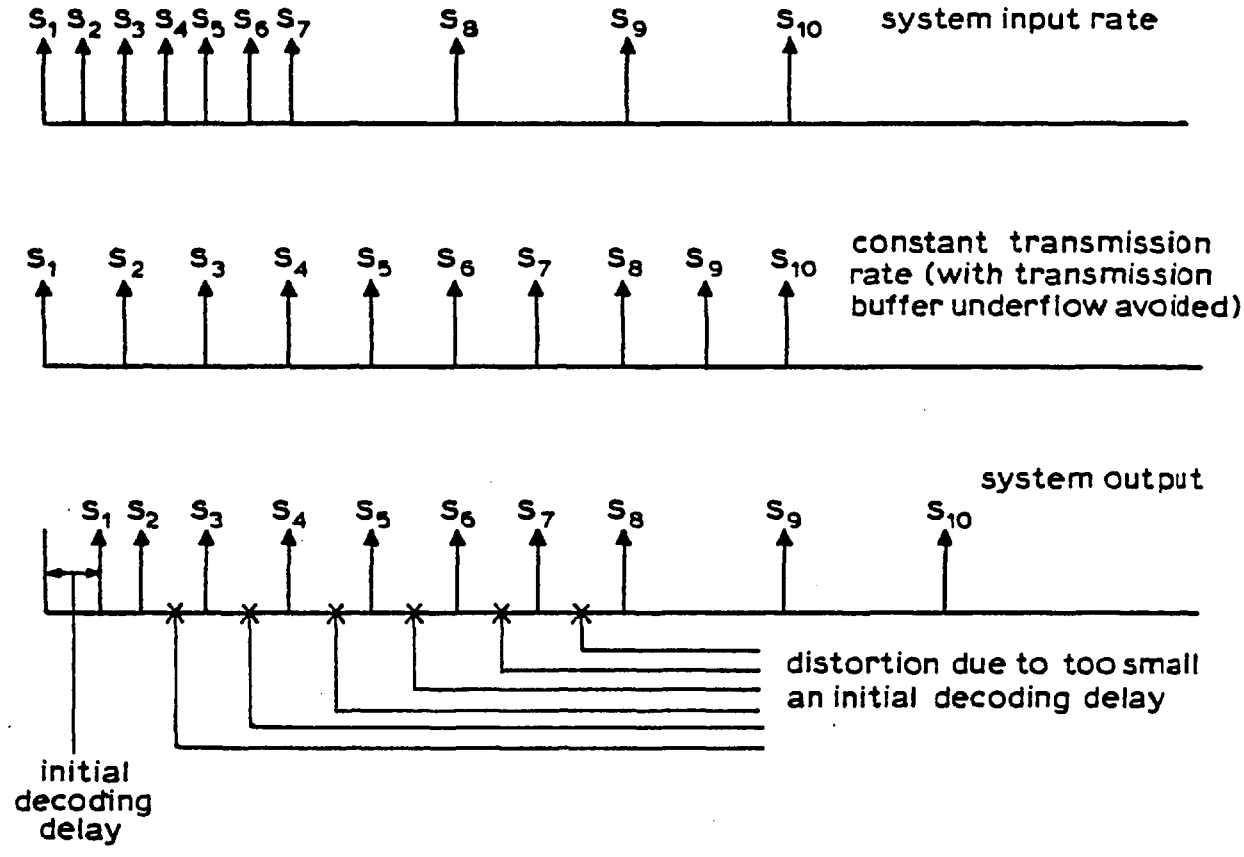


Figure 5.2. Illustration of output distortion due to insufficient decoding delay.

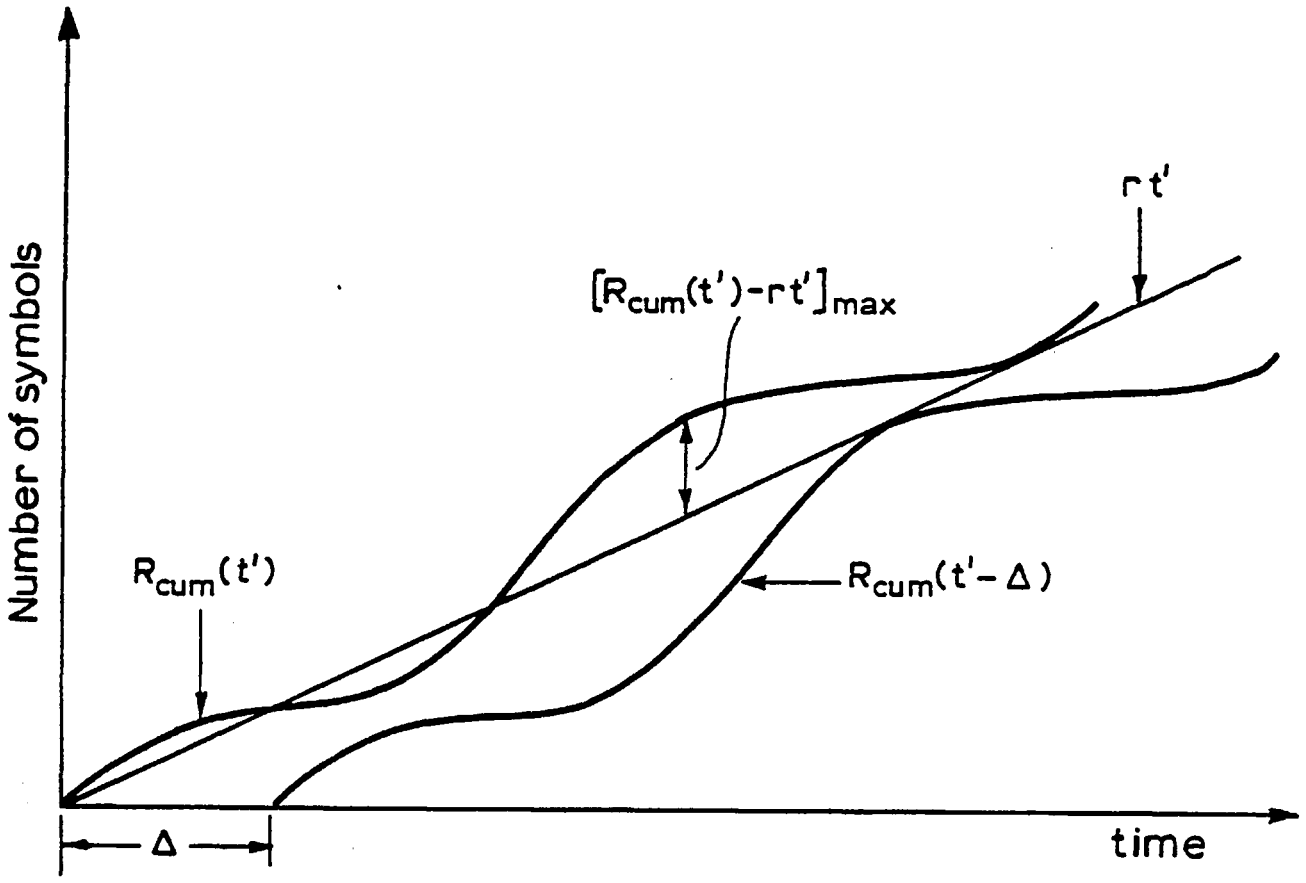


Figure 5.3. Delay at receiver to avoid output distortion.

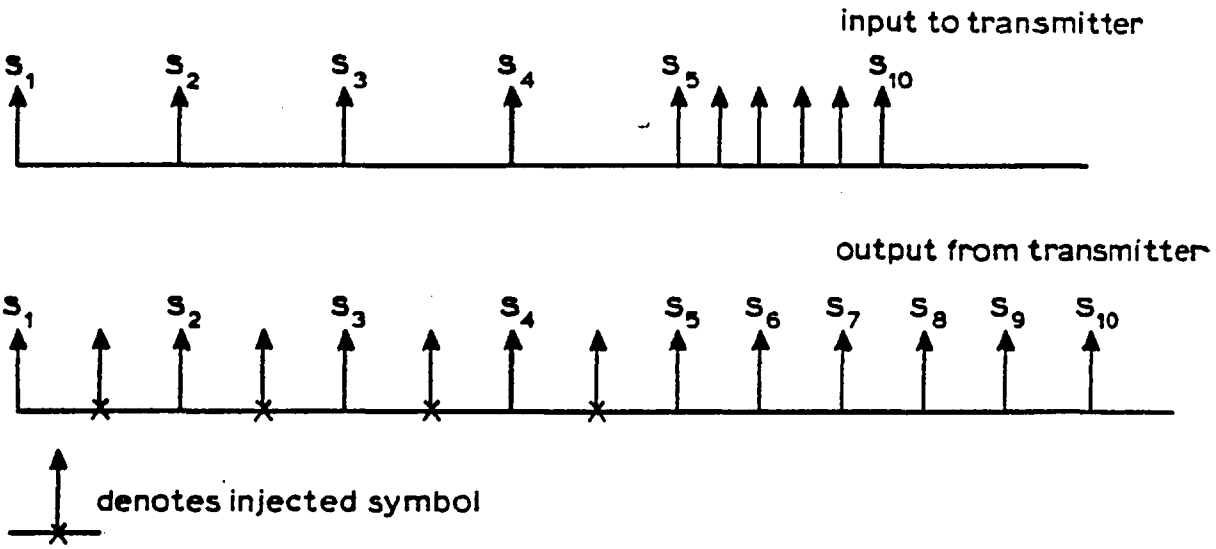


Figure 5.4. Illustration of symbol injection at transmitter.

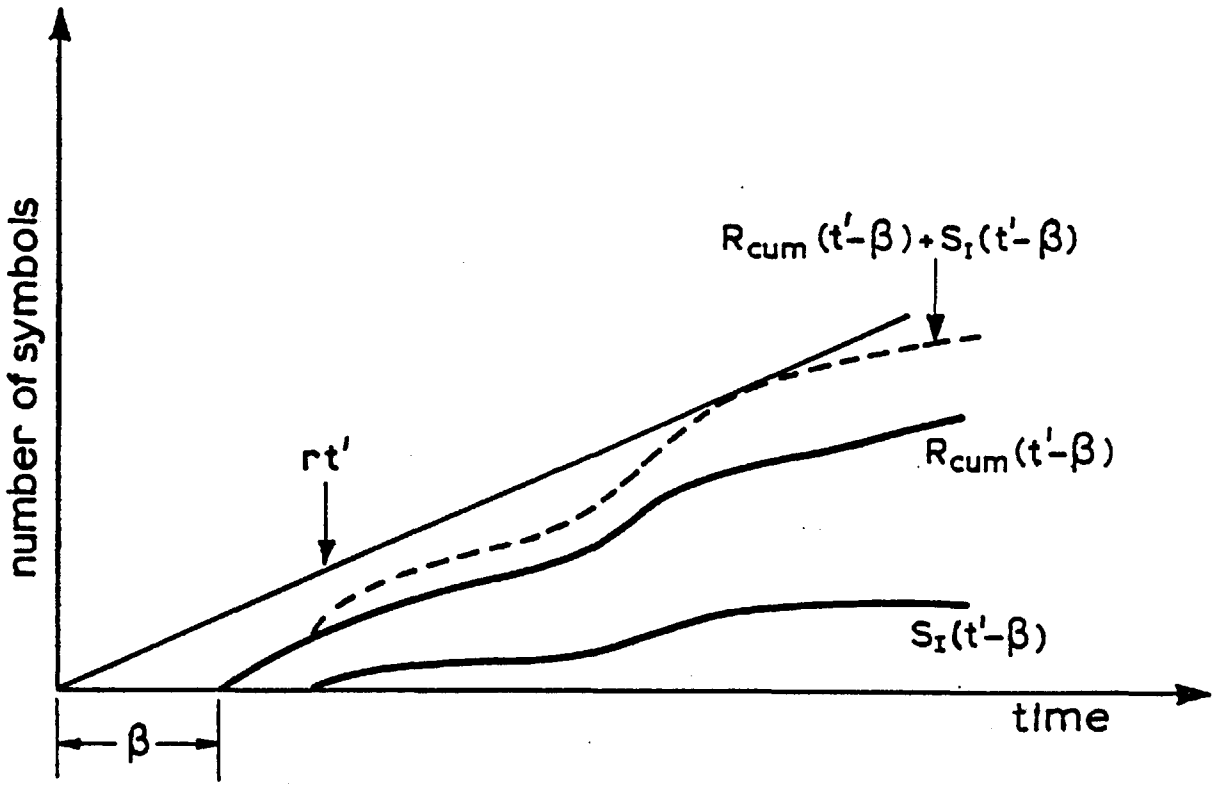


Figure 5.5. Illustration of delay necessary to avoid distortion in a symbol-injection system.

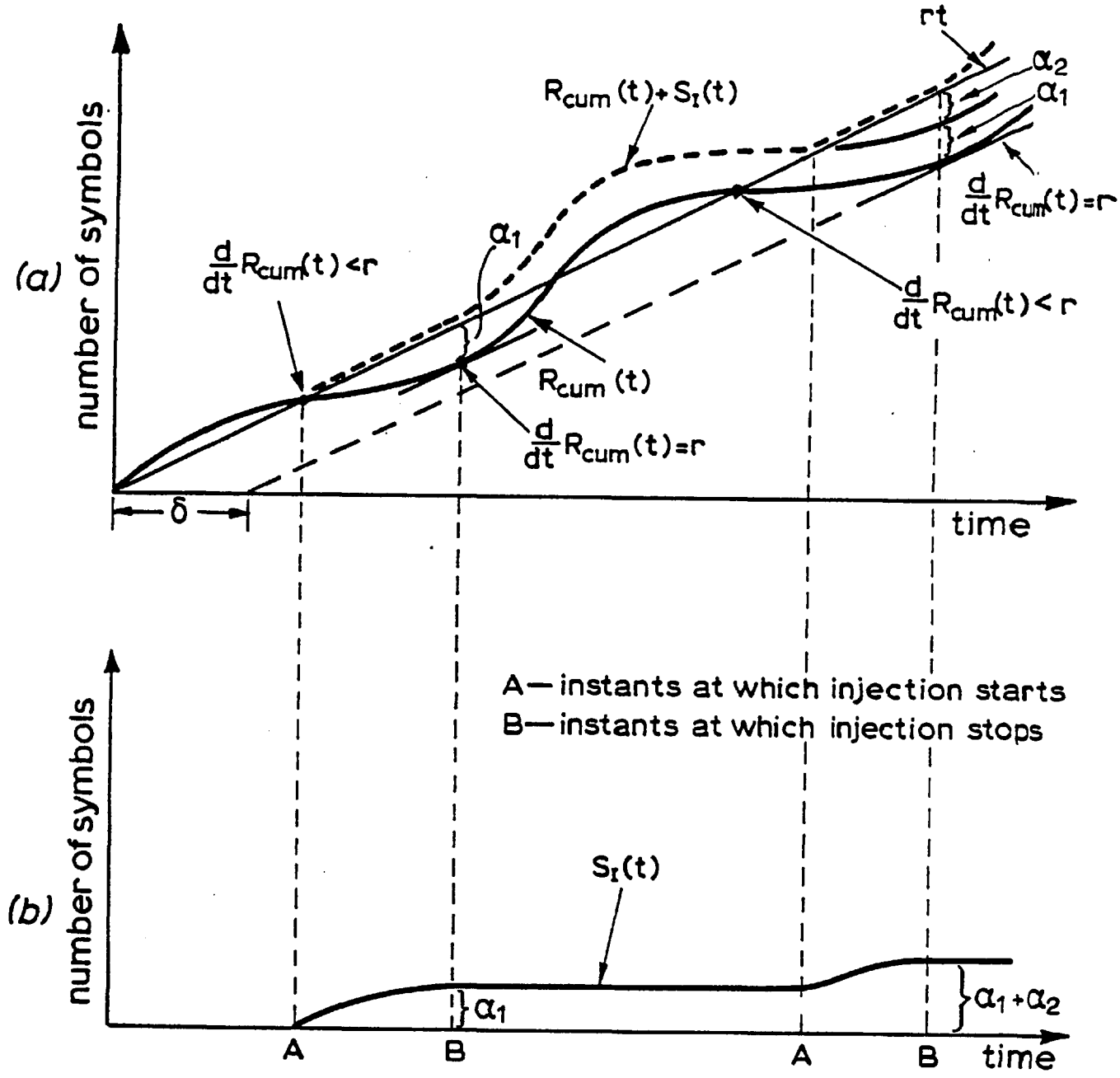


Figure 5.6. Illustration of symbol injection.

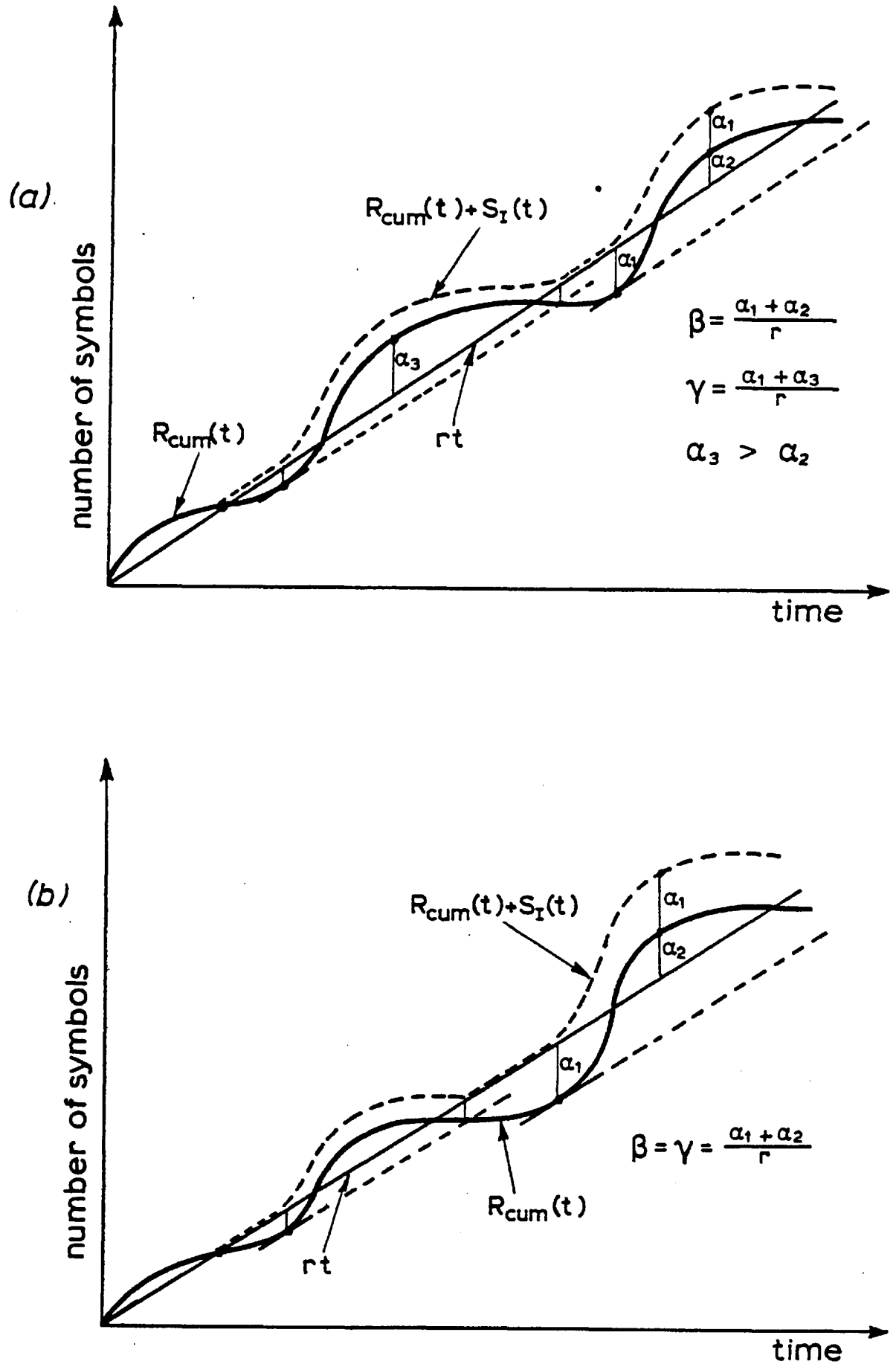


Figure 5.7. Illustration of overall system delays involved in the delay (γ) and symbol-injection (β) systems. (a) $\beta \neq \gamma$, (b) $\beta = \gamma$.

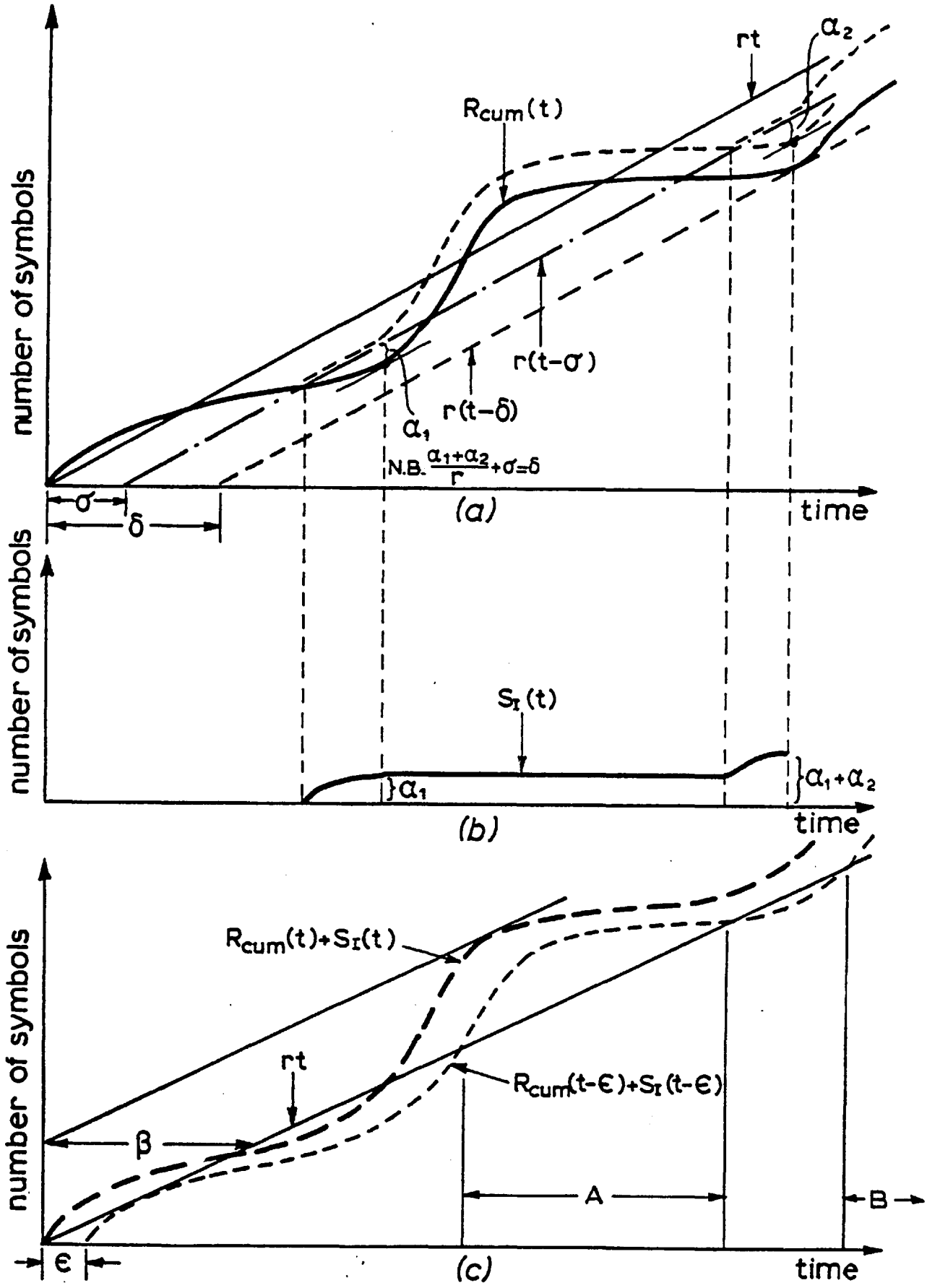


Figure 5.8. Illustration of distortion in systems with insufficient delays.

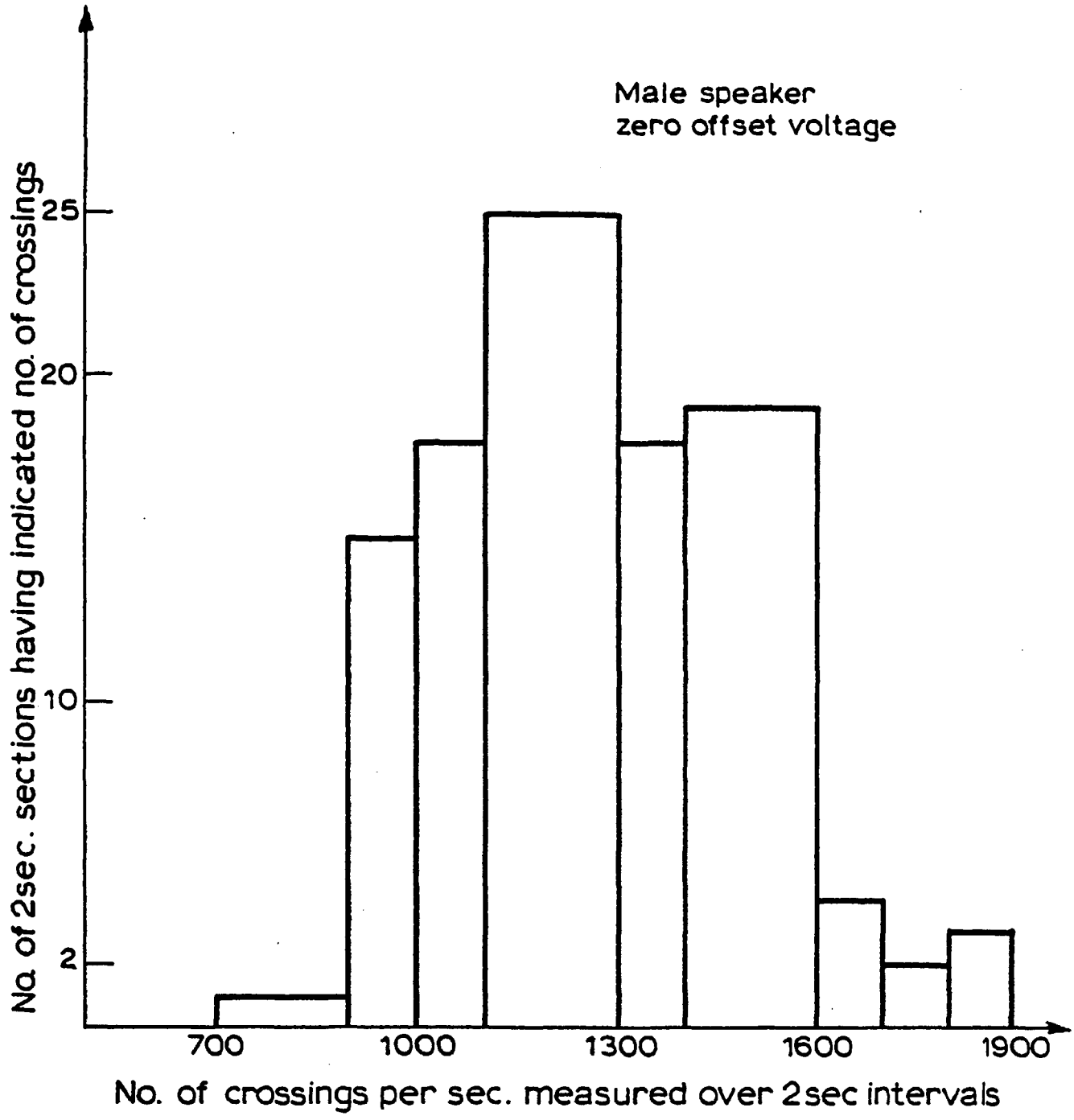


Figure 5.9. Histogram relating number of crossings in 2 sec intervals of speech (male speaker; zero offset voltage).

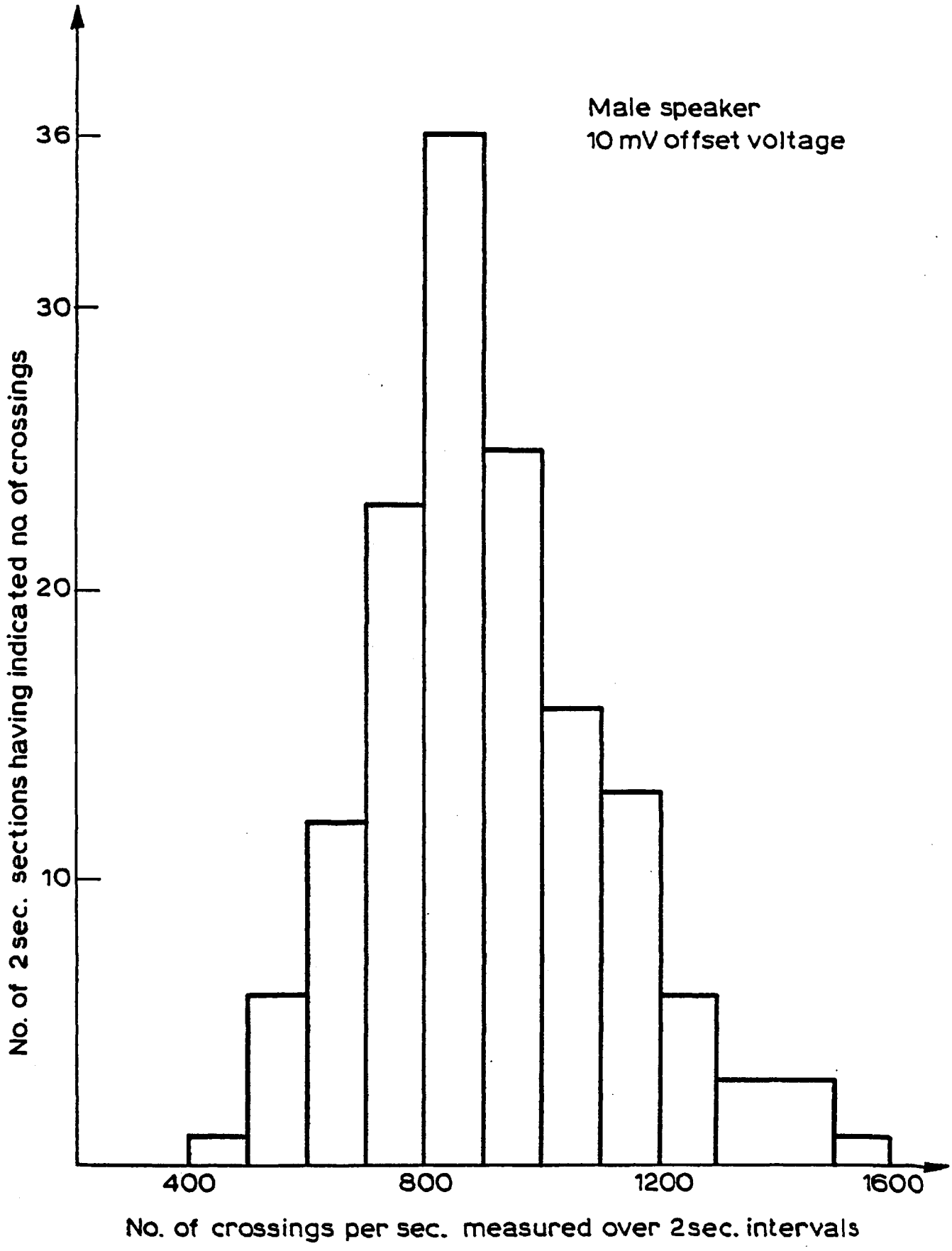


Figure 5.10. Histogram relating number of crossings in 2 sec intervals of speech (male speaker; 10 mV offset voltage).

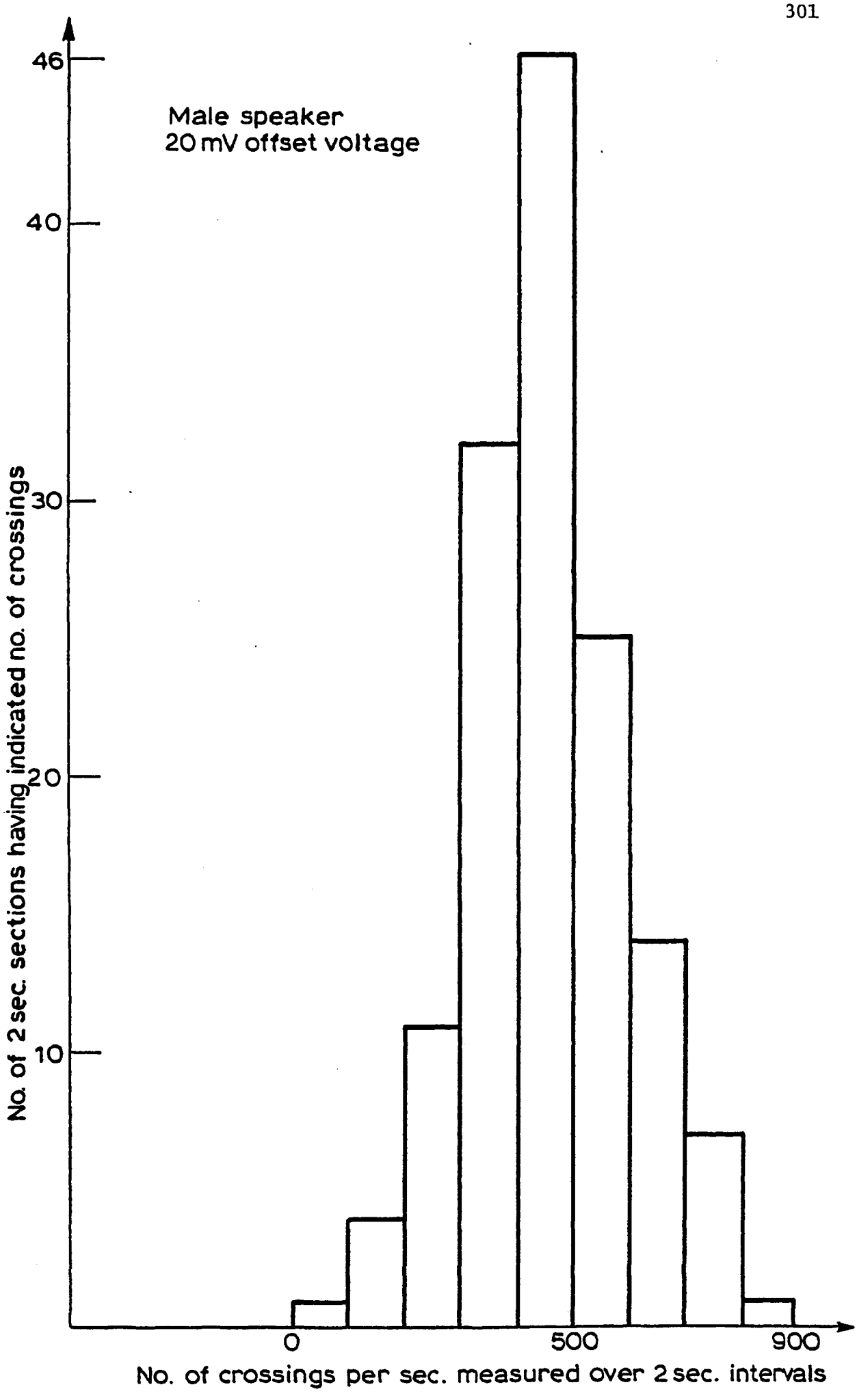


Figure 5.11. Histogram relating number of crossings in 2 sec intervals of speech (male speaker; 20 mV offset voltage).

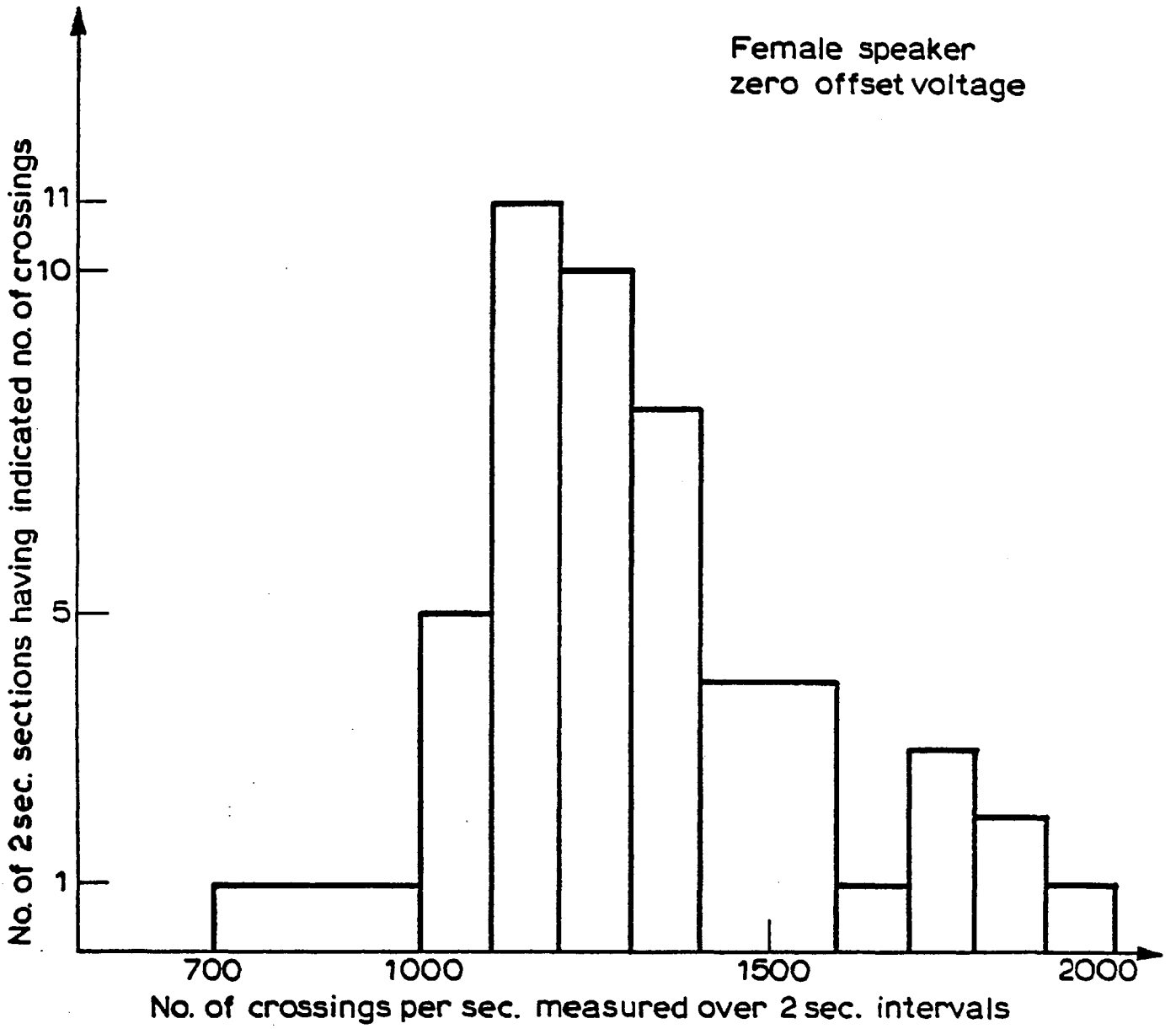


Figure 5.12. Histogram relating number of crossings in 2 sec intervals of speech (female speaker; zero offset voltage).

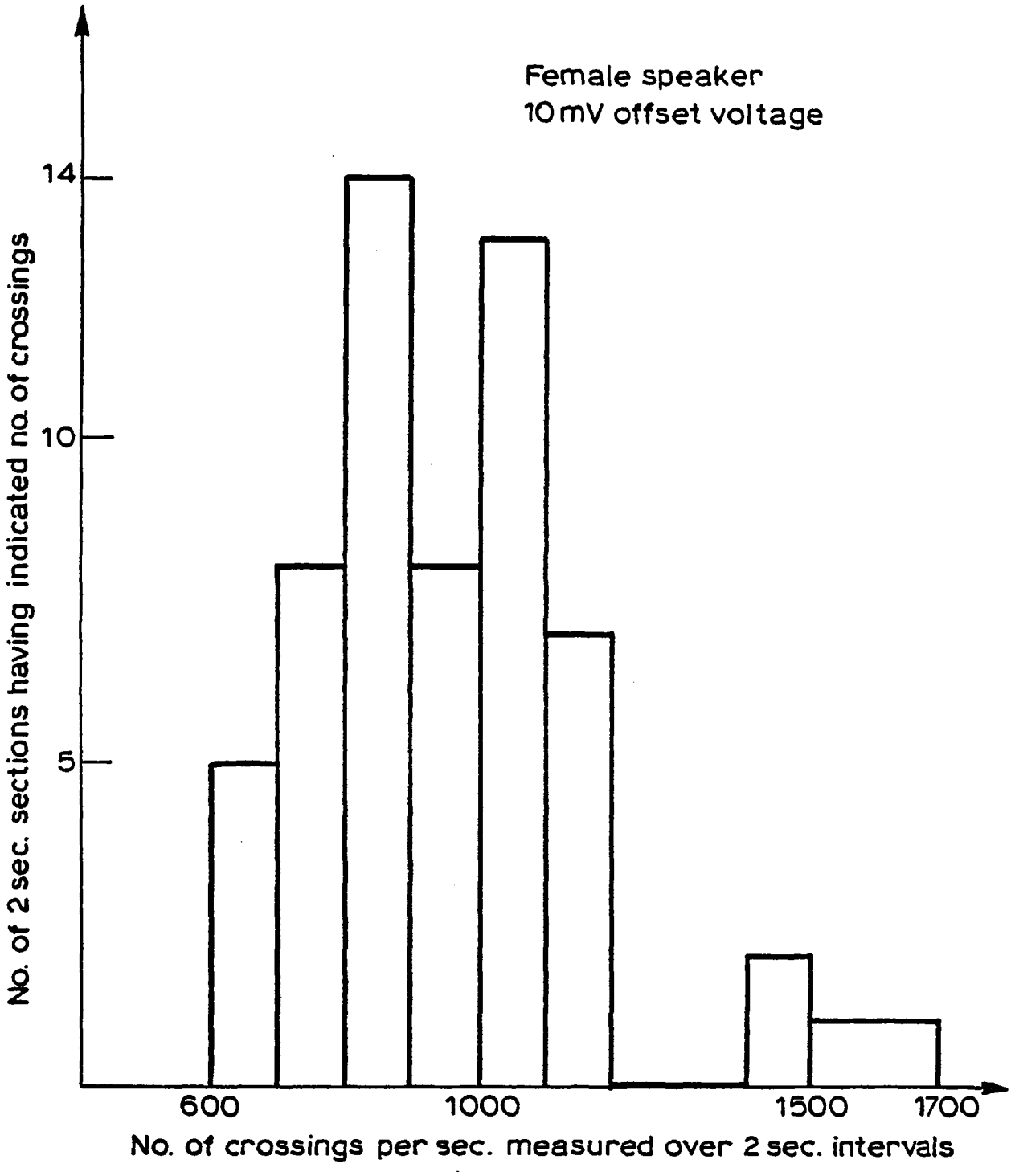


Figure 5.13. Histogram relating number of crossings in 2 sec intervals of speech (female speaker; 10 mV offset voltage).

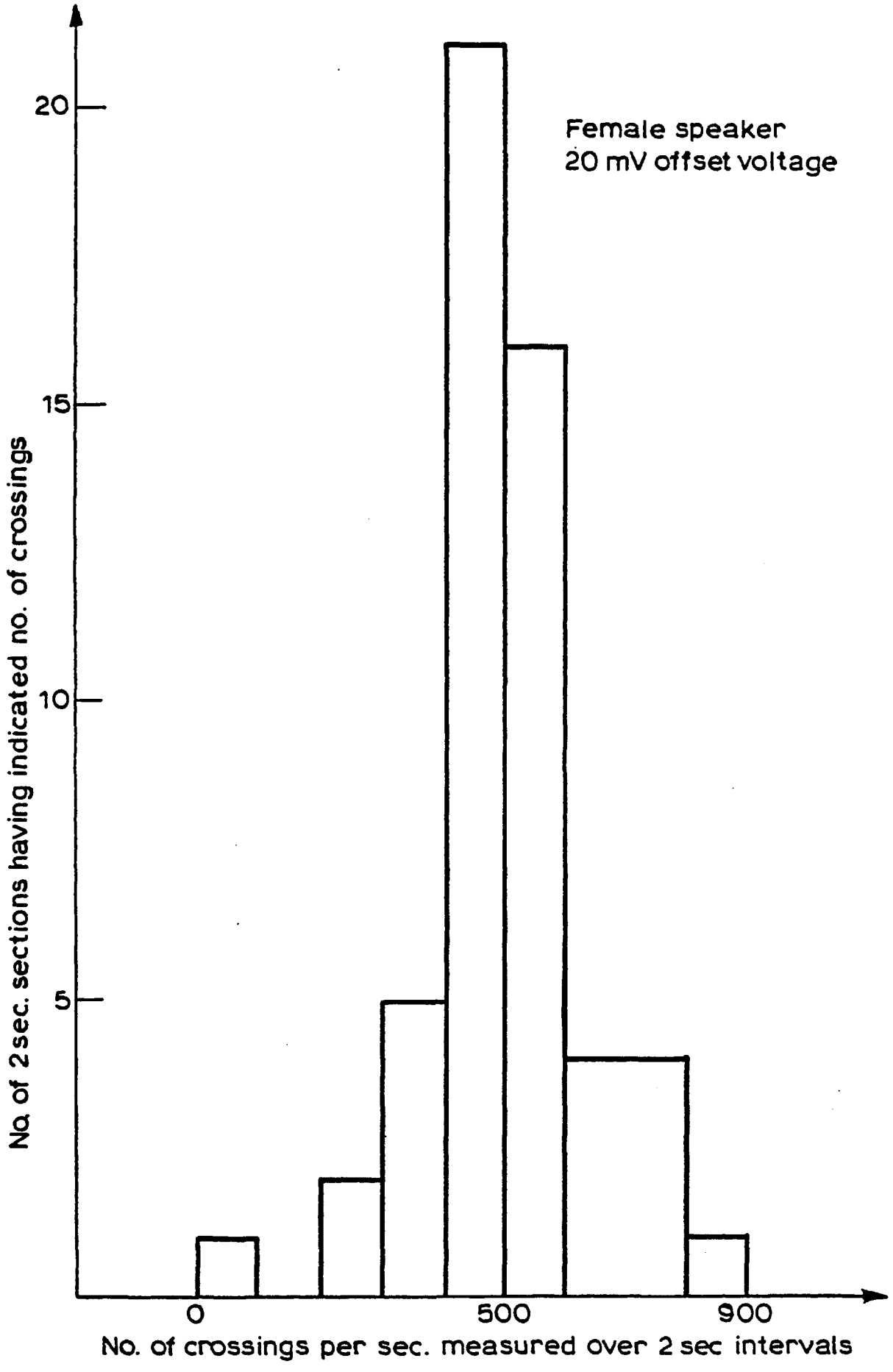


Figure 5.14. Histogram relating number of crossings in 2 sec intervals of speech (female speaker; 20 mV offset voltage).

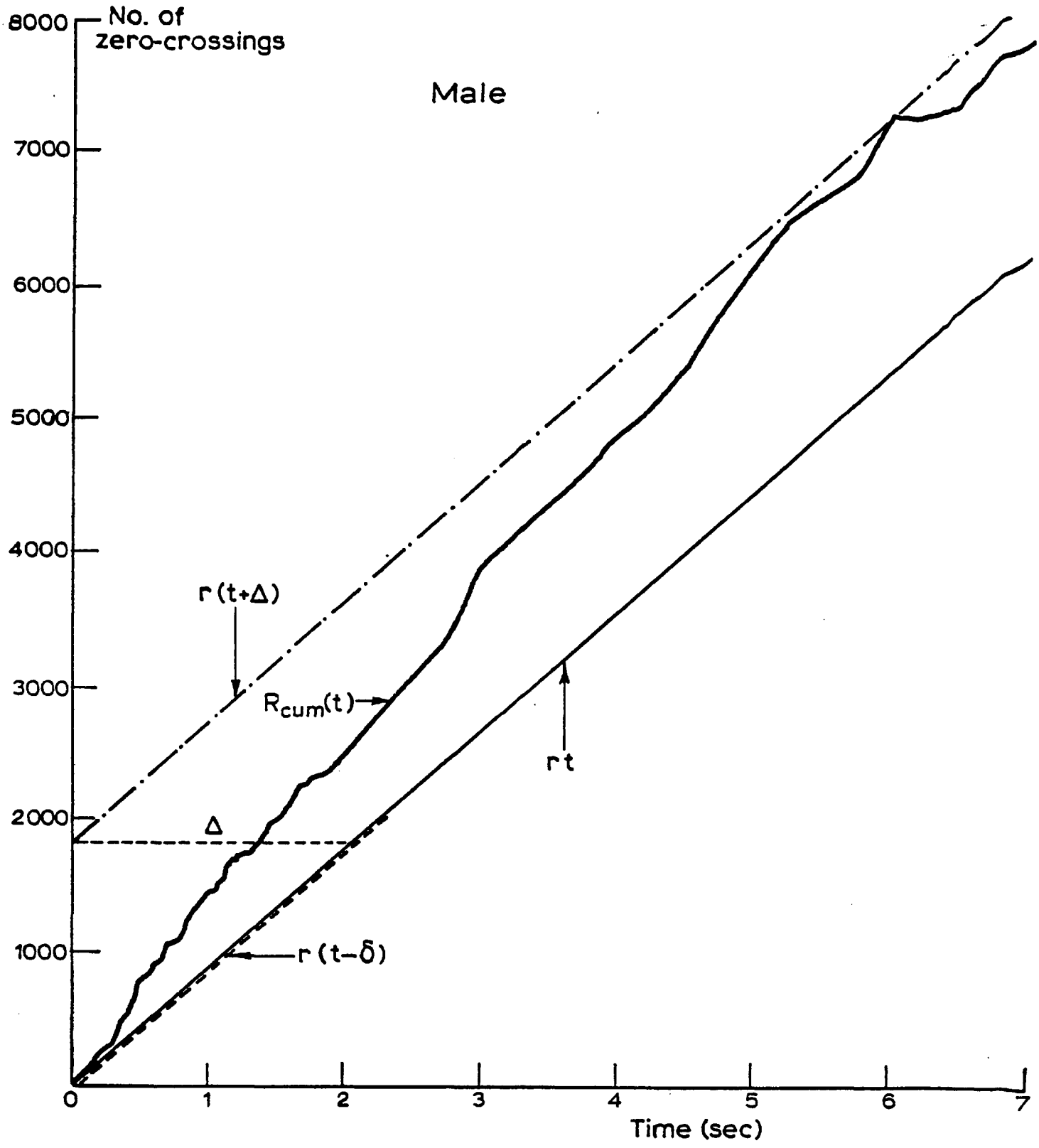


Figure 5.15. Variations in cumulative function (number of zero-crossings) over 7 sec interval of speech (male).

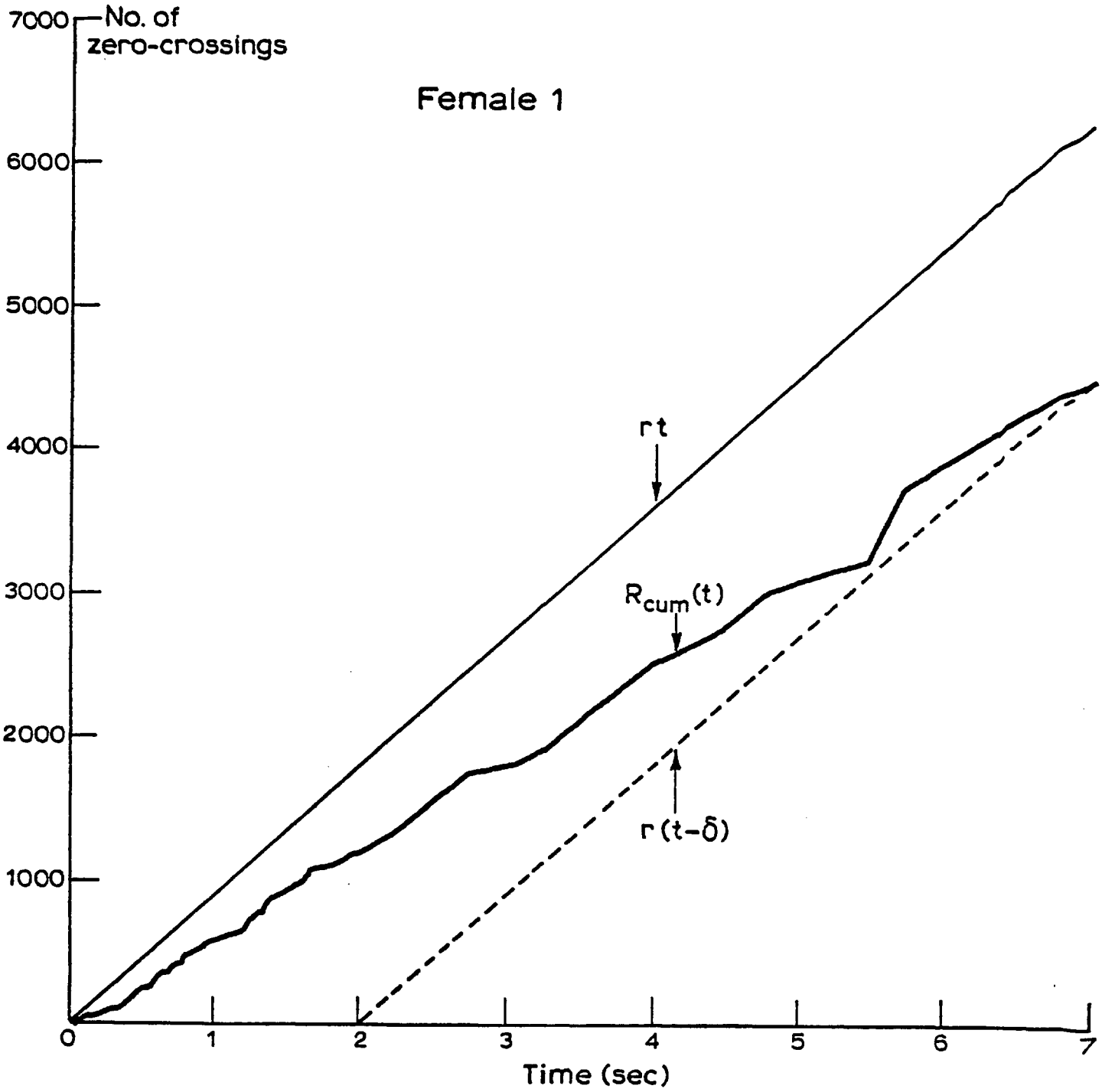


Figure 5.16. Variations in cumulative function (number of zero-crossings) over 7 sec interval of speech (female 1).

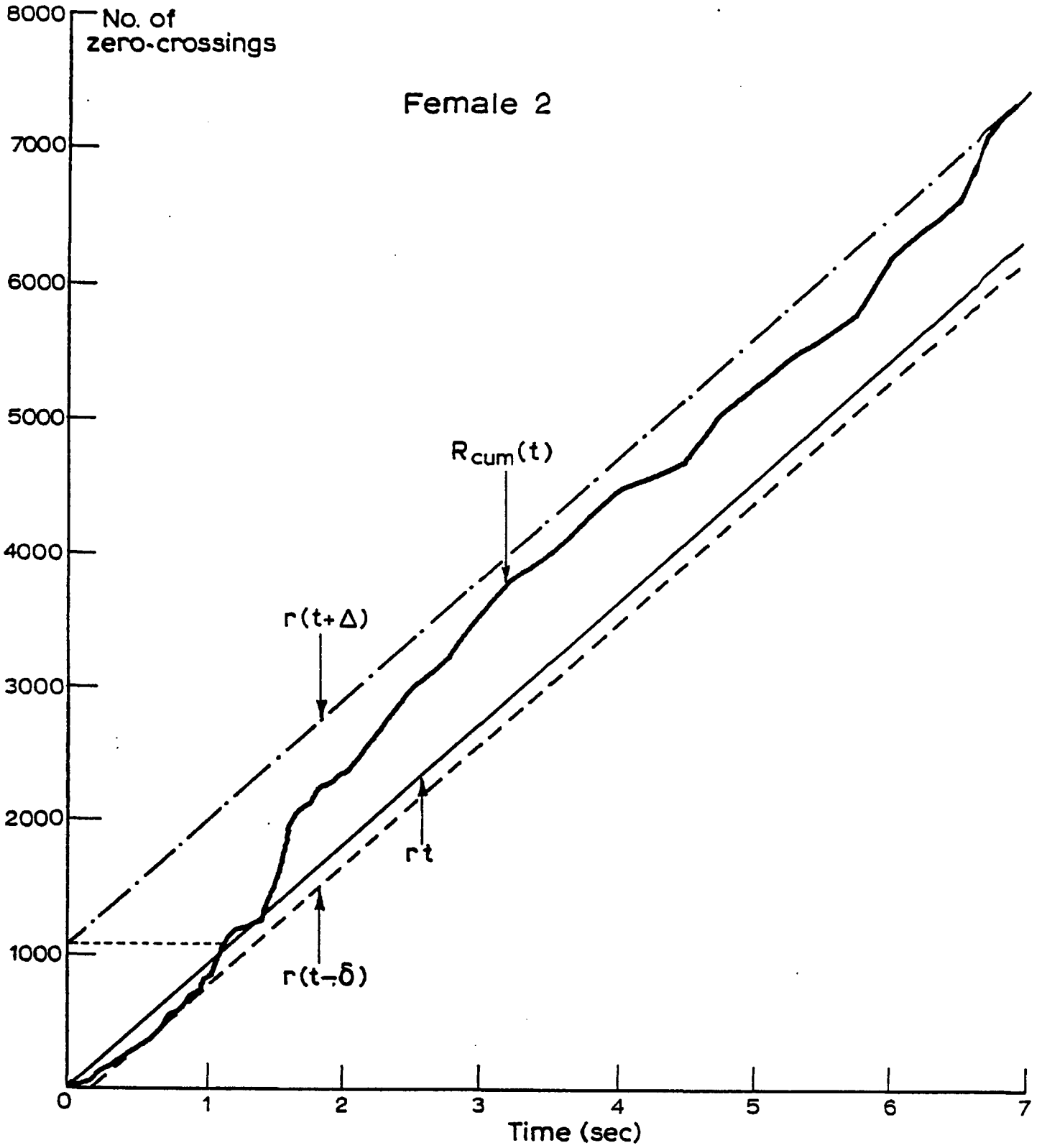


Figure 5.17. Variations in cumulative function (number of zero-crossings) over 7 sec interval of speech (female 2).

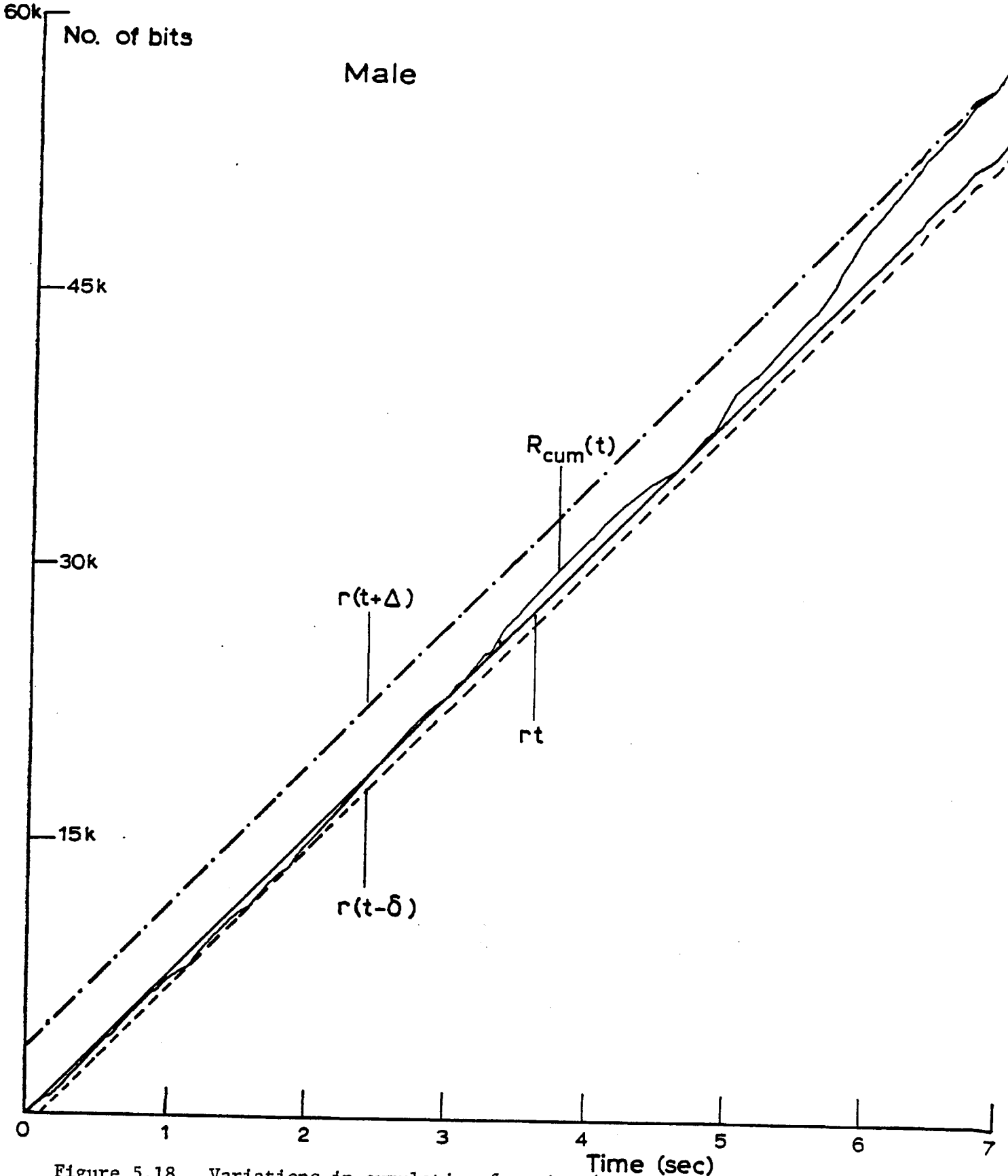


Figure 5.18. Variations in cumulative function (number of bits) for the proposed time coding scheme over 7 sec interval of speech (male).

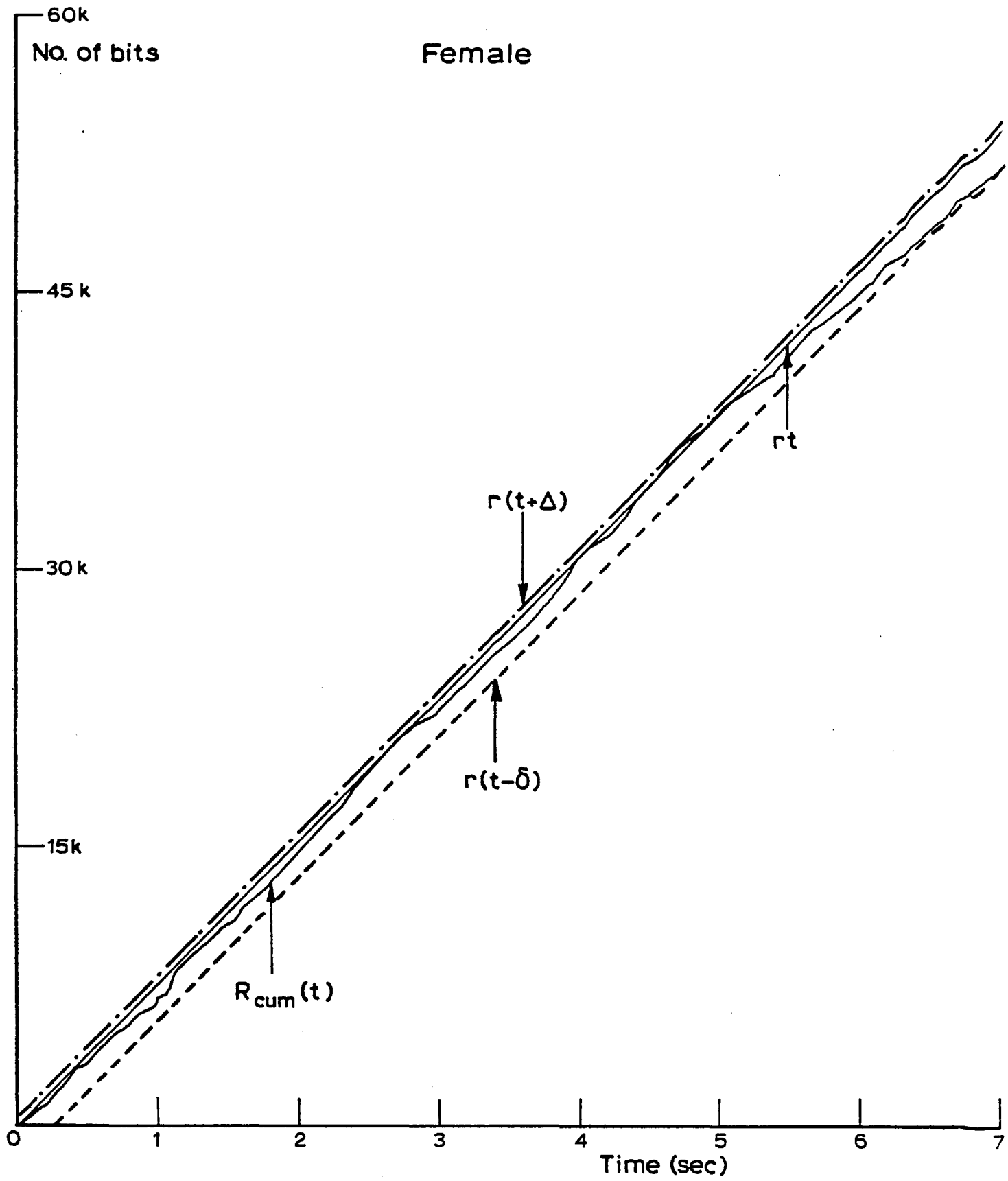


Figure 5.19. Variations in cumulative function (number of bits) for the proposed time coding scheme over 7 sec interval of speech (female).

CHAPTER 6

CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

CHAPTER 6

CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

6.1. Conclusions

The possibilities of exploiting two different approaches to the low-bit-rate coding of speech signals have been investigated in this thesis. The results of this investigation were presented in Chapters 3 and 4. In Chapter 5, the results of a study of various factors related to the performance of variable-source-rate to constant-transmission-rate schemes of data compression were presented, with particular attention being given to time-domain speech coding techniques.

The methods described in Chapter 3 were examined with a view to improving the speech quality obtained with Hadamard transform coding at low-bit-rates (below 6 Kbits/sec). It was found that the use of double Hadamard speech coding led to a small improvement of speech quality for bit-rates of 3.5, 4.625, and 5.75 Kbits/sec. The improvement results from the removal of residual redundancy that still remains between the dominant Hadamard coefficients. It was also clear that the speech quality obtained with the double Hadamard method of coding is better than that obtained with the Hadamard-Haar coding scheme. In terms of system complexity, the double Hadamard

technique is simpler than the Hadamard-Haar scheme, and is slightly more complex than the basic Hadamard coding scheme. However, the improvement of speech quality obtained with the double Hadamard method of speech coding is not sufficient to justify the increased processing operation that is necessary.

The use of oversampling in the speech waveform processed with the Hadamard coding technique at low-bit-rates (below 6 Kbits/sec.) was found to produce a small improvement in speech quality. The improvement was obtained by removing the effect of spectrum aliasing in the Hadamard coded speech. However, the inherent degradation associated with Hadamard coding was found to be more significant than the distortions due to spectrum aliasing. For this reason, the net improvement of speech quality obtained with the use of oversampling in decoding is limited. When oversampling is combined with waveform smoothing the resultant speech quality is noticeably better than that obtained from the original Hadamard technique. The smoothing technique functions as an adaptive low-pass filter, producing a waveform which is closer to that observed in natural speech and which constitutes a smoothed version of the Hadamard representation of the original speech. The improvement in performance obtained by the combined use of oversampling and smoothing in decoding does not result in speech of commercial quality. However, the results obtained for the Hadamard coded speech indicate that combined oversampling and smoothing may be useful when applied to other coding schemes which

produce similar signal regeneration characteristics. Of course, the extent to which it will be possible to use the oversampling-smoothing method efficiently will depend on the degree of distortion initially introduced by the original system.

In Chapter 4 a new method for low-bit-rate time coding of speech signals was presented. Consideration was given to time coding techniques which treat voiced and unvoiced portions of the speech waveform separately, and which exploit the quasi-periodic nature of the voiced portions of the waveform. For voiced speech, the following parameters are transmitted: (1) amplitudes [A(NP1) and A(NP2)] and locations (NP1 and NP2) of two adjacent pitch peaks; (2) amplitudes and locations of local maxima of the waveform intrapitch (between NP1 and NP2); and (3) amplitudes and locations of the other pitch peaks. For unvoiced speech, the following parameters are transmitted: (1) zero-crossing locations; (2) rms amplitude information; and (3) average distance between two successive instants associated with the alternate extremes of the waveform.

The new time-domain pitch period detection (TDPD) algorithm which was described in Chapter 4 forms the basis of the approach to the encoding of voiced speech. The attractiveness of the TDPD algorithm is based on the following factors: (1) the operations involved are simple and compatible with those employed in time coding; (2) pitch period markers are provided directly on the speech waveform; (3) basically, only three simple tests are required to identify the

pitch markers; (4) the accuracy of the pitch estimates and voiced-unvoiced classification compares favourably with those obtained with the modified autocorrelation (MAUT) method [149] of pitch extraction; (5) the TDPD algorithm is suitable for real-time implementation with currently available microprocessors because the numerical complexity of the algorithm is relatively low (the basic operations involved are additions and comparisons); and (6) the algorithm appears to be simpler than other time-domain pitch detection schemes[143,144] since only a small number of simple tests are needed in the TDPD algorithm to obtain the pitch estimate.

A series of experiments on the voiced portions of the speech waveform were performed with a view to effecting bit-rate reduction. The results of the experiments led to the development of a simple procedure for the analysis and reconstruction of voiced sounds. This procedure was used in the complete time coding scheme. It was found that in order to obtain good quality regeneration of voiced speech it is sufficient to use only the local maxima of the waveform contained in one pitch period and to recirculate this information to other pitch periods. The experiments also revealed that it is necessary to take small variations of pitch period (i.e., extraction of the amplitudes and locations of pitch peaks for every period) into account.

A simple method of clipped unvoiced speech regeneration was described in Chapter 4. The method was found to produce good quality regeneration of unvoiced speech and was used in the complete time coding scheme.

The results of the time coding technique described in Chapter 4 were presented in terms of waveforms of original and processed speech, spectrograms, and recorded utterances. It was observed that the waveform information is in general well preserved. The spectrograms revealed that for the first and second formant frequency range, the processed speech closely resembles the original speech. In addition, the important features of voiced-unvoiced discrimination and pitch period are well preserved. The spectrograms also revealed that the third formant region is, in most cases, either attenuated or missing. It is believed that this is due to the low-pass effect introduced by the smooth interpolation process.

It has been found that, for a bit-rate of 7.6 Kbits/sec, a good quality of decoded speech is produced by the time coding method described in Chapter 4. In informal listening tests, male and female speech coded with this method (at 7.6 Kbits/sec) were compared with the same speech material processed with the Hadamard technique (at 9.6 Kbits/sec). It was observed that more natural speech regeneration is obtained with the time coding scheme. In addition, unlike the results obtained with the Hadamard scheme, no significant differences were observed between male and female speech processed with the time coding technique.

It should be noted that although the time coding technique has not been refined to the point where it is possible to obtain speech of commercial quality (equivalent in quality to 56 Kbits/sec PCM speech),

it clearly shows promise at low-bit-rates (below 8 Kbits/sec). The time coding scheme is simpler than other schemes of intermediate complexity (such as adaptive predictive coding and adaptive transform coding) which also show promise at low-bit-rates. Although it is slightly more complex than the Hadamard coding technique, one must bear in mind the fact that the Hadamard coding technique is not successful[12] at bit-rates below 6 Kbits/sec. Furthermore, the results presented in Chapter 3 suggest that good quality speech cannot be obtained with the Hadamard technique at these low-bit-rates, even if special strategies are employed in the Hadamard domain. It is also important to note that potentially the time coding technique can operate at bit-rates below 1 bit per sample. This fact contrasts with waveform coding techniques such as PCM, DPCM, and DM, which are incapable of operating at such low rates. Even for the lower limit of 1 bit per sample, waveform coding techniques require more than 2 F bits/sec (F being the higher cut-off frequency of the speech signal), since it is necessary to sample at a rate above the Nyquist rate.

In addition to providing good performance at a bit-rate of 7.6 Kbits/sec the time coding method is computationally efficient. The estimates of numerical complexity given in Chapter 4 indicate that the most complex part of the scheme (analysis of voiced speech) can be implemented in real-time using currently available microprocessors.

One shortcoming of time coding in particular (and other variable-information-rate-source schemes of data compression in general) is that buffer storage at the transmitter and/or the receiver is necessary in order to transmit the data at a constant rate. Theoretical factors related to this problem and estimates of overall system delays were discussed in Chapter 5. When operating at a constant transmission bit-rate of 7.6 Kbits/sec, the maximum value of the estimated delays in the case of the proposed coding system was 0.58 sec. In the case of zero-crossings - based speech coding methods, using a constant rate of 900 zero-crossings per second for transmission, the maximum value of the estimated delays was 2 sec. It has also been found[14] that, when operating at a constant transmission bit-rate of 15 Kbits/sec, a delay of approximately 1 sec is required in the case of extremal coding of speech. Thus, the indications are that the delays involved in the time coding method described in Chapter 4 are significantly less than those involved in other methods of this kind such as "extremal coding"[14] and "time-encoded speech" type[18] methods.

The speech coding scheme presented in Chapter 4 belongs to a class of systems which have been generally termed 'time coders'. The time coding approach represents the speech signal as a sequence of features extracted directly from the speech waveform. Since the features correspond to simple time-domain measurements such as zero-crossings and extremes of the waveform, this approach is attractive because the required digital processing is simple to implement. In

addition, the features occur at a rate well below the Nyquist rate. For this reason, bit-rate reduction can be achieved. The new time coding scheme discussed in this thesis operates at 7.6 Kbits/sec, and retains most of the information which is important in preserving the intelligibility and naturalness of speech.

6.2. Suggestions for Further Research.

The adaptive transform coder (ATC), proposed by Zelinski and Noll[10] in 1977, is an interesting and important development in the field of low-bit-rate speech coding. The method utilises a discrete cosine transform (DCT), chosen on the basis of the minimised signal-to-noise ratio as compared with PCM.[10] This criterion, however, is not necessarily indicative of superior subjective performance of the DCT as compared to other transformations used in either the adaptive or non-adaptive modes. It would be interesting to compare the DCT with the Hadamard transform when used in adaptive transform coding. Since the Hadamard transform is simpler to implement, the use of the Hadamard transform in ATC would make the system simpler and more attractive. It may also be possible to obtain good performance of ATC (when the Hadamard transform is used), if the oversampling-smoothing technique presented in Chapter 3 is applied to the decoded speech.

In order to obtain a relatively simple speech coding system which gives good quality speech at bit-rates below 8 Kbits/sec the time coding approach appears to be an attractive choice. An interesting

area for future study would be the refinement of the new time coding method discussed in this thesis. For this purpose the investigation of various factos which may lead to a more efficient coding scheme is recommended. One such factor is the determination of the extent to which the errors introduced by an automatic pitch period detector are responsible for the audible noise accompanying the processed speech. The application of a non-linear smoothing technique[156] to the pitch contour produced by the TDPD algorithm described in Chapter 4 is one possibility which may lead to an improved performance of the new time coder discussed in this thesis. Another possibility would be the consideration of a more efficient description of the waveform intrapitch. The use of a better fitting technique of the waveform between local maxima with respect to the original speech may improve the performance in comparison with the use of a smooth interpolation. Since the zero-crossings rate of speech is, in general, considerably lower for the voiced portions of the waveform, a strategy of the " time-encoded speech" type[18] constitutes an interesting possiblity for coding the waveform intrapitch.
A further aspect worth considering is the use of more sophisticated techniques (such as differential or adaptive techniques) of encoding the pitch peak amplitudes and the local maxima of the waveform intrapitch. This may lead to significant reductions in transmission bit-rates, since the number of bits allocated to these parameters is a major fraction of the total bit-rate.

In the case of time coding techniques, the delays involved in the conversion of the variable-rate-source data into constant-transmission-rate data may sometimes be sufficiently long to impede two-way conversation. Thus, another interesting problem worth examination relates to the perceptual effects on the received speech of any distortion introduced by the use of delays shorter than the necessary minimum value to avoid distortion. The investigation of these effects for different speech coding schemes would provide useful information concerning the performance and applicability of the schemes.

REFERENCES

REFERENCES

1. HOUSE, A.S. : "On Vowel Duration in English",
J. Acoust. Soc. Amer., Vol. 33, No. 9, September, 1961,
pp. 1174-1178.

2. FANT, C.G. : "On the Predictability of Formant Levels and
Spectrum Envelopes from Formant Frequencies", in For Roman
Jakobson, Mouton and Co., The Hague, 1956, pp. 109-120.

3. FLANAGAN, J.L. : "Speech Analysis Synthesis and Perception",
Springer-Verlag, Berlin-Heidelberg - New York, 1972.

4. DUDLEY, H. : "Remaking Speech",
J. Acoust. Soc. Amer. Vol. 11, No. 2, October, 1939,
pp. 169-177.

5. DUDLEY, H. : "The Vocoder",
Bell. Lab. Record, Vol. 18, No. 4, December, 1939,
pp. 122-126.

6. GOLD, B. :"Digital Speech Networks",
Proc. IEEE, Vol. 65, No. 12, December, 1977,
pp. 1636-1658.

7. JAYANT, N.S. : "Digital Coding of Speech Waveforms : PCM, DPCM
and DM Quantizers",
Proc. IEEE, Vol. 62, No. 5, May, 1974, pp. 611-632.

8. CROCHIERE, R.E., WEBBER, S.A., and FLANAGAN, J. L. :
"Digital Coding of Speech in Sub-Bands",
Bell. Syst. Tech. J., Vol. 55, No. 8, October, 1976, pp. 1069-1085.

9. ATAL, B.S., and SCHROEDER, M.R. : "Adaptive Predictive Coding
of Speech Signals",
Bell. Syst. Tech. J., Vol. 49, No. 8, October, 1970, pp. 1973-1986.

10. ZELINSKI, R., and NOLL, P. : "Adaptive Transform Coding of Speech
Signals",
IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP - 25,
No. 4, August, 1977, pp. 299-309.

11. SHUM, Y.Y., ELLIOTT, A.R., and BROWN, O.W. :
"Speech Processing with Walsh-Hadamard Transforms",
IEEE Trans. on Audio and Electroac., Vol. AU - 21, No. 3,
June, 1973, pp. 174-179.

12. FRANGOULIS, E. : "Orthogonal Transform Methods of Speech Coding",
Ph.D. Thesis, Imperial College, London, 1978.

13. DAVID, Jr. E.E., and McDONALD, H.S. :"Note on Pitch-Synchronous Processing of Speech",
J. Acoust. Soc. Amer., Vol. 28, No. 7, November, 1956, pp. 1261-1266.

14. MATHEWS, M.V. : "Extremal Coding for Speech Transmission",
IRE Trans. on Inf. Theo., Vol. IT - 5, No. 3, September, 1959,
pp. 129-136.

15. AINSWORTH, W.A. : "Relative Intelligibility of Different Transforms of Clipped Speech",
J.Acoust. Soc. Amer., Vol. 41, No. 5, 1967, pp. 1272-1276.

16. STOVER, W.R. : "Time-Domain Bandwidth-Compression System",
J. Acoust.Soc. Amer.,Vol. 42, No. 2, 1967, pp. 348-359.

17. SOBOLEV, V.N., and TELEPNEV, V.N. : "Simple Methods of Clipped Speech Regeneration",
Telecommunications, Vol. 23, No. 3, 1969, pp. 37-44.

18. KING, R.A., and GOSLING, W. : "Time-Encoded Speech",
Electronics Letters, Vol. 14, No. 15, July, 1978, pp. 456-457.

19. JAYANT, N.S., and CHRISTENSEN, S.W. : "Adaptive Aperture Coding for Speech Waveforms - I",
Bell. Syst. Tech. J., Vol. 58, No. 7, September, 1979, pp. 1631-1645.

20. JAYANT, N.S., and CHRISTENSEN, S.W. : "Adaptive Aperture Coding for Speech Waveforms - II",
Bell Syst. Tech. J., Vol. 59, No. 3, March, 1980, pp. 471-477.

21. SCHROEDER, M.R. : "Vocoders : Analysis and Synthesis of Speech",
Proc. IEEE, Vol. 54, No. 5, May, 1966, pp. 720-734.

22. SCHAFER, R.W. : "A Survey of Digital Speech Processing Techniques",
IEEE Trans. on Audio and Electroac., Vol. AU - 20, No. 1,
March 1972, pp. 28-35.

23. FLANAGAN, J.L., and RABINER, L.R. (Eds.) : "Speech Synthesis",
Dowen, Hutchinson and Ross, Inc., Stroudsburg, Pennsylvania, 1973.

24. ELLIOTT, A.R. : "Voice Processing in the Sequency Domain",
Proc. Symp. Appl. Walsh Func., Washington, D.C.,April, 1973, pp. 10-14.

25. FLANAGAN, J.L., SCHROEDER, M.R., ATAL, B.S., CROCHIERE, R.E., JAYANT, N.S., and TRIBOLET, J.M. : "Speech Coding",
IEEE.Trans. on Comm., Vol. COM - 27, No. 4, April, 1979, pp. 710-736.

26. OLIVER, B.M., PIERCE, J.R., and SHANNON, C.E. :
"The Philosophy of PCM",
Proc. IRE, Vol. 36, No. 11, November, 1948, pp. 1324-1331.

27. DAVENPORT, W.B., Jr. : "An Experimental Study of Speech Wave Probability Distributions",
J. Acoust. Soc. Amer., Vol. 24, No. 4, July, 1952, pp. 390-399.

28. PANTER, P.F., and DITE, W. : "Quantisation Distortion in Pulse-Count Modulation with Non-Uniform Spacing of Levels", Proc. IRE, Vol. 39, No. 1, January, 1951, pp. 44-48.

29. SMITH, B. : "Instantaneous Companding of Quantized Signals", Bell Syst. Tech. J., Vol. 36, No. 3, May, 1957, pp. 653-709.

30. MAX, J. : "Quantizing for Minimum Distortion", IRE Trans. on Inf. Theo., Vol. IT - 6, No. 1, March 1960, pp. 7-12.

31. WOOD, R.C. : "On Optimum Quantization", IEEE Trans. on Inf. Theo., Vol. IT - 15, No. 2, March 1969, pp. 248-252.

32. GERSHO, A., and GOODMAN, D.J. : "A Training Mode Adaptive Quantizer". IEEE Trans. on Inf. Theo., Vol IT - 20, No. 6, November, 1974, pp. 746-749.

33. ALCAIM, A., and ALBUQUERQUE, J.P.A. : "Accelerated Learning Algorithms in Adaptive Quantization", Proc. 1978 IEEE Intern. Conf. on Commun., Vol. 2, Toronto, Canada, June, 1978, pp. 30.5.1 to 30.5.4.

34. WILKINSON, R.M. : "An Adaptive Pulse Code Modulator for Speech", Proc. 1971 IEEE Intern. Conf. on Commun., Montreal, Canada, June, 1971, pp. 1-11 to 1-15.

35. JAYANT, N.S. : "Adaptive Quantization with a One-Word Memory", Bell Syst. Tech. J., Vol. 52, No. 7, September , 1973, pp. 1119-1144.

36. McDONALD, R.A. : "Signal-to-Noise and Idle Channel Performance of Differential Pulse Code Modulation Systems - Particular Applications to Voice Signals ", Bell Syst. Tech. J., Vol. 45, No. 7, September, 1966, pp. 1123-1151.

37. O'NEAL, J.B., Jr. and STROH, R.W. : "Differential PCM for Speech and Data Signals", IEEE Trans. on Commun., Vol. COM - 20, No. 5, October, 1972, pp. 900-912.

38. PAPOULIS, A. : "Probability, Random Variables, and Stochastic Processes", McGraw-Hill Kogakusha, Ltd., 1965.

39. CUMMISKEY, P., JAYANT, N.S., and FLANAGAN, J.L. : "Adaptive Quantization in Differential PCM Coding of Speech", Bell Syst. Tech. J., Vol. 52, No.7, September, 1973, pp. 1105-1118.

40. JAGER, F. de : "Deltamodulation, a Method of PCM Transmission Using the 1-Unit Code", Philips Res. Rep., Vol.7, No. 6, December, 1952, pp. 442-466.

41. STEELE, R. : "Delta Modulation Systems", Pentech Press, London, 1975.

42. ABATE, J.E. : "Linear and Adaptive Delta Modulation", Proc. IEEE, Vol. 55, No. 3, March, 1967, pp. 298-308.

43. GREEFKES, J.A., and JAGER, F. de : "Continuous Delta Modulation", Philips Res. Rep., Vol. 23, No. 2, April, 1968, pp. 233-246.

44. GREEFKES, J.A., and RIEMENS. K. : "Code Modulation with Digitally Controlled Companding for Speech Transmission", Philips Tech. Rev., Vol. 31, No. 11/12, 1970, pp. 335-353.

45. JAYANT, N.S. : "Adaptive Delta Modulation with a One-Bit Memory", Bell Syst. Tech. J., Vol. 49, No. 3, March, 1970, pp. 321-342.

46. JAYANT, N.S.,and ROSENBERG, A.E. : "The Preference of Slope Overload to Granularity in the Delta Modulation of Speech", Bell Syst. Tech. J., Vol. 50, No. 10, December, 1971, pp. 3117-3125.

47. TRIBOLET, J.M., NOLL, P., McDERMOTT, B.J., and CROCHIERE, R.E., : "A Comparison of the Performance of Four Low-Bit-Rate Speech Waveform Coders", Bell. Syst. Tech. J., Vol. 58, No. 3, March, 1979, pp. 609-712.

48. TRIBOLET, J.M., and CROCHIERE, R.E., : "Frequency Domain Coding of Speech", IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP - 27, No. 5, October, 1979, pp. 512-530.

49. ATAL, B.S. and SCHROEDER, M.R. : "Predictive Coding of Speech Signals and Subjective Error Criteria", IEEE Trans. on Acoust., Speech and Sig. Proc., Vol. ASSP - 27, No. 3, June, 1979, pp. 247-254.

50. BOGERT, B.P. : "The Vobanc - A Two-to-One Speech Band-Width Reduction System", J. Acoust. Soc. Amer., Vol. 28, No. 3, May, 1956, pp. 399-404.

51. SCHROEDER, M.R., FLANAGAN, J.L., and LUNDRY, E.A. : "Bandwidth Compression of Speech by Analytic-Signal Rooting", Proc. IEEE, Vol. 55, No. 3, March, 1967, pp. 396-401.

52. VILBIG, F., and HAASE, K.H. : "Some Systems for Speech-Band Compression", J. Acoust. Soc. Amer., Vol. 28, No. 4, July, 1956, pp. 573-577.

53. DAVID, Jr., E.E. :"Naturalness and Distortion in Speech-Processing Devices", J. Acoust. Soc. Amer., Vol. 28, No. 4, July, 1956, pp. 586-589.

54. YAGGI, Jr., L.A.,and MASON, Jr., A.E. : "Polymodal Vocoder : A New Approach to Versatile and Reliable Voice Communications", J. Acoust. Soc. Amer., Vol. 35, No. 5, May, 1963, p. 806.

55. GOLD, B., and RADER, C.M. : "The Channel Vocoder", IEEE Trans. on Audio and Electroac., Vol. AU - 15, No. 4, December, 1967, pp. 148-161.

56. TIERNEY, J. : "Channel Vocoder with Digital-Pitch Extraction", J. Acoust. Soc. Amer., Vol. 35, No. 5, May, 1963, p. 805.

57. TIERNEY, J., GOLD, B., SFERRINO, V., DUMANIAN, J.A., and AHO, E. : "Channel Vocoder with Digital Pitch Extractor ", J. Acoust. Soc. Amer., Vol. 36, No. 10, October, 1964, pp. 1901-1905.

58. PETERSON, E., and COOPER, F.S. : "Peakpicker : A Band-Width Compression Device", J. Acoust. Soc. Amer., Vol. 29, No. 6, June, 1957, p. 777.

59. KRAMER, H.P., and Mathews, M.V., : "A Linear Coding for Transmitting a Set of Correlated Signals", Symp. on Inf. Theo., MIT, Cambridge, Massachusetts, September, 1956, pp. 41-46.

60. CROWTHER, W.R., and RADER, C.M. : "Efficient Coding of Vocoder Channel Signals Using Linear Transformation", Proc. IEEE, Vol. 54, No. 11, November, 1966, pp. 1594-1595.

61. SMITH, C.P. : "Vocal-Response Synthesizer", J. Acoust. Soc. Amer., Vol. 37, No. 1, January, 1965, pp. 170-171.

62. SCHAFER, R.W., RABINER, L.R., and HERMANN, O. : "Uniform and Non-Uniform Digital Filter Banks for Speech Analysis", 8th Int. Congress on Acoust., London, 1974, p. 242.

63. MUNSON, W.A., and MONTGOMERY, H.C. : "A Speech Analyzer and Synthesizer", J. Acoust. Soc. Amer., Vol. 22, No. 5, September, 1950, p. 678.

64. LAWRENCE, W. : "The Synthesis of Speech from Signals which Have a Low Information Rate", in Communication Theory, Butterworth and Co., Ltd., 1953, pp. 460-469.

65. FLANAGAN, J.L. : "Note on the Design of 'Terminal-Analog' Speech Synthesizers", J. Acoust. Soc. Amer., Vol. 29, No. 2, February, 1957, pp. 306-310.

66. FLANAGAN, J.L., COKER, C.H., and BIRD, C.M. : "Computer Simulation of a Formant-Vocoder Synthesizer", 15th Ann. Meeting Audio Eng. Soc., Preprint 307, 1963.

67. RABINER, L.R. : "Digital-Formant Synthesizer for Speech - Synthesis Studies", J. Acoust. Soc. Amer., Vol. 43, No. 4, 1968, pp. 822-828.

68. RABINER, L.R., JACKSON, L.B., SCHAFER, R.W., and COKER, C.H. : "A Hardware Realization of a Digital Formant Speech Synthesizer", IEEE Trans. on Comm. Tech., Vol. COM - 19, No. 6, December, 1971, pp. 1016-1020.

69. HOLMES, J.N. : "The Influence of Glottal Waveform on the Naturalness of Speech from a Parallel Formant Synthesizer", IEEE Trans. on Audio and Electroac., Vol. AU - 20, No. 3, June, 1973, pp. 298-305.

70. FLANAGAN, J.L. : "Automatic Extraction of Formant Frequencies from Continuous Speech", J. Acoust. Soc. Amer., Vol. 28, No. 1, January, 1956, pp. 110-118.

71. FLANAGAN, J.L. : "Evaluation of Two Formant-Extracting Devices", J. Acoust. Soc. Amer., Vol. 28, No. 1, January, 1956, pp. 118-125.

72. COKER, C.H. : "Computer-Simulated Analyzer for a Formant Vocoder", J. Acoust. Soc. Amer., Vol. 35, No. 11, November, 1963, pp. 1911-1912.

73. MARKEL, J.D. : "Digital Inverse Filtering - A New Tool for Formant Trajectory Estimation", IEEE Trans. on Audio and Electroac., Vol. AU - 20, No. 2, June, 1972, pp. 129-137.

74. STEAD, E.G., and WESTON, R.C. : "Sampling and Quantizing the Parameters of a Formant-Tracking Vocoder System", Proc. Speech Comm. Seminar, Stockholm, Vol. 2, August 29-September 1, 1962, paper G9.

75. COKER, C.H. : "Real-Time Formant Vocoder, Using a Filter Bank, a General Purpose Digital Computer, and an Analog Synthesizer", J. Acoust. Soc. Amer., Vol. 38, 1965, p. 940.

76. MAKHOUL, J. : "Linear Prediction : A Tutorial Review", Proc. IEEE, Vol. 63, No. 4, April, 1975, pp. 561-580.

77. SORENSON, H.W. : "Least-Squares Estimation : From Gauss to Kalman", IEEE Spectrum, Vol. 7, No. 7, July, 1970, pp. 63-68.

78. MARKEL, J.D., and GRAY, A.H. : "Linear Prediction of Speech", Springer-Verlag, Berlin-Heidelberg - New York, 1976.

79. ITAKURA, F., and SAITO, S. : "Analysis Synthesis Telephony Based on the Maximum Likelihood Method", Proc. 6th Int. Congr. on Acoust.,Tokyo, August, 1968, Paper C-5-5, pp. C17-20.

80. ITAKURA, F., and SAITO, S :"On the Optimum Quantization of Feature Parameters in the PARCOR Speech Synthesizer", Proc. IEEE Conf. on Speech, Comm., and Process., New York, 1972, pp. 434-437.

81. ATAL, B.S., and HANAUER, S.L. : "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", J. Acoust. Soc. Amer., Vol. 50, No. 2 (Part 2), 1971, pp. 637-655.

82. HASKEW, J.R., KELLY, J.M., KELLY, Jr., R.M., and McKINNEY, T.H. : "Results of a Study of the Linear Prediction Vocoder", IEEE Trans. on Comm., Vol. COM - 21, No. 9, September, 1973, pp. 1008-1014.

83. SAMBUR, M.R., ROSENBERG, A.E., RABINER, L.R., and McGONEGAL, C.A. : "On Reducing the Buzz in LPC Synthesis", J. Acoust. Soc. Amer., Vol. 63, No. 3, March, 1978, pp. 918-924.

84. SAMBUR, M.R. : "An Efficient Linear-Prediction Vocoder", Bell Syst. Tech. J., Vol. 54, No. 10, December, 1975, pp. 1693-1723.

85. KELLER, L.O., CLEMENT, G.L., STRONG, W.J., and PALMER, E.P. : "Two Preliminary Studies of the Intelligibility of Predictor-Coefficient and Formant Coded Speech", IEEE Trans. on Acoust. Speech, and Sig. Proc. Vol. ASSP - 24, No. 5, October, 1976, pp. 429-432.

86. WONG, D.Y., and MARKEL, J.D. : "An Intelligibility Evaluation of Several Linear Prediction Vocoder Modifications", IEEE Trans. on Acoust., Speech, and Sig. Proc., Vol. ASSP - 26, No. 5, October, 1978, p. 424-435.

87. DAVID, Jr., E.E., SCHROEDER, M.R., LOGAN, B.F., and PRESTIGIACOMO, A.J.: "New Applications of Voice-Excitation to Vocoders", Proc. Speech Comm. Seminar, Stockholm, Vol. 2, August 29-September 1, 1962, paper G3.

88. GOLD, B., and TIERNEY, J. : "Digitized Voice-Excited Vocoder for Telephone - Quality Inputs, Using Bandpass Sampling of the Baseband Signal", J. Acoust. Soc. Amer., Vol. 37, No. 4, April, 1965, pp. 753-754.

89. FLANAGAN, J.L. : "A Resonance-Vocoder and Baseband Complement: A Hybrid System for Speech Transmission", IRE Trans. on Audio, Vol. AU - 8, No. 3, May-June, 1960, pp. 95-102.

90. WEINSTEIN, C.J. : "A Linear Prediction Vocoder with Voice Excitation", IEEE Electronics and Aerospace Systems Convention, EASCON' 75, Washington, D.C., 1975, pp. 30A - 30G.

91. OPPENHIEM, A.V., and SCHAFER, R.W. : "Homomorphic Analysis of Speech", IEEE Trans. on Audio and Electroac., Vol. AU - 16, No. 2, June, 1968, pp. 221-226.

92. OPPENHEIM, A.V., SCHAFER, R.W., and STOCKHAM, Jr. T.G. : "Nonlinear Filtering of Multiplied and Convolved Signals", Proc. IEEE, Vol. 56, No. 8, August, 1968, pp. 1264-1291.

93. OPPENHEIM, A.V. : "Speech Analysis - Synthesis Systems Based on Homomorphic Filtering", J. Acoust. Soc., Amer., Vol. 45, No. 2, 1969, pp. 458-465.

94. WEINSTEIN, C.J., and OPPENHEIM, A.V. : "Predictive Coding in a Homomorphic Vocoder", IEEE Trans. on Audio and Electroac., Vol. AU - 19, No. 3, September, 1971, pp. 243-248.

95. DUDLEY, H. : "Phonetic Pattern Recognition Vocoder for Narrow-Band Speech Transmission", J. Acoust. Soc. Amer., Vol. 30, No. 8, August, 1958, pp. 733-739.

96. DUDLEY, H., and BALASHEK, S. : "Automatic Recognition of Phonetic Patterns in Speech", J. Acoust. Soc. Amer., Vol. 30, No. 8, August, 1958, pp. 721-732.

97. SMITH, C.P. : "Voice-Communications Method Using Pattern Matching for Data Compression", J. Acoust. Soc. Amer., Vol. 35, No. 5, May, 1963, p. 805.

98. SMITH, C.P. : "Perception of Vocoder Speech Processed by Pattern Matching", J. Acoust. Soc. Amer., Vol. 46, No. 6 (Part 2), 1969, pp. 1562-1571.

99. FANO, R.M. : "Short-Time Autocorrelation Functions and Power Spectra", J. Acoust. Soc. Amer., Vol. 22, No. 5, Septbember, 1950, pp. 546-550.

100. BIDDULPH, R. : "Short-Term Autocorrelation Analysis and Correlatograms of Spoken Digits", J. Acoust. Soc. Amer., Vol. 26, No. 4, July, 1954, pp. 539-541.

101. HUGGINS, W.H. : "A Note on Autocorrelation Analysis of Speech Sounds", J. Acoust. Soc. Amer., Vol. 26, No. 5, September, 1954, pp. 790-792.

102. SCHROEDER, M.R. : "New Approach to Time Domain Analysis and Synthesis of Speech", J. Acoust. Soc. Amer., Vol. 31, No. 6, June, 1959, pp. 852-853.

103. SCHROEDER, M.R., and CRYSTAL, T.H. : "Auto-Correlation Vocoder", J. Acoust. Soc. Amer., Vol. 32, No. 7, July, 1960, p. 913.

104. SCHROEDER, M.R. : "Correlation Techniques for Speech Bandwidth Compression", J. Audio Eng. Soc., Vol. 10, No. 2, April, 1962, pp. 163-166.

105. PIROGOV, A.A. : "A Harmonic System for Compressing Speech Spectra", Telecommunications, No. 3, 1959 , pp. 229-242.

106. KULYA, V.I. : "Application of Laguerre Functions to Parametric Coding of Speech Signals", Telecommunications, No. 7, 1962, pp. 34-42.

107. KULYA, V.I. : "Analysis of a Chebyshev-Type Vocoder", Telecommunications, Vol. 17, No. 3, March, 1963, pp. 23-33.

108. FLANAGAN, J.L. and GOLDEN, R.M. : "Phase Vocoder", Bell Syst. Tech. J., Vol. 45, No. 9, November, 1966, pp. 1493-1509.

109. PORTNOFF, M.R. : "Implementation of the Digital Phase Vocoder Using the Fast Fourier Transform", IEEE Trans. on Acoust. Speech, and Sig. Proc., Vol. ASSP - 24, No. 3, June, 1976, pp. 243-248.

110. BÖßWETTER, C. : "Analog Sequency Analysis and Synthesis of Voice Signals", Proc. Symp. Appl. Walsh Func., Washington, D.C., 1970, pp. 220-229.

111. CAMPANELLA, S.J., and ROBINSON, G.S. : "Digital Sequency Decomposition of Voice Signals", Proc. Symp. Appl. Walsh Func., Washington, D.C., 1970, pp. 230-237.

112. CAMPANELLA, S.J., and ROBINSON, G.S. : "A Comparison of Orthogonal Transformations for Digital Speech Processing" IEEE Trans. on Comm. Tech., Vol. COM - 19, No. 6, December, 1971, pp. 1045-1049.

113. CAMPANELLA, S.J., and ROBINSON, G.S. : "A Comparison of Walsh and Fourier Transformations for Application to Speech", Proc. Symp. Appl. Walsh Func., April, 1971, pp. 199-202.

114. GETHÖFFER, H. : "Speech Processing with Walsh Functions", Proc. Symp. Appl. Walsh Func., March, 1972, pp. 163-168.

115. LACKEY, R.B. : "The Wonderful World of Walsh Functions", Proc. Symp. Appl. Walsh. Func., March, 1972, pp. 2-7.

116. GETHÖFFER, H. : "Transform Coding of Speech Using Walsh Functions", Proc. Symp. Appl. Walsh Func., April, 1973, pp. 194-201.

117. FRANGOULIS, E., and TURNER, L.F. : "Hadamard-Transformation Technique of Speech Coding : Some Further Results", Proc. IEE , Vol. 124, No. 10, October, 1977, pp. 845-852.

118. FRANGOULIS, E., and TURNER, L.F. : "Speech Coding by the Hadamard-Haar Transform", IEE J. on Computers and Digital Techniques, Vol. 1, No. 3, August, 1978, pp. 106-112.

119. TRIBOLET, J.M., and CROCHIERE, R.E. : "A Vocoder-Driven Adaptation Strategy for Low-Bit Rate Adaptive Transform Coding of Speech", Proc. Int. Conf. on Digital Sig. Proc., Florence, Italy, September, 1978, pp. 638-642.

120. ZELINSKI, R., and NOLL, P. : "Approaches to Adaptive Transform Speech Coding at Low Bit Rates", IEEE Trans. on Acoust. Speech, and Sig. Proc.,Vol. ASSP - 27, No. 1 , February, 1979, pp. 89-95.

121. MANLEY, H.J. : "Analysis-Synthesis of Connected Speech in Terms of Orthogonalized Exponentially Damped Sinusoids", J. Acoust. Soc. Amer., Vol. 35, No. 4, April, 1963, pp. 464-474.

122. HARMUTH, H.F. : "Transmission of Information by Orthogonal Functions", Springer-Verlag, Berlin-Heidelberg-New York, 1972.

123. AHMED, N., and RAO, K.R. : "Orthogonal Transforms for Digital Signal Processing", Springer-Verlag, Berlin-Heidelberg - New York, 1975.

124. HUANG, J.J.Y., and SCHULTHEISS, P. : "Block Quantization of Correlated Gaussian Random Variables", IEEE Trans. on Comm. Syst., Vol. CS - 11, No. 3, September, 1963, pp. 289-296.

125. NITADORI, K. : "Linear Transformation Coding and Predictive Coding - Two Methods of Digital Encoding for Continuous Sources with Discrete Parameters", Electr. Comm. Japan, Vol. 53 - A, No. 2, 1970, pp. 37-45.

126. WALSH, J.L. : "A Closed Set of Normal Orthogonal Functions", American Journal of Mathematics, Vol. 45, 1923, pp. 5-24.

127. KENNETT, B.L.N.: "A Note on the Finite Walsh Transform", IEEE Trans. on Inf. Theo., Vol. IT -16, No. 4, July, 1970, pp. 489-491.

128. TURNER, L.F. : "Some Properties of the Hadamard Transformation Relating to its Application in Data Compression", IERE Conf. on Sig. Proc. in Commun. Syst., October, 1974.

129. SHUM, Y.Y., and ELLIOTT, A.R. : "Computation of the Fast Hadamard Transform", Proc. Symp. Appl. Walsh.Func., Washington, D.C., March, 1972, pp. 177-179.

130. BERAUER, G. : "Fast 'In Place' Computation of the Discrete Walsh Transform in Sequency Order", Proc. Symp. Appl. Walsh Functions, Washington, D.C., March 1972, pp. 272-275.

131. ELLIOTT, A.R., and SHUM, Y.Y. : "A Parallel Array Hardware Implementation of the Fast Hadamard and Walsh Transforms", Proc. Symp. Appl. Walsh Func., Washington, D.C., March, 1972, pp. 181-183.

132. AHMED, N., NATARAJAN, T., and RAO, K.R. : "Discrete Cosine Transform", IEEE Trans. on Computers, Vol. C - 23, No. 1, January, 1974, pp. 90-93.

133. HARALICK, R.M. : "A Storage Efficient Way to Implement the Discrete Cosine Transform", IEEE Trans. on Computers, Vol. C - 25, No. 7, July, 1976, pp. 764-765.

134. RABINER, L.R., and GOLD, B. : "Theory and Application of Digital Signal Processing", Prentice-Hall, Englewood Cliffs, N.J., 1975.

135. SHORE, J.E. : "On the Application of Haar Functions", IEEE Trans. on Comm., Vol. COM - 21, No. 3, March, 1973, pp. 209-216.

136. LICKLIDER, J.C.R. : "Effects of Amplitude Distortion upon the Intelligibility of Speech", J. Acoust. Soc. Amer., Vol. 18, No. 2, October, 1946, pp. 429-434.

137. LICKLIDER, J.C.R. and POLLACK, I.: " Effects of Differentiation, Integration, and Infinite Peak Clipping upon the Intelligibility of Speech", J. Acoust. Soc.Amer., Vol. 20,No.1, January, 1948, pp. 42-51.

138. LICKLIDER, J.C.R. : "The Intelligibility of Amplitude-Dichotomized, Time-Quantized Speech Waves", J. Acoust. Soc. Amer., Vol. 22, No. 6, November, 1950, pp. 820-823.

139. KOCK, W.E. : "Audio Problems in Space", IEEE Trans. on Audio Vol. AU - 13, No. 6, November/December, 1965, pp. 132-134.

140. SPOGEN, L.R., SHAVER, H.N., BAKER, H.N., and BLOM, B.V. : "Speech Processing by the Selective Amplitude Sampling System", J. Acoust. Soc. Amer., Vol. 32, No. 12, December, 1960, pp. 1621-1625.

141. DOLANSKÝ, L.O. : "An Instantaneous Pitch-Period Indicator", J. Acoust. Soc. Amer., Vol. 27, No. 1, January, 1955, pp. 67-72.

142. GOLD, B. : "Computer Program for Pitch Extraction", J. Acoust. Soc. Amer.,Vol. 34, No. 7, July, 1962, pp. 916-921.

143. GOLD, B., and RABINER, L. : "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain", J. Acoust. Soc. Amer., Vol. 46, No. 2 (Part 2), 1969, pp. 442-448.

144. MILLER, N.J. : "Pitch Detection by Data Reduction", IEEE Trans. on Acoust. Speech, and Sig. Proc., Vol. ASSP - 23, No. 1, February, 1975, pp. 72-79.

145. GEÇKINLI, N.C., and YAVUZ, D. : "Algorithm for Pitch Extraction Using Zero-Crossing Interval Sequence", IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP - 25, No. 6, December, 1977, pp. 559-564.

146. NOLL, A.M. : "Short-Time Spectrum and 'Cepstrum' Techniques for Vocal-Pitch Detection", J. Acoust. Soc. Amer., Vol. 36, No. 2, February, 1964, pp. 296-302.

147. NOLL, A.M. : "Cepstrum Pitch Determination", J. Acoust. Soc. Amer., Vol. 41, No. 2, 1967, pp. 293-309.

148. SONDHI, M.M. : "New Methods of Pitch Extraction", IEEE Trans. on Audio and Electroac., Vol. AU - 16, No. 2, June, 1968, pp. 262-266.

149. DUBNOWSKI, J.J., SCHAFER, R.W., and RABINER, L.R. : "Real-Time Digital Hardware Pitch Detector", IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP - 24, No. 1, February, 1976, pp. 2-8.

150. MARKEL, J.D. : "The SIFT Algorithm for Fundamental Frequency Estimation", IEEE Trans. on Audio and Electroac.,Vol. AU - 20, No. 5, December, 1972, pp. 367-377.

151. ROSS, M.J., SHAFFER, H.L., COHEN, A., FREUDBERG, R., and MANLEY, H.L. : "Average Magnitude Difference Function Pitch Extractor ", IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP - 22, No. 5, October, 1974, pp. 353-362.

152. STEIGLITZ, K., WINHAM, G., and PETZINGER, J. : "Pitch Extraction by Trigonometric Curve Fitting", IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP - 23, No. 3, June, 1975, pp. 321-323.

153. WISE, J.D., CAPRIO, J.R., and PARKS, T.W. : "Maximum Likelihood Pitch Estimation", IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP - 24, No. 5, October, 1976, pp. 418-423.

154. AMBIKAIRAJAH, E., CAREY, M.J., and TATTERSALL, G. : "Estimating the Pitch Period of Voiced Speech", Electronics Letters, Vol. 16, No. 12, June, 1980, pp. 464-466.

155. McGONEGAL, C.A., RABINER, L.R., and ROSENBERG, A.E. : "A Semiautomatic Pitch Detector (SAPD)", IEEE Trans. on Acoust., Speech, and Sig. Proc., Vol. ASSP - 23, No. 6, December, 1975, pp. 570-574.

156. RABINER, L.R., SAMBUR, M.R., and SCHMIDT, C.E. : "Applications of a Nonlinear Smoothing Algorithm to Speech Processing", IEEE Trans. on Acoust., Speech, and Sig. Proc., Vol. ASSP - 23, No. 6, December, 1975, pp. 552-557.

157. RABINER, L.R., CHENG, M.J., ROSENBERG, A.E., and McGONEGAL, C.A. : "A Comparative Performance Study of Several Pitch Detection Algorithms", IEEE Trans. on Acoust. Speech, and Sig. Proc., Vol. ASSP - 24, No. 5, October, 1976, pp. 399-418.

158. McGONEGAL, C.A., RABINER, L.R., and ROSENBERG, A.E. : "A Subjective Evaluation of Pitch Detection Methods Using LPC Synthesized Speech", IEEE Trans. on Acoust. Speech, and Sig. Proc., Vol. ASSP - 25, No. 3, June, 1977, pp. 221-229.

159. ATAL, B.S., and RABINER, L.R. : "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition", IEEE Trans. on Acoust. Speech, and Sig. Proc.,Vol. ASSP - 24, No. 3, June, 1976, pp. 201-212.

160. MORRIS, L.R. : "The Role of Zero Crossings in Speech Recognition and Processing", Ph.D. Thesis, Imperial College, London, 1970.

161. HEERING, J. : "The Intel 8086, the Zilog Z8000, and the Motorola MC 68000 Microprocessors", Int. J. of Microprocessing and Microprogramming, Vol. 6, No. 3, May, 1980.

162. MC68000 16-Bit Microprocessor User's Manual, Motorola Inc., Second Edition, January, 1980.

APPENDIX

APPENDIX

DESCRIPTION OF ATTACHED RECORDED MATERIAL

The utterances included on the attached recorded material will be referred to in this description by numbers, as follows:

1 - " The box was thrown beside the parked truck"
2 - " Glue the sheet to the dark blue background"
3 - " These days a chicken leg is a rare dish"
4 - " It's easy to tell the depth of the well"
5 - "Rice is often served in round bowls"
6 - "The juice of lemons makes fine punch".

Section 1. Comparison of Double Hadamard coded speech at 4.625 Kbits/sec with Hadamard and Hadamard-Haar techniques at 4.875 Kbits/sec.

(a) Hadamard
(b) Hadamard-Haar
(c) Double Hadamard.

Utterances 1, and 2, male speaker.

Section 2. Comparison of Double Hadamard coded speech at 3.5 Kbits/sec with Hadamard and Hadamard-Haar techniques at 3.25 Kbits/sec.

(a) Hadamard
(b) Hadamard-Haar
(c) Double Hadamard

Utterances 1, and 2, male speaker

Section 3. Double Hadamard coded speech at 5.75 Kbits/sec. Utterances 1,2,3,4,5, and 6, male speaker and utterances 1,2, and 3, female speaker.

Section 4. Effect of output oversampling the basic Hadamard transform coded speech using 3 dominant coefficients and 7 bits/coefficient amplitude.
(a) Basic Hadamard
(b) Introduction of output oversampling.
Utterances 1, 2, and 3, male and female speakers.

Section 5. Effect of output oversampling the basic Hadamard transform coded speech using 4 dominant coefficients and 4 bits/coefficient amplitude.
(a) Basic Hadamard
(b) Introduction of output oversampling.
Utterances 1, 2, and 3, male and female speakers.

Section 6. Effect of oversampling and smoothing the basic Hadamard transform coded speech using 3 dominant coefficients and 7 bits/coefficient amplitude.
(a) Basic Hadamard
(b) Introduction of oversampling and smoothing.
Utterances 1, 2, and 3, male and female speakers.

Section 7. Effect of oversampling and smoothing the basic Hadamard transform coded speech using 4 dominant coefficients and 4 bits/coefficient amplitude.
(a) Basic Hadamard
(b) Introduction of oversampling and smoothing.
Utterances 1, 2, and 3, male and female speakers.

Section 8. Pitch-synchronous time-domain processing of voiced speech.
Experiment I,
(a) Original speech
(b) Processed speech.
Utterances 1, 2, and 3, male speaker, and utterance 1, female speaker.

Section 9. Pitch-synchronous time-domain processing of voiced speech. Experiment II. Utterance 1, male speaker, and utterance 3, female speaker.

Section 10. Pitch-synchronous time-domain processing of voiced speech. Experiment III.

(a) D = max, I = 2.5 msec,

(b) D = ave, I = 2.5 msec,

(c) D = ave, I = 3 msec

Utterance 1, male speaker.

Section 11. Pitch-synchronous time-domain processing of voiced speech. Experiment IV.

(a) Using extremes of the waveform intrapitch.

(b) Using local maxima of the waveform intrapitch.

(c) Using the absolute extreme between each pair of adjacent zero-crossings.

Utterances 1, 2, and 3, male speaker.

Section 12. Pitch-synchronous time-domain processing of voiced speech. Experiment V.

(a) Original speech

(b) Processed speech.

Utterances 1, 2, and 3, male speaker.

Section 13. Clipped unvoiced speech regeneration.

(a) /ʃ/ from sheet

(b) /f/ from fine

(c) /s/ from served

(d) /p/ from parked

(e) /t/ from to.

Original speech followed by processed speech.

Section 14. Proposed time coding technique at a bit-rate of 7.6 Kbits/sec.
Utterances 1, 2, 3, 4, 5, and 6, male and female speakers.

Section 15. Comparison of the proposed time coding technique at 7.6 Kbits/sec with 64-point Hadamard coding at 9.625 Kbits/sec (using 8 dominant coefficients and a special technique for labelling 7 of these coefficients).

(a) Time coding
(b) Hadamard coding.

Utterances 1, and 2, male and female speakers.