FRACTAL BASED SPEECH RECOGNITION AND SYNTHESIS

By

Souhila Fekkai

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
AT
DE MONTFORT UNIVERSITY
LEICESTER, UNITED KINGDOM
OCTOBER 2002

DE MONTFORT UNIVERSITY
DEPARTMENT OF
COMPUTING SCIENCE AND ENGINEERING

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled " **Fractal Based Speech Recognition and Synthesis** " by **Souhila Fekkai** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy**.

Dated: October 2002

External Examiner: ______________________________
Prof. Costas Xydeas

Research Supervisor: ______________________________
Prof. Marwan Al-Akaidi

Examing Committee: ______________________________
Dr. Amar Aggoun

____

DE MONTFORT UNIVERSITY

Date: **October 2002**

Author: **Souhila Fekkai**

Title: **Fractal Based Speech Recognition and Synthesis**

Faculty: **Computing Science and Engineering**

Degree: **Ph.D.** Convocation: **June** Year: **2002**

Permission is herewith granted to De Montfort University to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

Signature of Author

To

My dearest father

My lovely and adorable mother

To Those who reside in my heart forever...

Table of Contents

List of Tables

List of Figures

Glossary of Symbols and Notations

TIMIT	Texas Instrument Massachusetts Institute of technology
$A(t)$	Amplitude envelope
CEM	Critical Exponent Method
D	Fractal Dimension
DARPA-ISTO	Defence Advanced Research Projects Agency-Information Science and Technology Office
DTMF	Dual Tone Multi-Frequency
EEG	electroencephalogram
f	Activation function
F_0	Fundamental Frequency
$F(t)$	Fractal Signal
$H(t)$	Hilbert Transform
IFS	Iterated Function System
KNN	K-nearest Neighbor
LFCC	Linear Frequency Cepstral Coefficient
MFCC	Mel Frequency Cepstral Coefficient
MIDI	Musical Instrument Digital Interface
NIST	National Institute of Standards and Technology
PSM	Power Spectrum Method
SI	Speaker Independent
$S_o(t)$	Synthetic Speech Word
β	Power Exponent

ϕ	phase of the speech signal
ϕ'	unwrap phase of the speech signal
w_{ji}^{k}	Connection Weight

Abstract

Transmitting a linguistic message is most often the primary purpose of speech communication and the recognition of this message by machine that would be most useful.

This research consists of two major parts. The first part presents a novel and promising approach for estimating the degree of recognition of speech phonemes and makes use of a new set of features based fractals. The main methods of computing the fractal dimension of speech signals are reviewed and a new speaker-independent speech recognition system developed at De Montfort University is described in detail. Finally, a Least Square Method as well as a novel Neural Network algorithm is employed to derive the recognition performance of the speech data.

The second part of this work studies the synthesis of speech words, which is based mainly on the fractal dimension to create natural sounding speech. The work shows that by careful use of the fractal dimension together with the phase of the speech signal to ensure consistent intonation contours, natural-sounding speech synthesis is achievable with word level speech. In order to extend the flexibility of this framework, we focused on the filtering and the compression of the phase to maintain and produce natural sounding speech. A 'naturalness level' is achieved as a result of the fractal characteristic used in the synthesis process. Finally, a novel speech synthesis system based on fractals developed at De Montfort University is discussed.

Throughout our research simulation experiments were performed on continuous speech data available from the Texas Instrument Massachusetts Institute of technology (TIMIT) database, which is designed to provide the speech research community with a standarised corpus for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition system.

Acknowledgements

I would like to express my deep gratitude and address my thanks to:

- My supervisor: Professor Marwan Al Akaidi for his invaluable advice, help, trust and support throughout my PhD studies.

- Dr Amar Aggoun and Prof Malcolm McCormick for their help and support.

- Miss Dina alnsour and Mr James Wheeler for their endless help, kindness and encouragements.

- The support staff at DeMontfort University for their constant help.

- My sister Dr Zakia Fekkai and my brother in law Dr Nazir Mustapha for their patient love, encouragement and support.

- Thanks also go to: Miss Amal Mehanna, Miss Meriem Mazri, Dr.Mahmoud Shafik, Dr.Sandra Sampaio, Mr Osama Yousef, Mr Omar Oalani, Dr.Khamis Algharbi, Miss Eva Torres, Dr.Silvia Cirstea, Dr.Marcian Cirstea, Miss Oana Spulber, Dr.and Mrs. Andrei and Anca Dinu, Mr. Paolo Fioravanti, Dr. Omar Farooq for their precious friendship, kindness and support.

- My parents, for their endless love and unfailing support.

Leicester, United Kingdom
October, 2002

Souhila Fekkai

Chapter 1

Introduction

The human vocal tract and articulators are biological organs with non-linear properties, whose operation are not just under conscious control but also affected by factors ranging from gender to upbringing to emotional state. As a result, vocalisations can vary widely in terms of their accent, pronunciation, articulation, roughness, nasality, pitch, volume, and speed; moreover during transmission, our irregular speech patterns can be further distorted by background noise and echoes, as well as electrical characteristics (if telephones or other electronic equipment are used). All these sources of variability make speech recognition, even more than speech generation, a very complex problem [1].

Humans are so comfortable with speech that we would also like to interact with our computers via speech, rather than having to resort to primitive interfaces such as keyboards and pointing devices. A speech interface could support many valuable applications, for example, telephone directory assistance, spoken database querying for novice users, 'hands-busy' applications in medicine or fieldwork, office dictation devices, or even automatic voice translation into foreign languages. Such tantalizing applications have motivated research in automatic speech recognition since the 1950's. Significant progress has been made so far, especially since the 1970's, using a series of engineered approaches that include template matching, knowledge engineering, and statistical modelling. However, computers are still nowhere near the level of human performance, in terms of speech recognition, and it appears that further significant advances will require some new insights.

1.1 Speech Signals: Representation and Analysis

Continuous speech is a set of complicated audio signals, which makes producing them artificially difficult. Speech signals are usually considered as voiced or unvoiced, but in some cases, they are a combination of the two. Voiced sounds consist of a fundamental frequency (F_0) and a lot of harmonic components produced by vocal cords (vocal folds). The vocal tract modifies this excitation signal causing formant (pole) and sometimes antiformant (zero) frequencies [1]. Each formant frequency has an amplitude and bandwidth and it may be sometimes difficult to define some of these parameters correctly. The fundamental frequency and formant frequencies are probably the most important concepts in speech synthesis and also in speech processing in general.

With purely unvoiced sounds, there is no fundamental frequency in the excitation signal and therefore no harmonic structure either; the excitation can be considered as white noise. The airflow is forced through a vocal tract constriction, which can occur in several places between the glottis and mouth. Some sounds are produced with complete stoppage of airflow followed by a sudden release, producing an impulsive turbulent excitation often followed by a more protracted turbulent excitation. A discrete time model, which simulates this process, is depicted in Figure 1.1. The vibrations of the vocal cords are simulated by an impulse train generator, which produces pulses p at the pitch period. In the case of unvoiced speech, the airflow can be seen as white noise created by an appropriate random noise generator. This discrete sequence produced by either generator gets multiplied by a value representing the amplitude of the signal corresponding to the loudness of our speech. The vocal tract can be considered as a linear filter. Of course, this system cannot be considered static, but varies with time. Due to the slow movements of our vocal system, the model can be assumed static for short time intervals, e.g. 10ms [2].

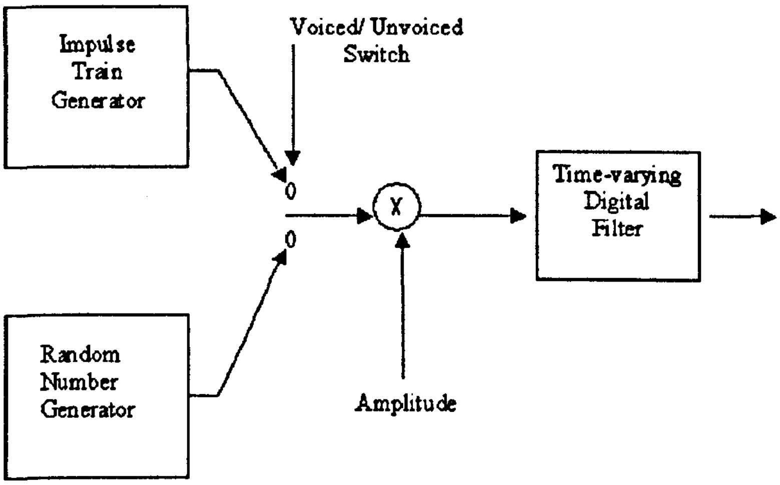

Figure 1.1: Model of Speech Production.

1.2 Speech Recognition

Speech recognition is an important subject that has been widely studied over the last decade. Until recently, classical mathematics and signal processing techniques have been used to develop recognition systems. In recent years, research has been carried out on the fractal nature of speech. As information technology becomes more important to us in many aspects of our daily lives, the problem of communication between humans and machines becomes more of an issue. The well-established forms of communication with computer keyboards, screens, and the mouse have many disadvantages, which could be overcome with a natural spoken language interface. However, the deceptively simple means of exchanging information by speech is in fact, extremely complicated.

Speech recognition is being used by thousands of people everyday. Systems such as calling cards and phone banking services use speech recognition by prompting the user to answer questions in voice rather than pressing digits on the phone pad to send Dual Tone Multi-Frequency ($DTMF$) signals. Speech Recognition is the ability to audibly detect human speech and parse that speech in order to generate a string of words or sounds to represent what a person has said. Speaker recognition is similar to that of speech recognition except that in addition to identifying the speech spoken, the system must also identify the individual who spoke [3].

The main problem areas for speech recognition are:

1. Continuous speech needs to be segmented in order to obtain the correct information.

2. Speech patterns vary not only between speakers but also within an individual speaker, even when identical words are spoken.
3. A word can vary in loudness, pitch, stress and pronunciation rate.
4. The geographical origin of the speaker is an important factor when words are pronounced.
5. Different words sound very similar.
6. Background noise and other interference can distort the original signal.
7. Individual elements tend to lose their identity in the speech process; for example, words merge into each other and phonemes suffer from co-articulation effects.

Due to these problems, research has mostly concentrated on solving specific tasks, such as speaker-dependent recognition and isolated word recognition. Isolated word recognition overcomes the problem of correctly segmenting continuous speech by demanding that appropriate pauses are inserted between each voiced word. Speaker dependent systems avoid problems such as regional accents and the sex of the speaker. The work reported in this thesis has focused on a speaker-independent isolated speech recognition system based on fractal techniques developed at De Montfort University.

The principal aims of the research have been to:

1. Investigate novel techniques for speech recognition systems based on fractal properties.
2. Improve the voice services in telecommunication systems through the development of new non-linear speech processing techniques.

3. Provide higher quality speech synthesis.
4. Improve speech recognition.
5. Improve speaker identification and verification.

The specific objectives have been:

1. Use different fractal signatures including the fractal dimension and information dimension to segment and uniquely quantify speech signals.
2. Compare the practical value of this approach against established techniques.
3. Consider the synthesis of a word in order to produce high quality, good intelligibility and natural sounding speech.

The research has delivered new technologies to provide more efficient speech recognition an higher quality speech synthesis.

1.3 Scope of the Thesis and Original Contributions

Fractal geometry is currently being used in many areas of physics and mathematics. Fractals have provided a framework for the characterisation and modelling of irregular and seemingly complicated structures found in nature. Classical geometry deals mainly with objects that cannot be recursively subdivided into self-similar components, while fractal geometry provides an approach to the analysis of self-similarity.

Since the advent of the new mathematical concept called fractals, fractal geometry has revolutionised the characterisation of self-similar structures and phenomena in nature [4]. The concept of fractals was introduced by Benoit Mandelbrot in the late 1970s and is used generally to refer to geometrical shapes in nature with self-similar structures. This new concept has stimulated renewed interest in the field of computing mathematics and simulation [5, 6].

Fractals have provided us with a new method of characterising seemingly complex and irregular structures in nature by means of the fractal dimension. It has been used extensively in modelling self-similar structures such as mountains, clouds, rocks, etc. and self-affinity found in thermal noise, human electroencephalogram (EEG) [7], music and more recently, vocal sounds [8]. Speech sounds, especially fricatives, contain some turbulence [9]. In the linear speech model this has been dealt with by having a white noise source exciting the vocal tract filter. It has been conjectured that the structures in turbulence can be modelled using fractals. This motivated Maragos [10] to use the short time fractal dimension of speech sounds as a feature to approximately quantify the degree of turbulence in them. Speech waveforms themselves are highly irregular patterns that can be quantified using fractal mathematics. It is worth noting at this point that the scope and accuracy of characterisation that can reasonably be expected using the fractal dimension (a real number between 1 and 2 for speech) essentially gives us a measure of the degree of roughness or irregularity of the object. The usefulness of fractal geometric as a model for characterising speech is inherently limited by the accuracy to which a fractal dimension can safely be calculated. A speaker identity is strongly dependent of the physiological and behavioral characteristics of the speech production system. The first step of a basic speech recognition system is to extract from the speech samples a 'good' parametric representation. These parameters must be, as much as possible, representative of a speaker, presenting low variability for the speaker's speech samples, and great difference when used with others speakers' speech samples [11].

There are many difficulties in communications and signal processing algorithms, whose linear techniques have failed to be addressed satisfactorily. It is a generally held belief, that these problems may however, have solutions in the growing field of non-linear signal processing. The recent rise in neural network concepts is, for example, largely fuelled by this promise. In addition, the past decade has seen a remarkable growth in the theory of the dynamics of non-linear systems. One cause of this interest has been the realisation that deterministic mathematical models with few degrees of freedom can generate extremely complex behaviour. Thus, complicated physical systems may be well modelled by relatively simple non-linear models such as fractals. This research project has aimed to advance non-linear signal processing techniques for speech in telecommunication systems by investigating non-linear models for solving speech recognition problems.

The thesis encompasses the following novel developments and original contribution:

- A Novel Spectral Estimation of the Fractal Dimension for Phonetic Elements in Speech.
- Evaluation of the fractal dimension techniques to reduce the recognition error.
- Fractal Dimension Segmentation For Isolated Speech Recognition [12].
- A Novel approach to measure the Fractal Dimension of Speech Phonemes.
- A Novel Word Recognition Scheme Based on Fractal Properties [13].

- New Features to Improve Fractal Speech Recognition.
- A New Neural Network Model Based On a Parametric representation for Isolated Speech Recognition [14].
- Hybrid Techniques for speech Recognition based on the Fractal Dimension.
- A New Natural Sounding Speech Synthesis Based on Fractals [15].

1.4 Thesis Outline

This thesis is organised as follows:

Chapter 2 introduces the concepts of fractal geometry and the fractal dimension.

Chapter 3 describes the main approaches to speech recognition and synthesis procedure. Various existing methods and algorithms are also discussed.

Chapter 4 discusses a new speech recognition system based on fractals. It is shown that the algorithms developed add new features, which are reported for the first time. The simulation results are discussed in detail in this chapter.

Chapter 5 proposes a new algorithm for natural sounding speech synthesis. The algorithm has two components:

1. Phase compression based on the unwrapped phase of the speech signal.
2. Fractal synthesis based on the fractal dimension.

The results obtained from the simulation are assessed and compared in this chapter.

Chapter 6 formulates the conclusions, and possible future work is suggested.

1.5 References

[1] I.Witten, *Principles of Computer Speech*, Academic Press, San Fransisco, 1982.

[2] K.Kleijn and K.Paliwal (Editors), *Speech Coding and Synthesis*, Elsevier Science B.V., The Netherlands, 1998.

[3] M.Machowski, "Speech recognition and natural language processing as highly effective means of human-computer interaction," University of Colorado, Department of Computing Sciences, June 3 1997.

[4] B. B. Mandelbrot, *Fractal Geometry of Nature*, Freema, San Francisco, 1982.

[5] J. Feder, *Fractals*, Plenum Press, 1988.

[6] R. F. Voss, *The science of fractal images*, Springer, Berlin, edited by h.o.peitogen and d. saupe edition, 1988.

[7] I. Dvorak and J. Siska, "On some problems encountered in the estimation of the correlation dimension of eeg," *Phys. Lett.*, vol. A 118, no. 63, 1986.

[8] C. A. Pickover and A. Khorassani, "fratctal characterisation of speech waveforms graphs," *Computers and*, , no. 1082, 1985.

[9] P. Maragos, "Modulation and fractal models for speech analysis and recognition," *Proceeding of COST-249 Meeting, Porto, Portugal*, Feb. 1998.

[10] P. Maragos and J. F. Quatieri, "Speech nonlinearities, modulations, and energy operators," *in Proc.IEEE ICASSP-91, Toronto, Canada*, pp. 421–424, May 1991.

[11] A. Petry and Barone, "Fractal dimension applied to speaker identification," *ICASSP2001, Salt Lake city, Utah*, 2001.

[12] S.Fekkai, M.Al-Akaidi, and J.Blackledge, "Fractal dimension segmentation: Isolated speech recognition," *IEE Electronic & Communications Event on 'Speech Coding Algorithms For Radio Channels', Savoy place, London*, April 2000.

[13] S.Fekkai and M. Al-Akaidi, "Words recognition based on fractal properties," *International Conference on Image Science Systems & Technology (CISST'99), Las Vegas, USA*, 28/6-1/7 1999.

[14] S.Fekkai and M. Al-Akaidi, "Neural network techniques for speech recognition speaker independent," *Euromedia 2001, SCS, Spain*, May 2001.

[15] S.Fekkai and M. Al-Akaidi, "A new speech synthesis based on fractal," *EUSIPCO 2002, XI European Signal Processing Conference, Toulouse, France*, September 2002.

Chapter 2

Fractal Geometry and the Fractal Dimension

2.1 Introduction

The introduction of the word “fractal” to the scientific community and the world at large can be attributed to one man, Benoit Mandelbrot [1]. Since the introduction of its concept in his classical 1975 paper, ‘Fractal objects: Form, chance and Dimension’ [2], fractal models have been successfully applied to describe and understand the geometry of countless natural phenomenon and geometries ranging from particle trajectories and hydrodynamic flow to landscape structures and biological studies.

The purpose of this chapter is to address the fundamental questions frequently posed by the uninitiated when first encountering the subject:

1. What is fractal?

2. What is fractal dimension?

3. How can fractal dimensions be calculated and what use are they?

In answering these, the question of how fractals can be applied to speech science must be addressed.

2.2 Description of Fractal Geometry

The word "fractal" is relatively new to the world. It describes a new system of mathematics, which is so powerful that it can actually describe the structure of mountain, coastlines, galaxies and other such natural phenomena. Mathematicians and scientists generally believed that such complex natural phenomena were almost beyond rigorous description. Fractals represent objects or patterns that appear to be *self-similar*, that is, no matter what scale is used to view the pattern, the magnified portion of the fractal shape looks similar to the original pattern. Benoit Mandelbrot [2] studied fractal structures and succeeded not only in describing them as for the first time, but also in showing that they are related to one another. He coined the word "fractal" to encompass his new generalisations of complex shapes.

The name *'fractal'*, came from the Latin word *'fractus'* which means broken and was given to highly irregular sets by Mandelbrot in his foundational essay in 1975. Since then, fractal geometry has attracted widespread, and sometimes controversial attention. The subject has grown on two fronts: on one hand many 'real fractals' of science and nature have been identified. On the other hand, the mathematics that is available for studying fractal sets, much of which has its roots in geometric measure theory, has developed enormously with new tools emerging for fractal analysis [3]. A very important characteristic of fractals that is useful for their description and classification is their fractal dimension D. Intuitively, D measures the degree of irregularity over multiple scales [4].

In order to characterise one object from another, the so-called *fractal dimension* D, of the object can be calculated. The fractal dimension is a real number, which in general, falls in the range 0 to 5 and can be calculated in a number of ways. In short, D gives a measure of the degree of irregularity or roughness for an object [5]. The fact that such complicated structures can be characterised by single numbers has led to work being carried out in the area of acoustic and speech science. Speech waveforms themselves are highly irregular patterns that can be quantified using fractal mathematics [6].

The field of fractal geometry was initially created by Mandelbrot [1, 7] and has experienced almost explosive growth and refinement in the past several years. At the most fundamental level, it is a system for describing the shapes of objects of the real world, rather than the abstract or ideal structures that are the focus of the more traditional. Given this difference, it is not surprising that fractal and Euclidean geometries differ at their most basic levels [8]. In general, a common educational background leads one to consider natural structures in Euclidean terms. Understanding fractal geometry, therefore, requires a significant reconceptualisation of the way the world is.

A small demonstration can help, tear the corner off a sheet of paper (Do not crease it before hand just tear it), a roughly triangular piece of paper results and two of its edges-the untorn ones- represent straight lines. The torn edge however, is likely to be very irregular. However, how can the erratic, disorderly shape of the torn edge be characterized? Traditional mathematics attempted to solve this problem by making the assumption that if the irregular edge were magnified sufficiently it would appear as a series of (extremely short but perfectly straight) simple line segments, each of which could be easily described by traditional geometry. Since a straight line is a one-dimensional shape, the implication of this method is that a complex irregular and disorderly line can be described as a string of very much shorter (but individually orderly) one-dimensional structures. Looking at the torn edge with ever-stronger magnifying glasses quickly shows the problem with this assumption: Every time the irregular line is magnified, more irregularity appears. In fact, no matter how much it is enlarged, the torn edge of a real piece of paper can *never* be reduced to a set off perfectly straight lines. (In the formal language of fractal geometry, the edge is said to be 'self-similar' at all scalings). This implies that it cannot be a one-dimensional shape. On the other hand, the irregular edge obviously cannot have a dimension of as much as two, which is the dimension of a plane surface. However, counterintuitive it may seem at first that logic demands that the irregular edge must have a dimension that is between one and two, that is, its dimension must be fractal [8].

Fractal geometry is rapidly being assimilated into many diverse fields of physics and mathematics. Whereas man-made objects are well defined in Euclidean geometry, natural objects can often best be modelled by fractal geometry. Fractal geometry is the geometry of the broken-up, the pitted and pocked, the tangled and twisted, the turbulent and the chaotic. Central to fractal geometry is the concept of self-similarity in which an object appears to look similar at different scales - an obvious concept when observing naturally occurring features, but one that has only relatively recently started to be developed mathematically and applied to various branches of science and engineering. This concept can be applied to systems of varying physical size depending on the complexity and diversity of the fractal model that is considered. Ultimately, it is of philosophical interest to view the universe itself as a single fractal, the self-similar parts of which have yet to be fully categorized; those naturally occurring objects for which fractal models abound, being smaller subsets of a larger whole. This view is closely related to the concept of a chaotic universe in which the dynamical behavior of a system cannot necessarily be pre-determined. Such systems exhibit self-similarity when visualized and analyzed in an appropriate way (i.e. an appropriate phase space). In this sense, the geometry of a chaotic system may be considered to be fractal [1].

Self-similarity is a very general term, there are two distinct types of self-similar objects (as illustrated in Figure 2.1 and described in the following subsections):

2.2.1 Deterministic self-similarity

A deterministic fractal is composed of distinct features, which resemble each other in some way at different scales (feature scale invariance).

Deterministic fractals are usually generated through some Iterated Function System (IFS) remarkable for the complexity that can be derived through the simplest of these centering

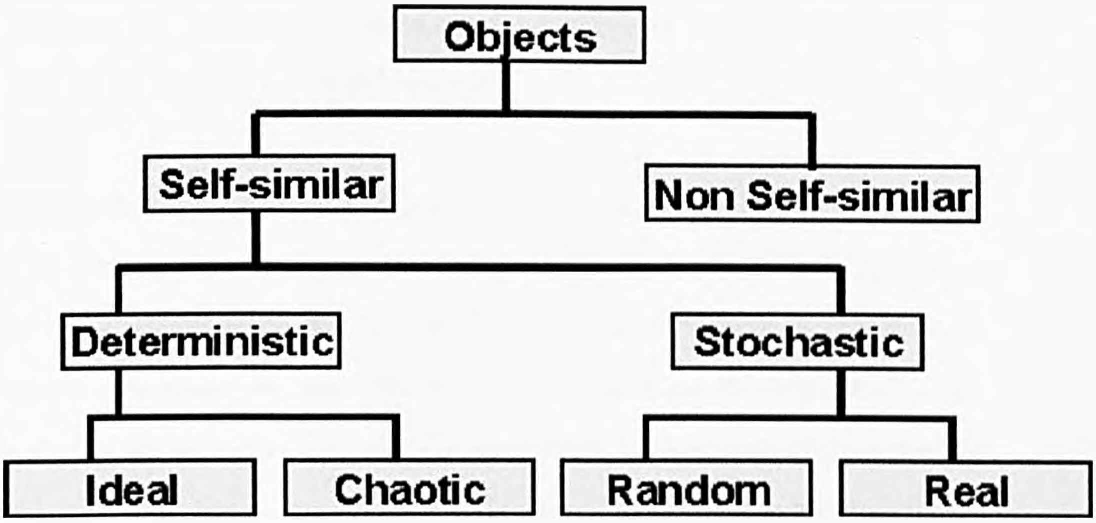

Figure 2.1: Fractal Types.

iterated systems. The way in which the output from these systems is viewed graphically and interpreted geometrically changes substantially from one fractal to another but the overall principal remains the same.

The Von Koch Curve

The construction of the Koch curve starts with a line segment of unit length $L(1) = 1$. This starting form is called the initiator and may be replaced by a polygon. The initiator is the 0-th generation of the Koch curve. The construction of the Koch curve proceeds by replacing each segment of the initiator by the generator shown as the curve marked $n = 1$ as shown in Figure 2.2. Thus, we obtain the first generation, which is a curve of 4 line segments each of length 1/3. The length of the curve is now $L(1/3) = 4/3$. The next generation is obtained by replacing each line segment by a scaled-down version of the generator. Thus in the second generation we have a curve consisting of $N = 4^2 = 16$ segments each having length $\delta = 3^{-2} = 1/9$, and the length of this curve is $L(1/9) = (4/3)^2 = 16/9$.

By applying a reduced generator to all segments of a generation of the curve, a new generation is obtained. According to Mandelbrot's definition of a fractal, "*A fractal is by definition a set for which the Hausdorff-Besicovitch dimension strictly exceeds the topological dimension, so that, the topological dimension is always an integer whilst the fractal dimension is not*" [2]. Curves for which D exceeds the topological dimension 1 are called *fractal curves.* Since the Hausdorff-Besicovitch dimension D for the Koch curve exceeds its topological dimension D_T, the Koch curve is a *fractal* set with the fractal dimension $D = \ln 4/\ln 3$.

2.2.2 Statistical self-similarity

The features of a statistical self-similar fractal may change at different scales but their statistical properties at all scales are the same (statistical scale invariance).

Statistically self-similar fractals are those used to model a variety of naturally occurring objects (background noise, clouds, landscapes, coastlines etc.). They can be generated through a variety of different stochastic modelling techniques. They can also be considered to be the solution to certain classes of stochastic differential equations of fractional order.

The most commonly associated measure with a self-similar object is its fractal (or similarity) dimension. If we consider a bounded set A in a Euclidean n dimensional

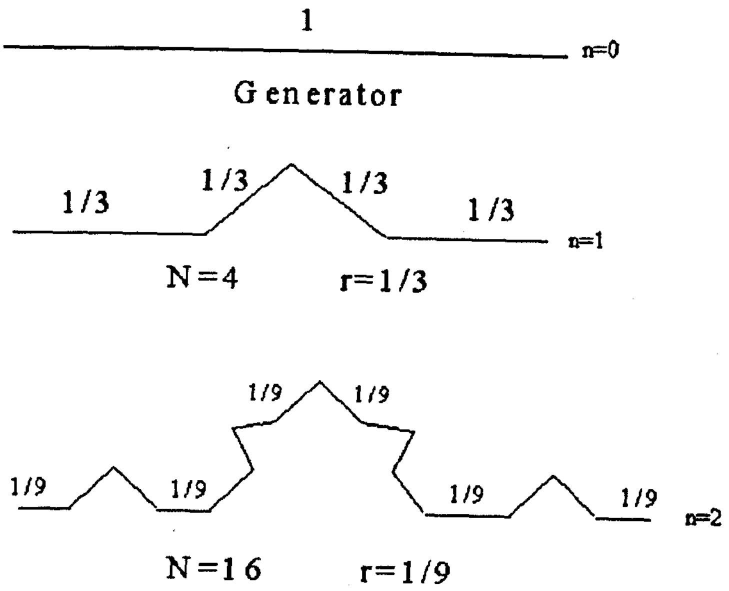

Figure 2.2: The Von Kock Curve.

space, then the set A is said to be self-similar if A is the union of N distinct (non-overlapping) copies of itself, each of which has been scaled down by a ratio $r < 1$ in all coordinates. The fractal is described by the relationship:

$$Nr^D = 1; \quad D = -\frac{\ln N}{\ln r} \tag{2.1}$$

where D is the fractal dimension. The ranges of value of D characterizes the type of fractal as shown in Table 2.1.

In each case, the fractal may be deterministic or random. In the latter case, the fractal is taken to be composed of N distinct subsets each of which is scaled down by a ratio $r < 1$ from the original and is the same in all-statistical respects to the scaled original. The fractal dimension in this case is also given by Eqn.2.1.

Fractal Dimension	Fractal Types
$0 < D < 1$	Fractal dust
$1 < D < 2$	**Fractal Signals and Curves [as used for speech]**
$2 < D < 3$	Fractal Images and Surfaces
$3 < D < 4$	Fractal Volumes
$4 < D < 5$	Fractal Time

Table 2.1: Fractal types and associated ranges of the fractal dimension.

The scaling ratios need not to be the same for all the scaled down copies. Certain fractal sets are composed of the union of N distinct subsets each of which is scaled down by a ratio r_i , $1 \leq i \leq N$ from the original in all coordinates. The fractal dimension is given by a generalization of Eqn.2.1, namely

$$\sum_{i=1}^{N} r_i^D = 1 \tag{2.2}$$

Finally, there are self-affine fractal sets, which are scaled by different ratios in different coordinates. For example, consider the curve $f(x)$ which satisfies

$$f(\lambda x) = \lambda^{\alpha} f(x) \quad \forall \lambda > 0$$

where λ is a scaling factor and α is the scaling exponent. This equation implies that a scaling of the x-coordinate by λ gives a scaling of the f-coordinate by a factor λ^{α}, which is an example of self-affinity. A special case occurs when $\alpha = 1$ when we have a scaling of x by λ producing a scaling of f by λ which is an example of self-similarity. Random fractal signals are, in general, examples of self-affine records [5].

Naturally occurring fractals also differ from the strictly mathematically defined fractals in that they do not display statistical or exact self- similarity over all scales. Rather, they display fractal properties over a limited range of scales.

2.3 Self-similarity and the Fractal Dimension

The property of self-similarity or scaling is one of the central concepts of fractal geometry. It means that some types of mainly naturally occurring objects look similar at different scales. It is closely connected with the intuitive notion of fractal dimension [9]. An object normally considered as one-dimensional, a line segment, for example, also possesses a similar scaling property. It can be divided into N identical part, each of which is scaled down by the ratio from the whole.

$$r = 1/N^1 \tag{2.3}$$

Similarly, a two-dimensional object, such as a square area in the plane, can be divided into N self-similar parts, each of which is scaled down by a factor

$$r = 1/N^{1/2} \tag{2.4}$$

Three-dimensional objects like a solid cube may be divided into N smaller cubes, each of which is scaled down by a ratio

$$r = 1/N^{1/3} \tag{2.5}$$

With self-similarity the generalisation to fractal dimensions is straightforward. D-dimensional self-similar object can be divided into N smaller copies of itself, each of which is scaled down by a factor r where

$$r = 1/N^{1/D} \tag{2.6}$$

Conversely, given a self-similarity object of N parts scaled by a ratio r from the whole, its fractal or similarity dimension is given by

$$D = \log(N)/\log(1/r)$$

which quantifies roughness.

A signal with D close to 1 looks smooth whereas a signal with D approaching 2 looks rough [10]. This is illustrated with the example given in Figure 2.3, where we can clearly observe how changes to D affect the signal.

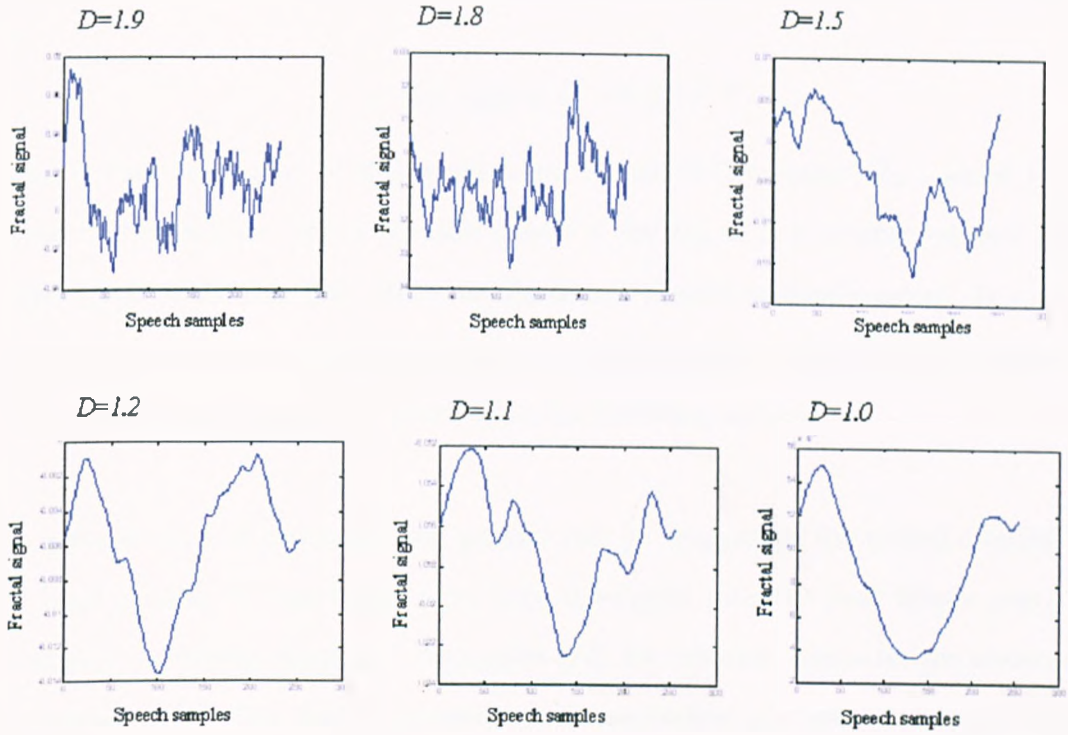

Figure 2.3: Fractal signals.

The dynamic airflow during speech production may often result in some smaller or larger degree of turbulence during the production of speech sounds by human vocal-tract system [11]. The geometry of the speech turbulence as reflected in the fragmentation of the time signal that can be quantified by using fractal models [7]. The fractal dimension is an important characteristic of fractals, which contains information about their geometrical structures at multiple scales [11].

Let $s(t)$ represent a continuous real-valued function where $0 \leq t \leq T$, and its graph is

$$S = (t, s(t)) \in R^2 : 0 \leq t \leq T$$

The Fractal Dimension of S is equal to its Hausdorff Dimension D_H , which is defined by Mandelbrot, and S is called fractal if the D_H of S is strictly exceeds 1 (its topological dimension) [12]. However,D_H is only a mathematical concept. In reality, its value is very hard to compute and can only be closely related by other methods with smaller complexities as described in the following section.

In general, there is no unique and general rule for computing the fractal dimension. A large number of algorithms have been developed over the past fifteen years to compute the fractal dimension. Nakagawa [12], for example, estimates the fractal dimensions of self-affine data with power spectra according to a power law based on the moment exponent. This method is called critical exponent method (CEM). The next section describes four methods to calculate the fractal dimension of speech phonemes.

2.4 Fractal Dimension Techniques

2.4.1 Box Counting Method (BCM)

The Box counting dimension $D_B(S)$ of S is defined as

$$D_B(S) \equiv \lim_{s \to 0} \frac{\ln(N(S))}{\ln(1/s)} \tag{2.7}$$

Where $N(s)$ is the number of squares that intersect S when partioned by a grid of squares with size ε [13], the principle is illustrated in Figure 2.4. Assuming the digital speech signal $S_1, S_2, ..., S_T$ is discretized from a real speech signal $S(t)(0 \leq t \leq T)$ by sampling the computation of D_B via Eqn.2.7 we have

$$D_B(S) = \frac{\left\{J \cdot \left(\sum_{j=1}^{J} \ln(1/\varepsilon_j) \cdot \ln(N(\varepsilon_j))\right) - \left(\sum_{j=1}^{J} \ln(1/\varepsilon_j)\right) \cdot \left(\sum_{j=1}^{J} \ln(N(\varepsilon_j))\right)\right\}}{\left\{J \cdot \sum_{j=1}^{J} (\ln(1/\varepsilon_j))^2 - \left(\sum_{j=1}^{J} \ln(1/\varepsilon_j)\right)^2\right\}} \tag{2.8}$$

where, $\varepsilon_j(1 \leq j \leq J)$ are J computation of resolutions, $\varepsilon_{min} \leq \varepsilon_1 < \varepsilon_2 < ... < \varepsilon_3 < \varepsilon_J < \varepsilon_{max}$ (ε_{min} and ε_{max} represent the maximum and the minimum resolution of computation). $N(\varepsilon_j)$ is the number of squares that intersect S when the grid of square of size ε_j. Hence, from Eqn.2.8, the D_B is the slope of the line $\ln(\varepsilon_j) - \ln(\varepsilon_j))$ obtained via the least-squares error principle.

2.4.2 Continuous Box Counting Method (CBCM)

Functions never fold back on themselves, and as box counting can be thought of as only looking at functions, boxes can be counted in and between columns. The CBCM method gives a more accurate value for the fractal dimension as it does not look only at the points within each column along the curve, as the dimension is calculated but also to the relationship between columns [14].

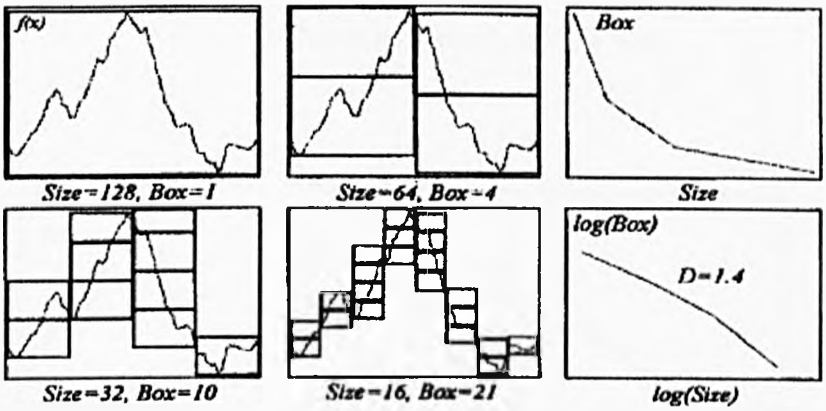

Figure 2.4: Illustration of the box counting method for computing the fractal dimension D of a signal showing 4 iterations and the least squares fit [5].

2.4.3 Walking- Divider Method (WDM)

The WDM makes use of a chord length (*step*) and measures the number of chord lengths (*length*) needed to cover a fractal curve. This technique is based on the principle of taking smaller and smaller rulers of size *step* to cover the curve and counting the number of ruler lengths required in each case [5] (see Figure 2.5). It is a recursive process in which the *step* is decreased (typically halved) and the new length calculated. The input signals are taken to be of size N where N is a power of 2 because of the recursive nature of the method. A least squares fit to the ln ln plot of *length* against step gives where ($D = -\beta$):

$$\beta = \ln\left[\text{total length}\right] \text{ Vs} \ln\left[\text{step size}\right] \tag{2.9}$$

The fractal dimension is then given:

$$D = -(\ln[\text{total length}] \text{ Vs} \ln[\text{step size}]) \tag{2.10}$$

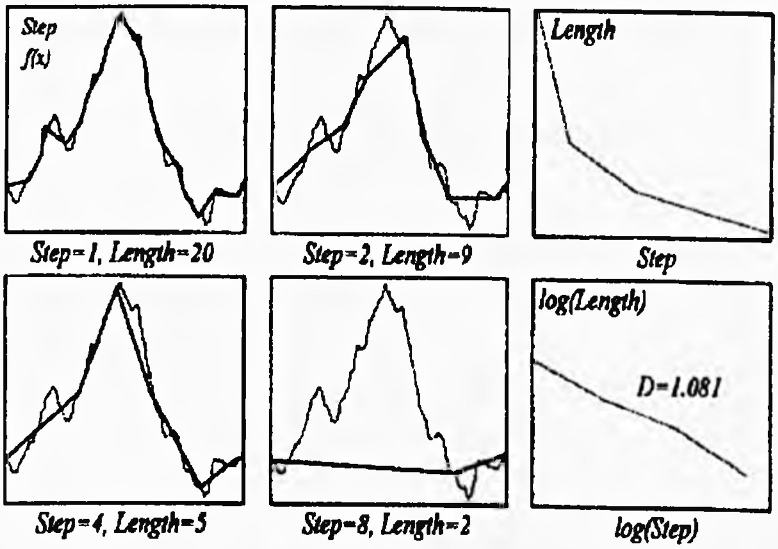

Figure 2.5: Illustration of the line walking-divider method for computing the fractal dimension D of a signal showing 4 iterations and the least squares fit [5].

2.4.4 Power Spectrum Method (PSM)

Let us investigate the property of scaling law by answering the following question : Why is c/k_i^β a power spectrum of a signal which is self-affine?

The scaling law implies that we can model signal in term of the full equation (assuming analogue signal) [15]:

$$f(x) = \frac{n(x)}{k_i^\beta}$$

where $n(x)$ is a white noise (i.e. noise whose PSDF is a constant) and $1/k_i^\beta$ is a filter of frequencies k_i, its Fourier Transform is given by:

$$F_k = \int_{-\infty}^{\infty} f(x)\exp(-ikx)dx$$

Using this definition, the inverse Fourier Transform is given by:

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{(ki)^\beta} N(k)\exp(ikx)dk$$

where $N(k)$ is the Fourier transform of $n(x)$. Application of the convolution theorem allows us to write this result in the form:

$$f(x) = \int_{-\infty}^{\infty} h(x-y)n(y)dy$$

where h is given by

$$h(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\exp(ikx)}{(ik)^\beta}dk$$

Substituting p for ik, $h(x)$ can be written in terms of the inverse Laplace transform of p^β. Since

$$\hat{\mathcal{L}}\{x^\beta\} = \frac{\Gamma(\beta+1)}{\beta^{\beta+1}}, \quad \beta > -2, \quad Re(p) > 0$$

where $\hat{\mathcal{L}}$ is taken to denote the laplace transform and Γ is the Gamma function.

$$\Gamma(\beta) = \int_{0}^{\infty} t^{\beta-1}\exp(-t)dt$$

we can write

$$x^\beta = \hat{\mathcal{L}}\left\{\frac{\Gamma(\beta+1)}{p^{\beta+1}}\right\}$$

or

$$\hat{\mathcal{L}}^{-1}\left\{\frac{1}{p^{\beta}}\right\} = \frac{1}{\Gamma(\beta)}x^{\beta-1}$$

Thus,

$$f(x) = \frac{1}{\Gamma(\beta)}\int\limits_{-\infty}^{x}\frac{n(y)}{(x-y)^{1-\beta}}dy$$

This is the Liouville-Riemann transform and is an example of a fractional integral. Is this transform consistent with the concept of statistical self-affinity? Let's Consider the case where

$$f'(x) = \frac{1}{\Gamma(\beta)}\int\limits_{-\infty}^{x}\frac{n(\lambda y)}{(x-y)^{1-\beta}}dy$$

where λ is a scaling parameter. Substituting $z = \lambda y$, we obtain,

$$f'(x) = \frac{1}{\lambda^{\beta}}\frac{1}{\Gamma(\beta)}\int\limits_{-\infty}^{\lambda x}\frac{n(z)}{(\lambda x - z)^{1-\beta}}dz = \frac{1}{\lambda^{\beta}}f(\lambda x)$$

Now both $f'(x)$ and $f(\lambda x)$ are stochastic functions of the same type but over different scales (n(x) being white noise at any scale) and so, although $f'(x) \neq f(\lambda x)$,

$$Pr[f'(x)] = \frac{1}{\lambda^{q}}Pr[f(\lambda x)]$$

which describes a statistically self-affine signal.

The power spectrum of a fractal signal is by definition given by the fundamental scaling law [15]

$$\hat{P}_i = \frac{c}{\mid k_i \mid^{\beta}} \tag{2.11}$$

where k_i is the frequency in Hz and c is a constant of proportionality. Taking the natural logarithm of this equation yields

$$\ln \hat{P}_i = C - \beta \ln \mid k_i \mid \tag{2.12}$$

where c is a constant and $C = \ln c$. Suppose we construct a norm in the form of a least-squares error given by

$$e = \| \ln P_i - \ln \hat{P}_i \|_2^2 \tag{2.13}$$

Substituting Eqn.2.12 into Eqn.2.13 yields:

$$e = \sum_{i=0}^{n} [\ln P_i - (C - \beta \ln \mid k_i \mid)]^2 \tag{2.14}$$

The error, e , is a function of both C and β . The values of C and β which minimise e are therefore the values for which the estimated curve $\hat{P}_i$ provides a best fit (in a least-squares sense) to the data P_i. This occurs when

$$\frac{\partial e}{\partial \beta} = 2\left(-C\sum_{i=0}^{n} \ln \mid k_i \mid + \beta \sum_{i=0}^{n} (\ln \mid k_i \mid)^2 + \sum_{i=0}^{n} (\ln P_i \mid k_i \mid \right) = 0 \tag{2.15}$$

and

$$\frac{\partial e}{\partial C} = 2\left(C\sum_{i=0}^{n} 1 - \beta \sum_{i=0}^{n} \ln \mid k_i \mid - \sum_{i=0}^{n} (\ln P_i) \right) = 0 \tag{2.16}$$

Solving for β and C, we have

$$\beta = \frac{(n+1)\sum_{i=0}^{n} (\ln P_i) \ln \mid k_i - \sum_{i=0}^{n} \ln \mid k_i \mid \sum_{i=0}^{n} \ln P_i}{(\sum_{i=0}^{n} \ln \mid k_i \mid)^2 - (n+1)\sum_{i=0}^{n} \ln \mid k_i \mid} \tag{2.17}$$

and

$$C = \frac{\sum_{i=0}^{n} (\ln P_i) + \beta \sum_{i=0}^{n} \ln \mid k_i \mid}{(n+1)} \tag{2.18}$$

Using the relationship,

$$\beta = 5 - 2D \tag{2.19}$$

provides a non-iterative formula for computing the fractal dimension from the power spectrum of a signal.

The implementation of the PSM consists of applying the FFT to the speech signal in order to obtain a spectral representation of the phoneme. A pre-filter step is then used to adjust the estimated values of the fractal dimension to fit within the range 1 and 2. The power spectrum of the pre-filtered signal is computed and then the least squares approach is applied to calculate the power exponent β and the fractal dimension D.

It is important to mention that without the pre-filtering step, the values of the fractal dimension do not satisfy the range of the fractal model. However, the use of the pre-filter $1/k$ has the effect of conforming the speech data to fit the range of the fractal dimension for speech signal which lies between the range 1 and 2. Figure 2.6 illustrate the power spectum method for the speech word /open/.

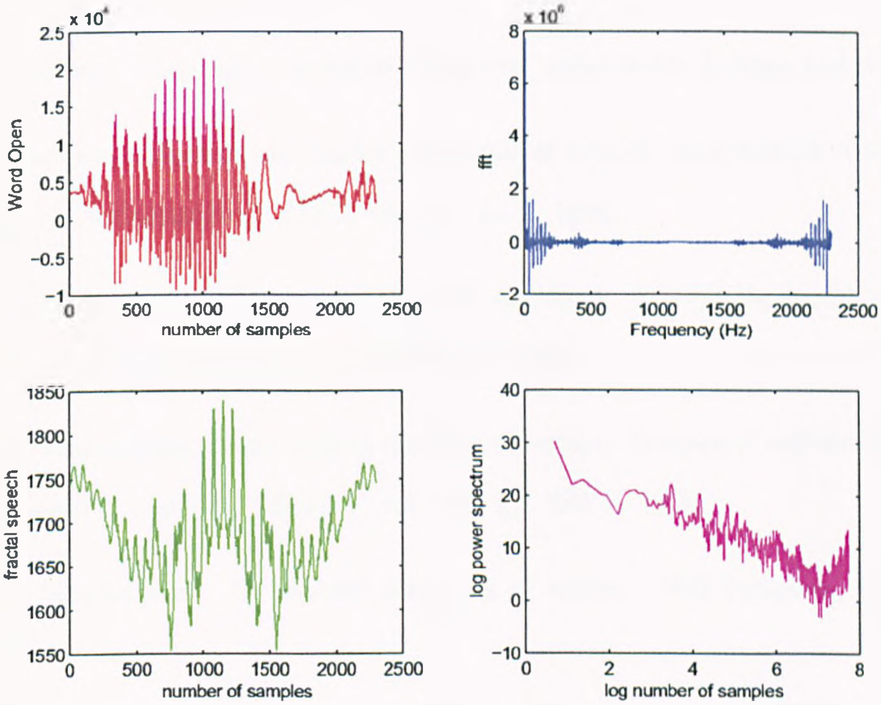

Figure 2.6: Illustration of the power spectrum method for computing the fractal dimension D showing the power spectrum of the speech word /Open/.

2.5 References

[1] B. B. Mandelbrot, *Fractals: Form, Chance and Dimension.*, W.H.Freeman, San Fransisco, CA, 1977.

[2] B. B. Mandelbrot, *Les objets Fractals: forme, Hazard et Dimension*, Flamirion, Paris:, 1975.

[3] K. Falconer, *Techniques in fractal geometry*, John Wiley & Sons Ltd, UK, 1997.

[4] P. Maragos, "Measure the fractal dimension of signals: morphological covers and iterative optimisation," *IEEE*, vol. 41, no. 1, 1993.

[5] M. J. Turner, J.M.Blackledge, and P.R.Andrews, *Fractal Geometry in Digital Imaging*, Academic Press, San Fransisco, 1998.

[6] B. B. Mandelbrot, "How long is the cost of britain: Statistical self-similarity and fractional dimension," *Science*, vol. 156, pp. 636–8, 1978.

[7] B. B.Mandelbrot, *The fractal geometry of nature*, WH Freeman, New York, 1983.

[8] R.J.Baken, "Irregularity of vocal period and amplitude: A first approach to fractal analysis of voice," *Journal of Voice*, vol. 4, no. 3, pp. 185–197, 1990.

[9] F.Michael, *The science of fractal images*, pp. 28–29, Springer-verlag, New York Inc., 1986.

[10] W.Tetschner, *Voice processing*, pp. 17–20, artech house Inc., Narwood, 1991.

[11] K.Kumar and S.K.Mullick, "Nonlinear dynamical analysis of speech," *J. Acoustic. Soc. Amer.*, vol. 100, no. 1, pp. 615–629, 1996.

[12] M.Nakagawa, "A critical exponent method to evaluate fractal dimensions of self-affine data," *J. of the Physical Society of Japan*, vol. 62, no. 12, 1993.

[13] F.Wang, F.Zheng, and W.Wu, "A c/v segmentation method for mandarin speech based on multi-scale fractal dimension," *International Conference on Spoken Language Processing, Beijing, China*, vol. 62, no. 12, pp. IV–648–651, oct. 16-20 2000.

[14] M.Al-Akaidi, J.M.Blackledge, and S.Mikhailov, "Random fractal coding techniques," *Euromedia 98, DeMontfort University, Leicester, Uk*, pp. 14–18, 1998.

[15] J.M.Blackledge, *On the Synthesis and Processing of Fractal Signals and Images*, Springer-Verlag, Berlin, 1993.

Chapter 3

Speech Recognition and Synthesis

3.1 Categories of Speech Recognition

Speech recognition tasks can be classified according to the following categories:

- Isolated word recognition
- Connected word recognition
- Continuous speech recognition
- Speech understanding
- Word spotting
- Speaker identification and verification
- Language identification

In isolated word recognition: The words are spoken in isolation, pauses between words simplify recognition because they make it relatively easy to identify endpoints (i.e., the start and end of each word), and they minimise co-articulation effects between words. In addition, isolated words tend to be pronounced somewhat more carefully, since the need to pause between words impedes fluency, which would otherwise tend to encourage a more natural and hence more careless pronunciation. Isolated words are adequate for many applications but are far from being a natural way of communication.

In connected word recognition: The spoken input is a sequence of isolated words from a specified vocabulary and the recognition is based on recognising isolated words.

The recognition of continuous speech is an attempt to transcribe naturally spoken utterances (i.e. without artificial pauses between phonemes, syllables, words, or sentences) in accordance with the rules of language orthography. This implies the need for some form of segmentation of the speech into linguistic units. The fluency of speech in natural speech imposes co-articulation between adjacent phonemes and words in a phrase. This leads to neglecting some phonemes in a phrase, especially between words, which makes the recognition process very difficult to achieve.

The goal of a speech understanding system is to identify the meaning of the speech without constraining the speaker's sentence structure. In such a system, traditional speech recognition techniques are integrated with artificial intelligence techniques to give the extra power needed to deal with natural continuous speech. High-level knowledge sources (i.e. morphological, syntactic, semantic, and pragmatic) are incorporated in this system.

In word spotting: The speech recognition deals with detecting the occurrence of a given word in continuous speech. In this case, all the speech is ignored until a keyword is spoken. Therefore the system is tuned to recognise words, which have high correlated to one of the pre-specified keywords.

In speaker identification and verification: The aim of the speech recognition here is not to recognise what has been said but actually to highlight differences between speakers.

In speaker identification: An unknown speaker is to be recognised from a previously specified group of speakers, while in speaker verification the speech recognition technique is used in addition to other identification systems (such as a magnetic card reader) to verify the identity of the speaker.

In language identification: The speech understanding techniques are used to form some sort of linguistic chains from the phonetic transcription of speech and these are used as a means of discrimination between different languages.

Two terms, which are frequently used to describe a speech recognition system are speaker-dependent and speaker-independent. In a speaker-dependent system, the system is to be trained to the speech of each new speaker for the entire vocabulary. In a speaker-independent or multi-speaker system, no training is required for the new speaker. Actually, for a large vocabulary system and for continuous speech recognition, instead of full training, the system can adapt to a new speaker by some relatively simple restricted procedures using a few words or sentences. The latter case is often called speaker adaptation.

3.2 Feature Measurement

A speech signal is a highly redundant signal, it carries linguistic messages as well as information about speakers, regarding their physiology, psychology, etc. Feature measurement, some times called feature extraction, is basically a data reduction technique. The digitised speech signal is transformed into a smaller set of features, which faithfully describe the salient properties of the acoustic waveform. Data reduction rates (or compression ratios) of 10 to 100 are generally practical.

A number of different feature sets have been proposed ranging from simple sets such as energy and zero-crossing rates to complex representation such as:

- Short-time spectrum (DFT or filter bank)
- Linear predictive coding
- Cepstral parameters (homomorphic model)
- Articulatory parameters
- Auditory model

The motivation for choosing one feature set over another is often dependent on the imposed on the system in terms of cost, speed, and recognition accuracy.

3.2.1 Linear Prediction Parameters

Linear prediction coding coefficients can model the spectral envelope well, and are widely used. The basic idea behind LPC is that a given speech sample can be approximated as a linear combination of past speech samples [1]. For each sample, a prediction error e(n) is defined as follows:

$$e(n) = s(n) - \overline{s}(n) \tag{3.1}$$

where

$$\overline{s}(n) = \sum_{i=1}^{p} a(i)s(n-1) \tag{3.2}$$

with

$$H(z) = \frac{1}{1 - \sum_{i=1}^{p} a(i)z^{-1}} \tag{3.3}$$

Here, $\overline{s}(n)$ is the linearly predicted sample, $s(n)$ is the actual sample, p is the degree of the LPC model filter, and a(i) where i=1, 2,...p are the filter predictor coefficients. By minimizing the mean-square prediction error e(n), over a finite interval, a unique set of predictor coefficients can be determined. The LPC coefficients give good short-time spectral estimation of the linear time varying system. $H(z)$ in Eqn.3.3 represents the z-transform of the transfer function of the vocal tract (all pole model).

For a short interval (M samples of speech), the LPC coefficients are computed to yield an N-dimensional feature vector, where N equals p (the model's degree) which is usually taken to be between 8 to 14 [2]. The time variation of these feature vectors defines a pattern for the speech utterance.

Formant frequencies and their bandwidths can be extracted from the transfer function of the vocal tract by a peak picking procedure. Computing the FFT over the set of LPC parameters and taking the inverse of the result, yields the transfer function of the vocal tract Eqn.3.3. Another way to find the Formant Frequencies and their bandwidths is to solve the inverse of Eqn.3.3 and find its roots (complex pole-pairs) [2].

3.2.2 Filter Bank Parameters

A popular set of features used in many speech recognition systems is the output of a bank of filters. The speech signals are passed through a bank of band pass filters covering the speech bandwidth. The energy at the output of each channel is estimated from the output of each particular filter [3]. The set of energy values at each interval of time (frame) constitutes an N-dimensional feature vector. The time variation of these features vectors defines a pattern for the speech utterance. In general, the band pass filters are linearly spaced at low frequencies (below 1000 Hz) and logarithmically spaced at high frequencies. It has been found [4], that 13 filters spaced along a critical-band frequency scale (or bark scale), are enough for high recognition accuracy, and using 15 filters spaced uniformly in frequency give the same result as critical-band filters in a template matching approach.

3.2.3 Cepstral Parameters

Three types of cepstral parameters have been used in speech recognition systems (homomorphic model) [5], namely the linear frequency cepstral coefficient (LFCC) [6] the mel-frequency cepstral coefficients (MFCC) [7], and the LPC-derived cepstral Coefficients (LPCC) [8].

The LFCCs are computed from the log-magnitude discrete Fourier transform (DFT) directly as follows:

$$LFCC_i = \sum_{k=0}^{K-1} Y_k \cos(\frac{\pi i k}{K}) \tag{3.4}$$

where k=0,1,2,...,K-1 and K is the number of DFT log-magnitude coefficients Y_k, i=1,2,...,N and N is the number of cepstral coefficients employed.

In mel-frequency scale, the DFT magnitude spectrum is frequency-warped to follow a critical band scale (mel-scale) [7, 9] and amplitude-warped (logarithmic scale), before computing the inverse DFT parameters. Therefore, Q band pass filters are used to cover the required frequency range, and the MFCCs are computed as follows:

$$MFCC_i = \sum_{k=0}^{Q} X_k \cos \pi[i(k - \frac{1}{2})\frac{\pi}{Q}] \tag{3.5}$$

where i= 1,2,...,N and N is the number of cepstral coefficients used, k= 0,1,2,...,Q, and Q is the number of band pass filter used,and X_k represents the log-energy output of the k^{th} filter.

The LPCCs are obtained from the LPC parameters directly as follows:

$$LPCC_i = LPC_i + \sum_{k=1}^{i-1} \frac{k-1}{i} LPCC_{i-k} \times LPC_k \qquad (3.6)$$

where i= 1,2, ..., N and N is the number of cepstral parameters, k the LPC model's order. For i greater than the order of the LPC model, LPC_i is taken to be equal to zero.

The set of N parameters (LFCCs, MFCCs, or LPCCs) constitutes an N-dimensional feature vector. The time variation of these feature vectors defines a pattern for the speech utterance.

3.2.4 Articulatory Parameters

Another set of features for describing speech sounds are the parameters giving the position of the tongue, lips, jaws and the velum as functions of time. These parameters can be estimated from the speech signal [10]. A new speech production theory based on distinctive regions along the vocal tract has been introduced [11, 12], which provides a new concept in the acoustic-articulatory-phonetic relation. By performing acoustic-articulatory inversion, the area function can be used as an articulatory parameter for speech recognition.

3.2.5 Auditory Model Parameters

Another approach for feature measurements is the use of the auditory model [13]. The psychophysical aspects of critical bandwidth, loudness, timbre, and subjective duration have been used as feature measures [14]. Another design, which tries to capture the time-varying nature of the auditory model by combining the psychophysical critical-band, and loudness estimation with a firing-rate model, has improved the accuracy of the speech recognition compared to previous filter-bank feature measures [15].

3.3 Recognition

The recognition of single isolated words is usually approached by the method of template matching, this is illustrated in Figure 3.1. The incoming speech is pre-processed (start and end points detected and amplitude normalized) and, if the recogniser is in the training mode, a suitable representation of the word is extracted and stored as a template for the selected word. If the recogniser is operating in the recognition mode, the input word will be compared to each of the stored templates using a suitable distance metric to determine the best match. If the best match exceeds a decision threshold then the match is considered to have been found, but if the match is less than the decision threshold the input word is not considered to be any of the stored templates and consequently is not recognized.

The features that are commonly extracted from the speech waveform for recognition purposes are the spectral components and their energies during time-frame intervals of typically 20 to 50 ms. These are frequently LPC (linear predictive coding) coefficients, the FFT (Fast Fourier Transform) coefficients [16], the output of a bank of filters is illustrated in Figure 3.2 or Mel-based cepstral coefficients [17]. The banks of band pass filters usually covers the whole range of significant frequency components

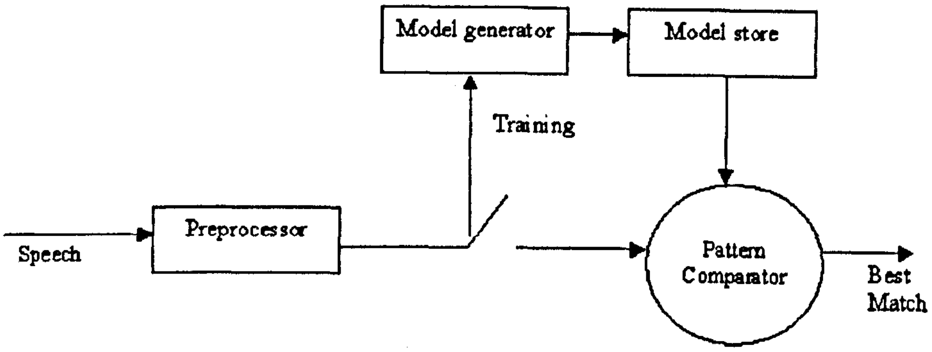

Figure 3.1: A Typical template-matching technique for isolated word recognition.

for speech (100 to 8000 Hz) and typically up to 20 filters are used. The output of these filters is sampled over the frame interval and the energy of each output averaged over the frame interval to produce a single energy intensity value for each filter output. Templates are often stored in this format as they occupy less storage space than the sampled time waveform and pattern comparison is faster.

3.3.1 Creating Reference Templates

In the recognition system, a training phase is assumed before an actual recognition can take place. The simplest speaker-dependent systems employ causal training, in which each speaker utters every word in the vocabulary one or more times and a reference template is created. Since speakers tend to pronounce a given word differently at different times or in different contexts (because of different articulatory structure for different speakers), a few repetitions of each word are often used in training. Most speaker-dependent systems use 1-3 templates per word, while speaker-independent systems (multi-speakers) use 10-12 [19].

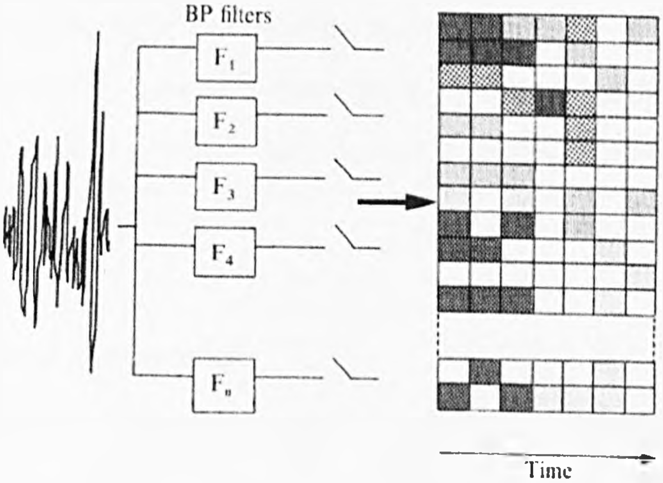

Figure 3.2: Filter bank analysis of speech. The shaded boxes represent the average energy output from each filter during a fixed sample period: the darker the box the higher the energy [18].

Reducing the number of templates for each word to a reasonable number (as mentioned above) is necessary to reduce confusions and storage requirements in speech recognition systems. Two methods are used for creating reference templates, namely averaging and clustering. In averaging, all the occurrences of a given word are averaged together, after some form of time alignment. This gives a single reference template for a speaker-dependent system [20]. For a speaker-independent system, averaging can create an unrepresentative pattern if the templates differ substantially.

In a speaker-independent system, at least 100 speakers must provide multiple training for each word, which implies that substantial clustering is necessary to merge the tokens to a representative set of 10-12 templates for efficiency. The K-means clustering method [21], and the unsupervised K-means clustering without averaging method [22], has been used. In clustering, the N templates of each vocabulary word are grouped together to form M clusters, using the nearest neighbour rule. For each such cluster, a single template is created using averaging technique over the tokens of that cluster.

3.3.2 Neural Networks

In recent years, the advent of new learning procedures and the availability of high speed parallel supercomputers, have given rise to a renewed interest in parallel distributed processing models known as Artificial Neural Networks or simply Neural Nets. These models attempt to achieve good performance via dense interconnections of simple computational elements. The Neural Nets are particularly interesting for cognitive tasks that require massive constraint satisfaction, i.e. the parallel evaluation of many clues and facts and their interpretation in the light of numerous interrelated constraints. Cognitive tasks such as a vision, speech, and language processing, are also characterised by high degrees of uncertainty and variability and it has proven difficult to achieve good performance of these tasks using standard sequential programming methods. In general, such constraints are too complex to be easily programmed and require the use of automatic learning strategies, which are now available [23]. Learning or adaptation is a major focus of Neural Nets research. The ability to adapt and continue learning is essential in areas such as speech recognition, where training data is limited and new talkers, new words, new dialects, new phrases, and new environments are continuously encountered.

Experiments on using NN for speaker-independent recognition yield 95% for 20 isolated words [24], and 98% accuracy for 10 isolated words [25], using different NN implementations. These results suggest that appropriately designed Artificial Neural Networks are well-suited for speaker-independent recognition tasks.

3.4 Matching Techniques

Two principle methods have been used to match the sampled word with existing templates:

- The Least Squares Method (LSM);
- The Neural Network Method (NNM).

The LSM is a standard fitting and matching technique that is used in many areas of signal and image processing. The LSM matching technique works when the words compared are of the same length, and when corresponding times in separate utterances of a word represent the same phonetic features. In practice, speakers vary their speed of speaking and often do so non-uniformly so that different voicing of the same word can have the same total length but may differ in the middle.

3.5 Introduction to Speech Synthesis

Speech synthesis and the automatic generation of speech waveforms have been under development for several decades [26, 27]. Recent progress in speech synthesis has produced synthesizers with very high intelligibility but the sound quality and naturalness still remain a major problem. However, the quality of present products has reached an adequate level for several applications, such as multimedia and telecommunications. With some audiovisual information or facial animation (talking head) it is possible to increase speech intelligibility considerably [28]. Some methods for audiovisual speech have been recently introduced by [26, 29, 30]. Some milestones of speech synthesis development are shown in Figure 3.3.

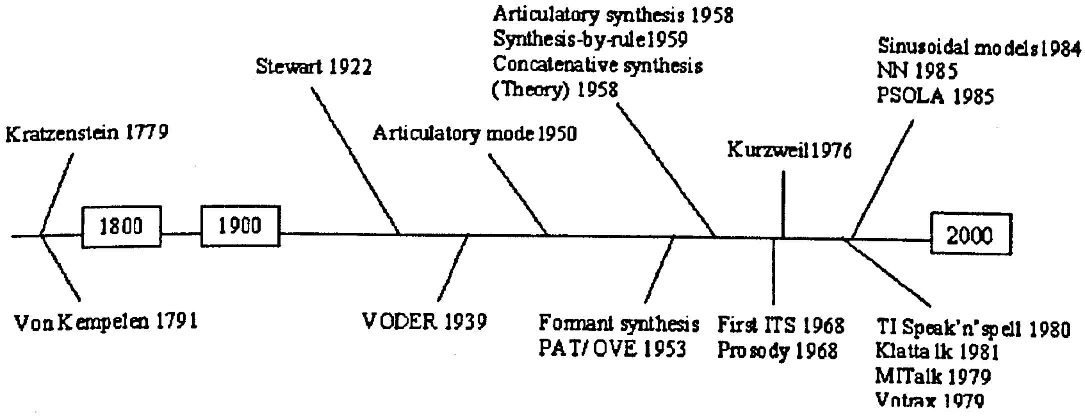

Figure 3.3: Some milestones in speech synthesis [28].

3.6 Methods, Techniques, and Algorithms

Synthesized speech can be produced by several different methods. The methods are usually classified into three groups:

- Articulatory synthesis, which attempts to model the human speech production system directly.

- Formant synthesis, which models the pole frequencies of the speech signal or transfer function of the vocal tract based on a source-filter-model.
- Concatenative synthesis, which uses different length pre-recorded samples derived from natural speech.

The formant and concatenative methods are the most commonly used in present synthesis systems. The formant synthesis was dominant for a long time, but today, the concatenative method is becoming more and more popular. The articulatory method is still too complicated for high quality implementations, but may arise as a potential method in the future.

3.6.1 Linear Prediction Based Methods

Linear predictive methods are originally designed for speech coding systems, but may also be used in speech synthesis. In fact, the first speech synthesizers were developed from speech coders. Like formant synthesis, the basic LPC is based on the source-filter-model of speech. The digital filter coefficients are estimated automatically from a frame of natural speech.

The basis of linear prediction is that the current speech sample $y(n)$ can be approximated or predicted from a finite number of previous p samples $y(n-1)$ to $y(n-k)$ by a linear combination with a small error term $e(n)$ called a residual signal. Thus,

$$y(n) = e(n) + \sum_{k+1}^{p} a(k)y(n-k) \tag{3.7}$$

and

$$e(n) = y(n) - \sum_{k+1}^{p} a(k)y(n-k) = y(n) - \widetilde{y}(n) \tag{3.8}$$

where $\widetilde{y}(n)$ is a predicted value, p is the linear predictor order, and $a(k)$ are the linear prediction coefficients which are found by minimizing the sum of the squared errors over a frame. Two methods, the covariance method and the autocorrelation method, are commonly used to calculate these coefficients and only with the autocorrelation method is the filter guaranteed to be stable [27, 31].

In the synthesis phase, the excitation is approximated by a train of impulses for voiced sounds and by random noise for unvoiced. The excitation signal is then gained and filtered with a digital filter for which the coefficients are $a(k)$. The filter order is typically between 10 and 12 at 8 kHz sampling rate, but for higher quality at 22 kHz sampling rate, the order needed is between 20 and 24 [27, 32]. The coefficients are usually updated every 5-10 ms.

The main deficiency of the ordinary LP method is that it represents an all-pole model, which means phonemes that contain anti-formants such as nasals and nasalized vowels are poorly modelled. The quality is also poor with short plosives because the time-scale events may be shorter than the frame size used for analysis. With these deficiencies, the speech synthesis quality with standard LPC methods is generally considered poor, but with some modifications and extensions for the basic model, the quality may be increased.

Warped Linear Prediction (WLP) takes advantages of the human hearing properties and the order of filter needed is reduced significally from orders 20-24 to 10-14 with 22 kHz sampling rate [32, 33]. The basic idea is that the unit delays in a digital filter are replaced by the following all-pass sections

$$\tilde{z}^{-1} = D_1(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}} \tag{3.9}$$

where λ is a warping parameter between -1 and 1 and $D_1(z)$ is a warped delay element and with Bark scale it is $l = 0.63$ with sampling rate of 22 kHz. WLP provides better frequency resolution at low frequencies and worse at high frequencies. However, this is very similar to the human hearing properties [32].

Several other variations of the linear prediction method have been developed to increase the quality of the basic method [34, 35]. With these methods, the excitation signal is different from the ordinary LP method and the source and filter are no longer separated. These kind of variations are for example Multi-Pulse Linear Prediction (MLPC) where the complex excitation is constructed from a set of several pulses, Residual Excited Linear Prediction (RELP) where the error signal or residual is used as an excitation signal and the speech signal can be reconstructed exactly, and Code Excited Linear Prediction (CELP) where a finite number of excitations used are stored in a finite codebook [36].

3.6.2 Sinusoidal Models

Sinusoidal models are based on a well-known assumption that the speech signal can be represented as a sum of sine waves with time-varying amplitudes and frequencies

[27, 37, 38]. In the basic model, the speech signal $s(n)$ is modeled as the sum of a small number L of sinusoids

$$s(n) = \sum_{l=1}^{L} A_l \cos(w_l n + \phi_l) \tag{3.10}$$

where A_l and ϕ_l represent the amplitude and phase of each sinusoidal component associated with the frequency track ω_l. To find the parameters A_l and ϕ_l, the DFT of windowed signal frames is calculated and the peaks of the spectral magnitude are selected from each frame (see Figure 3.4). The basic model is also known as the McAulay/Quatieri Model and has also some modifications such as ABS/OLA (Analysis by Synthesis / Overlap Add) and Hybrid / Sinusoidal Noise models [38].

While the sinusoidal models are very suitable for representing periodic signals, such as vowels and voiced consonants, the representation of unvoiced speech becomes problematic [38].

Sinusoidal models are also used successfully in singing voice synthesis [38, 39]. The synthesis of singing differs from speech synthesis in many ways. In singing, the intelligibility of the phonemic message is often secondary to the intonation and musical qualities. Vowels are usually sustained longer in singing than in normal speech, and naturally, easy and independent controlling of pitch and loudness is also required. The best-known singing synthesis system is the LYRICOS, which was developed at Georgia Institute of Technology. The system uses sinusoidal-modeled segments from an inventory of singing voice data collected from a human vocalist maintaining the characteristics and perceived identity. The system uses a standard MIDI interface

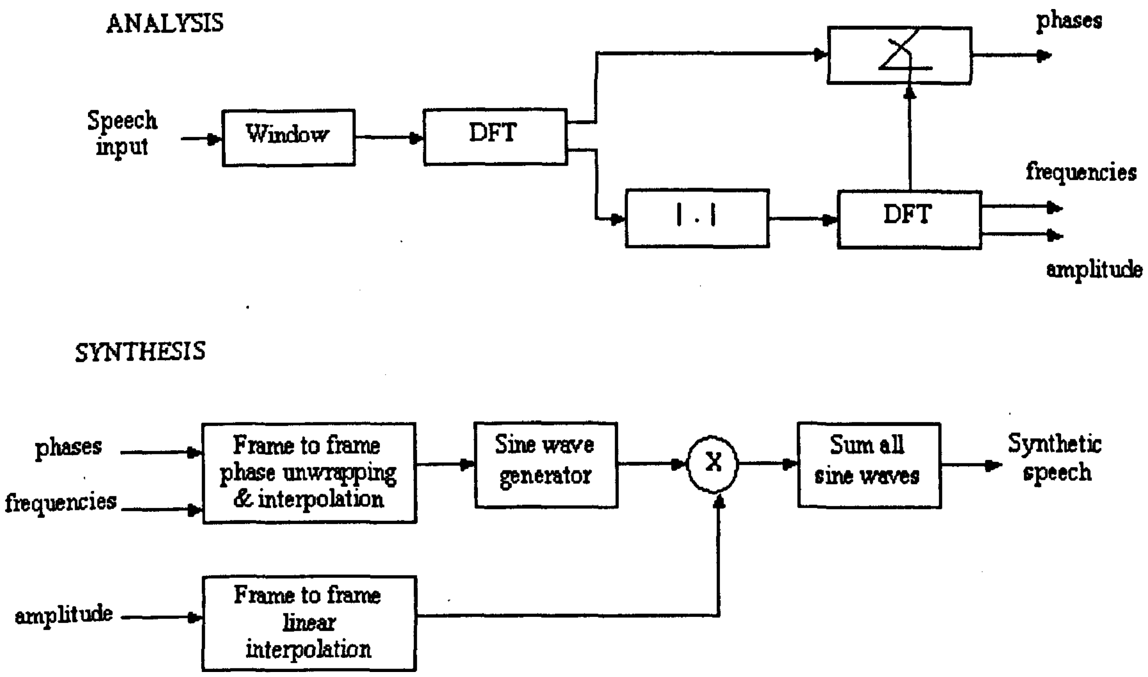

Figure 3.4: Sinusoidal analysis / synthesis system [38].

where the user specifies a musical score, phonetically spelled lyrics, and control parameters such as vibrato and vocal effort [39].

3.7 References

[1] J.Makhoul, “Linear prediction: A tutorial review,” *Proc.IEEE*, vol. 63, pp. 561–580, April 1975.

[2] J.D. Markel and A.H.Gray, *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.

[3] B.A.Doutrich, L.R. Rabiner, and T.Martin, “On the effect of varying filter bank parameters on isolated word recognition,” *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-31, pp. 793–806, Aug. 1983.

[4] A.V. Oppenheim and R.W.Schafer, *Digital Signal Processing*, Prentice-Hall, 1975.

[5] L. R. Rabiner and R. W .Schafer, *Digital Processing of Speech Signal*, Prentice-Hall, 1978.

[6] S.B.Davis and P.Mermelestein, “Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentence,” *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-28, pp. 357–366, Aug. 1980.

[7] S.B.Davis and P.Mermelestein, “Subdivision of the audible frequency range into critical bands,” *J.Acoust. Soc.Am.*, vol. 33, pp. 248, Feb. 1961.

[8] G.White and R.B. Neely, “Speech recognition experiments with linear presiction , bandpass filtering, and dynamic programing,” *IEEE Trans. Acoust. Speech Siganl Processing*, vol. ASSP-27, pp. 183–188, April 1979.

[9] E.Zwicker and E.Terhadt, "Analytical expression for critical-band rate and critical bandwidth as a function of frequency," *J.Acoust. Soc. Am.*, vol. 68, pp. 1523–1525, Nov. 1980.

[10] K.Shirai and M.Honda, "Estimation of articulatory parameters from speech waves," *Electronic and Communication in Japan*, vol. 61, no. 5, pp. 1–8, May 1978.

[11] M.Mrayati, R.Carre, and B.Guerin, "Distinctive regions and modes: A new theory of speech production," *Speech Communication*, vol. 7, pp. 257–286, Oct. 1988.

[12] R.Carre and M.Mrayati, "New concept in acoustic-articulatory-phonetic relation perspectives and application," *Proc. IEEE ICASSP-89*, vol. 7, pp. 231–234, May 1989.

[13] M.Blomberg, R.Carlson, K.Elenius, and B.Grantsrom, "Auditory models in isolated word recognition," *Proc. IEEE ICASSP-84*, vol. 7, pp. 17.9.1- 4, May 1984.

[14] Z.Wicker and E.Terhadt, "Automatic speech recognition using psychoacoustic models," *J.Acoust. Soc. Am.*, vol. 65, pp. 478–489, Feb. 1979.

[15] J.Cohen, "Application of an adaptive auditory model to speech recognition," *J.Acoust. Soc. Am.*, vol. 78, pp. s50(A), Feb. 1985.

[16] R.C.Gonzalez and P.Wintz, *Digital Image Processing*, Addison-Wesley, Reading, Mass, 2nd edition, 1987.

[17] S.Furui, "Cepstral analysis technique for automatic speech verification," *IEEE Transactions on Acoustics Speech .and Signal Processing*, vol. 29, pp. 254–272, 1981.

[18] C. Rowden, *Speech Processing*, McGraw-Hill Book Company, UK, 1992.

[19] L.R.Rabiner and J.G.Wilpon, "A simplified robust training procedure for speaker-trained isolated word recognition system," *J. Acoust. Soc. Am*, vol. 68, pp. 1271–1276, november 1990.

[20] Y.I.Liu, "On creating averaging templates," *Proc. IEEE ICASSP-84*, pp. 9.1.1–4, 1984.

[21] S.Levinson, L.R.Rabiner, A.Rosenberg, and I.Wilpon, "Interactive clustering techniques for selecting speaker-independent reference templates for isolated word recognition," *IEEE Trans. Acoust Speech Signal Processing*, vol. ASSP-27, pp. 134–141, April 1979.

[22] L.R.Rabiner and I.Wilpon, "Considerations in applying clustering technique to speaker-independent word recognition," *I. Acoust Soc. Am.*, vol. 66, pp. 663–673, September 1979.

[23] P.Lippmann, "An introduction to computing with neural nets," *IEEE ASSP Magazine*, pp. 4–22, April 1987.

[24] A.Kranse and H.Hackbarth, "Scaly artificial neural networks for speaker-independent recognition of isolated words," *IEEE Proc. ICASSP-89*, pp. 21–24, 1989.

[25] H.Sakoe, R.Isotani, K.Yoshida, K.Iso, and T.Watanabe, "Speaker- independent word recognition using dynamic time programminig neural networks," *IEEE Proc. ICASSP-89*, pp. 29–32, 1989.

[26] J.Santen, J.Olive, and J.Hirschberg, *Progress in Speech Synthesis*, Springer-Verlag New York Inc, 1997.

[27] K.Kleijn and K.AliwalP, *Speech Coding and Synthesis*, Elsevier Science B.V, The Netherlands, 1998.

[28] J.Beskow, M.Dahlquist, B.Granstrm, and M.Lundeberg, "Disability, feasibility, and intelligibility," *Proceedings of Fonetik97*, 1997.

[29] A.Breen, E.Bowers, and W.Welsh, "An investigation into the generation of mouth shapes for a talking head," *Proceedings of ICSLP 96*, vol. 4, 1996.

[30] J.Beskow, "Talking heads - communication, articulation and animation," *Proceedings of Fonetik*, pp. 53–56, 1996.

[31] I.Witten, *Principles of Computer Speech*, Academic Press Inc, 1982.

[32] M.Karjalainen, T.Altosaar, and M.Vainio, "Speech synthesis using warped linear prediction and neural networks," *Proceedings of ICASSP*, 1998.

[33] U.Laine, M.Karjalainen, and T.Altosaar, "Warped linear prediction (wlp) in speech synthesis and audio processing," *Proceedings of ICASSP94*, vol. 3, pp. 349–352, 1994.

[34] R.Donovan, *Trainable Speech Synthesis*, Phd. thesis, Cambridge University Engineering Department, Englnd, 1996.

[35] D.Childers and H.Hu, "Speech synthesis by glottal excited linear prediction," *Journal of the Acoustical Society of America, JASA*, vol. 4, pp. 2026–2036, 1994.

[36] N.Campbell, "Chatr: A high-definition speech re-sequencing system," *Acoustical Society of America and Acoustical Society of Japan, Third Joint Meeting*, Dec 1996.

[37] R.McAulay and T.Quatieri, "Speech analysis-synthesis based on sinusoidal representation," *Proceedings of ASSP-34*, vol. 4, 1986.

[38] M.Macon, *Speech Synthesis Based on Sinusoidal Modeling*, Phd thesis, Georgia Institute of Technology, 1996.

[39] M.Macon, L.Jensen-Link, J.Oliverio, M.Clements, and E.George, "A singing voice synthesis system based on sinusoidal modeling," *Proceedings of ICASSP97*, 1997.

Chapter 4

Fractal Speech Recognition

4.1 Introduction

Speech is man's most natural channel of communication, so it is only natural that it should be the subject of much work in speech recognition. Recognising speech is difficult because of the nature of the speech communication process, which carries many messages. Transmitting a linguistic message is most often the primary purpose of speech communication and it is the recognition of this message by machine that would be most useful. The first step of a basic speech recognition system is to extract from the speech samples a 'good' parametric representation.

In this chapter a speech recognition system is performed using a combination of Mel-Frequency Cepstral Coefficients MFCCs with a nonlinear dynamic invariant 'fractal dimension' (D). This combination leads to a more accurate speech recognition system with more accurate results. The best results are obtained when the fractal dimension is combined with the Mel-Frequency Cepstral Coefficients MFCCs. It is shown that the suggested method add a new feature which is reported for the first time.

The corpus used is the speech data performed on the Texas Instruments Massachusetts Institute of Technology (TIMIT) database (see Appendix B).

4.2 Speech Parameters

A speech signal is a highly redundant signal, which carries linguistic messages as well as other information about the speaker, regarding their physiology, psychology, etc.. Feature measurement, some times called feature extraction, is basically a data reduction technique, whose features should meet the following criteria [1]:

- Insensitive to extraneous variables (i.e. emotion, state of talker etc.)
- Stable over long periods of time
- Frequently occurring
- Easy to measure
- Not correlated with other features

In general, it is impossible to find features that meet all of these requirements at once, and compromises are inevitable. The features are usually selected intuitively. A feature is judged by how-well it separates recognition classes from one another. In fact, the selection of the best parametric representation of the acoustic data is an important task in the design of any speech recognition system.

Several parametric representations of speech signal, such as LPC parameters, filter bank parameters and cepstral parameters, have been of interest to many researchers. The objective of these parametric representations are to compress the speech data by eliminating information not pertinent to the phonetic analysis of the data and to enhance those aspects of the signal that contribute significantly to the detection of phonetic differences. In this work, the fractal dimension (D) and the Mel-frequency cepstral coefficients ($MFCCs$) are chosen as the new parametric representation of the speech signal.

4.3 Application of the fractal dimension to speech recognition

The dynamics of speech airflow might create small or large degrees of turbulence during the production of speech sounds by the human-tract systems. Static airflow and acoustic characteristics of turbulent speech, e.g. fricative and stop sounds with aspiration, have been studied by several researchers; references and related discussion can be found in [2, 3, 4]. While the majority of work in this area has mainly associated turbulence in speech with consonants, it is also possible to have vowels uttered with some amount of aspiration which adds some small degree of turbulence to them [5]. Most approaches to modelling speech turbulence at the speech-waveform level have focused on the random nature of the corresponding signal component. Another important aspect of speech sounds that contain frication or aspiration is the high degree of geometrical complexity and fragmentation of their time waveform; due to lack of a better approach, this has been left unmodeled and treated in the past as noise [5].

In this research, we use the theory of fractals [6] to model the geometrical complexity of speech waveforms via their fractal dimension and related fractal parameter, which quantifies the degree of signal fragmentation. In the next sub-section, we provide some motivation and justification from the field of speech aerodynamics for using the fractal dimension to quantify the degree of turbulence in speech signals.

4.3.1 Speech Aerodynamics and Fractals

Conservation of momentum in the airflow during speech production yields the Navier-Stokes governing equation .[7]:

$$\rho(\frac{\partial \mathrm{u}}{\partial t} + \mathrm{u} \cdot \nabla \mathrm{u}) = -\nabla p + \nabla \mu^2 \mathrm{u}, \tag{4.1}$$

where ρ is the air density, p is the air pressure, u is the (vector) air-particle velocity, and μ is the (assumed constant) air-viscosity coefficient. It is assumed that flow compressibility is negligible [valid since in speech flow (Match numbers)^2 $<<$ 1], and hence $\nabla \cdot \mathrm{u} = 0$. An important parameter characterising the type of flow is the Reynolds number $Re = \frac{\rho UL}{\mu}$, where U is a velocity scale for u and L is a typical length scale, e.g. the tract diameter. For the air we have very low μ, and hence high Re. This causes the inertia forces [in the left-hand of 4.1] per unit volume to have a much larger order of magnitude than the viscous forces $\mu \nabla^2 \mathrm{u}$. While μ is low and may not play an important role for the speech airflow through the interior of the vocal tract, it is essential for the formation of boundary layers along the tract boundaries and for the creation of vortices. A *vortex* is a region of similar (or constant) vorticity ω, where $\omega = \nabla \times \mathrm{u}$. Vortices in the airflow have been experientally found above the glottis by [8, 9], and theoritically predicted by [10, 8, 11], using simple geometries. There are several mechanisms for the creation of vortices: (1) velocity gradients in boundary layers,(2) separation of flow, which can easily happen at cavity inlets due to adverse pressure gradients (see [10, 8] for experimental evidence of separated flow during speech production), and (3) curved geometry of tract boundaries, where due to the dominant inertia forces the flow follows the curvature and develops rotational components. After a vortex has been created, it can propagate downstream as governed by the vorticity equation [7]:

$$\frac{\partial \omega}{\partial t} + \mathrm{u} \cdot \nabla \omega = \omega \cdot \nabla \mathrm{u} + \nu \nabla^2 \omega, \nu = \mu \rho \qquad (4.2)$$

the term $\omega . \nabla \mathrm{u}$ causes vortex twisting and stretching, whereas $\nu \nabla^2 \omega$ produces diffusion of vorticity. As *Re* increases (e.g., in fricative sounds or during loud speech), all these phenomena may lead to instabilities and eventually result in *turbulent flow*, which is a "state of continuous instability" [7] characterised by broad-spectrum rapidly varying (in space and time) velocity and vorticity. The transition to turbulence during speech production may occur for lower *Re* closer to the glottis because there is an air jet flowing out from the vocal cords, and for jets, turbulence starts at a much lower *Re* than for flows attached to walls (as is the case downstream in the vocal tract). Modern theories that attempt to explain turbulence [7] predict the existence of eddies (vortices with characteristic size λ) at multiple scales. According to the energy-cascade theory, energy produced by eddies with large size λ (of the order of the boundary-layer thickness) is transferred hierarchically to the small-size eddies, which actually dissipate this energy due to viscosity. The result is compounded by the Kolmogorov law

$$E(k, r) \propto r^{2/3} k^{-5/3}, \qquad (4.3)$$

where $k = \frac{2\Pi}{\lambda}$ is the wave number in a finite nonzero range, r is the energy-dissipation rate, and $E(k, r)$ is the velocity wave spectrum, i.e. Fourier transform of spatial correlations. This multiscale structure of turbulence can in some cases be quantified by *fractals* [5].

All the above theoretical considerations, and the fact that the speech signal is produced by a non linear dynamical system, which often generates small or large degrees of turbulence, motivated our study of its fractal aspects. One of the main quantitative ideas that we focused on is the fractal dimension of speech signals, because it can quantify their graph's roughness (fragmentation).

As shown later in this chapter, the fractal dimension 'D' as a single feature is able to achieve a good recognition when a small number of phoneme classes is used. However, the problem of class limitation is solved by adding 13 $MFCCs$. It is also important to mention that the Mel Frequency Cepstral coefficients 'MFCCs' method uses short time Fourier Transform for feature extraction and for some class of phonemes like fricatives and plosives it can not extract good features as these latter are non stationary speech signals. However, the fractal dimension 'D' is not sensitive to that and can process stationary as well as non-stationary speech signals. Therefore, the fractal dimension 'D' provides different information than supplied with the $MFCCs$ and the combination of 13 $MFCCs$ with one fractal dimension 'D' (as shown later) shows an improvement in the recognition of speech phonemes.

4.4 Evaluation of the Fractal Dimension

In order to test the validity and accuracy of the different methods for computing the fractal dimension, a speech word 'close' with a known fractal dimension D was used. From Table 2.1 in chapter 2, it has been shown that speech signals have a fractal dimension within the range 1 and 2. These values of 'D' have been used as known fractal dimensions for the word 'close' in our evaluation process. For each value of 'D' within this range a fractal signal is computed, then used with the four methods mentioned later to evaluate the new fractal dimension 'D' for comparison purpose. This procedure is called the 'inverse solution '[12] and can be summarized in two steps as follow:

1. Given D compute f

2. Given f compute D

Where D is the fractal dimension of the speech signal and f its fractal signal.
The fractal dimensions of the speech word 'close' have been then evaluated using the Walking- Divider, the Box Counting, the continuous box counting and the Power Spectrum Method (PSM). The results are given in Table 4.1.

From Table 4.1, it is clear that the Power Spectrum Method provides the most consistently accurate results throughout the range 1 to 2. The box counting method provides good results for fractal dimensions with a value below 1.4. After this, the fractal dimension is below the original value of the speech word used; for a value of 2.0, the box counting method returns a value of 1.615. The Walking-Divider Method provides a good approximation of the fractal dimension for values below 1.5, returning results that are slightly higher than those produced by the Box Counting Method and the continuous Box Counting Method. For an original value of 2.0 the Walking-Divider Method returns a value of 1.665. Of the four methods tested, the PSM is the fastest as it is based on a non-iterative approach based on the least square estimate which relies on the use of an FFT.

Original Value of the fractal Dimension	Walking-Divider Method	Box Counting Method	Continuous Box Counting Method	Power Spectrum Method
1.0	1.220	1.088	1.178	1.025
1.1	1.268	1.178	1.259	1.130
1.2	1.268	1.163	1.269	1.214
1.3	1.351	1.229	1.276	1.301
1.4	1.370	1.263	1.393	1.400
1.5	1.432	1.348	1.419	1.495
1.6	1.571	1.397	1.470	1.599
1.7	1.552	1.459	1.516	1.696
1.8	1.638	1.513	1.574	1.771
1.9	1.639	1.572	1.621	1.915
2.0	1.665	1.615	1.647	1.968

Table 4.1: Evaluation and comparison of fractal dimensions.

The accuracy, efficiency and the versatility of the PSM lead naturally to its use in many areas of signal processing [12]. The test discussed above provides confidence in the realization that the PSM is the most appropriate technique for applications to speech processing. Thus, in the following sections the fractal dimension computation is obtained from the PSM only.

4.4.1 Fractal dimension of speech signals

Four techniques for computing the fractal dimension have been used in this research as discussed in Chapter 2. However, the best of the four and the most used one , as seen previously, is the Power Spectrum Method (PSM) because it is easy to implement, computationally less time consuming and is not based on iterative procedures. PSM is generalisable and potentially more accurate computationally among the three other methods. One of its main advantage is that the computation of the fractal dimension D is based on an explicit formula ($\beta = 5 - 2D$).

4.4.1.1 Power Spectrum Method

Direct application of the PSM for computing the fractal dimension of arbitrary speech signals leads to a wide range of values, many of which lie outside the natural range [13, 14]. This is not surprising, since many speech waveforms will not conform to patterns, which are statistically self-affine with a single spectral signature of the type $1/k^q$. It is expected that parts of speech are fractal in nature while other parts are not. In other words, like any other model for signal analysis, a fractal model cannot be assumed to be applicable to all aspects of speech. As with any other signal, one should expect it to be composed of both fractal and non-fractal components, particularly with highly non-stationary waveforms such as those observed in speech. A new step in the power spectrum method is therefore required in order to force a speech signal or component waveform to conform to a spectral signature of a fractal type, in particular, some appropriate low-pass filter. Of all the possible low-pass filters, a filter of the type $1/k$ is the most appropriate as it conforms to the case for q = 1.

The implementation of the PSM consists of applying the FFT to the speech signal in order to obtain a spectral representation of the phoneme. A pre-filter step is then used to adjust the estimated values of the fractal dimension to fit within the range 1 and 2. The power spectrum of the pre-filtered signal is computed and the least square approach is applied for the calculation of the power exponent β Eqn.2.17, which yields to the computation of the fractal dimension D Eqn.2.19.

4.5 Experiments on Computing Fractal Dimension of Speech Signals

The speech waveforms of Figure 4.1 show three phonemes /f/,/u/ and /o/ spoken by a male and a female speaker. Their fractal dimension is shown in Figure 4.2 where we can notice that the fractal dimensions for female are higher than male's, this is due to the fact that the speech waveforms of female speakers are rougher which means more *zero crossing* and in turn means more irregularities, hence a higher fractal dimension. In fact, in term of texture a signal with a fractal dimension close to one looks smoother that a signal with D closer to two.

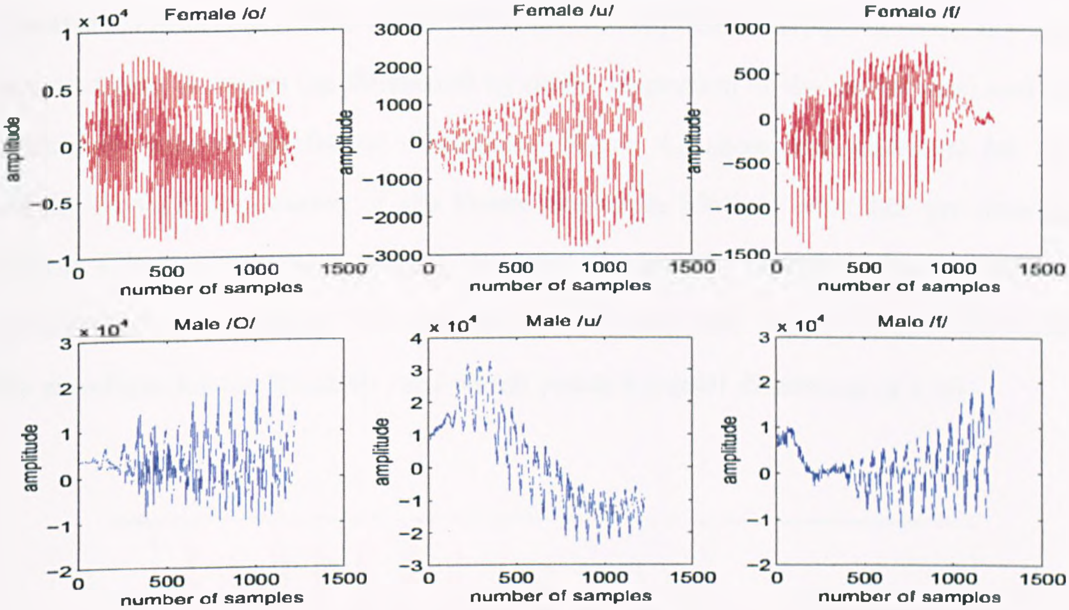

Figure 4.1: Phonemes Waveforms.

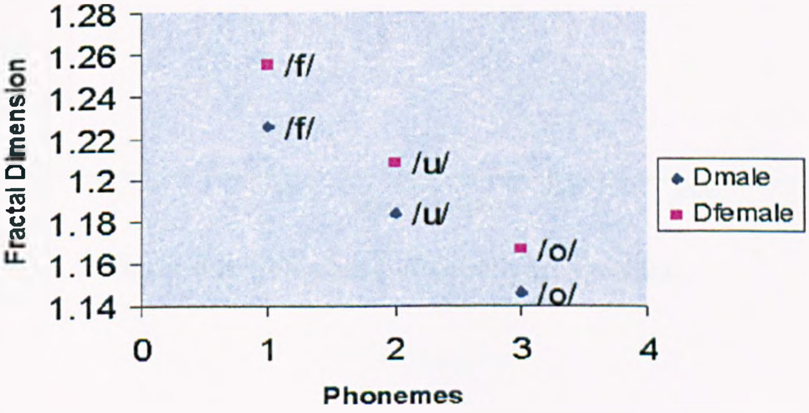

Figure 4.2: Fractal Dimension values.

The following results are based on a single male speaker of the vowel /o/ and the fricatives /z/ and /sh/. The results illustrate the expected increase in frequency content and zero crossings (as illustrated by direct inspection of the waveforms) and the increasing value of the fractal dimension. Figure 4.3 shows the waveform for /o/, which, through application of the Power Spectrum Method with $1/k$ pre-filtering, returns a fractal dimension of 1.20. Figure 4.4 shows the waveform for the fricative /z/ which is characterised by a fractal dimension of 1.38. Finally, Figure 4.5 shows the waveform for the fricative /sh/ which yields a fractal dimension of 1.58.

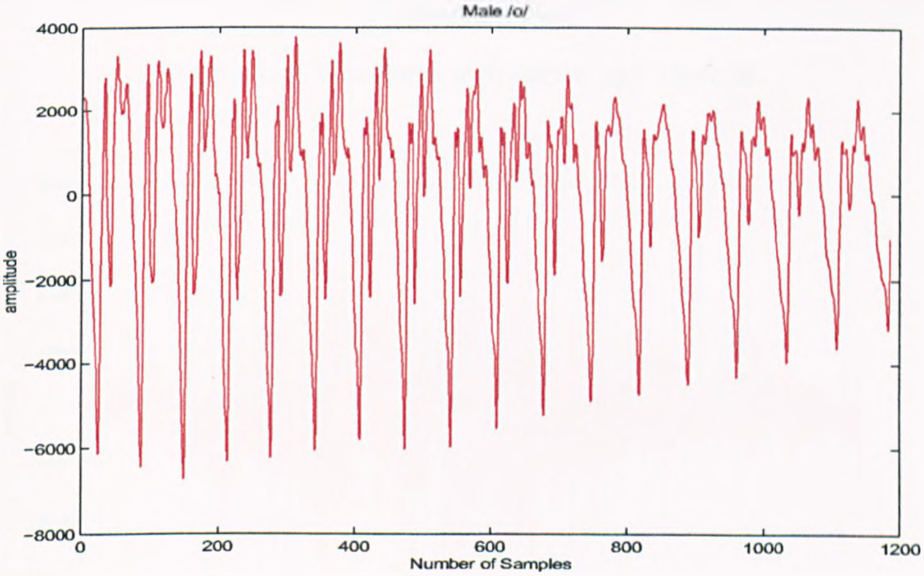

Figure 4.3: Waveform of vowel /o/; D=1.20.

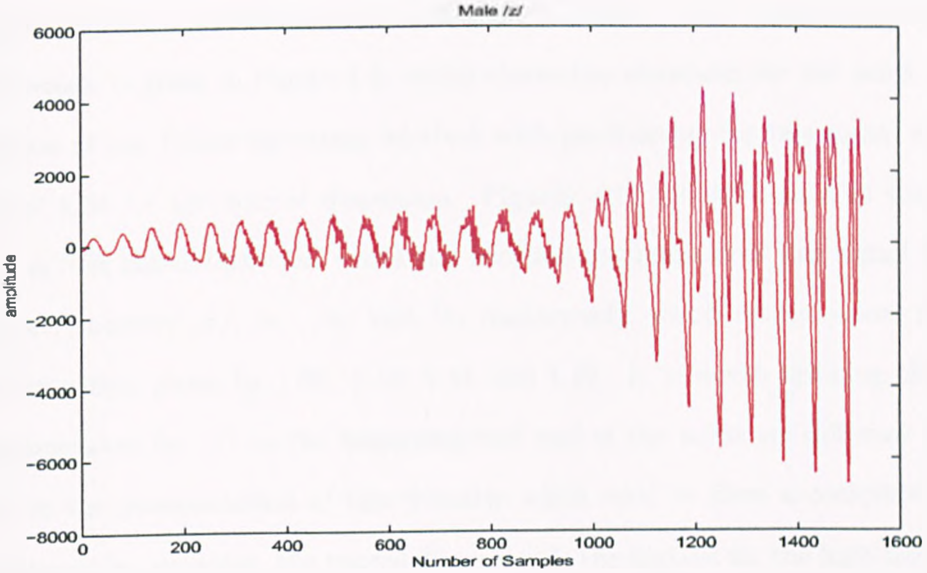

Figure 4.4: Waveform of fricative /z/; D=1.38.

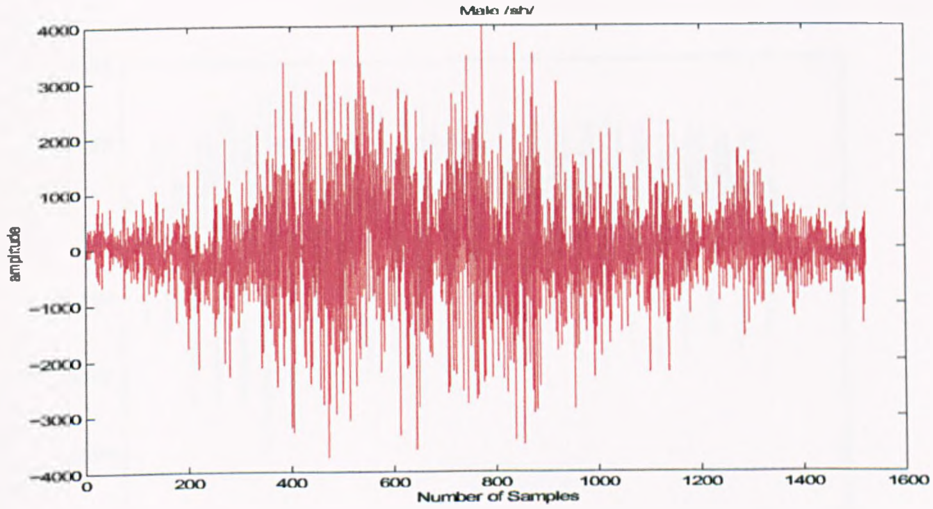

Figure 4.5: Waveform of fricative /sh/; D=1.58.

Another example of computing the fractal dimension for a single isolated word and a female speaker is given in Figure 4.6, which shows the waveform for the word /test/ Application of the Power Spectrum Method with pre-filtering for this signal returns a value of 1.34 for the fractal dimension. Figures 4.7, 4.8, 4.9 and 4.10 show the waveforms that characterize the vowel and fricative components of this signal for the word /test/, namely /t/, /e/, /s/ and /t/ respectively, which return values for the fractal dimension given by 1.06, 1.36, 1.41 and 1.19. It is worth noticing that the fractal dimension for /t/ at the beginning and end of the word are different due to changes in the pronunciation of this fricative when used to form a complete word. Also as should be expected, the fractal dimension is the highest for the high frequency fricative /s/.

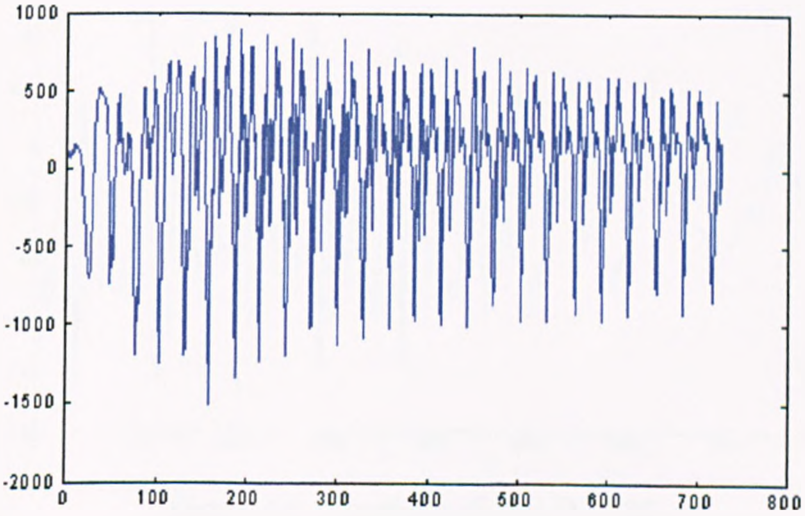

Figure 4.6: Waveform of word /test/; D=1.34.

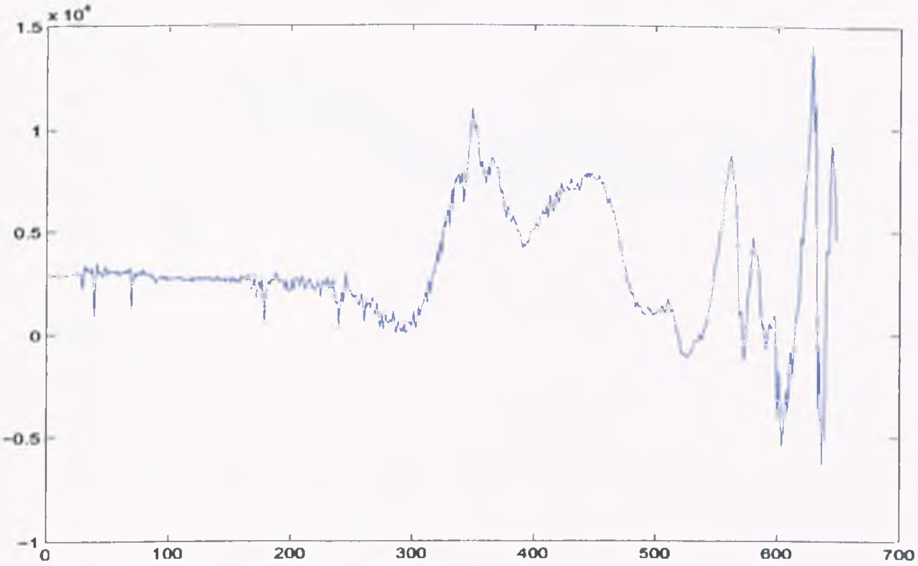

Figure 4.7: Waveform of /t /; D=1.06.

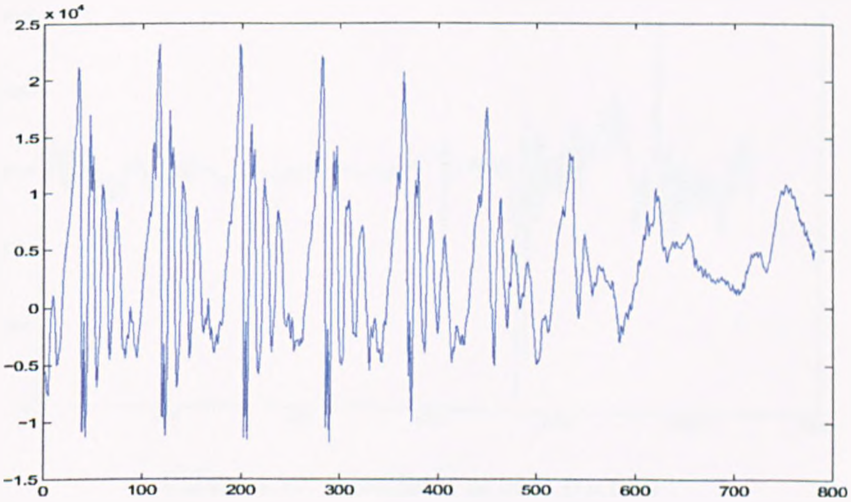

Figure 4.8: Waveform of /e/; D=1.36.

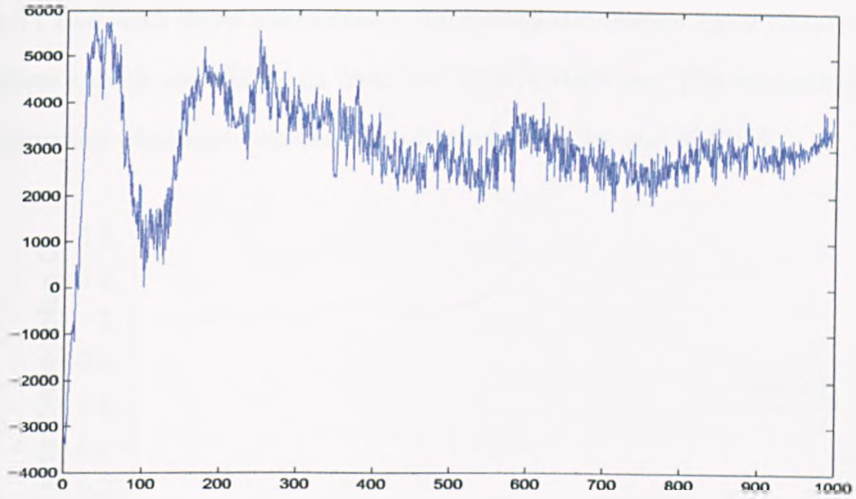

Figure 4.9: Waveform of /s/; D=1.41.

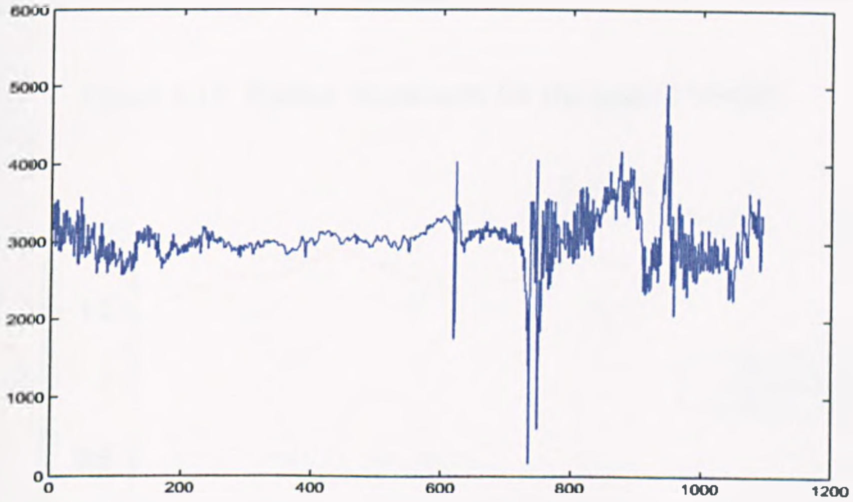

Figure 4.10: Waveform of /t/; D=1.19.

Figures 4.11 and 4.12 show respectively the fractal dimension for 9 classes of vowels and fricatives speech sounds taken from the TIMIT datbase. The computation of the fractal dimension obtained has been conducted with the use of PSM.

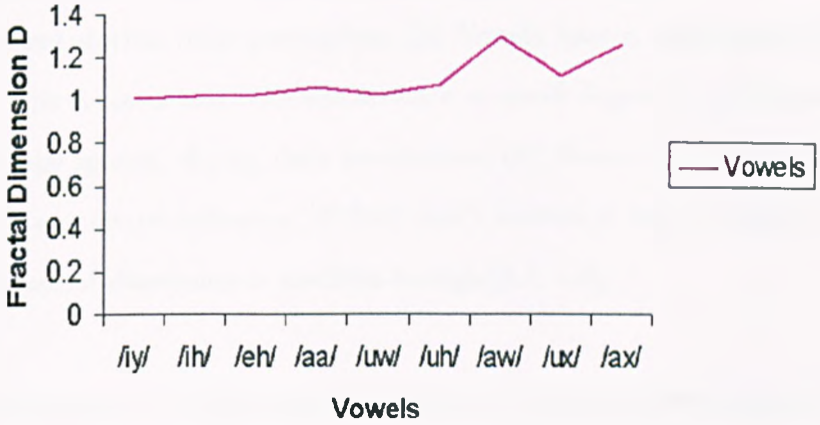

Figure 4.11: Fractal dimensions for the case of Vowels.

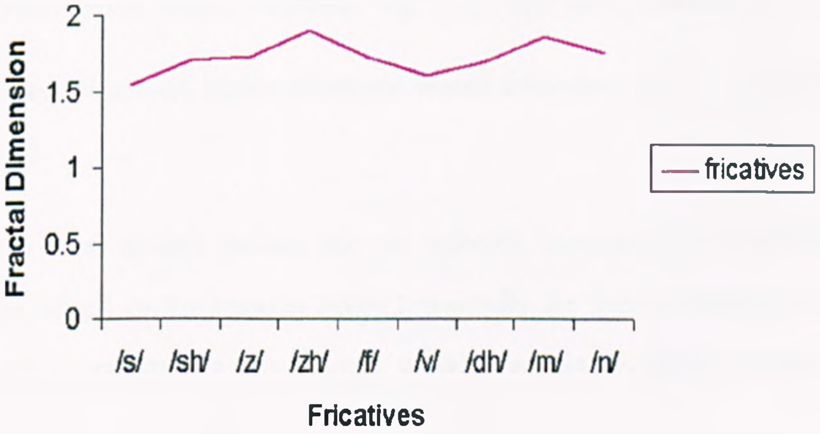

Figure 4.12: Fractal dimensions for the case of Fricatives.

We have conducted many experiments similar to the one shown in Figures 4.11 and 4.12 respectively, from which we concluded the following: (i) Unvoiced fricatives (/f/, /th/, /s/), affricates, stops (during their turbulent phase), and some voiced fricatives like /z/ have a high fractal dimension $\in [1.6, 1.9]$, consistent with the turbulence phenomena present during their production; (ii) Vowels have a small fractal dimension $\in [1, 1.3]$. This is consistent with the absence or small degree of turbulence (e.g. for loud or breathy speech) during their production; (iii) Some voiced fricatives like /v/ and /th/ have a mixed behavior. If they don't contain a fully developed turbulent state, their fractal dimension is medium-to-high [1.3, 1.6].

Thus, we have found that the fractal dimension can roughly distinguish three classes of speech sounds:

- Vowels have a small fractal dimension D.
- Low-turbulence voiced fricatives, e.g. /v/, /th/ have a medium D value.
- Unvoiced fricatives, high-turbulence voiced fricatives, stops, and affricates have large D.

However, for loud speech (where the air velocity increases, and hence turbulence occurs more often) or for breathy voice (especially for female speakers), the fractal dimension of several speech sounds, e.g. vowels may significantly increase [15].

The fractal dimensions discussed above are the global measure of self-affinity for different geometrical objects; however, such global measure can not represent all the fractal characteristics in different levels of complex objects. For digital speech processing, the global fractal dimension is just a reference feature and must be used with other features because the information it supports is limited [15]. To overcome the information limitation of a single global fractal dimension, 13 Mel-Frequency Cepstral Coefficients (MFCCs) are added, which we will introduce later in this chapter.

4.6 Fractal Dimension for phoneme Segmentation

A word is composed of phonemes which are different from one another. Both word segmentation and endpoint detection are based on an approach similar to the convex hull method [16]. If the fractal dimension 'D' vs. window number plot shows a dip, an endpoint or a phoneme boundary is confirmed. The fractal dimension can be used for endpoint detection and therefore segmentation of speech [17]. To obtain the fractal dimension variations across speech waveforms, a moving window with a finite width was used. The moving window of size n, is taken to have s=2^n samples. The fractal dimension in each window is obtained using the Power spectrum method discussed in Chapter 2. The segmentation method is based on the criteria of detecting the peaks in the fractal dimension plot, which means that if the fractal dimension vs. window number plot shows a dip a phoneme boundary is confirmed. Figure 4.13 shows the fractal dimension variation across the word */greasy/*.

From Figure 4.13 it can be noticed that some dips appear only at phonemes boundaries within a specific number of windows. Each window is 512 samples wide. For

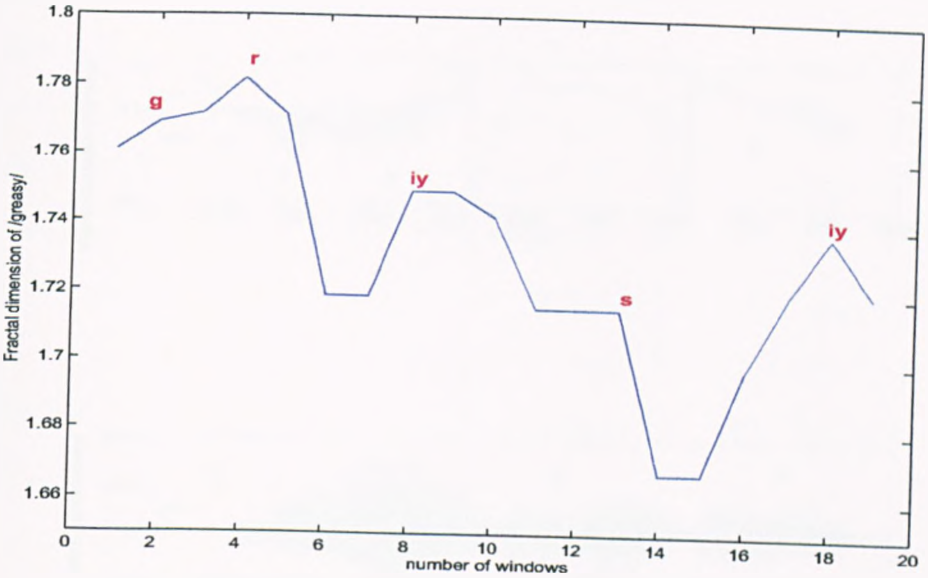

Figure 4.13: Fractal segmentation of word /greasy/.

example the boundary or end point of the phoneme /*iy*/ is detected at window number 8 and since each window has 512 samples, this means that the phoneme's /*iy*/ boundary is detected around 4096 samples, which correspond approximately to its number of samples in Figure 4.14.

A comparison of the fractal segmentation method performance with the ground truth in the TIMIT database has been undertaken to confirm its validity and the results are shown in Figure 4.14, where the red line in the bottom Figure show the segmentation of the word /*greasy*/ as provided by the TIMIT database and the top Figure shows its fractal segmentation. It is clearly seen from Figure 4.14 that the fractal segmentation of each phoneme in the word /*greasy*/ coincide with the segmentation provided by the TIMIT database. It is also worth to mention that the fractal dimension vs.

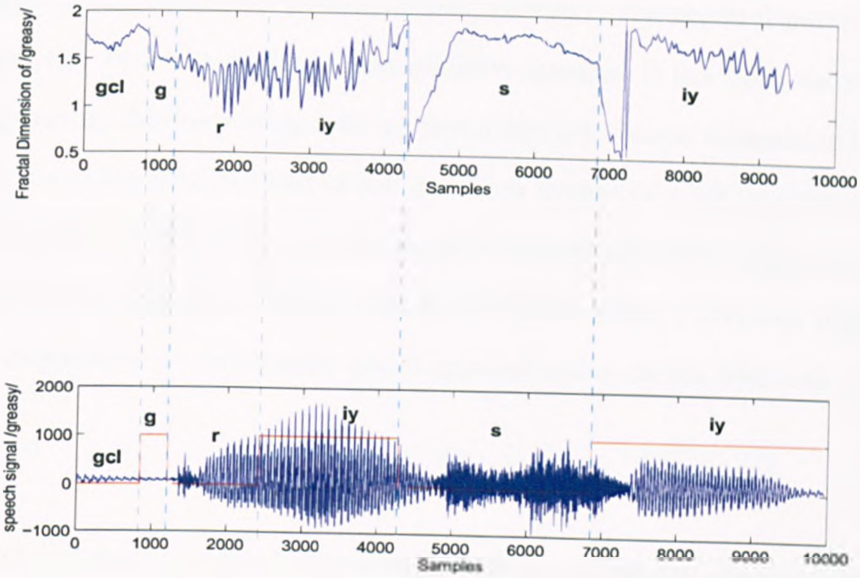

Figure 4.14: Fractal dimension across the word /greasy/.

window number plot as shown in Figure 4.13 satisfy the criteria mentioned previously. To obtain a more accurate segmentation using fractal dimension, the window size can, for example, be reduced to 8 which corresponds to 256 samples.

Finally, segmentation based on fractals as presented is simpler and appears to be more effective than most of the existing methods, which for example, make use of the average zero crossing rate and mean square amplitude [5].

4.7 Mel-Frequency Cepstral Coefficients

Many systems currently use Mel-cepstral coefficients (*Mel-frequency Cepstral Coefficients-MFCCs*) for speech recognition. Mel-cepstral analysis is increasingly replacing the traditional forms of cepstral parameters utilization although cepstrum coefficients at first glance seem to be a good parameter set. However, the cepstral parameters do not consider the structure of the human auditory system. It has been shown in [18] that incorporating this knowledge into the parameter set used in recognition improves performance considerably. Instead of using a linear frequency scale, we can use a logarithmic scale just as that human auditory system does. One of the popular scales is the Mel scale (Appendix A). This scale is almost linear below 1 kHz and logarithmic at higher frequencies. A technically useful approximation to the Mel scale is of the form [19]:

$$y = k \log(1 + f/1000) \tag{4.4}$$

where f is the frequency in Hz, k is a constant. The constant k is computed with the consideration that a tone with a frequency of 1000 Hz is defined as having a pitch of 1000 mels.

Computing the MFCCs makes use of a filter bank, which is placed according to Mel scale. At the output of each filter a log energy coefficient is obtained representing the energy in that band. An inverse DCT is then performed to get back to the cepstral domain and obtain Mel-cepstral coefficients. This procedure is summarised in Figure 4.15. The i^{th} *MFCC* is given as:

$$MFCC_i = \sum_{k=1}^{K} E_k \cos\left[i(k - 1/2)\frac{\pi}{K}\right] \tag{4.5}$$

where $i = 1, 2, ..., N$ and $k = 1, 2, ..., K$. N is the number of required MFCC and K is the number of filters in the filter bank which cover the frequency range of the input speech signal. E_k represents the logarithm of the energy output of the k^{th} filter.

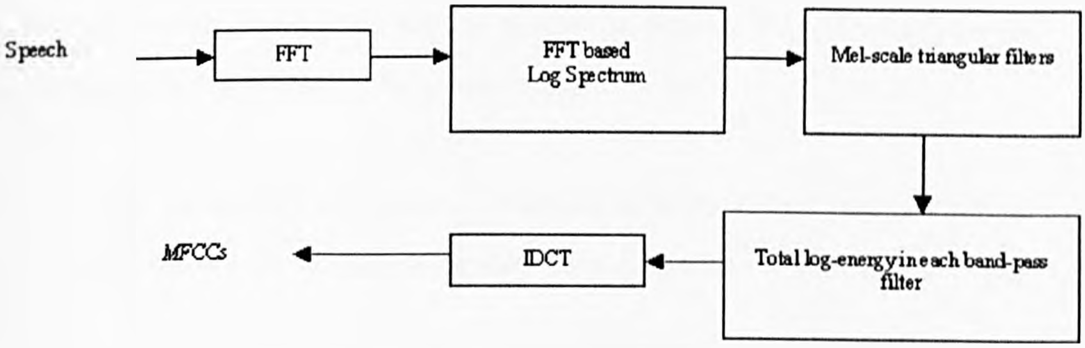
Speech
FFT
FFT based Log Spectrum
Mel-scale triangular filters
Total log-energy in each band-pass filter
IDCT
MFCCs

Figure 4.15: *MFCCs* Computation.

Some of the results of the *MFCC* computation are displayed in Figure 4.16. Note the effect of smoothing performed by the Discrete Cosine Transform (DCT) in Figure 4.16 (d).

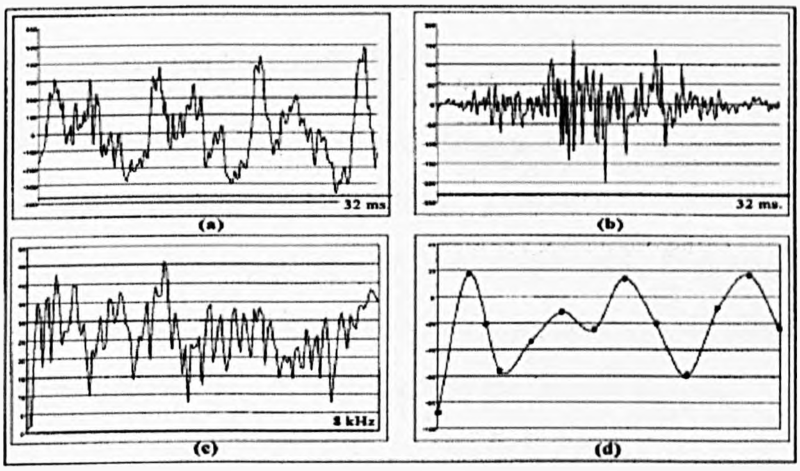

Figure 4.16: Speech waveform of a fragment of phoneme æ (*a*), after pre-emphasis and Hamming windowing (b), power spectrum (c) and MFCC (d)

The DCT has the ability to produce highly uncorrelated features; therefore, the stochastic characterization of the feature process is simpler [20]. The number of MFCC is generally lower than 14 in speech recognition [21].

In this research, 13 *MFCCs* are combined together with the fractal dimension D in order to form a feature set characterising each phoneme, which is used for the recognition.

4.8 System Description

In this section, we have discussed the two matching techniques used for the recognition of phonemes for the case of speaker independent.

4.8.1 Template Matching Techniques

The basic idea involved in template matching is that each word is divided into phonemes and each phoneme is represented by a template (in some cases, more than one), which is a reference pattern created from speech data. Each input phoneme to be recognised is then compared with the stored template and identified as an instance of that phoneme whose template best matches the unknown input.

The matching techniques used in this research are:

- The Least Square Error Method (LSEM)
- The Neural Network Method (NNM)

4.8.1.1 The Least Square Error Method (LSEM)

The LSEM compares the unknown input signal S_i with the standard reference template T_j. The difference between the matched input against the reference template is evaluated as an error function Eqn.4.6 and obviously, the smallest the error is, the more accurate the recognition.

$$E = \sum_{i,j}^{n,m}(S_i - T_j)^2 \tag{4.6}$$

where n is the number of inputs, m is the number of templates, S_i the unknown input and T_j is the standard reference template. The proposed algorithm used for the recognition of speaker-independent speech phonemes is illustrated in Figure 4.17.

This algorithm makes use of the following steps:

Step1: The data are taken from the TIMIT database, which has been subdivided into suggested training and test sets following the criteria that roughly 20 to 30% of the corpus should be used for testing purposes, leaving the remaining 70 to 80% for training.

Step2: Two sentences spoken by all speakers both male and female from 2 major dialect divisions of the United States are used to extract phonemes. Once the phonemes are extracted they are stored in order to be used for testing and training. The training set of the TIMIT is used for template creation and the test set is used as unknown input in the matching process.

Step3: One Fractal dimension 'D' and 13 Mel Frequency Cepstral Coefficients 'MFCCs' are computed and added together forming a new set of feature 'DMFCC' for each phoneme in the test set and training set respectively.

Step4: The unknown phoneme, which is represented by 14 coefficients is compared with the stored template and the error is computed as follows:

$$E = \sum_{i,j}^{n,m} (S_i - T_j)^2 \tag{4.7}$$

Step5: The error is then compared to a threshold, which has been set up by experiment to 0.3 since there is no empirical theory that defines it.

Step6: Finally the recognition performance of the input class is computed as the total number of error below threshold divided by the total number of input phonemes within this class multiplied by 100.

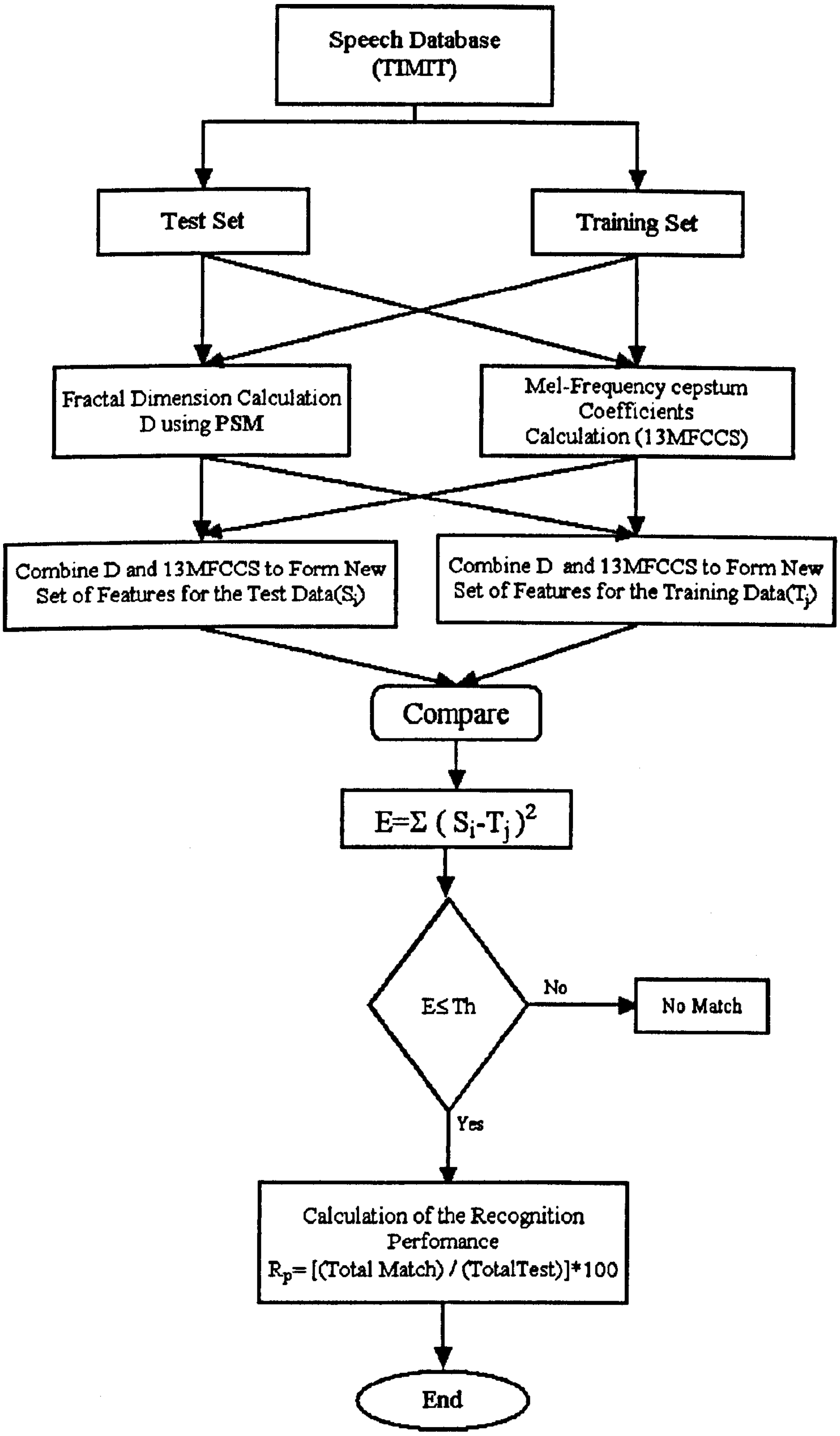
Speech Database
(TIMIT)
Test Set
Training Set
Fractal Dimension Calculation D using PSM
Mel-Frequency cepstum Coefficients Calculation (13MFCCS)
Combine D and 13MFCCS to Form New Set of Features for the Test Data(S_i)
Combine D and 13MFCCS to Form New Set of Features for the Training Data(T_j)
Compare
$E=\Sigma (S_i-T_j)^2$
$E\leq Th$
No
No Match
Yes
Calculation of the Recognition Perfomance
R_p= [(Total Match) / (TotalTest)]*100
End

Figure 4.17: LSE Matching Algorithm.

4.8.1.2 Neural Network Approach

Neural networks are usually used to perform static pattern recognition, that is, to statically map complex inputs to simple outputs, such as an N-array classification of the input patterns [22]. Moreover, the most common way to train a neural network for this task is via back-propagation, whereby the network's weights are modified in portion to their contribution to the observed error in the output unit activations (relative to desired outputs). One of the main advantages of neural networks is the massive parallelism they offer [23].

The neural network model used for the recognition of speaker-independent speech phonemes is shown in Figure 4.18. This makes use of the new set of features, which consist of the combination of the fractal dimension (D) and 13 Mel frequency cepstral coefficients ($MFCCs$).

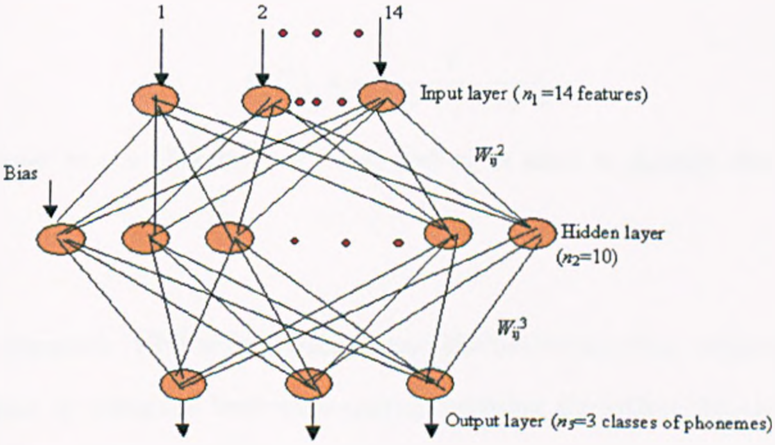
1
2
14
Input layer (n_1=14 features)
W_{ij}^2
Bias
Hidden layer
(n_2=10)
W_{ij}^3
Output layer (n_3=3 classes of phonemes)

Figure 4.18: The configuration of the three-layer NN.

Algorithm Used

First layer (k=1)

I_j^k= input of the j^{-th} node at layer k. $j = 1, ...n_k$

Second layer (k=2) and Third layer (k= 3)

$$I_j^k = \sum_{i=0}^{n_{k-1}} w_{ji}^k o_i^{k-1} \tag{4.8}$$

$j = 1, ..., n_k$, where o^k is a bias node of layer k, w_{ji}^k is the connection weight from the i^{th} node at layer k-1 to the j^{th} node at layer k, and n_k is the number of nodes at k^{th} layer.

The output of node j at layer k=1, 2, 3 is given by

$$o_j^k = f(I_j^k) \tag{4.9}$$

($j = 1, ...n_k$) and f is an activation function, which is the sigmoid function as

$$f(I_j^k) = \frac{1}{1 + e^{-(I_j^k + n_j)/n_0}} \tag{4.10}$$

where n_j serves as the threshold or bias, and n_0 is used to modify the shape of the sigmoid.

The Neural Network (NN) is first trained and the net connection weights w_{ij}^k Eqn.4.8 are determined by using the back-propagation learning algorithm [24] and then stored in order to be used in the testing part of the recognition. The input to the neural network is a vector of combined features (13 *MFCCs* + *D*) extracted from each phoneme chosen from the test set of the TIMIT database.

4.9 Computer Simulation Set up

Experiments for both techniques described in the previous sections were performed on continuous speech data from the Texas Instrument Massachusetts Institute of Technology (TIMIT) database, which were directly digitised at a sample rate of 20 kHz using Digital Sound Corporation DSC 200 with the anti-aliasing filter at 10 kHz. The speech was then filtered, debiased and down sampled at 16 kHz. The TIMIT database has been subdivided into suggested training and test sets following the criteria that roughly 20 to 30% of the corpus should be used for testing purposes, leaving the remaining 70 to 80% for training.

In this research, all sentences spoken by male and female speakers from dialect regions one and two of the TIMIT database of the United States were used. These dialect regions are geographically close, DR1 corresponds to New England and DR2 to the Northern US. The total number of speakers in DR1 was 49 of which 27% were female. In DR2 the total number of speakers was 102 of which 30% were female giving an overall number of speakers equal to 151 including 50 female speakers. The total number of speech utterances of varying length in DR1 and DR2 respectively is equal to 472 and 1040, which numbered a total of 1512 utterances. These latter were extracted from the database and recursively searched to find all instances of each phone used in all its possible contexts. Since the speech was originally 16 kHZ bandlimited it was segmented into 32ms sections corresponding to 512 samples per signal.

Two classes of phonemes have been used and are referred to as 'vowel' and 'fricatives'. For each phoneme in these two classes , a set of features, which consist of the combination of one fractal dimension D together with 13 *MFCCs*, are computed. The training set of the TIMIT is used for template creation and the test set is used as unknown input in the matching process. The simulations were based on MATLAB version 5.3.

4.10 Results and Discussion

In this section, recognition results and analysis of the Least Square and the Neural network methods are clearly presented.

The Least Square method based on the new set of features 'DMFCC' is applied to two classes of speech data 'Vowels' and Fricatives'.

For comparative purposes, an identical operation was carried out on both 'MFCC' and fractal dimension feature 'D' and classification performance similarly obtained.

The first dataset contained the three vowels /iy/,/aa/,/ah/ corresponding (according to Figure 4.19) to front-,back-, and mid-voiced sounds.

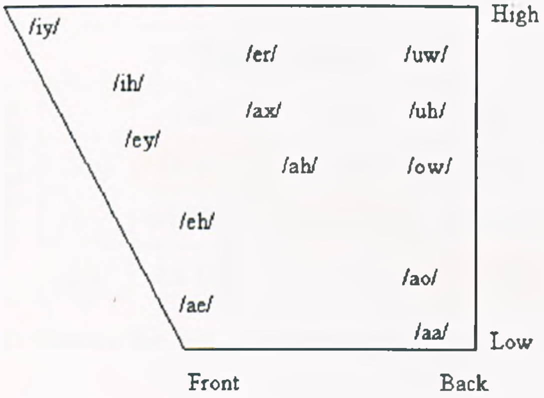

Figure 4.19: Diagram showing distances between several vowel sounds according to the position of the tongue bulk.

Figure 4.20 and Tables 4.2, 4.3 and 4.4 illustrate the recognition performances and the confusion matrix obtained for this case of vowel, respectively.

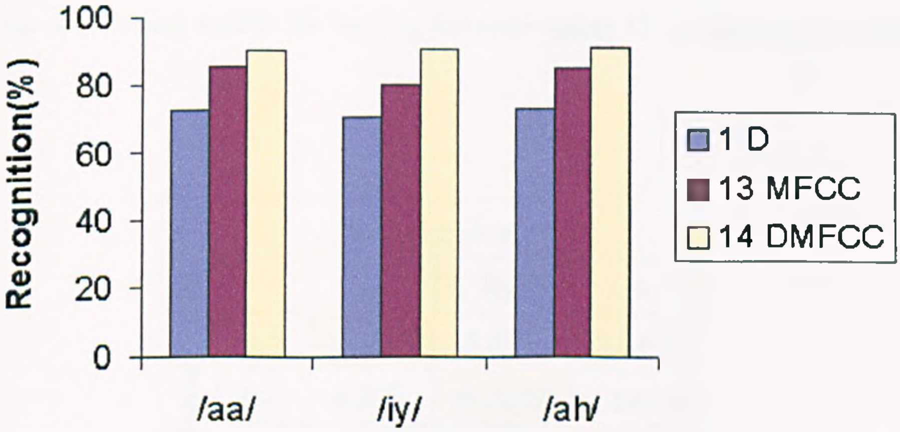

Figure 4.20: Vowels recognition performance for SI.

		Training class		
Testing class		/aa/	/iy/	/ah/
	/aa/	72.42%	17.82%	9.76%
	/iy/	18.875%	70.3125%	10.8125%
	/ah/	12.12%	15.15%	72.73%

Table 4.2: Confusion matrix for testing datasets using 'D' as a single parameter

		Training class		
Testing class		/aa/	/iy/	/ah/
	/aa/	85.29%	8.88%	5.83%
	/iy/	12.6875%	79.6875%	7.625%
	/ah/	6.07%	9.09%	84.84%

Table 4.3: Confusion matrix for testing datasets using 13 coefficients of 'MFCC'

		Training class		
Testing class		/aa/	/iy/	/ah/
	/aa/	90.18%	5.88%	2.94%
	/iy/	6.25%	90.625%	3.125%
	/ah/	3.04%	6.06%	90.90%

Table 4.4: Confusion matrix for testing datasets using a combination of 'D' and 13 coefficients of 'MFCC'

The confusion matrix for the three classes of vowels cited above, shows that $/iy/$ and $/aa/$ have higher confusion with each other than with $/ah/$ while of the two, $/ah/$ has higher confusion with $/iy/$. This is probable that overlap has occurred due to the range of pitch intonation likely to exist between speakers, particularly those of different gender. However, improvement is marked in the recognition performance when the fractal dimension 'D' and the Mel Frequency Cepstral Coefficients 'MFCC' are combined together as seen in Figure 4.20 and Table 4.4.

The second dataset contained the three classes of fricatives /f/,/T/, and /s/. This set of fricatives are generated by creationg a turbulent airflow at some point of constriction in the vocal tract. *Labiodental* as in /f/ causes the sound by creating friction between the top teeth and the lower lip. Forcing airflow between the top teeth and the tip of the tongue as in the 'th' sound of thing (/T/) is known as interdental and where articulation takes place between the tip of the tongue and the gum is called alveolar, an example of which is /s/ as in sing.

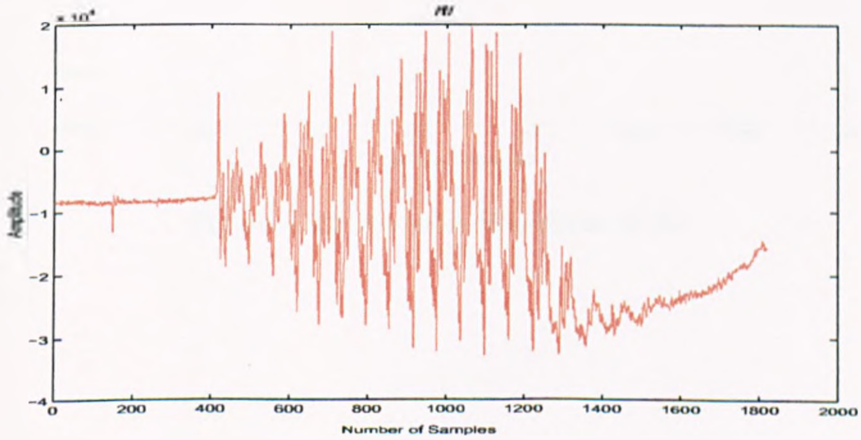

Figure 4.21: Acoustic Waveform of /f/

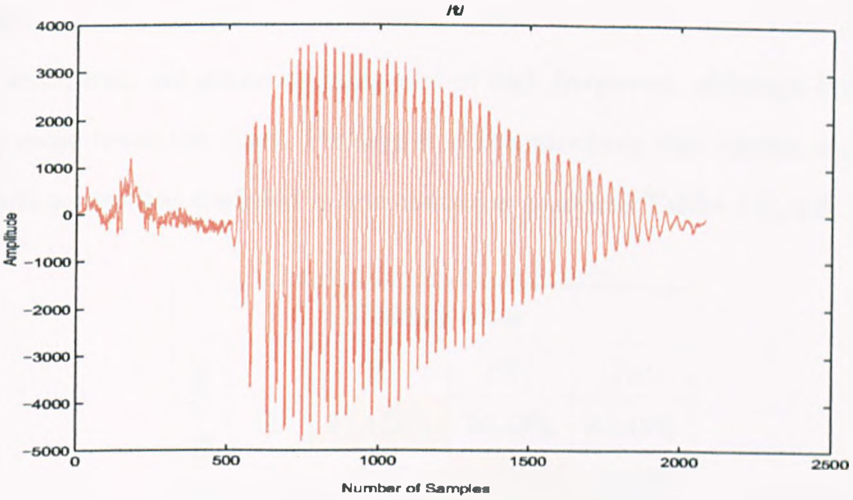

Figure 4.22: Acoustic Waveform of /T/

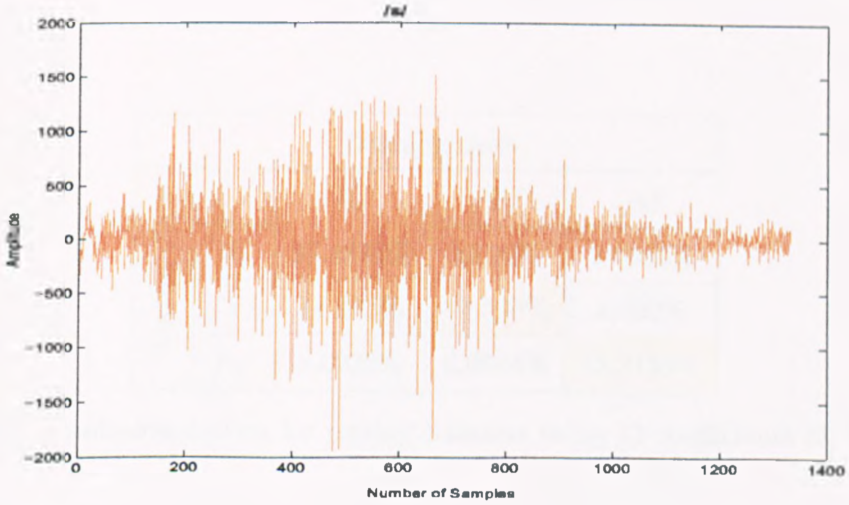

Figure 4.23: Acoustic Waveform of /s/

The main characteristic observable from Figures 4.21, 4.22, and 4.23, is that the acoustic waveforms are generally noisy and of high frequency, although both /f/ and /T/ have more burst-like start. Perhaps it is this attribute that causes more /f/ and /T/ sounds to become confused in the confusion matrices Tables 4.5, 4.6, and 4.7.

		Training class		
Testing class		/f/	/T/	/s/
	/f/	75.125%	16.43%	8.445%
	/T/	19.4%	74.91%	5.69%
	/s/	10.7%	12.89%	76.41%

Table 4.5: Confusion matrix for testing datasets using 'D' as a single parameter

		Training class		
Testing class		/f/	/T/	/s/
	/f/	87.1835%	10.12%	2.6965%
	/T/	11.143%	84.125%	4.732%
	/s/	5.6223%	6.0644%	88.3133%

Table 4.6: Confusion matrix for testing datasets using 13 coefficients of 'MFCC'

		Training class		
Testing class		/f/	/T/	/s/
	/f/	95.11%	3.15%	1.74%
	/T/	4.26%	94.23%	1.51%
	/s/	1.03%	3.88%	95.09%

Table 4.7: Confusion matrix for testing datasets using a combination of 'D' and 13 coefficients of 'MFCC'

From Figure 4.24, it can be notice that the recognition performance is greatly improved when D and 13 $MFCCs$ are combined together and used as a new single set of features.

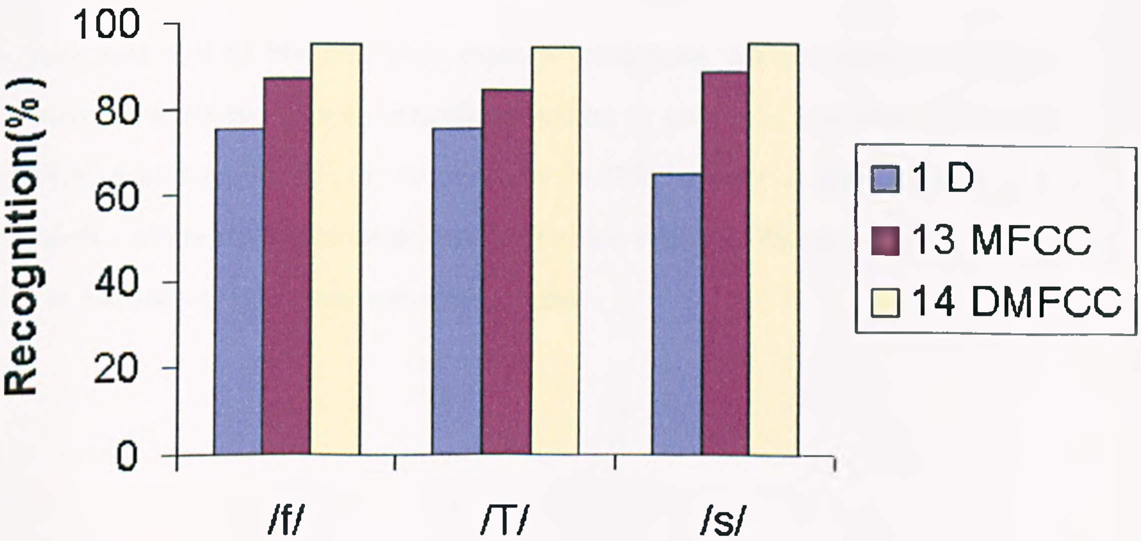

Figure 4.24: Fricatives recognition performance for SI.

From the previous confusion matrix it can be seen that the recognition performance achieved by using fractal dimension is very good for the three phoneme classification problem. This is due to the fact that the phonemes chosen in each class have very different articulator positions giving rise to different degree of randomness in the signal. This is very successfully captured by the fractal dimension 'D', hence it shows high recognition performance. Since this feature is uncorrelated to the 'MFCC' features therefore use of one fractal dimension along with 13 'MFCCs' features shows substantial improvement. However, 'D' cannot be used all alone for the phoneme recognition in a practical scenario as the number of phonemes is large this will give very poor recognition. But it can be used along with 13 'MFCCs' features and will improve the recognition performance as compared to the system using 13 'MFCCs' features alone.

The combination of 13 Mel frequency cepstral coefficients 'MFCCs' with the fractal dimension 'D' leads the speech recognition system to give even better results when using NN, achieving 98.21% (fricatives) and 96.75% (vowels) accuracy, which is a good result considering the amount speech used for training. These results are elaborated in the tables and histograms drawn below.

	/s/	/f/	/T/
14 DMFCC	98.21%	96.3%	95.25%
13 MFCC	90.80%	88.6%	86.21%
1 D	77.30%	76.94%	75.22%

Table 4.8: Fricatives recognition rates

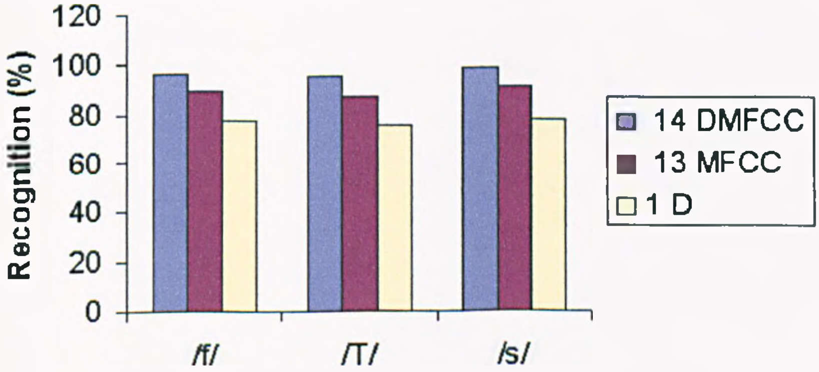

Figure 4.25: Neural network recognition rates for fricatives.

	/aa/	/iy/	/ah/
14 DMFCC	96.75%	94.77%	95.22%
13 MFCC	86.33%	83.30%	85.43%
1 D	74.11%	71.25%	73.01%

Table 4.9: Vowels recognition rates

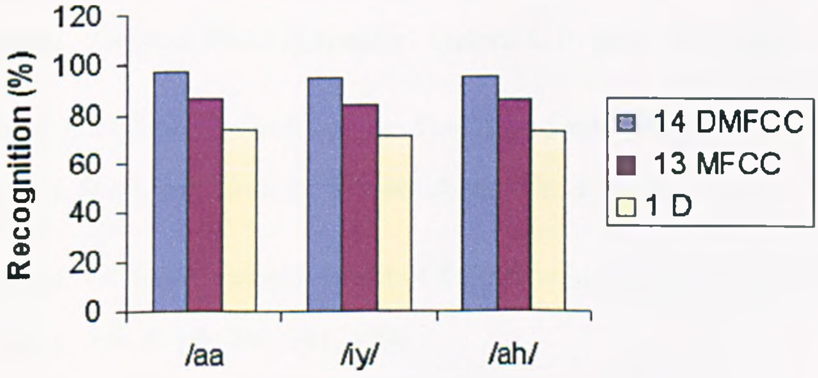

Figure 4.26: Neural network recognition rates for Vowels.

4.11 References

[1] R.J.Baken, "Irregularity of vocal period and amplitude: A first approach to fractal analysis of voice," *Journal of Voice*, vol. 4, no. 3, pp. 185–197, 1990.

[2] G.Fant, *Acoustic theory of speech production (Mouton, Hague)*, Springer-Varlag, 1970.

[3] J.Flanagan, *Speech Analysis, Synthesis, and Perception*, pringer-Verlag, Berlin-Heidelberg-New York, 1972.

[4] K.N.Stevens, "Airflow and turbulence noise for fricative and stop consonants: Static considerations," *J.Acoust. Soc. Am. 50, no.4*, vol. 2, 1971.

[5] P. Maragos and A.Potamianos, "Fractal dimension of speech sounds: Computation and application to automatic speech recognition," *J.Acoust. Soc. Am. 150*, vol. 3, 1999.

[6] B. B. Mandelbrot, *Fractal Geometry of Nature*, Freema, San Francisco, 1982.

[7] D.J. Tritton, *Physical Fluid Dynamics*, Oxford U.P, New York, 1988.

[8] S.M Teager H.M Teager, *Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract*, vol. 55 of *D*, Kluwer Academic, Boston, 1990.

[9] T.J Thomas, "A finite element model of fluid flow in the vocal tract," *Comput. Speech Lang.*, vol. 1, pp. 131–151, 1986.

[10] J.F Kaiser, "Some observations on vocal tract operation from a fluid flow point of view," *Biomechanics Acoustics and Phonatory Control*, pp. 358–386, 1983.

[11] R.S McGowan, "An aeroacoustic approach to phonation," *J.Acoust.Soc.Am.83*, pp. 696–704, 1988.

[12] J.M.Blackledge, *On the Synthesis and Processing of Fractal Signals and Images*, Springer-Verlag, Berlin, 1993.

[13] B.B.Mandelbrot, *Fractals: Form, Chance and Dimension*, W.H. Freeman, san fransisco,ca edition, 1977.

[14] J.Srinivas, "Fractals classification, generation and application," *IEEE*, vol. 2, pp. 1024–1027, 1992.

[15] M.Al-Zabibi, *An Acoustic-Phonetic Approach in Automatic Arabic Speech Recognition*, Phd thesis, Loughborough University, 1990.

[16] P. Mermelstein, "Automaic segmentation of speech into syllabic units," *Acoustic Soc Amer. 58*, pp. 880–883, 1975.

[17] E.L.J. Bohez and T.R. Senevirathne, "Speech recognition using fractals," *Pattern Recognition Society*, pp. 2227–2243, 2001.

[18] J.Tebelskis, *Speech recognition using neural network*, Phd thesis, University of Pennsylvania, 1995.

[19] G.Fant, "Speech sounds and features," Tech. Rep., The MIT Press, 1973.

[20] M.Shelberg, "The development of a curve and surface algorithm to measure fractal dimensions," Msc thesis, Ohio State University, 1982.

[21] R.P.Lippman, "Pattern classification using neural networks," *IEEE Communications Magazine*, pp. 54–66, 1989.

[22] M.Robinson, H.yoneda, and E.Sanchez-Sinencio, "A modular cmos design of hamming networks," *IEEE transaction on neural network*, vol. 3, no. 3, May 1992.

[23] S.Fekkai, M. Al-Akaidi, and J.Blackledge, "Novel technique using fractal dimension for speech recognition," *the proceeding SCSC'98,Arlington*, pp. 679–683, 1998.

[24] C.Becchetti and L. P. Ricotti, *Speech Recognition*, John Willey & Sons, UK, 1999.

Chapter 5

Fractal Speech Synthesis

5.1 Introduction

In an ideal world, speech synthesiser should be able to synthesise any arbitrary word sequence with complete intelligibility and naturalness. The trade-off schematic in Figure 5.1 illustrates how current synthesisers have tended to strive for flexibility of vocabulary and sentences at the expense of naturalness (i.e. arbitrary words can be synthesised, but do not sound very natural). This applies to articulatory, rule-based and concatenative methods of speech synthesis [1, 2, 3, 4].

An alternative strategy is one which seeks to maintain naturalness by operating in a constrained domain. There are potentially many applications where this mode of operation is perfectly suitable. In conversational systems for example, the domain of operation is often quite limited, and is known ahead of time [5].

Past work by others have examined how unit selection algorithms can be formulated, and what constraints must be maintained [1, 3, 4].

In this work, we develop a framework for natural-sounding speech synthesis using fractal. Our objective was to place naturalness as a paramount goal. Our research follows the bottom curve of Figure 5.1 where we view naturalness as the highest priority.

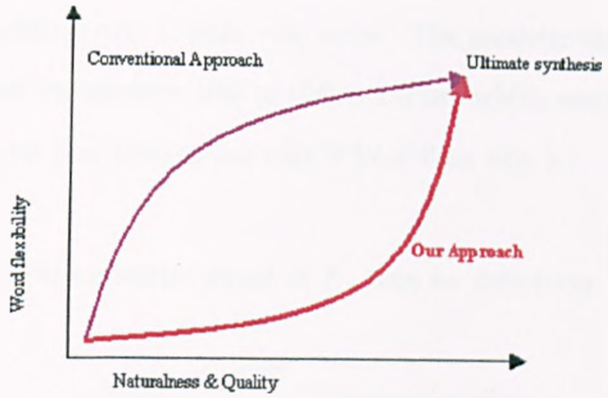

Figure 5.1: Schematic trade-off synthesis development.

In signal analysis (where the independent variable is usually time), a real valued signal can be represented in terms of the so-called analytic signal, which will be explained in the next section. The analytic signal is important because it is from this signal that the amplitude, phase and frequency modulations of the original real valued signal can be determined.

5.2 The Analytic Signal and Phase Unwrapping

If $f(x)$ is a real valued signal with spectrum $F(k)$, then $f(x)$ can be computed from $F(k)$ via the inverse Fourier transform

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(k) \exp(ikx) dk \tag{5.1}$$

This involves integrating over k from $-\infty$ to ∞. The analytic signal is obtained by integrating only over the positive half of the spectrum, which contains the physically significant frequencies (i.e. integrating over k from 0 to $+\infty$).

If s is used to denote the analytic signal of f , then by definition

$$s(x) = \frac{1}{\pi} \int_{0}^{+\infty} F(k) \exp(ikx) dk \tag{5.2}$$

From Eqn.5.2 it is possible to obtain an expression for s in terms of f which is done by transforming s into Fourier space and analysing the spectral properties of the analytic signal.

a) Important Result

Theorem: The analytic signal is given by

$$s(x) = f(x) + iq(x) \tag{5.3}$$

where $q(x)$ is the Hilbert transform of $f(x)$.

Proof: In Fourier space, the analytic signal can be written as

$$S(K) = 2U(k)F(k) \tag{5.4}$$

where S and F are the Fourier transforms of s and f respectively and $U(k)$ is the unit step function given by

$$U(k) = \begin{cases} 1, & k \geq 0 \\ 0, & k < 0 \end{cases}$$

We now employ a simple but useful analytical trick by writing the step function in the form

$$U(k) = \frac{1}{2} + \frac{1}{2}\mathrm{sgn}(k) \tag{5.5}$$

where

$$\mathrm{sgn}(k) = \begin{cases} 1, & k > 0 \\ -1, & k < 0 \end{cases}$$

The inverse Fourier transform of this function can then be written as

$$\begin{aligned} u(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{2}\exp(ikx)dk + \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{2}\mathrm{sgn}(k)\exp(ikx)dk & (5.6) \\ &= \frac{1}{2}\delta(x) + \frac{i}{2\pi x} & (5.7) \end{aligned}$$

where,

$$\delta(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(ikx)dk$$

and

$$\int_{-\infty}^{\infty} \frac{1}{2}\text{sgn}(k) \exp(ikx)dk = \frac{i}{\pi x}$$

since

$$2U(k)F(k) \Longleftrightarrow 2u(x) \otimes f(x)$$

Therefore we have

$$s(x) = 2u(x) \otimes f(x) \tag{5.8}$$

substituting Eqn.5.7 into Eqn.5.8 gives:

$$s(x) = f(x) \otimes (\delta(x) + \frac{i}{\pi x}) = f(x) + \frac{i}{\pi x} \otimes f(x) \tag{5.9}$$

or

$$s(x) = f(x) + iq(x) \tag{5.10}$$

where $q(x)$ is the Hilbert transform of $f(x)$, i.e.

$$q(x) = \frac{1}{\pi x} \otimes f(x) \tag{5.11}$$

From the last result it is clear that the analytic signal associated with a real valued function f can be obtained by computing its Hilbert transform to provide the quadrature component. This process is called quadrature detection.

The analytic signal is a complex function and therefore contains both amplitude and phase information. The important feature of the analytic signal is that its spectrum (by definition) is zero for all values of k less than zero. This type of spectrum is known as a single sideband spectrum because the negative half of the spectrum is zero. An analytic signal is therefore a single sideband signal. This provides another way of computing the Hilbert transform of a function $f(x)$:

- Compute the Fourier transform of $f(x)$
- Set the component of the complex spectrum $F(k)$ in the negative half space to zero.
- Compute the inverse Fourier transform, which will have real and imaginary parts $f(x)$ and $q(x)$ respectively.

5.2.1 Attributes of the analytic signal

From the original speech signal we compute the analytical signal (by Hilbert Transform) from which we can derive the amplitude envelop and the phase. The amplitude and phase can then be processed as required for the purpose of fractal speech synthesis as shown in this section.

As with any other complex function, the behavior of the analytic signal can be analysed using an argand diagram and may be written in the form

$$s(x) = A(x)\exp[i\theta(x)] \tag{5.12}$$

where

$$A = \sqrt{f^2 + q^2} \tag{5.13}$$

$$\theta = \tan^{-1}\left(\frac{q}{f}\right) \tag{5.14}$$

The parameter A describes the average dynamical behavior of the amplitude modulations of f. For this reason, it is sometimes referred to as the amplitude envelope. The parameter θ measures the phase of the signal at an instant in time and is therefore known as the instantaneous phase.

Note: Because the arctangent function is periodic, this parameter is multivalued. Hence, strictly speaking, the analytic function should be written as

$$s = A\exp[i(\theta + 2\pi n)]; \quad n = 0, \pm1, \pm2, ... \tag{5.15}$$

If we confine the value of the phase to a fixed period (i.e. we compute the phase using only one particular value of n), then it is referred to as the wrapped phase. In this case, there is only one unique value of the phase within a fixed period. However, any other interval of length 2π can be chosen. Any particular choice, decided upon in advance, is called the principal range. The value of the phase within this range, is called the principal value.

5.3 Phase Unwrapping

An alternative way of defining the phase can be obtained by taking the natural logarithm of the analytic signal, and this yields the equation

$$\ln s = \ln A + i(\theta + 2\pi n) \tag{5.16}$$

where,

$$\text{Real}[\ln s] = \ln A$$

$$\text{Im}[\ln s] = (\theta + 2\pi n)$$

Another important property of the analytic signal is its instantaneous frequency. This parameter (denoted by ψ) measures the rate of change of phase $d\theta/dx$ and from the previous expression can be written as

$$\psi = \frac{d\theta}{dx} = \text{Im}\left(\frac{d}{dx}\ln s\right) = \text{Im}\left(\frac{1}{s}\frac{ds}{dx}\right) \tag{5.17}$$

The instantaneous frequency provides a quantitative estimate of the frequency of the real valued signal f at any instant in time.

The phase θ can be obtained as follow

$$\int_0^x \psi(x)dx = \int_0^x \frac{d\theta}{dx}dx$$

Using the initial conditions

$$\theta(x=0) = \theta_0$$

we get

$$\theta(x) = \theta_0 + \int_0^x \psi(x)dx$$

This phase function is called the unwrapped phase. It is not multi-valued and therefore the problem of choosing a principal range to compute the phase does not occur.

5.4 System Description

5.4.1 Phase Compression Method

As explained in section 5.2, a real valued signal can be represented in terms of the so-called analytic signal, which is important because it is from this signal that the amplitude, phase and frequency modulations of the original real valued signal can be determined.

The phase compression method makes use of two parameters related to the speech signal namely the fractal dimension D and the phase ϕ. The fractal dimension of the speech word is computed using the power spectrum method (Chapter 2), then it is used to generate the fractal signal F, the idea is to use the fractal properties of the word in the calculation of the amplitude envelope in order to obtain more natural speech synthesis.

To avoid the problem of choosing a principal range to compute the phase ϕ, we use the function unwrap phase, which we note by ϕ'. This phase is then compressed and used with the amplitude envelope within a specific algorithm to produce the synthetic speech. The method is illustrated in Figure 5.2.

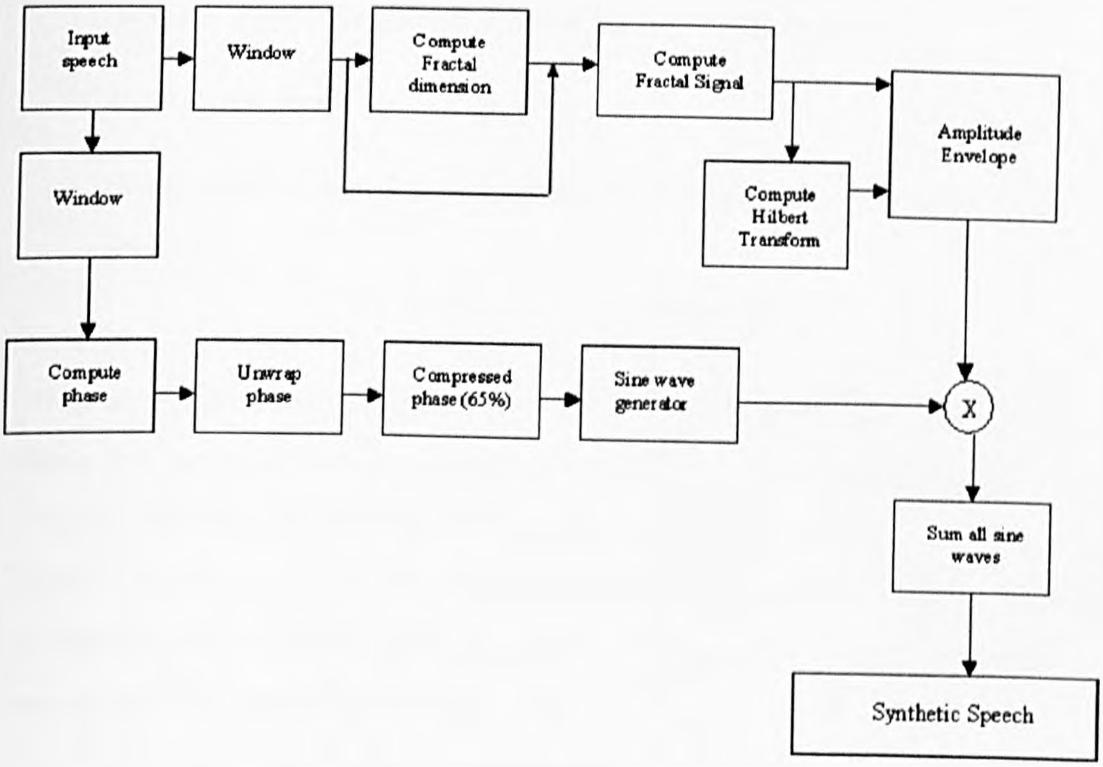

Figure 5.2: Phase compression algorithm.

The simulation process consists of the following steps:

The speech signal is divided into frames using a window size of 512 samples. For each j^{th} window of the speech word the following steps are undertaken:

Step 1: Compute the fractal dimension D of the speech word $S_j(t)$.

Step 2: Compute the fractal signal $F_j(t) = Re\{IFFT(FFT(S_j(t)) \times \frac{1}{k_i^{\beta}})\}$ of the speech signal $S_j(t)$, using the power spectrum method discussed in Chapter 2.

Step 3: Compute the Hilbert transform $H_j(t)$ of the fractal signal F_j , ($j = 1, 2, ..., N$, where N is the number of samples of S_j) as follows:

1. Take the Fourier Transform $F_j(k)$ of the fractal signal $F_j(t)$
2. Multiply the result by $[-i \text{ sgn } (k)]$
3. Compute the inverse Fourier Transform to obtain the Hilbert transform as follows:

$$H_j(t) = F^{-1}\{(-i \text{ sgn } (k)F_j(k))\}$$

Step 4: Compute the amplitude envelop $A'_j(t) = \sqrt{F_j(t)^2 + H_j(t)^2}$
Step 5: Compute the phase $\phi_j(t) = tan^{-1}\left(\frac{Im(S_j(t))}{Real(S_j(t))}\right)$ of the speech signal
Step 6: Compute the unwrapped phase $\phi'_j(t) = unwrap(\phi_j(t))$
Step 7:Compress $\phi'_j(t)$ by 65% from it's original size with the use of the Discrete Cosine Transform (DCT), which has the role of retaining the low frequencies and section the high frequencies to zero.

Step 8: From the standard complex representation $A(t)e^{i\phi(t)}$ of which the real part $S_o(t) = Re[A(t)e^{i\phi(t)}] = A(t)cos\phi(t)$ is taken to be the synthesised word, the synthetic speech signal is then reproduced as follows:
$S_o(t) = \sum\limits_{j=1}^{N} A'_j(t)\cos[\phi'_j(t)]$, where $A'_j(t)$ is the amplitude envelop cited previously and N the number of samples in $S_j(t)$.

The texture of the fractal signal $F_j(t)$ as shown in Figures 5.3 and 5.4, for example, looks smoother than the input speech signal $S_j(t)$, which means that the fractal signal requires less samples to reconstruct the synthetic speech word, therefore together with the compression of the phase $\phi_j(t)$ it provides a method of synthesising a speech signal from very limited data. The principal information related to the speech to be synthesised comes from the phase of the original signal. However, the realistic texture or naturalness of the synthetic speech waveform is related directly to the fractal field $F_j(t)$. This is because the fractal field has the same textural properties as the original signal as compounded in the computation of the fractal dimension D. As with the approach of fractal geometry to compute graphics in which the fractal dimension defines random fractal constricted with natural object (e.g. cloud, mountains and other natural surfaces) so, this approach to speech synthesis provides an equivalent acoustic effect.

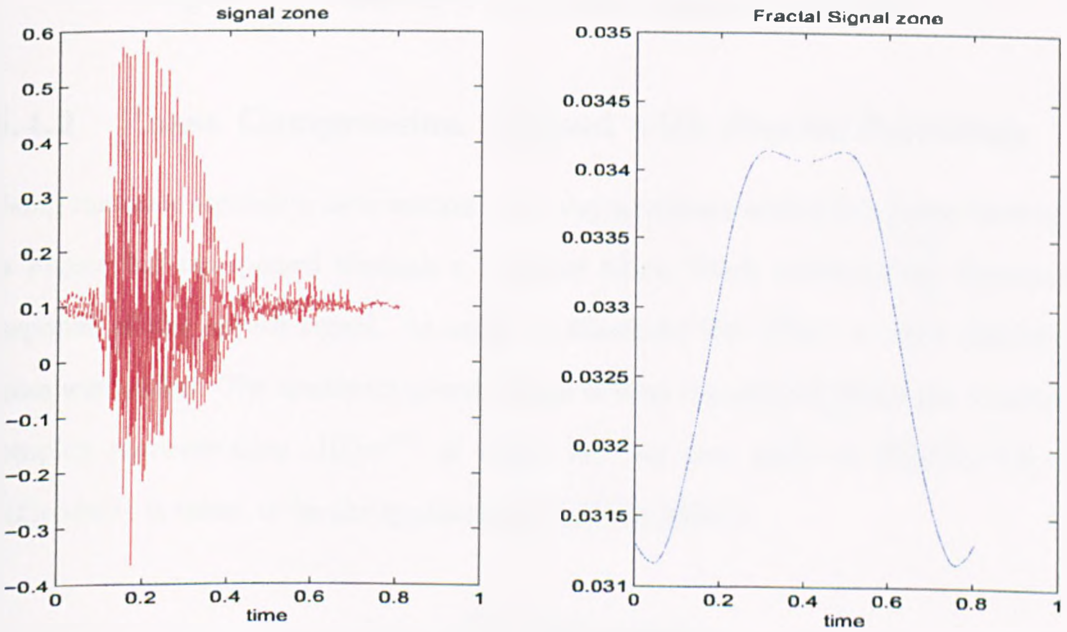

Figure 5.3: Waveform of word /zone/ and its fractal signal.

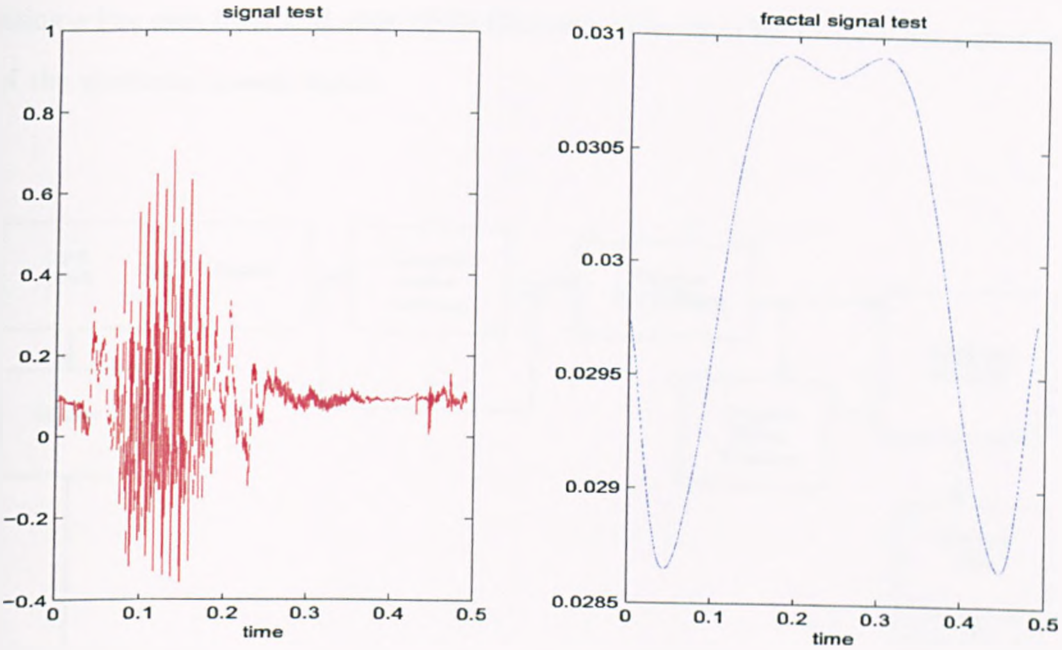

Figure 5.4: Waveform of word /test/ and its fractal signal.

5.4.2 Phase Compression Method with Fractal Synthesis

Using the same algorithm as in section 5.4.1, the amplitude of this simulation as shown in Figure 5.5 was passed through a low pass filter, which distorted the frequency response of the speech signal. In order to eliminate this effect, a white Gaussian noise was added. The synthetic speech signal is then reproduced (from the standard complex representation $A(t)e^{i\phi(t)}$ of which the real part $S_o(t) = Re[A(t)e^{i\phi(t)}] = A(t)cos\phi(t)$ is taken to be the synthesised word) as follows:

$$S_o(t) = \sum_{j=1}^{N} A'_j(t) \cos[\phi'_j(t)]$$

This method adds novelty to the synthesis speech as it doesn't not only uses the fractal dimension D to compute amplitude envelop but also compresses the amplitude by using a low pass filter and adds white Guassian noise in order to show the robustness of the synthetic speech signal.

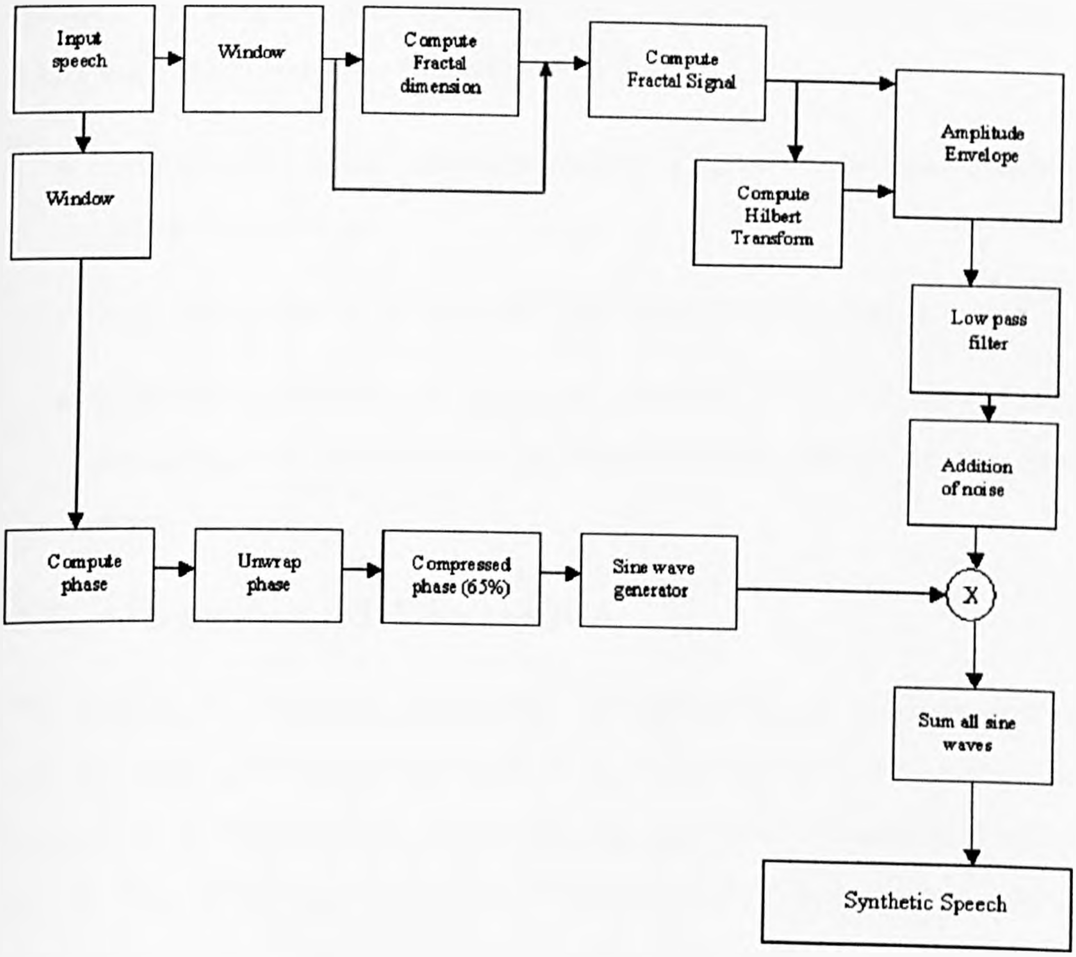

Figure 5.5: Fractal Synthesis algorithm.

5.5 Computer Simulation Set up

Three experiments were conducted in the simulation process involving four different words namely “test”, “best”, “open” and “zone” spoken by different male and female speakers. The speech words were recorded and directly digitized at a sample rate of 8 kHz using the digital audio editor Cool Edit 96.

- In the first experiment, only the unwrapped phase of the word was used along with the fractal signal.
- In the second one, the phase is 65% compressed from the original.
- In the third experiment, the amplitude envelope is low pass filtered and the remaining signal is replaced by a white Guassian noise as discussed in the previous section.

5.6 Results and Discussion

The three experiments gave good quality and intelligibility of the synthetic words and they all sounded very natural as shown in the subjective evaluation subsection; however, of the three, the best natural sounding speech was enhanced when the amplitude of the speech signal was low pass filtered and white Gaussian noise added.

It is worth mentioning here, that the use of the fractal properties in the energy of the reconstructed speech signal has the effect of controlling the naturalness of the synthetic speech as mentioned in subsection 5.4.1.

The results are illustrated in Figures 5.6, 5.7, 5.8 and 5.9 respectively where the unwrapped phase; the synthetic amplitude envelope, the original word and its reconstructed word are plotted. We clearly notice the similarities between the waveforms of the input speech word with the synthetic waveform, which confirm the accuracy and efficiency of the new algorithm in the synthesis of a speech word.

Essentially , the method involves retention of the phase (with or without compression) and detection of the signal amplitude envelope which is reconstructed from knowledge of the fractal dimension of the original signal.

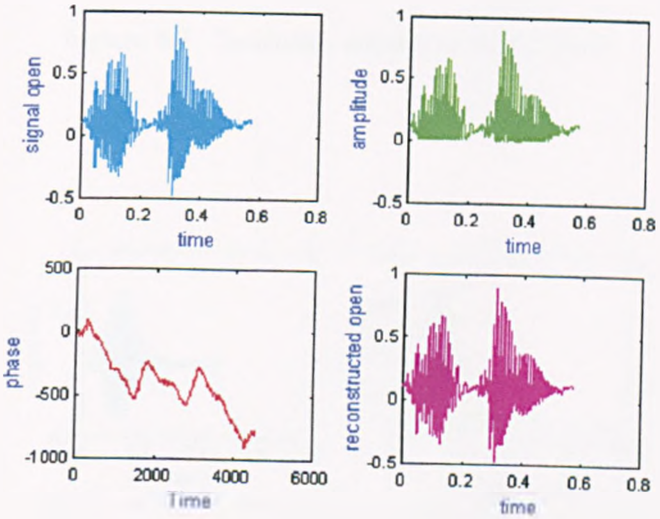

Figure 5.6: Synthesis results of word 'open'.

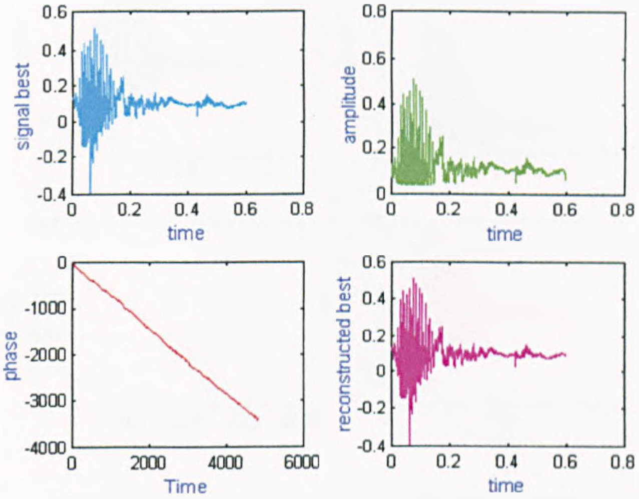

Figure 5.7: Synthesis results of word 'best'.

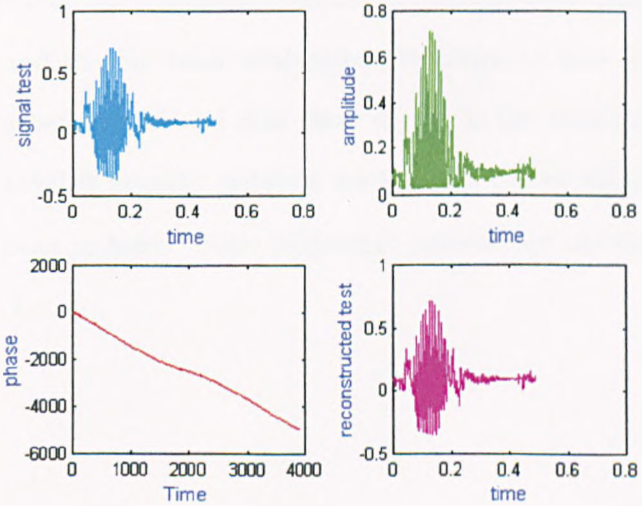

Figure 5.8: Synthesis results of word 'test'.

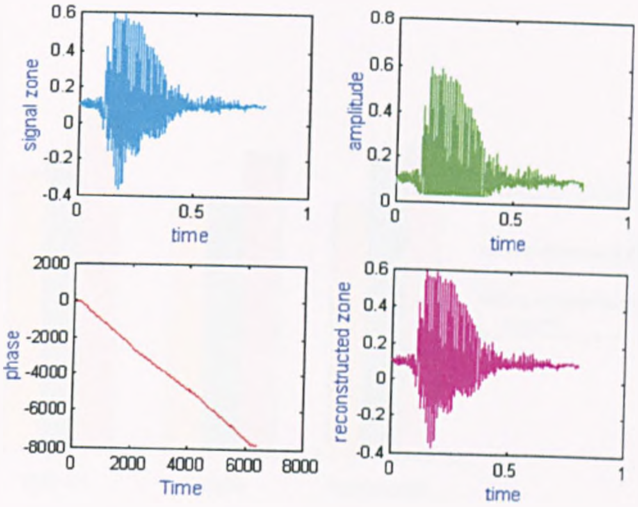

Figure 5.9: Synthesis results of word 'zone'.

5.6.1 Subjective Evaluation

To test the validity of the simulation results, 30 people were asked to listen to the synthetic words and to give their evaluations in terms of how good or bad is the quality of the synthetic word and how clear or not is the intelligibility of the word and also how natural it sounds, without having any earlier information about the word they were going to hear. These subjective evaluations are elaborated in Figure 5.10 and Table 5.1.

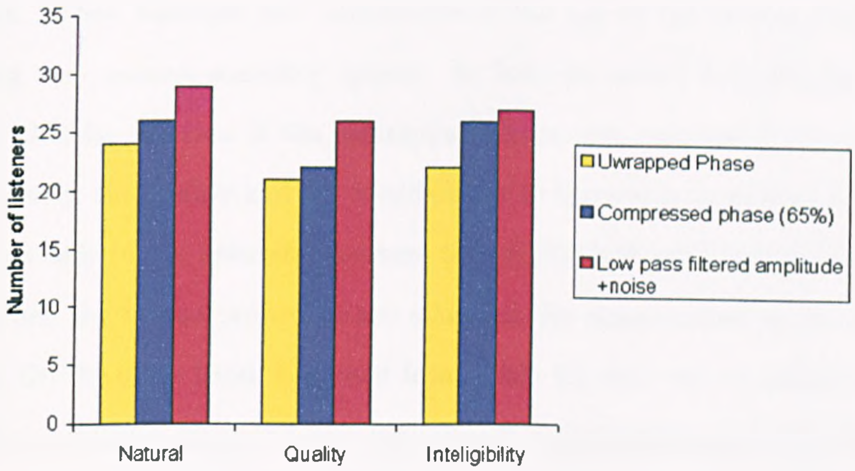

Figure 5.10: Subjective of the Synthesis Evaluation.

	Quality	Naturalness	Intelligibility
Unwrapped phase	Good	natural	clear
Unwrapped compressed phase (65%)	Good	natural	clear
Unwrapped compressed phase with fractal synthesis	very good	very natural	Very clear

Table 5.1: Subjective evaluations

The audible synthesised words were considered to be satisfactory in the subjective evaluation. They highlight the importance of the use of the fractal dimension in generating very natural sounding speech. In fact, we notice from Figure 5.10 and Table 5.1 that for the case of the unwrapped phase and compressed phase methods simultaneously, the quality and the intelligibility of the words is good and the difference occurs only in the naturalness where the synthetised word is found to be more natural when the fractal properties are added in the computation of the amplitude envelope. On the other hand, it is clear from Table 5.1 that the compression method with fractal synthesis, gives the best results with a high quality and intelligibility of the word and very natural sounding words synthesis when compared to the synthesis that uses only the unwrapped phase or the unwrapped compressed phase.

5.7 References

[1] N.Campbell, "Chatr: A high-definition speech re-sequencing system," *Acoustical Society of America and Acoustical Society of Japan, Third Joint Meeting*, Dec 1996.

[2] X.Huang, A.Acero, J.Adcock, H.Hon, J.Goldsmith, J.Liu, and M.Plumpe, "Whistler: A trainable text-to-speech system," *in Proc. ICSLP, Philadelphia, PA*, pp. 2387–2390, Oct 1996.

[3] A.J.Hunt and A.W.Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *in Proc. ICASSP, Atlanta, GA*, pp. 373–376, May 1996.

[4] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of nonuniform synthesis units," *in Proc. ICASSP, New York, NY*, pp. 679–682, April 1988.

[5] R.W.Y.Jon and R.J.Glass, "Natural-sounding speech synthesis using variable-length units," *ICLSP98*, 1998.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

A new spectral representation of speech phonemes based on fractals has been presented and the use of pre-filtering in the Power Spectrum Method described. This new spectral technique for the determination of the fractal dimension of speech phonemes appears to be a practical procedure of acceptable validity. In fact, one of its main advantage is that the computation of the fractal dimension D is based on an explicit formulae $D = (5 - \beta)/2$. This gives a good computation of the fractal dimension of various speech phonemes as they all obey the law of fractal dimension when applied to fractal speech and curves.

New ideas to construct a more robust and more reliable speaker recognition system has been highlighted. A new set of features for phoneme speech recognition under the template matching technique was described in detail. Experimental results from using this technique on TIMIT continuous speech showed significantly improved recognition accuracy. Chapter 4 explained and discussed the simulation results for the recognition research of this thesis. It is clearly noticeable from Figures 4.20 and

4.24 respectively that the recognition performance has been greatly improved when D and the $MFCCs$ were combined together and used as a single set of features characterizing the specific use of phoneme. In particular, the recognition rates increased significantly for vowels and fricatives respectively from 71% to 90% and from 75% to 95% when $MFCCs$ were added to the fractal dimension D. This shows that when the new hybrid features (D and $MFCCs$) are combined together, a better performance of the speech phoneme recognition can be achieved.

It has also been shown that the combination of the Mel-frequency cepstral analysis with non-linear features lead to a speaker recognition system with better results when a Neural Network algorithm is used as shown in Figures 4.25 and 4.26 respectively achieving 98.21% (fricatives) and 96.75% (vowels) of accuracy, which is considered to be a good and promising result.

A new algorithm based on fractals has been used for the synthesis of speech words. The synthesis process involved three cases of experiments, which increased the quality and the intelligibility of the synthesised speech. It is worth mentioning here, that the texture of the fractal signal $F_j(t)$ as shown in Figures 5.3 and 5.4, for example, looks smoother than the input speech signal $S_j(t)$, which means that the fractal signal requires less samples to reconstruct the synthetic speech word, therefore together with the compression of the phase $\phi_j(t)$ it provides a method of synthesising a speech signal from very limited data. The principal information related to the speech to be synthesised comes from the phase of the original signal. However, the realistic texture or naturalness of the synthetic speech waveform is related directly to the fractal field

$F_j(t)$. This is because the fractal field has the same textural properties as the original signal as compounded in the computation of the fractal dimension D. As with the approach of fractal geometry to compute graphics in which the fractal dimension defines random fractal constricted with natural object (e.g. cloud, mountains and other natural surfaces) therefore, this approach to speech synthesis provides an equivalent acoustic effect.

The naturalness level (paramount in our work) was achieved as a result of the fractal characteristic used in the synthesis process. Despite the small size of vocabulary used, the naturalness is very high and, as the pursuit of naturalness dominates, human listening provided the best feedback.

6.2 Future Work

There are many difficulties in communications and signal processing algorithms, which linear techniques have failed to address satisfactorily. It is generally held belief that these problems may however have solutions in the growing field of non-linear signal processing. The recent rise in neural network concepts and fractals are, for example, largely fuelled by this promise. In addition, the past decade has seen a remarkable growth in the theory of the dynamics of non-linear systems such as fractals. One cause of this interest has been the realisation that deterministic mathematical models with few degrees of freedom can generate extremely complex behavior. Thus complicated physical systems may be well modelled by relatively simple non-linear models.

The research work presented in this thesis can be enhanced through further investigations and practical implementation, which may be achieved by taking into consideration that:

- **Speech coefficients**: Speech recognition based only on fractals analysis will not be possible. Fractals will however be good additional features to be combined with other approaches like, for example, LP-derived cepstral coefficients which are obtained from the predictor coefficients, or wavelets. This could extract new informations that specifically distinguish different speakers and improve recognition. Non-linear techniques will allow us to merge feature extraction and classification problems and to include the dynamics of the speech signal in the model. This is likely to lead to significant improvements over current methods, which are inherently static.

- **Multifractal Analysis**: Speech processing (e.g., characterisation, compression, recognition, and analysis) depends heavily on processing nonstationary parts of the signals considered (e.g., *consonants, vowels transitions, consonant-vowel transitions)* because the transient parts of speech often carry most of speech information. Multifractal (or singularity) processing of speech is capable of providing important processing aspect (decomposition, representation and spectrum characterisation). The multifractal approach can decompose speech into various segments and is also used in characterising speech through singularity spectrum, which can be used to develop a better accuracy speech recognition schemes.

- **Speaker recognition application :** A simple recognition system can be envisaged where each class (which could be composed of phones, diphones, etc) can be characterised by a non-linear model. Then, given an input frame of speech, it will be possible to use the sum of the error residual from the predictor over the frame to decide which class the input speech belongs to. Thus, the feature extraction and the classification problem are merged together and solved by one unit. Further, the dynamics of the speech signal may be included in the non-linear model. For speaker recognition applications it has been shown that the residual signal of a linear analysis contains enough information to enable humans to identify people. Thus, there is relevant information that is ignored with a linear analysis. Non-linear techniques allow us to merge feature extraction and classification problems and to include the dynamics of the speech signal in the model. It is likely to lead to significant improvements over current methods, which are inherently static. Work on detecting such features and using them in recognition systems is very promising.

- **Speech application systems:** Such as recognisers and synthesisers, require some parametric representation of the signal. These parameters should reflect our understanding of speech production and speech perception mechanisms. For example, actual recognition systems are not robust in terms of noise and variations of voice quality and speaker. Synthesisers suffer for being characterised by poor and unnatural quality of speech, and by a lack of flexibility in terms of changes in voice gender and reflection of emotional states.

The development of new applications using speech technology has sparked interest in determining when and how to incorporate speech for user input and is leading to the design of efficient dialogue generation. Along these lines, class grammars may provide better modelling of sentence, topic and discourse structure. In addition, an accurate prosodic model would be useful for choosing between parses in natural language as well as for producing natural speech for both language learning and instruction.

- **Analysis of the acoustic signal**: One way for gaining insight in how speech sounds are structured is to analyse the acoustic signal in a very detailed manner. This procedure allows the definition of a number of acoustic attributes, which characterise, in a significant way, the speech units of a language. However, as is well known, the acoustic attributes of a given sound vary as a function of many factors, among which the context in which the sound is embedded and the speaker have the strongest effects. Therefore, a careful analysis of the speech signal requires sophisticated experiments in which a large number of tokens of a given sound are recorded to form the experimental database. Accurate measurements of the acoustic parameters must be performed on all collected data, and their significance must be evaluated. However, even when doing so, a variety of factors such as the speaking rate and emotional state may be neglected. It should be noted, that the acoustic attributes also depend upon the language under examination and thus, the findings obtained on one language can hardly be extended to other languages. All these aspects make the analysis complex and time-consuming, a feature that is often not compatible with the timing of speech application systems development.

- **Speech synthesis technology**: It plays an important role in many aspects of man machine interaction, particularly in telephony applications. New telecommunication services include the capability of a machine to speak with a human in a 'natural way'; to this end, a lot of work must be done in order to improve the actual voice quality. This will involve constructing models, which operate in the state space domain, such as neural network architectures and fractal models. The speech synthesised by these methods will be more natural-sounding than linear concatenation techniques because the low dimensional dynamics of the original signal are learnt. In addition to generating high quality speech, other associated tasks will also be addressed. The most important of these is to examine techniques for natural parameterisation that can be linked into the non-linear model.

Speech synthesis has been developed steadily over the last decades and it has been incorporated into several new applications. For most applications, the intelligibility and comprehensibility of synthetic speech have reached an acceptable level. However, in prosodic, text preprocessing and pronunciation fields, there is still much work and improvements to be done to achieve more natural sounding speech.

Several normal speech processing techniques may be used also with synthesised speech. For example, by adding some reverberation it may be possible to increase the pleasantness of synthetic speech afterwards. Other effects, such as digital filtering, chorus, etc., can also be used to generate different voices. However, using these kind of methods may increase the computational load. Most information of the speech signal is focused at a frequency range less than 10 kHz. However, by using higher sampling rates than necessary, the speech may sound slightly more pleasant.

As long as speech synthesis needs to be developed, the evaluation and assessment play one of the most important roles. Before performing a listening test, the method used should be tested with smaller listener groups to find out possible problems and the subjects should be chosen carefully. The development of speech synthesis is going forward steadily and in the long run, the technology seems to make progress faster than we can imagine. Thus, when developing a speech synthesis system, we may use almost all resources available, because in a few years time today's high resources will be available in every personal computer. Regardless of how fast the development process will be, speech synthesis, whenever used in low-cost calculators or state-of-the-art multimedia solutions, has probably the most promising future. If speech recognition systems someday achieve a generally acceptable level, we may develop for example a communication system that will first analyse the speakers' voice and its characteristics, transmit only the character string with some control symbols, and finally synthesise the speech with individual sounding voice at the other end. Even interpretation from a language to another may became feasible.

Appendix A

Mel Scale and Critical Bands

A.1 Threshold of Hearing

The absolute sensitivity of the human ear is measured as the smallest Sound Pressure (SP), which leads to the sensation of hearing. The threshold depends on the frequency of the sound. Figure A.1 displays this threshold as a function of frequency for a typical young adult [1].

The human ear is most sensitive between 1000 and 3000 Hz with the threshold rising from lower to higher frequencies. If a threshold at 1000 Hz is taken as a reference (see Figure A.1), the signal is to be increased a hundred times to reach a threshold at 100 Hz and 15,000 Hz, and a thousand times to reach a threshold at 18,000Hz. The threshold of pain occurs more or less uniformly at sound intensities equal to 140dB.

The frequency limits of hearing are generally considered to lie between 20 and 20,000 Hz.

A.2 Pitch and Mel Scale

Pitch is the subjective attribute of a sound, which corresponds to the physical attribute of frequency. Although the pitch of a pure tone is monotically related to its frequency, a linear relationship does not hold. The unit of pitch is the 'mel'. The mel scale has been constructed on the basis of subjective pitch evaluations. This involves the determination of the frequency corresponding to halving and doubling of the pitch and equal increments of the pitch by nave listeners. A tone with a pitch of 500 mels sounds half as high as one with a pitch of 1000 mels. However, its frequency will be 400 Hz. Similarly, a tone with a pitch of 2000 mels will sound twice as high as one with a pitch of 1000 mels, yet its frequency will be 3000 Hz rather than 2000 Hz. Figure A.2 illustrates the relationship between the pitch scale and the frequency scale for pure tone of 40db intensity [1].

The mel scale is essentially linear at low frequencies and logarithmic at higher frequencies. A useful approximation to the mel scale is of the form [2]

$$y = k \log(1 + \frac{f}{1000}) \tag{A.1}$$

where f is the frequency in Hz and k is a constant. The constant is computed with the consideration that a tone with a frequency of 1000 Hz is defined as having a pitch of 1000 mels. Thus, k is equal to 3322. A conversion of a frequency to a mel scale is roughly identical with an estimate of the spatial position of the corresponding point of maximum excitation on the basilar membrane in the cochlea (in the inner ear).

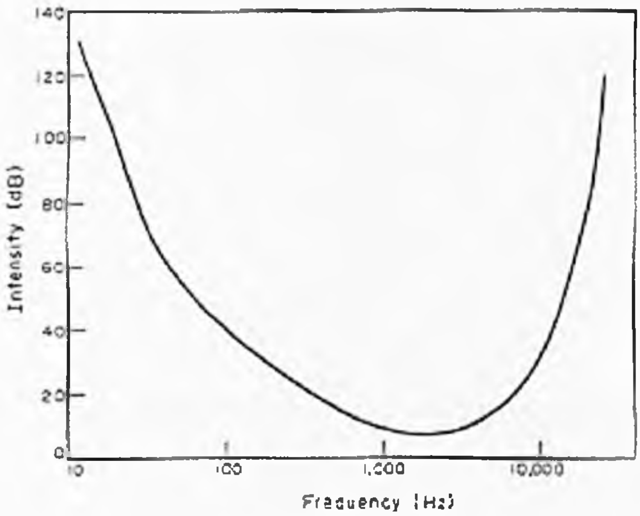

Figure A.1: Threshold of hearing as a function of frequency.

The cochlea, a liquid-filled tube located in the inner ear, performs a continuous broad-band analysis of the sound, which enters the ear, and transmits the results to the brain through the neural fibre outputs of the cochlea. The basilar membrane in the cochlea, which performs the spectral analysis, has a different frequency response along its length. Each location along the basilar membrane has a characteristic frequency; at which it vibrates maximally for a given frequency at the input of the cochlea is that of a band pass filter with almost constant Q (fixed ratio of centre frequency to bandwidth). Because of this constant-percentage bandwidth, frequency resolution along the basilar membrane is best at low frequencies. For every input frequency, there is a point on the basilar membrane of maximal vibration.

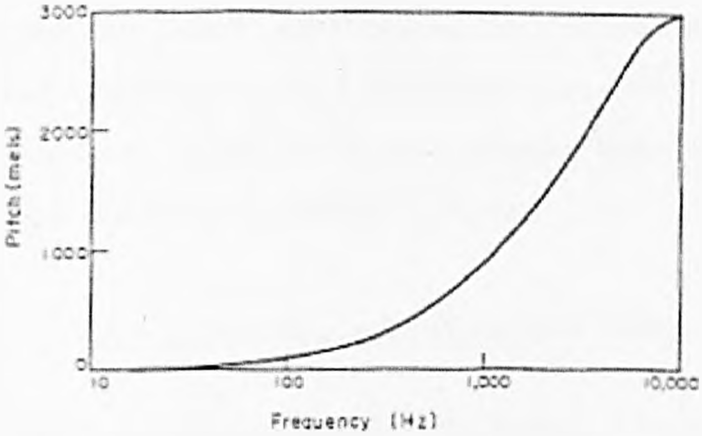

Figure A.2: Relationship between pitch scale and frequency.

According to the mel scale, the frequency range over which the human ear is able to perceive sounds can be divided into a bank of band pass filters. These filters are linearly spaced below 1000 Hz and logarithmically spaced above 1000 Hz. The filters under 1000 Hz have fixed bandwidths and are taken to be equal to 100 Hz. The filters at and above 1000 Hz follow a logarithmic distribution according to Eqn.A.1, and these filters are assumed to have constant Q given by

$$Q = \frac{f_0}{BW} \tag{A.2}$$

where f_0 and BW are the filter's centre frequency and bandwidth respectively. From Eqn.A.1, the frequency f is given as a function of its value y on the mel scale as follows:

$$f = 10^3(10^{y/k} - 1) \tag{A.3}$$

For a band pass filter 1000 Hz =1000 mels and for a bandwidth equal to 100 mels, Q is computed by substituting Eqn.A.3 into Eqn.A.2 to yield $Q = 7.3$. In order to have a flat composite spectrum over the whole frequency range of the filter bank, the centre frequency of a filter i is computed as follows:

$$f_i = f_{i-1} + BW_{i-1} = f_{i-1}(1 + 1/Q) = 1.137 f_{i-1} \tag{A.4}$$

For $f_{i-1} = 1000$ Hz, the centre frequency of the following filter is at 1137 Hz. Table A.1 illustrates values of the centre frequency of a bank of 22 filters covering the range 50-4980 Hz [3], where the centre frequencies above 1000 Hz follow Eqn.A.4

Filter No.	Center Frequency (Hz)	Bandwidth(Hz)
1	100	100
2	200	100
3	300	100
4	400	100
5	500	100
6	600	100
7	700	100
8	800	100
9	900	100
10	1000	118
11	1137	146
12	1292	166
13	1469	189
14	1671	215
15	1899	244
16	2159	316
17	2455	359
18	2791	408
19	3173	464
20	3607	527
21	4662	599

Table A.1: Filter bank centre frequencies and bandwidths (mel scale).

A.3 Critical Bands

When a weak tone is heard in the presence of an adjacent tone, the threshold for hearing the first tone is raised. This phenomenon is known as 'masking'. It was found that the threshold is raised only when the tones are close to each other in the frequency. If they are more than a critical distance apart, the second tone (whose intensity is above the hearing threshold) has no effect on the threshold for hearing the first tone [1]. This has led to the concept of the critical band. Signals within the critical band influence the perception of each other.

Critical bands are measured throughout the frequency range of hearing by listening to tones mixed with band-limited noise. The tone is set at the centre frequency of the band of noise. As the bandwidth of the noise is increased, the intensity at which the tone was just perceived is also increased until the bandwidth of the noise is equal to the critical band. Thereafter, the intensity for hearing the tone remains constant. It has been found the critical bandwidth increases as the centre frequency is raised. The critical bandwidth for a centre frequency of 200 Hz is found to be about 100 Hz, and for 5000 Hz about 1000 Hz [1].

In the cochlea, the point of maximum vibration moves along the basilar membrane as the frequency of excitation is increased. The critical bandwidths correspond approximately to fixed spacing (1.5 mm spacing) along the basilar membrane, suggesting that a set of 24 band pass filters would model the basilar membrane well. A perceptual measure, called the 'Bark' scale [4] or 'critical-band rate', relates acoustical frequency to perceptual frequency resolution, in which one Bark covers one critical bandwidth over the whole frequency range, and corresponds nearly to a pitch interval of 100 mels. Table B.2 gives the values for preferred frequencies defining the limits of auditory critical bands [4].

An analytical expression [5] mapping the frequency f into critical -band rate Z, and another expression for critical bandwidth CB are given as follows:

$$Z_i = 13\arctan(0.76f) + 3.5\arctan(f/7.5)^2 \tag{A.5}$$

$$CB_i = 2575(1 + 1.4f^2)^{0.69} \tag{A.6}$$

Where f is taken in KHz. These expressions approximate the tabulated data with an accuracy of ±10%. From Table A.2, we notice that the critical bandwidth is constant at low frequencies but increases with the logarithm of frequency at high frequencies. Also the critical-band rate is proportional to frequency at low frequencies, but at medium and high frequencies it is proportional to the logarithm of frequency. The critical bands have a certain width, but their position on the frequency scale is not fixed.

Critical Band Rate Bark	Center Frequency (Hz)	Critical Bandwidth (Hz)
1	50	100
2	150	100
3	250	100
4	350	100
5	450	110
6	570	120
7	700	140
8	840	150
9	1000	160
10	11701	190
11	1370	210
12	1600	240
13	1850	280
14	2150	320
15	2500	380
16	2900	450
17	3400	550
18	4000	700
19	4800	900
20	5800	1100
21	7000	1300
22	8500	1800
23	10500	2500
24	13500	3500

Table A.2: Values of critical band rate and critical bandwidth as a function of frequency.

A.4 References

[1] W.A. Ainsworth, *Mechanism of Speech Recognition*, Pergamon Press, 1976.

[2] G.Fant, *Speech Sounds and Features*, The MIT Press, 1973.

[3] G.Ghazali, "Element of arabic phonetics," *Proc. Arab School on Science and Technology, Applied Arabic Linguistics and Signal & Information Processing, Rabat, Morocco*, Sept. 1983.

[4] B.H.Juang, "On the hidden markov model and dynamic time warping for speech recognition- a unified view," *AT& T B. Labs. Tech*, vol. 63, pp. 1213-1243, September 1984.

[5] P.Lippmann, "An introduction to computing with neural nets," *IEEE ASSP Magazine*, pp. 4–22, April 1987.

Appendix B

TIMIT Database

Since the wish reported in this thesis uses speech data from the TIMIT database, we now present a brief overview of this database. TIMIT is an acoustic-phonetic speech corpus designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of speech processing systems [1]. It is prepared by the National Institute of Standards and Technology (NIST) with sponsorship from the Defence Advanced Research Projects Agency-Information Science and Technology Office (DARPA). TIMIT consists of a total of 6300 sentences, 10 sentences spoken by each of 630 male and female speakers from 8 major dialect regions of the United States. The speech data in TIMIT is divided into two broad groups: train and test for training and testing purposes. Each group is further subdivided into eight dialect groups. There are four files associated with each sentence data: a wave file (.wav), a text file (.txt), a word file (.wrd) and a phone file (.phn). The wave file consists of waveform speech data with a header. The speech waveforms are digitized at the sampling rate of 16 kHz and are stored in binary format. The text file contains the associated orthographic transcriptions of the words in a sentence. The word file is composed of the time-aligned word transcriptions while the phone file consists of the time-aligned phonetic transcription. A more detailed description of the TlMIT phonetic lexicon can be found in [1]. In Tables 2.4 and 2.5 we present the TIMIT phonetic transcription that is used consistently throughout in this thesis.

Phone type	Symbol	Example word	Phonetic transcription
Stops	b	bee	BCL B iy
	d	day	DCL D ey
	g	gay	GCL G ey
	p	pea	PCL P iy
	t	tea	TCL T iy
	k	key	KCL K iy
	dx	muddy,dirty	m ah DX iy, dcl d er DX iy
Affricatives	jh	Joke	DCL JH ow kcl k
	ch	choke	TCL CH ow kcl k
Fricatives	s	sea	S iy
	sh	she	SH iy
	x	Zone	Z ow n
	zh	Azure	ae ZH er
	f	fin	F ih n
	th	thin	V ae n
	dh	Van	DH e n
		then	
Nasals	m	Mom	M aa M
	n	noon	N uw N
	ng	sing	S ih NG
	em	Bottom	Baa tcl t EM
	en	button	B ah q EN
	eng	washington	W aa sh ENG tcl t ax n
	nx	winner	w ih NX axr

Table B.1: Phonetic transcription used in the TIMIT database for Stops, Affricates, Fricatives and nasals.

Semivowels	l	lay	L ey
	el	bottle	Bcl b aa tcl t EL
	r	ray	R ey
	w	way	W ey
	y	yacht	Y aa tcl t
Aspiration	hh	hay	HH ey
	hv	ahead	AxHV eh dcl d
Vowels	iy	Beet	bcl b IY tcl t
	ih	Bit	bcl b IY tcl t
	eh	Bet	bcl b EH tcl t
	ey	Bait	bcl b AE tcl t
	ae	Bat	bcl b AE tcl t
	aa	Bott	bcl b AA tcl t
	aw	Bout	bcl b AW tcl t
	ay	Bite	bcl b AY tcl t
	ah	But	bcl b AH tcl t
	ao	Bought	bcl b AO tcl t
	oy	Boy	bcl b OY
	ow	Boat	bcl b OW tcl t
	uh	Book	bcl b UH kcl t
	uw	Boot	bcl b UW tcl t
	ux	toot	tbcl t UX tcl t
	er	Bird	bcl b ER dcl d
	ax	about	AX bcl b aw tcl t
	ix	debit	dcl d eh bcl b IX tcl t
	axr	Butter	bcl b ah dx AXR
	ax-h	suspect	s AX-H s pcl p eh kcl k tcl t

Table B.2: Phonetic transcription used in the TIMIT database for semivowels, Aspiration and vowels.

B.1 References

[1] National Institute of Standards and Technology (NIST), "The darpa timit acoustic-phonetic continuous speech corpus," Oct. 1990.

Appendix C

List of Publications

1. S.Fekkai, M. Al-Akaidi, "A New Fractal Word Synthesis", EUREUMEDIA'2002, April 2002, Modena, Italy.

2. S. Fekkai, M. Al-Akaidi, J.Blackledge, "Speaker Independent Phoneme Recognition Based on Fractal Dimension D_f and the Mel-Frequency Cepstral Coefficients Features", The IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol.VI, 7-11May, 2001, Salt Lake City, Utah, USA.

3. S.Fekkai, M. Al-Akaidi, "Neural Network Techniques for Speech Recognition Speaker Independent",EUREUMEDIA'2001, SCS, Spain, May 2001.

4. S.Fekkai, M. Al-Akaidi, "New Features for Speaker Independent Speech Recognition", 3rd Middle East Symposium on Simulation and Modelling, Jordan, Spetember 3-5, 2001.

5. S.Fekkai, M. Al-Akaidi, J Blackledge, "Fractal Dimension Segmentation: Isolated Speech Recognition", ICASPAT2000, Dallas, USA, 16-18 October 2000.

6. S. Fekkai, M. Al-Akaidi, J.Blackledge, "Fractal Dimension Segmentation: Isolated Speech Recognition", IEE Electronic & Communications Event on 'Speech

Coding Algorithms For Radio Channels', Savoy place, London, 17 April 2000.

7. S.Fekkai, M Al-Akaidi and J Blackledge, "A Comparative Study of Fractal Dimension Estimation for Speech, ESM 2000, 14th European Simulation Multiconference, Ghent, Belgium, pp676-680, May 23-26, 2000.

8. S.Fekkai, M Al-Akaidi and J Blackledge, "Pattern Recognition of Speech Phonemes Based on Fractal Dimension", 2^{nd} Middle East Symposium on Simulation and Modelling, Jordan, August 28-30, 2000.

9. S.Fekkai, M Al-Akaidi, "New Feature to Improve Fractal Speech Recognition", International DSP Conference'2001, Boston, USA, September 10-11, 2001.

10. M. Al-Akaidi & S.Fekkai, "Words Recognition Based on Fractal Properties", International Conference on Image Science Systems & Technology (CISST'99), 28/6 1/7, 1999, Las Vegas, USA.

11. M. Al-Akaidi, S. Fekkai & J.Blackledge, "Simulation of Fractal-Based Words Recognition", The Society for Computer Simulation, Summer Computer Simulation Conference, SCSC31st Annual, July11-15, 1999, Chicago, Illinois, USA.

12. S. Fekkai, M. Al-Akaidi, "A Novel Fractal Speech Synthesis", International Symposium on *Performance Evaluation of Computer and Technology,* SPECT 2002, California, USA.

13. S.Fekkai, M. Al-Akaidi, "A New Speech Synthesis Based on Fractal", EUSIPCO 2002, XI European Signal Processing Conference, Toulouse, France, September 2002.