ON THE GENERATION OF TEXT AND SPEECH FOR DATABASE APPLICATIONS -

Ramon Castillo-Ocampo

A Thesis

in

The Department

of

Computer Science

Presented in Partial Fulfillment of the Requirements

for the degree of Master of Computer Science at

Concordia University

Montreal, Quebec, Canada

June 1981

ABSTRACT

ON THE GENERATION OF TEXT AND SPEECH FOR DATABASE APPLICATIONS

Ramon Castillo-Ocampo

Man machine communication from computers to end-users in the form of spoken utterances is the central theme of this thesis. A solution to this problem is considered at two stages: generation of sentences in a natural language as the database response to a user's query is one stage, and generation of speech from such texts or an ordered sequence of sentences is the second stage. The natural language considered in this thesis is Spanish. In general, there is a need to generate natural sounding speech from texts for communicating with the end-users. The use of syllables as a concatenative unit for speech synthesis is studied in this context. A procedure is described which produces a phonetic transcription of the input text. In this phonetic transcription, stress markers, intonation level markers and pause markers are incorporated on a syllabicated text. This would be useful in the production of natural sounding speech. Finally, a simple case study is used to illustrate the integration of sentence and speech generation in the context of a database. This is an interdisciplinary problem that involves areas of study such as text generation from Artificial Intelligence,

good-quality speech generation from signal processing and linguistics, and their integration into a database management system.

To my parents, Roberto and Graciela, for their moral support and devotion to their children.

To Rebeca -my friend and wife- for her time devoted to my studies and for the many lost weekends.

To the memory of our friend Sreeram Reddy.

ACKNOWLEDGMENTS

I am very much indebted to my supervisor Dr. Radhakrishnan. I greatly appreciate his assistance, advice and encouragement throughout the stages of this work.

I wish to thank Dr. Eric Regener, Mr. George Mack, Ms. Pauline Dubois and Mr. Cliff Grossner for their assistance and permission to use their software packages.

Financial support for my graduate studies has been provided by the Government of Mexico through the Consejo Nacional de Ciencia y Tecnologia (National Council of Science and Technology).

I also wish to thank the members of the Computer Science Department, the staff of the SEL Library, and the staff of the Computer Centre for their assistance during my studies at Concordia University.

My sincere gratitude to all my professors in Mexico City and to all my friends in Mexico and Montreal. I have got many of my best experiences with them.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

Figures

Tables

CHAPTER I

INTRODUCTION

1.1 Background:

In recent years, there has been a considerable growth of interest in speech processing by computers. Developments in large scale integration technology, electronics, microprocessors, memories, database systems, digital signal processing and the understanding of human speech, have all contributed to such a growth of interest. Depending on the sphere of application and goals of the end-user, speech processing involves the following:

a). Speech analysis - Analysis, extraction of parameters and features of speech, and their digital representation for computer processing or for speech transmission.

b). Speech synthesis - Generation of acoustic signals for the perception of speech by the human ear, models for speech synthesis and generation of speech from written texts.

c). Combination of analysis and synthesis procedures.

In this thesis, we are mainly concerned with speech

synthesis. Several approaches have been taken by researchers to produce synthetic voice. Mechanical analogs of the human-voice production mechanism were constructed by Kratzenstein in 1779 and by Von Kempelen of Vienna in 1791. Kratzentein's objective was to synthesize the five vowels in order to explain the physiological differences between them. As a next step in speech synthesis, Von Kempelen built and demonstrated a machine that produced connected utterances [Flan,72a]. A perceptual approach to speech synthesis was taken by scientists like Von Helmholtz (1854) and Stumpf (1926) in Germany, Richard Paget (1930) in England, and by D.C.Miller (1916) [Fant,68; Flan,72a].

The evolution of electrical technology contributed to further progress in speech synthesis research. Instead of mechanical analogs, electrical analogs were constructed by Stewart (1922) and Wagner (1936). Wagner was the first investigator who applied modern circuit theory concepts to speech synthesis [Fant,68]. Dudley, Riesz and Watkins [Dudl,39] developed the speech synthesizer called VODER (voice demonstration), based on electrical circuits for filters that produced continous speech.

Developments in computer technology have been the next important milestones that contributed to further growth in speech processing. Research in digital signal processing,

in language theories, and in linguistic and physiological characteristics of human-speech production systems have helped to promote the topic of speech processing to its present state. Results of the experiments in speech perception, conducted mainly by psychologists and phoneticians have been useful not only for synthesis but also for the evaluation of synthesized speech.

Speech recognition and speech synthesis render voice input to and voice output from computers respectively. Voice output from computer has opened several new areas of application to computers: reading machines for the blind [Alle,76], database interface to naive users through telephone lines [Witt,77], augmentation of CAI (Computer Aided Instruction) with voice output [Supp,79], issuing audible instructions to workers working in a place where reading of instructions is difficult or impossible [Flan,76], and educational talking-toys for children [Wigg,78] are but some examples.

1.2 Representation of Speech:

Speech is produced when air flows from the lungs through the mouth or the nostrils. The vocal tract (Fig.1.1), consists of the larynx, the pharynx, and the mouth. The nasal cavity is coupled to the vocal tract in

the production of certain sounds. In adult males, the average length of the vocal tract is about 17 cm. The diameter and shape of the vocal tract varies continuously during the speech production process, by the movement of the lips, the velum, the tongue, and other parts of the vocal tract [Malm,63].

Most speech sounds are produced using the expiratory phase of respiration [Lieb,67]. These sounds can be classified into two main categories, namely, vowel and consonant sounds. Vowel sounds are produced without any constriction in the vocal tract and with vibration of the vocal folds. The consonant sounds are produced with a constriction in the vocal tract. Consonant sounds can be classified into two main categories, namely voiced and unvoiced. Voiced speech sounds are produced when the flow of air from the lungs passes through the larynx forcing the vocal folds apart. The vocal folds are brought together due to their elasticity and the reduction in the pressure below them. Then, pressure below the vocal folds is built up again to repeat the cycle. The fundamental frequency of these vibrations is about 130 Hertz (Hz) for a male speaker, and about 200 Hz for a female speaker. The fundamental frequency of children's voice is higher than that of adults [Pott,50]. In the production of unvoiced sounds, there is no vibration of the vocal folds and the stream of air runs freely through the larynx.

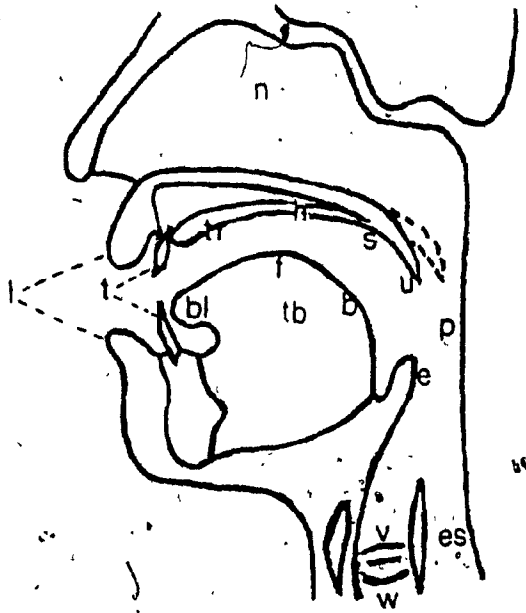


Figure 1.1 Vocal Tract and Nasal Cavity.

(Adapted from [Hill,79])

b - back of tongue	n - nasal cavity
bl - blade of tongue	p - pharynx
e - eppiglottis	s - velum
f - front of tongue	t - teeth
es - esophagus	tb - tongue body
h - hard palate	u - uvula
l - lips	v - vocal folds
tr - teeth ridge	w - larynx

The flow of air coming out from the glottis, that is the orifice between the vocal folds, undergoes several changes before coming out from the nostrils or the mouth. If the velum is lowered the air flow will pass through the nasal cavity and come out from the nostrils which produces nasal sounds. Non-nasal sounds are obtained when the velum is raised to block the nasal cavity, Another class of speech sounds is generated when a constriction is made in the vocal tract to cause turbulence. The speech sounds so produced are known as fricative sounds. Stop sounds are obtained when pressure behind a closure is abruptly released.

Speech is not made up of a sequence of discrete speech sounds. However, for analysis, it is convenient to consider speech as a concatenation of a finite number of distinguishable, mutually-exclusive sounds. These sounds, or linguistic units, can be thought of as minimal speech sounds which have the property that if one is replaced by another in a given utterance, the meaning of the utterance is changed or distorted. The acoustic realizations of such a basic liguistic unit are not always the same and they vary. These variations are considered to represent the same sound by a listener with competence in the language when he hears them. That is to say, in a given language the acoustio manifestations of a basic linguistic unit

signify the same linguistic element. These basic linguistic units are called phonemes. The different acoustic variations of a phoneme are referred to as allophones of that particular phoneme.

The nature and number of phonemes depend on the language under consideration. At one level of analysis phonemes are divided into vowel phonemes and consonant phonemes. The number of phonemes from each of these two categories varies among different languages. English consists of about eleven vowel phonemes and twenty four consonant phonemes; German has fourteen vowel phonemes and twenty three consonant phonemes; French has fifteen vowel phonemes and twenty consonant phonemes; and Spanish has five vowel phonemes and nineteen consonant phonemes [Dela,65].

Phonemes can be denoted by phonetic symbols. The set of phonetic symbols is called phonetic alphabet, and the process of representing speech sounds by phonetic symbols is known as phonetic transcription. In Table 1.1 one of the most widely used phonetic alphabets namely the International Phonetic Association (IPA) alphabet is presented.

Using a microphone as a transducer and a host of other

PHONEME	COMPUTER Repr. 2-Chars.	EXAMPLE	PHONEME	COMPUTER Repr. 2-Chars.	EXAMPLE
i	IY	beat	p	P	pet
I	IH	bit	t	T	ten
e	EY	bait	k	K	kit
ɛ	EH	bet	b	B	bet
æ	AE	bat	d	D	debt
ɑ	AA	Bob	g	G	get
ʌ	AH	but	h	HH	hat
ɔ	AO	bought	f	F	fat
o	OW	boat	θ	TH	thing
ʊ	UH	book	s	S	sat
u	UW	boot	ʃ	SH	shut
ə	AX	about	v	V	vat
ɨ	IX	roses	ð	DH	that
ɝ	ER	bird	z	Z	zoo
aʊ	AW	down	ʒ	ZH	azure
aɪ	AY	buy	ʧ	CH	church
ɔɪ	OY	boy	ʤ	JH	judge
y	Y	you	ʍ	WH	which
w	W	wit	l̩	EL	battle
r	R	rent	m̩	EM	bottom
l	L	let	n̩	EN	button
m	M	met	flapped t	DX	batter
n	N	net			
ŋ	NX	sing			

TABLE 1.1
IPA and ARPABET symbols for representing the phonemes of English (Adapted from [Lea,80b] p.127)

instruments, the acoustic characteristics of speech can be represented by a set of relations between time, frequency, and amplitude of the transduced electrical signals. The commonly used relationship are:

a).Time vs. amplitude, a two dimensional plot of the amplitude variations of a given speech

signal (Fig. 1.2).

b). Time vs. frequency vs. energy, a three dimensional plot known as a sound spectrogram. The energy (Fig. 1.3) content in this plot is indicated by the degree of darkness on the graph at that frequency and time. The x-axis and y-axis correspond to the time-scale and frequency-scale respectively.

c). Frequency vs. amplitude, a plot in which the amplitude of the harmonic component at each frequency is shown.

d). Time vs. frequency, a two dimensional plot of the formant trajectories. Formants refer to those frequencies of the speech sound in which the energy is concentrated into packets. It has been found [Fant,60] that the first three formants are sufficient for an adequate perception of speech signals. Usually the formants are denoted by F1, F2, F3 and so forth in the increasing order of their frequency values.

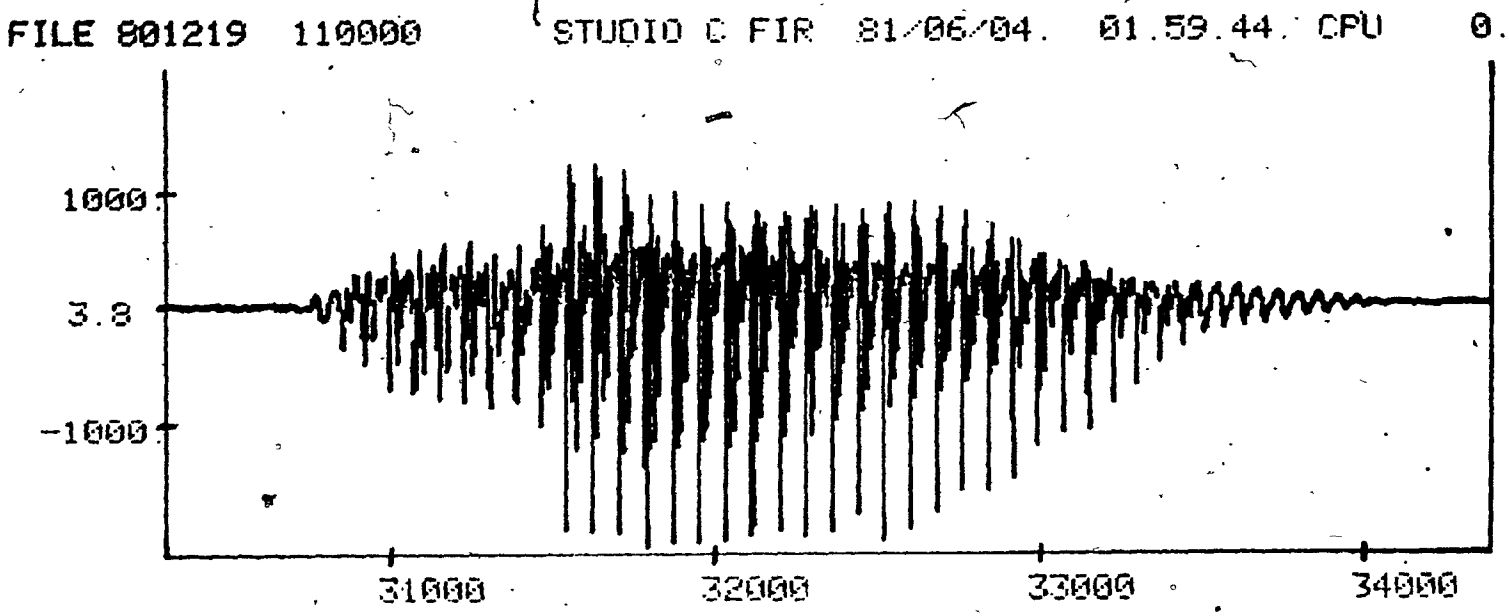


FIGURE 1.2 SPEECH SIGNAL (TIME VS. AMPLITUDE)

SRATE 10000 TTIM 3.030 #DISP 4096(4) AVER 0 +-(),<>[],GO,END

LSAMP 34396 SAMP# 30300 CON 0.0500 DISPL SIG,ENV,SPEC .. YNN

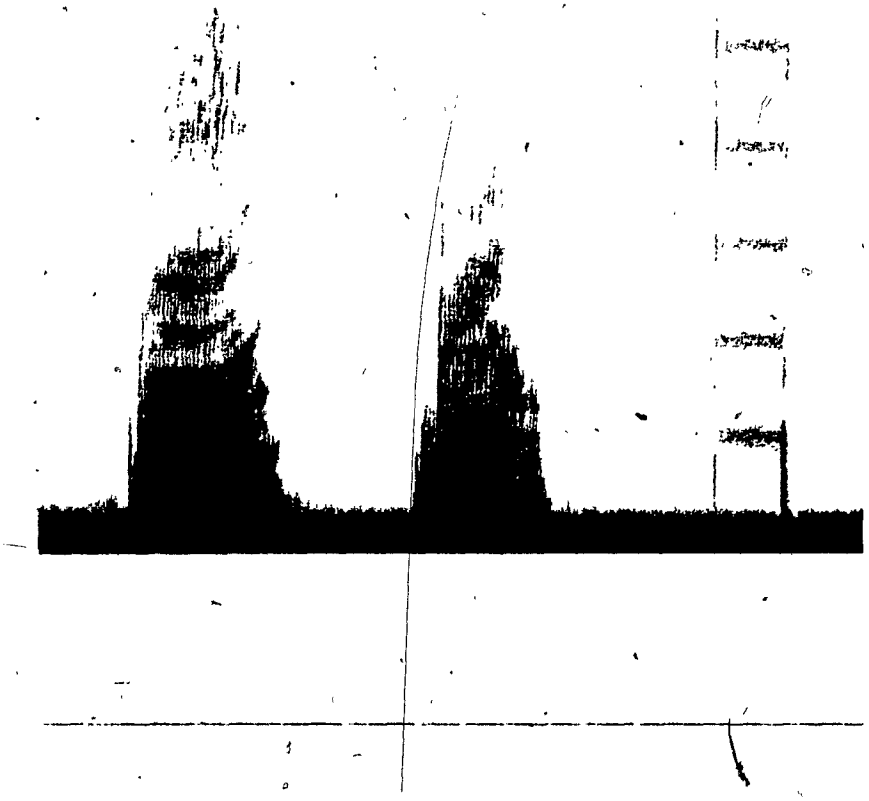


Figure 1.3 Speech Signal(Time vs. Amplitude vs. Frequency)

1.3 Speech Interface to Databases:

The modern trend in data processing is toward the collection of data into one centralized structure known as database. Such a database may be accessed for data retrieval or for data updating by remotely located users who are connected to the central site through communication lines. With the decreasing cost of electronics hardware, there is a proliferation of home computers, modified home-TV sets, and push-button telephones that function as computer terminals. There is a growing trend towards "public information utilities" which are collectively known as "videotex" or "viewdata". There have been several field trials of such information utilities with Prestel in England [d'Ag,79], Telidon in Canada, Bildschirmtext in Germany, and Antiope in France [Ball,80]. These utilities are designed to offer a variety of information services to the public such as messaging, teleshopping, telebanking, teleconferencing, electronic mail, and interest matching.

As a consequence of the introduction of the public information utilities, more and more people with no programming background have a need to access a database and to understand the output from the database management system (DBMS). By far the most common mode of output from

a DBMS is the text printed on a hardcopy terminal or displayed on a softcopy terminal. This mode of output necessitates the computer users to be physically present near the terminal and be able to read and interpret the printed output and messages. While this mode of output is well accepted by time-sharing users of a computer system, it is not suitable for the output from a monitoring system where the time of occurrence of the output is unpredictable [Fall,78]. Also, visually handicapped users will not benefit from this mode of printed text output. It is in this context that we feel speech output from databases will be beneficial. Moreover, for communicating with the naive users of public information utilities, either speech output or speech output combined with text output would be useful.

There are two distinct aspects in the speech interface to databases:

a).Voice input to DBMS and speech recognition.

b).Voice output from DBMS and speech synthesis.

Unrestricted speech recognition is still a research problem [Lea,80a]. We feel that it is not yet ready for use in public information utilities, as a source of input. However, speech synthesis has reached the commercial product stage and speech of acceptable quality can be synthesized from unrestricted texts in a cost-effective manner [Umed,76; Pinn,79]. When speech is used as a mode

of communication from a machine to a person, there are many acoustical cues available that could be used to emphasize the different aspects of the communicated message. Voice pitch, stress on utterances, pause, intonation contours, and repetition of phrases and sentences are some of the available acoustical cues. The DBMS or the voice-communicator can employ these cues to convey the message more effectively, perhaps in a "human-like" manner. The following is a hypothetical example.

Dialogue-1 Begin

Man : Can I pay $200 for my chargex bill this month?
(Implication - 200<=x; where x is the disposable amount in his account toward bills to be paid).

Machine : OK "short pause" You can pay it.

Another possible response might be:

Machine : No "pause" $200 is too much "pause" but "short pause" you may "stressed word" pay 50 dollars.

Dialogue-1 End

Suppose there is a software module that rewrites or transform a given input sentence into a format that can drive a "synthesizer unit" (SU) (Fig. 1.4). The commercially available SU's, in the form of LSI chips, are based on phonemes or on LPC (Linear Predictive Coding) parameters [Mark,76]. The software modules can be based on the text to speech rules reported in the literature

[Eloy,76; Alle,76]. Further, these modules will determine the place within a sentence for the acoustical cues and their types to be used in the generated speech.

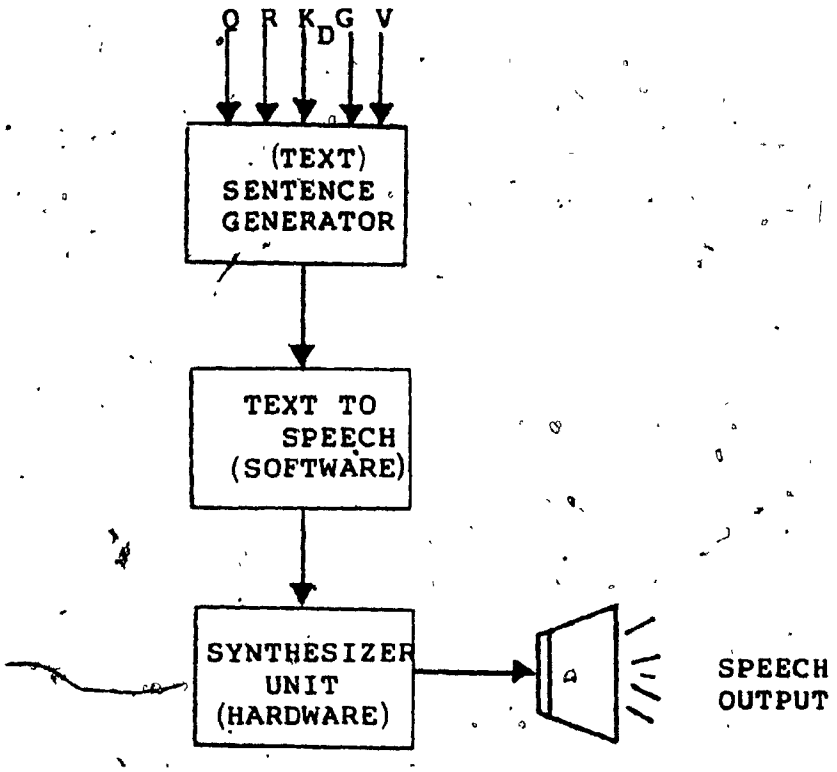


Fig. 1.4 A block diagram for voice output from a DBMS.

Q : From user's query.

R : From the results of the search from a database.

Kd : Functional and other dependencies, and synonyms of the data store in the database.

G : Grammar used for generating sentences.

V : Vocabulary used by the sentence generator.

1.4 Text Generation :

Text generation, in some sense, is a converse problem of natural language understanding. We use the terms text generation and sentence generation synonymously. Early work on automatic sentence generation emerged from the research in machine translation and from the machine to man responses in question answering systems. Random English sentences were generated by Yngve (1962) by means of a generative grammar. He produced a large number of sentences, many of which were not meaningful [Gold,75]. The earliest attempt to produce "meaning-oriented" sentence output was due to Klein (1965).

The formalization of transformational grammar by Noam Chomsky [Chom,65] had a strong impact on linguistic applications. A sentence generation program based on transformational grammars was developed by Friedman [Frie,69] as an aid to linguists. The sentences generated by his system were useful to test grammars. In Winograd's system [Wino,73], sentence generation was a by-product and was used for the man-machine dialogue. His system was mainly concerned with a restricted world of well defined physical objects (blocks) and their arrangements by a robot. Augmented Transition Network (ATN) grammars have been used by researchers like Woods [Wood,70] for sentence

generation. Similarly, Simmons and Slocum [Simm,72] have studied the generation of English sentences from "semantic networks".

In a database context, the problem of generating a sentential response can be broken down into two parts:

a). Selecting the information to be presented or selection phase.

b). Generating the sentence(s) from the output of (a) according to a grammar G, called sentence generation phase.

It is obvious that parts (a) and (b) are interrelated through the grammar. The selection phase will have input from many sources as shown in Fig. 1.4. For example, in the Dialogue-1 given in section 1.3, the noun "$200" and the verb "pay" are selected from the user's input query, the number "50" is selected from the response from the database search, the negation "not" is inferred from the database-knowledge-base and so on. Using the selected nouns, verbs, and modifiers, the sentence generator will create a sentence according to the grammar G.

In this thesis, it is our contention that a sentence generated in a database context can then be transformed into a symbolic input acceptable to a speech synthesizer unit (SU). The transformed sentence presented as input to

a SU might further contain "special markers" to indicate the location and the nature of the suprasegmental features within that sentence. Here, we are concerned with the special markers meant for pause, stress, and intonation level contours. It is expected that a SU will make use of these special markers to produce a "natural-sounding" speech for communication with the human users.

CHAPTER II

SPEECH AND SENTENCE GENERATION

2.1 Linguistic Aspects of Speech:

For speech recognition and speech synthesis, it is useful to have an adequate knowledge about the way in which information is encoded in the speech signal. Well-stablished areas of knowledge like Linguistics [Malm,68], Phonetics [Fant,68] and Physiology [Sone,68], among others, can provide invaluable details about diverse aspects of speech. These details can be used in the design of speech processing algorithms. In section 1.2, it was pointed out that speech is made up of a continuous stream of sounds with interspersed pauses or silence. However, for analysis it is adequate to regard speech as being produced by joining the basic linguistic units.

The study and description of speech sounds are in the domain of Phonetics which in turn is an integral part of Linguistics. Phonetics as a science, depends on methods and results from Linguistics, Physics, and Physiology of Speech [Malm,68; Fry,79; Sone,68]. Therefore, only as much Phonetics as is relevant to our work in understanding the

difficulties in speech synthesis is presented in this section. It is also in the interest of phoneticians to know which are the speech sounds that occur in the languages of the world, and how they are produced, modified, perceived and grouped. The speech sounds can be studied from several points of view. In Phonetics, there are at least three ways of studying the speech sounds which contribute to the three branches of Phonetics. In Articulatory Phonetics, the goal is to provide adequate descriptions on the diverse ways in which speech sounds are produced. Studies of the way a hearer perceives speech sounds is in the realm of Auditory Phonetics. Acoustic Phonetics is the branch of Phonetics that analizes the properties of the speech wave which emanates from the speaker to the listener in speech communication.

One of the fundamental concepts in Phonetics is that of the phoneme (for an extensive coverage of this topic, see [Jone,76]). A phoneme can be thought of as a minimal speech sound which has the property that when it is replaced in a given utterance by another phoneme, the meaning of the utterance is changed or distorted. Thus, the phoneme is considered as the minimal basic unit that can distinguish meaning. For example, consider the sequence of English words: bell, cell, dell, fell, gell, hell, tell, well, yell; these words are distinguished only

by the initial sound of each word. These distinctive units, phonemes, are commonly symbolized as: /b/, /s/, /d/, /f/, /j/, /h/, /t/, /w/, and /y/. The nature and number of phonemes is language dependent.

Most languages of the world can be characterized in terms of a finite set of phonemes [Ruhl,76]. In doing so, phonetic transcription is a useful tool for the description of speech. Usually only the significant articulations are recorded. Differences in articuation which do not affect the meaning of the utterance are not transcribed. For a given language, its phonemes can be broadly classified into vowels, diphthongs, semivowels, and consonants. English, in particular mid-western American English, has an inventory of eleven vowels, six diphthongs, four semivowels, and twenty consonants [Cohe,52]. Spanish, according to the analyses of Navarro [Nava,68], has five vowels, six falling diphthongs, eight rising diphthongs, four triphthongs, and nineteen consonants. In Tables 2.1 through 2.4 a list of the phonemes of English and Spanish is given.

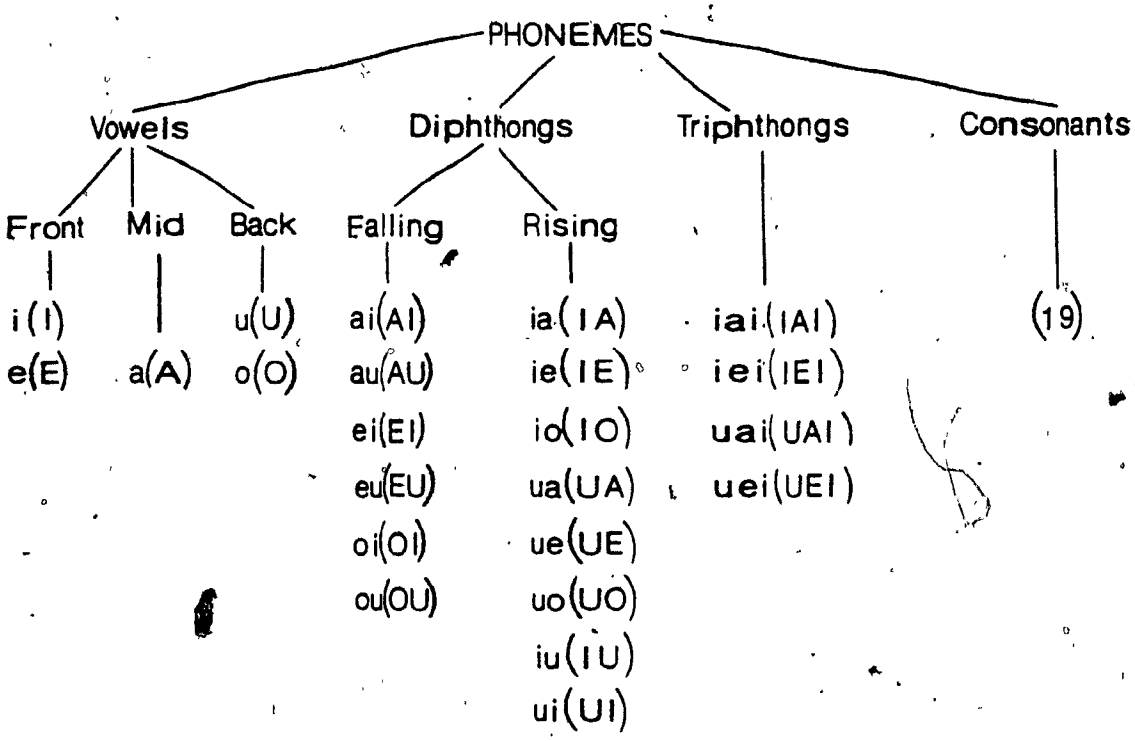


Table 2.1 The phonemes of Spanish

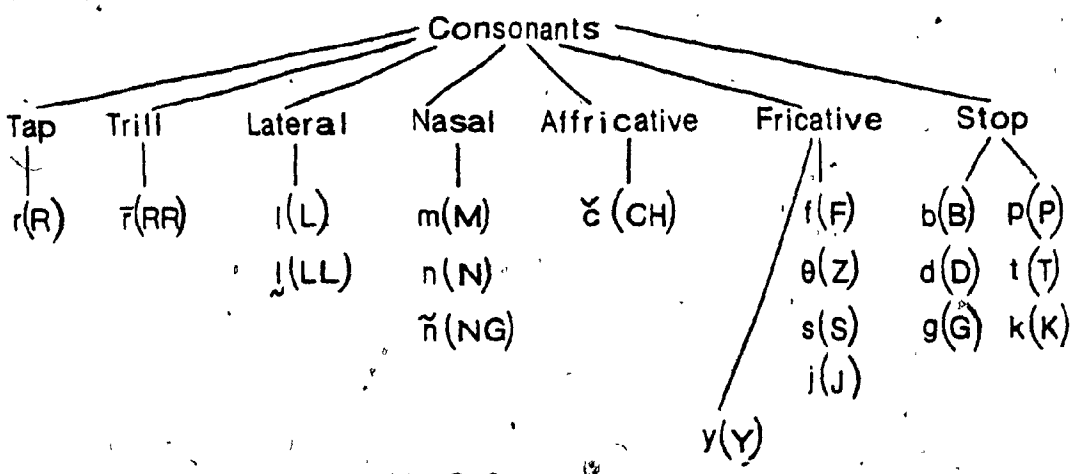


Table 2.2

Consonant phonemes of Spanish

Place of Articulation / Manner of Articulation	Bilabial	Labio - Dental	Inter - Dental	Alveolar	Palatal	Velar	Glottal
Stop	p b			t d		K g	
Fricative		f v	θ ð	s z	š ž		h
Affricate					č ǰ		
Nasal Resonant	m			n		ŋ	
Lateral Resonant				l			
Median Resonant				r			
Glides	w				y		(h)

Table 2.3 Consonant phonemes of English

	Front	Mid	Back
High	i	ɨ	u
Mid	e	ə	o
Low	æ	ɑ	ɔ

Table 2.4 Vowel phonemes of English

The corresponding sound of a phoneme does not occur in isolation, the consonant phonemes can barely be pronounced without a vowel before or after it. The linguistic unit next in size to the phoneme is the syllable. It is in the syllable that the phoneme achieves its physical reality. The concept of a linguistic unit above the phoneme and distinct from the morpheme or the word is not new. There have been several attempts to define the syllable, but still there is no commonly agreed definition [Bell,78; Malm,63; Pulg,70]. These efforts can be broadly classified into two categories. In one of them, the objective is to present a universal definition in phonetic terms; while in the other category, the aim is to propose a specific functional definition with reference to a particular language.

The prominence theory proposed by Grammont (1963), and the pulse theory developed by Stetson (1951) are examples of the phonetic approach [Malm,63]. The number of syllables in a word is equal to the number of "prominent" sounds in the same word, according to Grammont's theory. The problem in this theory is that it does not state where the syllable boundary is located. In the pulse theory, the number of syllables uttered is related to the number of "chest pulses" and the increase in air pressure. This theory suggests that the syllable but not the phoneme is

the basic unit of speech.

A definition of the syllable with reference to the structure of a particular language has some advantages. Under this approach the syllable would correspond to a unit of sound or stress. Also it would be possible to decide the place at which a syllabic division can be made. Identification of breath groups, that is the stream of speech between succesive pauses can be determined based on the number of syllables and word boundaries in a given text. In this context, a syllable might be defined as a sequence of sounds between two succesive points of relatively weak sonority, the nucleus of the syllable being a vowel sound which may or may not be accompanied by consonant sounds. In the next chapter a discussion on the characteristics of the syllable in Spanish will be given.

The syllabic boundary is linguistically relevant. A sequence of phonemes may mean different things depending on the place of the syllable boundary within the sequence. That is to say, a sequence of phonemes $p_1p_2 \dashv p_3$ where "$\dashv$" means syllabic boundary, has a different meaning than $p_1 \dashv p_2p_3$. Among the numerous examples given by D.Jones [Jone,76] for English, the following two are reproduced here: /mai$\dashv$'trein/ vs. /mait$\dashv$'rein/ (my$\dashv$ train vs. might rain) and /'nai$\dashv$treit/ vs. /'nait$\dashv$reit/ (nitrate

vs. night-rate). Some illustrations of similar situations in Spanish are as follows:

Spanish	Pronunciation	English
lei vs. ley	/lei/ vs. /le-'i/	law vs. I read
rey vs. reí	/rei/ vs. /re-'i/	king vs. I laughed

2.2 Speech Production Models and Methods:

Speech in humans is a complex process that involves activities at different levels such as semantics, syntax, phonology, motor control, articulatory movement, sound formation, and psychology. Understanding of the speech processes can be achieved by the use of models. In building models which can represent the speech processes, several assumptions are made.

Different models for the speech production process have been proposed by different researchers. Usually the articulatory or the acoustical aspects of speech are considered by such models. Mechanical analogs were used by Kratzenstein (1779), and Von Kempelen (1791) for the production of voice [Flan,72b]. Factors related to nonlinear properties of turbulent airflow in speech and to characteristics of unvoiced sounds are well tackled in mechanical analogs. Problems in the measurement of the

characterizing parameters of the human vocal tract have not been completely solved [Fant,59; Fant,68; Fry,79].

Acoustically, the vocal tract can be considered as a linearly separable system, modelled by a single tube with variable cross-sectional area. Then, a straight tube can approximate the L-shaped vocal tract since the transversal waves perpendicular to the tube surface can be ignored for the frequency range of speech signals (up to 4kHz). Further, the straight tube is approximated by several tube sections of constant length and area [Rabi,78]. This is a widely used speech-production model known as the acoustic tube model. The fundamental concept in acoustic-tube models is the simulation of the airflow or the air-pressure at different places of the vocal tract [Flan,72a; Rabi,78].

A geometrically oriented model of the human vocal tract was proposed by Coker [Coke,67]. The model is under the control of articulatory parameters which can be dynamically modified. The parameters represent the coordinate positions of the velum, the uvula, the tongue-body, the tongue-tip, and the position of the lips. The coordinates of the other vocal tract components are derived from the position of the tongue-body [Coke,76].

Another class of models are known as terminal-analogs

or formant synthesizers. In these models, the details of the vocal tract physiology have no relation to the actual synthesizer implementation. The synthesizer is implemented as a system whose transfer function approximates the actual vocal-tract transfer function [Oppe,78]. Basic studies of acoustic and perceptual phonetics have made use of digital simulations of this type of synthesizer. A block diagram of a general model for a terminal-analog synthesizer is given below [Rabi,78].

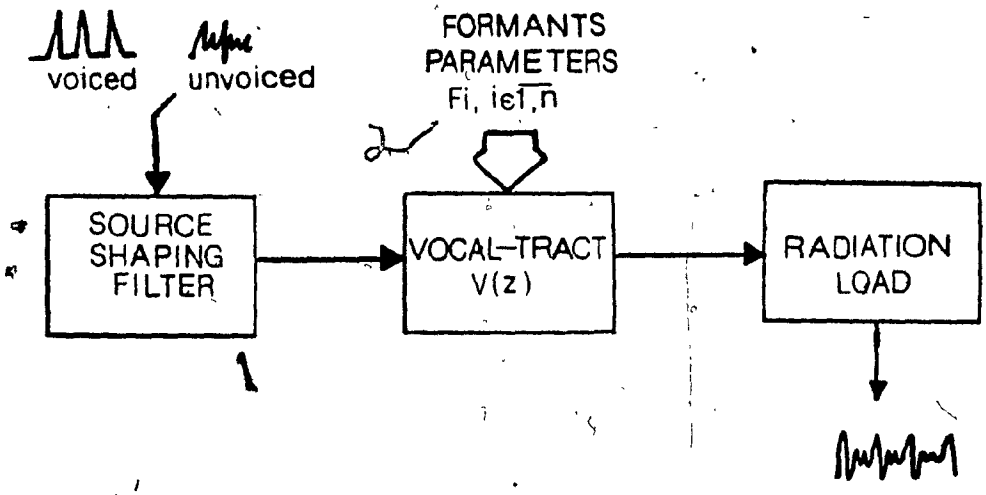


Figure 2.1 Model for a Terminal-Analog Synthesizer
(After Oppenheim [Oppe,78]p.124)

The precision and the simplicity in the control of parameters and the convenience in implementation make the terminal-analog synthesizers more popular [Oppe,78; Rabi,68; Rabi,78]. The fundamental concept in the

terminal-analog synthesizer is the representation of the speech production process, by a slowly time-varying linear system. This system is excited by an excitation signal whose basic nature changes from quasi-periodic pulses for voiced speech to random noise for unvoiced speech [Oppe,78].

The speech production process can be modelled in many other diverse ways [Fant,60; Flan,72a; Rabi,68; Coke,67; Oppe,78]. In doing so assumptions have to be made in order to maintain a non complex system. A block diagram representing a general discrete-time model for speech production is given in figure 2.2 [Rabi,78].

Today, speech synthesizers have a common place in speech processing laboratories. Their size, complexity and performance vary widely. LSI chips for the generation of synthetic speech are commonly available in the market [Wigg,78]. Current projections indicate that the anual market for speech synthesis chips range between $1.5 billion and $5 billion [Kapl,81].

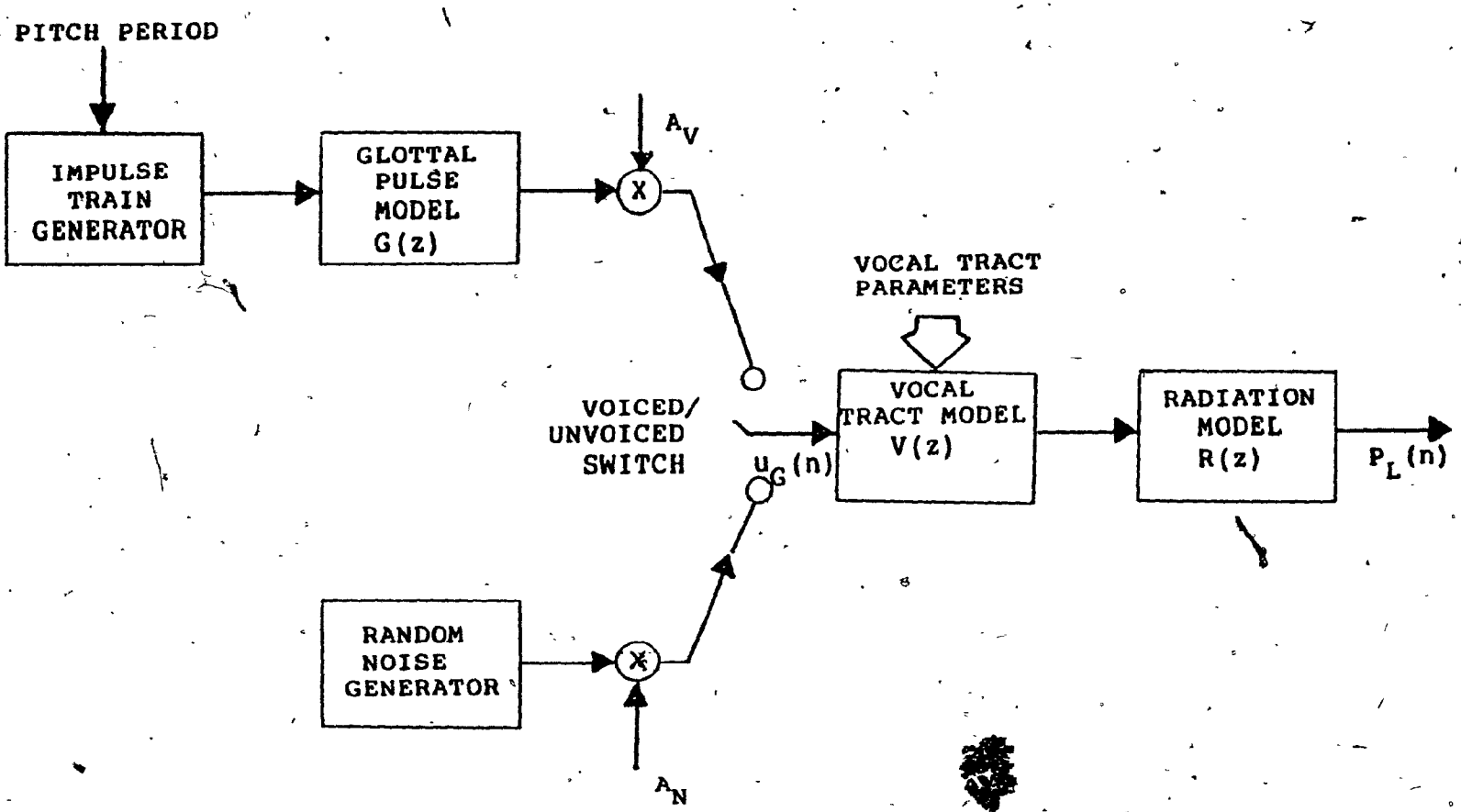


Figure 2.2 General Discrete-Time Model for Speech Production (after Rabiner and Schafer[Rabi,78])

The crucial aspect for the users of any synthesizer is the acquisition of adequate control parameters. Two major approaches for speech synthesis are: synthesis-by-analysis and synthesis-by-rule. In the former case, analysis of humanly produced utterances yield the control parameters which are used for synthesis. These parameters include: fundamental frequency and its amplitude, formant frequencies and their bandwidths, vocal-tract area, and the position of the articulators. Once the parameters are known they are stored, retrieved upon demand and used to generate utterances.

In speech synthesis-by-rule the objective is to generate voice from a symbolic input based on rules. The rules might operate on descriptions of formant-trajectories, phonetic representations, articulatory parameters, modified orthographic text, or even on regular orthographic text [Holm,64; Rabi,68; Flan,70; Alle,77]. The main difficulty in this method lies in the selection of data and rules which can reflect a general knowledge of speech production.

Speech synthesis-by-rule is a relatively new area in speech research. Synthesis-by-rule under computer control has been considered by Kelly and Gerstman [Kell,61]; Holmes, Mattingly, and Shearme [Holm,64]; Rabiner

[Rabi,68]; Allen [Alle,68; Alle,76]; Coker, Umeda, and Browman [Coke,73]; Elovitz, Johnson, McHugh, and Shore [Elov,76]; and many other researchers. The variety of their approaches have been influenced by developments in computer science, digital hardware, linguistics, and speech synthesis itself. In the next section, text-to-speech synthesis as a particular case of synthesis-by-rule is presented.

2.3 Text-to-Speech Synthesis for English and Other Languages:

In reaching for the remote goal of producing natural-sounding synthetic voice for unrestricted vocabulary, text-to-speech synthesis has been established as one of the most promising techniques [Flan,70; Alle,76]. Text-to-speech synthesis can be regarded as a particular case of synthesis-by-rule, where the symbolic input is the orthographic text and the rules are based on the conventions which describe the relationship between the graphemes (written symbols) and the phonemes of a given language.

Text-to-speech synthesis has attracted the attention of many investigators due to its general applications. Most of the work has been oriented towards automatic

orthographic text-to-speech synthesis. The main obstacles in orthographic text-to-speech synthesis are in the "knowledge" of the rules which establish the relation between a phonological system (sound system) and its corresponding writing system. In this respect, it should be made clear that a written language is only a symbolic representation of speech. Also we should be aware that changes in the spoken language occur more rapidly than in its orthographic representation.

A general strategy used by orthographic text-to-speech synthesis algorithms consists of extracting as much information as possible from the written text. The information obtained is complemented with the "knowledge" derived from linguistic and acoustical analysis. In order to attain naturalness in the synthetic voice output, proper stress, intonation levels and pauses should be incorporated in the output.

Work on orthographic text-to-speech synthesis has been carried on for diverse languages. For instance, Italian speech has been generated from text by Bertinetto et al. [Bert,77], and by Vivalda et al. [Viva,79]. Polish texts have been synthesized by Kielczewski [Kiel,78]. Also recently, Mangold and Stall used German texts to synthesize speech [Mang,78]. Speech in English has been synthesized

from text by Ainsworth [Ains,73], McILroy [McIl,74], Elovitz et al. [Elov,76], Allen [Alle,76], Sherwood [Sher,78], amongst others. Mandarin speech was generated by Suen [Suen,76]. Berdichevky et al. [Berd,79] and Sherwood [Sher,78] considered Spanish texts to produce synthetic speech. French texts have been used to synthesize speech by Rodet and Delatre [Rode,79]. The approaches taken in the above mentioned researches vary widely, mainly, due to the language under consideration and the hardware involved.

For unrestricted English text-to-speech synthesis, Allen [Alle,77] proposed and used structural models at several levels. These levels reflect articulatory, word formation, syntactic and semantic constraints. Essentially, the synthesis process is considered at two levels, namely, word-level and sentence-level. His thesis is based on the fact that many English words are made up of linguistic units called morphs. A morph is the basic unit of meaning within a word. Many words are also morphs such as: "house", "boat". These morphs include: prefixes, suffixes, infixes, and roots. After analyzing the abstract linguistic description of English texts, Allen devised rules for identifying the morphs of a given word and for generating their phonetic transcriptions.

There are several advantages in considering the morph as a basic unit. For instance, a fairly complete morph lexicon can be created since there are less than 12,000 morphs in English. The morph lexicon is sufficient to generate more than 1,200,000 words. Also, a morph lexicon provides an economical basis for representing most words in English. Perhaps one of the most notable advantages of this approach is the fact that new words can be generated rather easily using the elements of the morph lexicon. Besides, morphs are linguistic units which are more stable than words.

Once the constituent morphs of a word are identified, their pronunciation is obtained from the morph lexicon. If the morph is not in the lexicon "letter-to-sound" rules are applied. A phonetic transcription of the input word is obtained from these rules. A phonetic specification of a word can not be obtained by mere concatenation of the phonetic symbols. The phonetic symbols have to be adjusted properly to obtain the correct pronunciations for: plurals (busses vs. cats vs. donkeys), tenses (persuaded vs. walked vs. measured), and for the contextual effects of affixes (aggresion vs. rebellion, departure vs. failure). Also rules are used to predict the stress pattern of a word. Stress on an utterance is determined based on the lexical stress rules developed by Chomsky and Halle [Chom,68]. The

lexical stress rules are considered to be among the major achievements of modern Phonology. The output from the above linguistic preprocessing serves as input to the "phonemic-synthesis-by-rule-stage" where isolated words are synthesized.

In addition to the phonetic representation of words and stress patterns within words, further linguistic analysis is needed for synthesis at the sentence level. A parse of the sentence is one of the linguistic analyses needed. However, no parser is available to analyze all the possible sentences since there is no "complete" grammar of English. According to Allen [Alle,76], a procedure for the synthesis at the sentence level, should proceed from phrasal analyses to clause tests. His conclusion is that further work is needed in linguistics and parsing techniques before satisfactory sentence analysis becomes available.

A crucial element in synthesis at the sentence level is the relationship between acoustic and linguistic patterns. It has been found [Oliv,74] that in continuous speech the pitch parameter contributes the most towards the prosodics of speech. Although some adjustment is required, other acoustic parameters do not contribute substantially to make the synthesis speech sound natural [Oliv,74]. The

effect of the linguistic features in fundamental frequency patterns for English has been studied by O'Shaughnessy [O'Sh,79]. Sentence type, and syntactic construction were shown to influence the fundamental frequency values. These findings confirm that is impossible to predict the correct pitch contour of a sentence without reference to the context of the sentence [Alle,76].

Another approach taken towards unrestricted text-to-speech synthesis for English is due to Sherwood [Sher,78]. The strategy followed by Sherwood contrasts with the one taken by Allen [Alle,76]. The basic premise in Sherwood's work is that "simple" algorithms for unrestricted text-to-speech synthesis are possible only if "phonetic-text" is used [Sher,78].

The pronunciation of English has considerably changed since the seventeenth century, but the spelling has not been modified to that extent. Some authors [Simp,20] even consider that the English writing-system is far behind in comparision with the way English is spoken (comments on this subject can be found in [Wijx,66]). In this context, Sherwood considers English phonetic-text as the input text. The World English Spelling (WES) proposed by Dewey [Dewe,71], provides the alphabet for the English phonetic-text used in [Sher,78]. Once the English

phonetic-text is obtained, stress markers are manually inserted by Sherwood to indicate the stressed syllables. In the final stage, relatively easy letter-to-sound rules are applied to obtain a set of parameters to drive a phonemic synthesizer unit. The disadvantage of the above method is the manual rewriting of the regular English text into WES text.

2.4 Generation of English Sentences:

Several computer programs have been developed within the last decade to generate sentences in a natural language. Wong [Wong75] classifies them into two broad categories: (a) Based on the "Competence Model" due to Chomsky's theory of languages (b) Based on the "Performance Model" that depends on semantic representations. One of the differences between these models lies in the relative importance of the roles played by syntax and semantics. The competence model describes a language and it can not be a model of a speaker or a hearer. The performance model is advocated by works like Quillian [Quil,69], Simmons and Slocum [Simm,72], Goldman [Gold,75], etc. Their main concern is the understandability of the language produced. As a result the performance model has devices to limit the complexity of sentences whereas the competence model does not.

In the generation of a sentence, we are concerned with both grammar and form: Form determiners in English includes voice (active, passive..), state-form (progressive..), tense (past, present, future), aspect (simple,perfect..), and mood (affirmative, interrogative, exclamative..). Consider the following example from [Simm,72]:

- Input to sentence generator:

token (verb)	-	build
voice	-	passive
state-form	-	progressive
aspect	-	perfect
tense	-	future
mood	-	interrogative
subject	-	(John)
object	-	(house)

- Output from sentence generator under the control of a grammar:

"Will the house have been being built by John?"

Generation of a sequence of sentences that forms a discourse is referred to as "text generation". The problems involved in the generation of a multisentence discourse are coherence, inter sentential connections,

conditions on embeddings (complements and relative clauses), and rules for pronominalization, anaphora and ellipsis. The following example is reproduced from [Simm,72] to highlight some of the problems:

> "John saw Mary wrestling with a bottle at the liquor bar. He(John) went over to help her(Mary) with it(bottle). He(John) drew the cork and they(John and Mary) drank champagne together".

It has been found [Simm,72], [Wood,77], [Bate,78] that to support text generation a semantic network should include details such as explicit markers of the relative time-of-event for each verb structure, forms relations between sentences, and tools for pronominalization. A detailed illustration of a semantic network representation of this type can be found in [Simm,72]. Simmons and Slocum make use of the Augmented Transition Network Grammar (ATN) that was developed by Woods [Wood,70]. An excellent tutorial on ATN is found in [Bate,78]. In ATN a finite state model is augmented by the arbitrary testing and setting of a register associated with a node and by the nesting of transition networks one inside another. Transition from state to state in the ATN involves testing the deep structure and the information available in the registers being analyzed. For embedded structures, as in

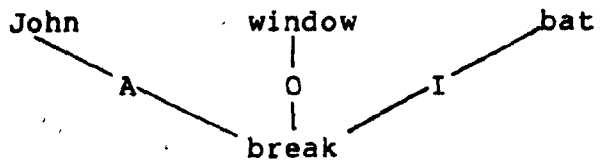
relative clauses, the grammar invokes other parts of the grammar to transform the embedded deep structures into surface sentences. Similarly, the necessary morphological transformation can be organized into procedures embedded in the grammar.

Another work of interest related to the generation of English sentences, particularly in the context of databases, originates from the University of Toronto [Mylo,75; Wong,75]. This project is known as TORUS (Toronto Understanding System). Representation of knowledge is central to any language processing system. In TORUS, the Case Grammar is used for this purpose. Fillmore in 1968 proposed a set of modifications to the theory of Transformational Grammar (TG) that resulted in the development of the Case Grammar [Fill,68]. According to Fillmore:

> "The sentence in its basic structure consists of a verb and one or more noun phrases, each associated with the verb in a particular case relationship. Each case relationship occurs only once (except noun phrase conjuction) in a simple sentence".

Further, he notes that each verb takes a particular set of

cases and the cases can be characterized suitably. For instance, in the sentence "John broke the window with a bat", the case frame pertinent to "break" would be as follows:



Where, A(agent) : the animate instigator of the action.

O(object) : the concept which is affected by the action.

I(instrument) : an inanimate object casually involved in the action.

It is important to note that the list of cases or case frames and their definitions are tentative. As the universe of discourse changes, the case list will have to be adjusted to include the new actions of the universe.

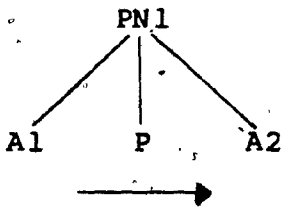
A semantic network, used in the project TORUS, is a directed graph in which both the nodes and edges are labelled. The nodes may be labelled as "concepts", "events", or "characteristics". Concepts are the basic entities of the system and they correspond to the nouns and adjectives in English. A concept node can denote either a

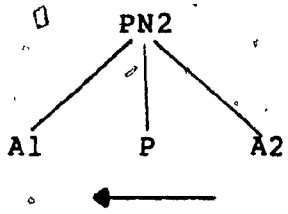
generic idea such as an aircraft or a particular instantiation of an idea such as DC-10. Event nodes most often correspond to verbs in English. In general, whenever we do not want to represent a situation too "deeply", characteristics are used. For example, sex of a person, his age, or address is represented as a characteristic. The arcs between nodes of a semantic network may be labelled by the symbols that denote cases such as agent, source, destination, location, instrument, instantiation of, etc.

A "selector" in TORUS takes a set of arguments from the question-answering system and copies part of the semantic network to form a "graph". This graph is transformed into a surface representation as a sentence, by means of a series of mappings. Detailed discussion of the mappings can be found in [Wong,75].

Yet another contribution from the University of Toronto can be found in [Kers,76]. In this work, a query stated in (pseudo) English form is analyzed, parsed and transformed into a set of RAP (Relational Associative Processor) primitives for the database search. Central to the knowledge representation in this work is a "predication". A predication represents a whole sentence such as an assertion, a command, or a question. It

consists of a predicate and either zero, one, or two arguments and may be viewed as a tree structure. For example let A1 = departments, A2 = parts, and P = supply. The two sentences "departments supply parts" and "parts are supplied by departments" would have the following predications PN1 and PN2 respectively. The predication structure is read from left to right. We use the verb's active voice if we traverse the arrow in its direction, and its pasive voice otherwise.





A number of different types of predications such as subordinate predications, downgraded predications, qualifying predications, and modifying predications are described in detail by Kerschberg et al. [Kers,76]. Also, they present a set of English sentences to describe a sample database of a department store. A set of synthetic English queries and an example sentence is given below:

Example query : "What are the names of items which are sold by departments that are located on floor equal 2?".

A research work that comes closest to the problem addressed in the present thesis is [Fall,78]. This system is designed to speak out the (abnormal) conditions that might arise in a computer-controlled water-supply network. An example sentence spoken out would be,

-Gorton Hill tower level is twenty nine point three percent

A grammar is used to produce English sentences which are then transformed into speech. The phonological component of the system is responsible for defining the way the isolated words of the utterance are to be concatenated and also for specifying the shape of the associated pitch contour. Each word to be spoken is digitally stored on a disk and a custom designed synthesizer unit (a set of cascaded filters) is used to produce the acoustic output. The prototype system has been implemented at the Cambridge University Engineering Department, England. The word oriented approach is appropriate for a such a dedicated application where the vocabulary size is not large and the vocabulary is invariant.

CHAPTER III

TEXT-TO-SPEECH SYNTHESIS FOR SPANISH

3.1 On Orthography and Phonology:

In an "optimal script", there is a one-to-one correspondence between the set of graphemes (symbols in the alphabet) and the set of phones (sounds in a language). Graphemes may represent single phonemes, as in European languages, or they may stand for syllables, as in Japanese and Tamil [Jone,76]. "Optimal scripts" or "optimal writing systems" are not common. Among the current European writing systems, none of them has an optimal writing system. However, some of the European languages possess fairly regular writing systems. For instance, Finnish and Czech writing systems are close to the ideal system. The Spanish writing system is also regular although the relation between graphemes and phones is not one-to-one [Seco,76].

According to the conventions established in [Real,74], one aspect of Orthography is the description of the relation between the phonological inventory (phonemes and their distribution) and the writing system of a language.

Also, a goal of Orthography is to set the conventions to represent stress and intonation patterns, at least partially, in the written text.

The alphabet used in the Spanish writing system is based on the Latin-Roman alphabet. The characteres a to z are part of the graphemes employed in the Spanish writing system. In addition to them, there are four more graphemes, the letter ñ (/ñ/) and the diagraphs (letter pairs) ch, ll, and rr. Other symbols used in this writing system are diacritical marks and punctuation marks. The diacritical marks in Spanish are known as orthographic accent (') and dieresis (¨). The orthographic accent can be associated only with the vowels (a, e, i, o, u), and its purpose is to identify where the stress should be applied. The letter u preceded by g and followed by e or i is silent except when u has dieresis [Seco,76]. The inverted question mark and the inverted exclamation mark are used at the begining of interrogative and exclamatory sentences respectively. The other punctuation marks , . : ! and ? are used in much the same way as in English.

According to Navarro [Nava,68], there are 42 phonemes in the Spanish phonological inventory. There are five phonemes corresponding to the vowels which give rise to

twenty allophones. Similarly, the nineteen phonemes that correspond to the consonants give rise to over thirty allophones. The nineteen consonant phonemes are:
/b/, /č/, /d/, /f/, /g/, /j/, /k/, /l/, /ĺ/, /m/, /n/, /ñ/, /p/, /r/, /r̄/, /s/, /t/, /y/, /z/;

In addition to the consonant phonemes and the vowel phonemes there are: six falling diphthongs /ai/, /au/, /ei/, /eu/, /oi/, /ou/; eight rising diphthongs /ia/, /ie/, /io/, /ua/, /ue/, /uo/, /iu/, /ui/; and four triphthongs /iai/, /iei/, /uai/, /uei/.

Although from the phonetic point of view diphthongs and triphthongs can be divided into vowels, semivowels, and semiconsonants, in Phonology they play the same role as single phonemes [Nava,68]. Some researchers disagree with this view [Gili,71].

We noted earlier that for Spanish, the mapping between the set of graphemes and the set of phonemes is many-to-many. For example, the phoneme /b/ is associated with the graphemes b and v; /k/ is associated with the graphemes c, qu, or k. The grapheme h has no associated phoneme. Depending on the context, the grapheme x may be associated with any of the phonemes /k/+/s/, /s/, /j/. For our experiments in text-to-speech, we have used the

letter-to-sound rules shown [Nava,63; Seco,76] below:

U = {a,b,c,ch,...,y,z}, Y∈U, A∈{a,e,i,o,u}

B∈{a,o,u}, C∈{b,c,ch,...,y,z}, E∈{e,i}

D∈U - {e,i}, # is a word boundary

guE --> /g/E, guB --> /gu/B, güY --> /gu/Y

gE --> /j/E, cE --> /z/E, cD --> /k/D

quE --> /k/E, xC --> /s/C, AxA --> A/ks/A

ch --> /c/, ll --> /l̲/, ñ --> /ñ/

#r --> /r̄/, rr --> /r̄/ r --> /r/

b --> /b/, v --> /b/, d --> /d/, f --> /f/,

j --> /j/, k --> /k/, l --> /l/, m --> /m/,

n --> /n/, p --> /p/, s --> /s/, t --> /t/

w --> /g/, y --> /y/, z --> /z/, a --> /a/,

e --> /e/, i --> /i/, o --> /o/, u --> /u/

The necessary information for the placement of suprasegmental features is mostly embedded in the text and in the discourse. A stressed syllable, the prominent syllable within a word, can be predicted for Spanish text by means of the following rules:

R1. Words ending in a vowel sound, /n/, or /s/ tend to be stressed on the next to the last syllable.

R2.Words ending in a consonant sound other than /n/, or /s/ tend to be stressed on the final syllable.

R3.An orthographic accent when present will take precedence over R1 and R2.

In Spanish, stress is phonemic, that is to say the meaning of an utterance can be changed by stress shifting. Therefore, it is very important to place the stress on the right syllable in a word. In many cases, the placement of stress distinguishes between present and past tense (estudio, estudió); between a noun and a verb (termino, terminó); or between adjective and verb (corto, cortó). Similarly, intonation enables a speaker/listener to differentiate among affirmative, interrogative, and exclamative sentences. Pauses are used in speech either for emphasis or for breathing (air intake). In the written text, pauses are indicated by punctuation marks. The syllable is the phonetic unit next in size to the "phonic group" or "breath group". A phonic group is the speech segment between two consecutive pauses meant for breathing. It has been estimated that in Spanish about eight to ten syllables constitute a phonic group [Seco,76; Nava,63; Nava,68]. However, smaller and larger phonic groups are not totally absent.

The purpose of this section of the thesis is to set up

the framework for presenting a text-to-speech synthesis-procedure for Spanish in the following sections. Further comments on the Spanish sound system will be given as they are required.

3.2 Concatenative Units:

For speech synthesis, words, morphs, syllables, demisyllables, diphones, phonemes and microphonemes have been considered as concatenative units by different researchers. In voice response systems that use a limited vocabulary, perhaps a word based approach is adequate. It had been observed that about 200 kilobytes of disc storage would be adequate to store approximately 1000 words, with an average duration of 2/3 seconds per word, when the adaptive DPCM method of coding is used [Rabi,76]. Allen [Alle,76], showed that for unrestricted English text-to-speech synthesis an approach based on morphs is more economical than a system based on words.

One or more syllables together make a word in a spoken language. From the articulatory point of view, the utterance of a syllable is a cyclic process that passes from onset to peak and peak to offset as the vocal tract moves from a more-closed to a more-open to a more-closed configuration [Matt,77]. Syllabication of words is one of

the problems in the synthesis of speech from texts using syllables. A great deal of research based on "syllables" is in progress at Haskins Laboratories [Matt,77], Bell Laboratories [Fuji,78], Bell-Northern Research [Hunt,80], and at other institutions [Bell,78]. It should be noted that the definition of a syllable is not the same for all of these groups.

Today, voice synthesizer units and LSI chips for phoneme based synthesis are commercially available, even for personal computer users. Methods for automatic translation of English text to phonetics by means of letter-to-sound rules have been discussed by Elovitz et al. in [Elov,76], and by McIlroy in [McIl,74]. For example, McIlroy's system contains more than 750 letter-to-sound rules which include 100 words, 580 word fragments, and 70 letters. The system also has a small 100 word exception dictionary. On the other hand, Elovitz's system is driven by a set of 350 letter-to-sound rules. Both systems produce phonetic transcriptions from the input text which are then mapped onto the codes that drive a phoneme based synthesizer unit.

In [Bert,77], Bertinetto et al. use diphones or the joint characteristics of a pair of phonemes, as concatenative units for the synthesis of speech from

written Italian texts. They have used about 150 diphones to generate the required sound in Italian. A favorable factor in the use of diphones is that the coarticulation characteristics at phoneme boundaries are included in the units themselves. The concept of demisyllable introduced by Fujimura [Fuji,78], has a close analogy to that of diphone but at the level of syllables.

For synthesizing speech from Polish texts, Kielczewski used concatenative units smaller than phonemes which he called "microphonemes" [Kiel,78]. According to him, microphonemes are short characteristic segments of audio signals (40 to 60 msec) which after a certain number of repetitions (1 to 20, depending on the type) enable correct perception of the phonemes they represent. He observes that the microphonemes of fricative phonemes are longer and that of voiceless stop consonants are a little shorter. By lengthening the duration of microphonemes, and changing the number of repetitions and the signal amplitude, he is able to produce the effect of stress and intonation.

In the foregoing paragraphs we referred to different concatenative units (CU). The transformations in those CUs when conacatenated have been studied by different researchers. For example, the tranformations of microphonemes when they are concatenated have been reported

in [Kiel,78], that of phonemes are discussed in [Mezz,74], and the problems associated with syllables and syllable boundaries are considered in [Fuji,78].

3.3 Synthesis of Speech Based on Phonemes:

In text-to-speech synthesis based on phonemes, the acoustic output is generated by concatenating the phonemes obtained from the transcription of the text. The transition between adjacent phonemes has considerable influence on the quality of the synthesized speech. In some cases linear interpolation is adequate to concatenate formant frequency contours of adjacent phonemes, but in other cases higher order interpolations are required. According to Flanagan et al [Flan,70], "to synthesize a continuous message, timing, pitch, and formant information must be generated". In a formant synthesizer, it is possible to independently control formant frequencies, pitch, timing, and fundamental frequency. Therefore, it is feasible to control the interpolation of formant frequency contours.

Two examples of formant synthesizers are the Norwegian made OVE-III and the CT-1 which used to be produced by Computalker Consultants. The latter is an inexpensive synthesizer which can be easily interfaced to a

microcomputer. In the CT-1 [Berd,79; Pinn,79] the synthesizer parameters can be updated at a rate of 100 times per second. This rate is adequate to obtain synthetic speech of high quality [Flan,72]. There are nine such synthesizer parameters: voicing source frequency, voicing source amplitude, F1-frequency, F2-frequency, F3-frequency, aspiration noise amplitude, fricative noise amplitude, fricative resonator frequency, and nasal resonator amplitude. Each parameter can be specified as an eight bit number.

One of the widely used terminal-analog synthesizers based on phonemes is VOTRAX [Gagn,78]. Phone commands serve as input to this unit which are operated upon by a series of built-in hardware units to generate acoustic output. Unfortunately, the user has no control over the transitions at phone boundaries. In VOTRAX there are sixty-four 8-bit phone commands, six bits determine the phone to be uttered and two bits indicate one of the four built in inflection levels. The VOTRAX VS-6 unit was optimized to synthesize Mid-Western American English therefore its performance in synthesizing other languages is not as acceptable as in English. Synthesis of Esperanto, Spanish, Italian, Russian, and English using a VOTRAX VS-6 was reported by Sherwood [Sher,78]. Suen [Suen,76] used a VOTRAX VS-6 to synthesize Mandarin.

Experimental systems for the synthesis of speech from text written in Spanish have been considered by Berdichevsky, Murillo, and Cutler [Berd,79], and by Sherwood [Sher,78]. The Compultalker CT-1 synthesizer and the VOTRAX VS-6 have been used respectively in the above experiments. In both cases phonemes have been considered as concatenative units. The objective of Berdichevsky, Murillo, and Cutler was to develop a low cost microcomputer-based reading-aid for the visually handicapped. Initially, they performed a study on acoustic parameters of Spanish speech. Formant frequencies and pitch contour were the acoustic parameters considered for their analyses [Muri,79]. The data so extracted were used as the synthesizer parameters to control the CT-1 synthesizer unit. The synthesized speech was considered to be of good quality [Berd,79].

The procedure described by Sherwood [Sher,78] has been built into the PLATO system which is a computer aided instruction system. In PLATO, synthetic speech from text is used to supplement the other educational materials. Sherwood concludes that the generated speech is of good quality, though it exhibits an English-like accent.

The use of VOTRAX VS-6 for Spanish-speech synthesis has certain drawbacks. First of all, VS-6 vowels phones

are similar to but not the same as the Spanish vowel phones. The Spanish phoneme /r̄/ (alveolar trill) is absent in the VS-6. Substitution of /r̄/ by the VS-6 phone /r/ [Sher,78] will cause confusion between word pairs such as pero (but) and perro (dog), enterar (to inform) and enterrar (to bury), or sentences like esta enterado (he is informed) vs. esta enterrado (he is buried). In other cases VS-6 phones are aspirated but in Spanish they are not. Aspirated phonemes might cause confusion in cases like tomar (to take) vs. domar (to tame). Depending on the application, VOTRAX VS-6 might be practical for synthesizing a language other than English or might present serious limitations.

It is our contention that the following factors are to be considered in the choice of concatenative units (CU) for synthesizing speech from written texts:

a). A definition or interpretation of the concatenative unit at an appropriate level.

b). The necessary modifications and preprocessing of the input text to suit the algorithm (set of rules) under consideration that maps on to the concatenative units.

c). Rules for the transcription of an input text into a sequence of concatenative units.

d). Transformations within CUs and between adjacent CUs when the units are concatenated to produce synthetic speech.

e). The complexities involved in the placement, and the realization of the effects of prosodic features so as to make the synthetic speech natural sounding.

f). Synthesizer units to suit the needs of the CUs.

3.4 The Syllabic Structure of Spanish:

The syllable structure of Spanish has been studied by several linguists such as Navarro [Nava,68], Gili-Gaya [Gili,72], Delattre [Dela,65], Malmberg [Malm,65]. In particular, the syllabic types in Spanish were studied by Navarro [Nava,68]. According to Navarro [Nava,68], there are nine syllabic types in Spanish that are shown below along with their frequencies in a sample of narrative texts:

where C stands for a consonant and V for vocalic nucleus.

1. CV	58.45
2. CVC	27.35
3. V	5.07
4. CCV	4.70
5. VC	3.31
6. CCVC	1.12
7. VCC	0.00
8. CVCC	0.00
9. CCVCC	0.00

Syllabic types

(after Navarro [Nava,68])

A comparative study on syllable frequencies and syllable structures of different languages with that of Spanish was carried out by Delattre [Dela,65]. In the study conducted by Delattre [Dela,65], the four most frequent syllabic types are reported to be:

1. CV	55.6
2. CVC	19.8
3. CCV	10.2
4. VC	3.1

Syllabic types

(after Delattre [Dela,65])

The absence of the syllabic type V in Delattre's analyses

which is the third most frequent type in the case of Navarro, is due to his syllabication criteria.

Spanish is known to be a syllable-timed language [Bros,70] that is a language in which all syllables "tend" to have more or less the same duration during non-emphatic speech. On the other hand, English is known as a stress-timed language, that is a language in which stressed syllables "tend" to be evenly spaced in time [Pike,65; Jone,76]. A syllable in Spanish always has a vowel as its nucleus which may or may not contain additional elements (consonant or semivowel) [Nava,63]. Based on samples taken from narrative text and dramatic text, Delattre [Dela,65] found the most frequent syllables in Spanish to be:

narrative material: /de/, /a/, /e/, /la/, /ba/, /do/, /ke/, /i/, /to/, /ra/

dramatic material: /de/, /te/, /no/, /do/, /ke/, /se/, /ka/, /a/, /to/, /da/

We have selected a sample of 2000 most frequent words in Spanish from [Juil,74] and obtained a syllable-frequency table. The 25 most frequent syllables are given below and the entire table can be found in Appendix I.

/de/, /el/, /a/, /la/, /ke/, /i/, /en/, /es/, /no/ /e/, /to/, /te/, /do/, /ko/, /ra/, /o/, /na/, /mo/, /ta/, /pa/, /kon/, /yo/, /su/, /si/, /por/

The underlined syllables are found also in the list reported by Delattre. Our main objective in this experiment had been to obtain a list of most frequent Spanish syllables that could be useful in the design of a syllable based synthesizer.

The total number of different phonological syllables in Spanish has been estimated at about 5520 by Saporta and Contreras [Sapo,68]. A speech synthesizer for Spanish based on syllables needs to have the facility to produce at least a subset of the most frequent syllables. The parameters of these syllables such as pitch, and formant contours can be extracted a priori and stored in a table. By feeding the stored parameters as input to a formant synthesizer, the corresponding sounds of the syllables can be generated. Such a system can further be supported by a phoneme based synthesis to account for the syllables whose parameters are not stored in the table.

3.5 Speech Synthesis Based on Syllables:

The syllable as the basic concatenative unit for speech synthesis has been considered by Mattingly [Matt,77] at Haskins Laboratories and by Fujimura [Fuji,78] at Bell Laboratories. Researchers at Haskins Laboratories consider that "speech is a code, and that the encoding unit is the phonetic syllable" [Matt,77]. For Mattingly's synthesis-by-rule schema the input is a phonetic transcription of an utterance. He makes use of syllable-features that can take binary values. These values determine a pattern of articulatory influences such as the vowels in adjacent syllables, the final consonants in the previous and current syllables, and the initial consonants of the current and following syllables. A set of parameters for the synthesizer is determined. Using these values he has used a software simulator of the synthesizer OVE-III [Lilj,68], to produce the acoustic output. Mattingly recognizes that the syllable plays a crucial role in stress and intonation [Matt,77].

Besides the findings of Mattingly and other researchers, we are motivated by the following factors in our choice of syllables as the primary concatenative unit for synthesizing Spanish speech:

a).Although there are exceptions [Seco,76], it is possible to identify syllable boundaries in a text.

b).Syllables tend to have more or less the same duration [Gili,71].

c).Syllables being higher order concatenative units than phonemes, for a given text the number of inter-unit boundaries will be small when syllables are used.

d).The syllable is the basic rhythm unit of Spanish [Malm,65].

e).Stress placement on a syllable within a word can be determined based on rules.

f).The formation of intonation contours over a sequence of words is simplified when syllables are considered as basic concatenative units.

g).Coarticulation effects are included in the syllabic unit.

The process by which a word or a breath group is broken down into its contituent syllables is known as syllabication or syllabification. Syllabication of Spanish text is possible through the application of a set of rules, although some exceptions can arise. The rules based on the work of Seco [Seco,76] and Navarro [Nava,63] are presented below: Let C denote a consonant sound, V a vowel sound, "-" a syllable boundary, and the set P be defined as:

$\{C_1C_2 \mid (C_1 \in (/p/, /b/, /f/, /g/, /t/, /k/)$ and $C_2 \in (/r/, /l/))$

OR $(C_1C_2 = /dr/)\}$

S1 The sequence VCV is divided as V-CV

S2 The sequence VC_1C_2V is divided as

a) $V\text{-}C_1C_2V$ if $C_1C_2 \in P$

b) $VC_1\text{-}C_2V$ otherwise

S3 The sequence $VC_1C_2C_3V$ is divided as

a) $VC_1\text{-}C_2C_3V$ if $C_2C_3 \in P$

b) $VC_1C_2\text{-}C_3V$ otherwise

S4 The sequence V_1V_2 is divided as

$V_1\text{-}V_2$ only if either V_1 or V_2 is a stressed /i/ or /u/;

otherwise the V_1V_2 pair is not separated.

S5 Three vowels in sequence are never separated

We use the term "orthographic-phrase" to denote a septence or part of a sentence that is terminated with a

pause while reading. In our case, an orthographic phrase is delimited by punctuation marks. The following procedure, READ-SPEAK, is proposed for synthesis of speech from text using syllables. Its goal is to extract as much information as possible from the orthographic text and then to "speak out the output-buffer contents".

In order to use syllables as concatenative units for speech synthesis from Spanish texts, we conducted some analyses of text and speech. Analyses of Spanish text was done at the word level. For this purpose, the two thousand most frequent Spanish words were syllabicated using the syllabication procedure described above. One of the results of these analyses is that 596 different syllables can generate 2000 words.

Analysis of speech was done at the syllable level. The sixty most frequent syllables were uttered consecutively by two different speakers under "noise-free" conditions and recorded on an audio tape. This time-domain signal was filtered at 4.5KHz and digitized at a 10KHz sampling rate with twelve-bit resolution per sample.

The digital signal of the consecutive syllables were stored on a magnetic disc. Using an interactive graphic terminal, the digital representation of each syllable was

```
PROCEDURE READ-SPEAK

1 Get an orthographic phrase;

2 Get a token from the orthographic phrase;

3 Repeat steps 4 thru 9 till the end of the data is
  reached;

4 Repeat steps 5 thru 7 till the end of the
  orthographic phrase is reached;

5 Case TOKEN of

  WORD :

      Generate phonetic transcription of the word,

      Syllabicate the phonetic word,

      Identify stressed syllables, and

      Place the result onto the output buffer;

  PUNCTUATION MARK :

      Place pause marker in the output buffer;

  EXCLAMATION OR QUESTION :

      Place sentence type marker in the output buffer

  End Case TOKEN of;

6 If there are more than 8 syllables in the phonic
  group then add a pause in the output buffer to
  indicate a breath group;

7 Get the next TOKEN from the orthographic phrase;

8 Speak out the buffer contents.

9 Get the next orthographic phrase;

End PROCEDURE READ-SPEAK
```

segmented and a dictionary of syllables was created.

Splicing of syllables was done "manually" in order to obtain words and short-sentences in digital form. Then, the digital signal of words and short sentences was converted to an analog signal which in turn was recorded on analog tape. The speech so obtained is intelligible but it is monotonous. This is mainly due to the fact that we dealt only with coding/decoding of speech and no attempt was done to modify the fundamental frequency. The objective of this experiment was to obtain an acoustic version of spliced syllables within a limited scope.

Spectrographic analyses of syllables were obtained using a sonograph. Narrow-band and wide-band spectrograms of syllables in isolation were produced. Some spectrograms of continuous speech as well as "casual" speech were also generated. These spectrograms can be useful to determine the interpolation function to apply at syllable boundaries when a formant synthesizer is available.

CHAPTER IV

QUALITY OF SPEECH AND SUPRASEGMENTAL FEATURES

4.1 The Need for the Good Quality Speech in Man - Machine Communication:

Historically, speech has dominated among all the modes of man-to-man communication. Ochsman and Chapanis compared the effectiveness of ten communication modes in problem solving [Ochs,74]. The gist of the comparisons is seen from the reproduction of their bar chart as shown in Figure 4.1. The five modes indicated on the left of this figure which involve a voice channel are significantly faster than the others. Speech as a mode of communication is natural and convenient. Hill [Hill,79] has extensively studied the advantages and disadvantages of this mode of communication and the following is a brief summary of his findings:

a). It leaves the human body free for other activities providing an opportunity for multimodal communication.

b). It is well suited for providing "alert" messages.

c). It is omnidirectional, so does not require a fixed operator position.

d).It is equally effective in the light or the dark, or under many conditions when visual or tactile communication would be impeded. It is well suited for the visually handicapped.

e).It allows limited security checking on the basis of voice characteristics.

f).It can be transmitted over the existing telephone network.

g).It is subject to interference by noise.

h).Random selection from a large inventory of items is difficult.

i).Speech input to machines seems to be expensive and less succesful as it stands today.

With the modern developments in speech synthesis, It is possible to add voice output, generated by the machine, to information utility systems such as the videotext projects. Since the man-machine communication in such cases will be to casual users of a computer system, the voice response should sound natural and it must be clear and less prone to misinterpretation. Another application area that would require the supplement of good quality speech is Computer Aided Instruction (CAI). Here again, the synthetic speech generated by the machine should sound

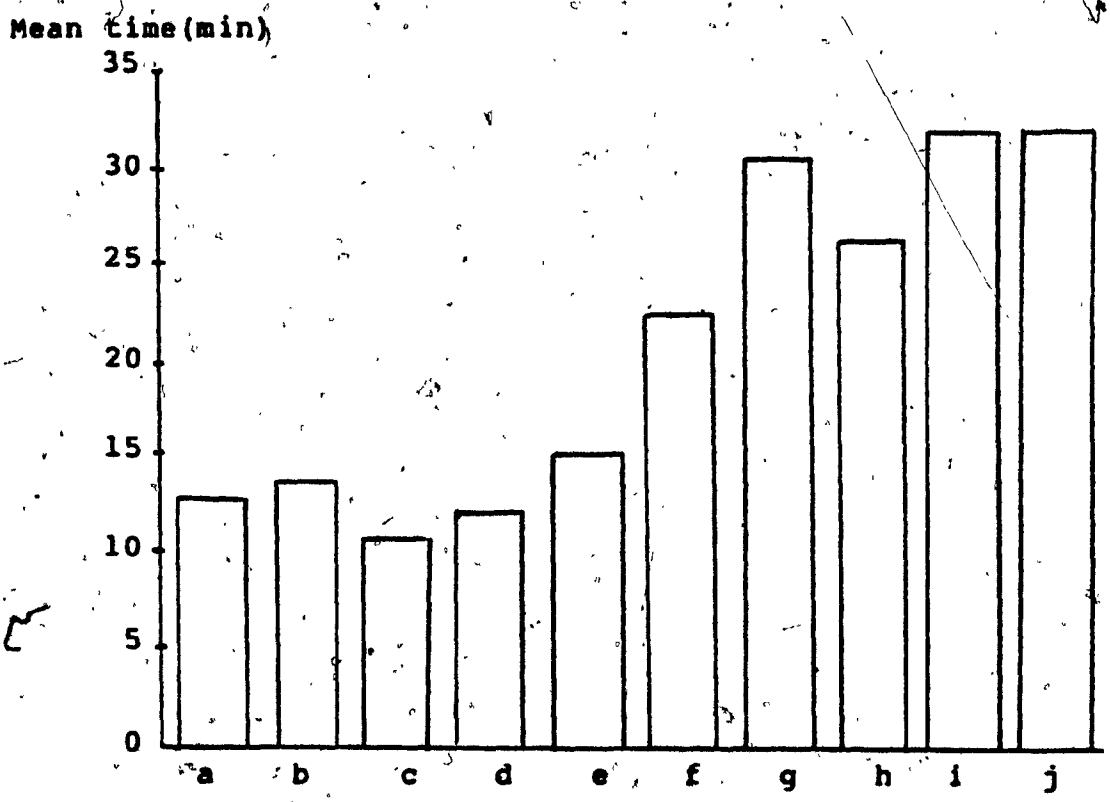


Figure 4.1 Mean-times to Solve Problems in Ten Different Communication Modes (Adapted from [Ochs,74])

a).Communication rich.

b).Voice and video.

c).Voice and handwriting.

d).Voice and typewriting.

e).Voice only.

f).Handwriting and video.

g).Typewriting and video.

h).Handwriting and typewriting.

i).Handwriting only.

j).Typewriting only.

natural and friendly in order to keep the learning spirit of the student as high as possible and allow the student to concentrate on the material being presented.

For the generation of synthetic speech of acceptable quality, we need to consider suprasegmental or prosodic features such as stress, stress patterns, intonation contours, rhythm, and pauses between utterances. Placement and control of the parameters require a good knowledge of the syntax of the language, the vocabulary, and the semantics of the underlying discourse.

4.2 Speech Quality:

There is a wealth of knowledge about the acoustic properties of the speech wave [Flan,72a; Oppe,78; Rabi,78; Jako,63; Fant,73]. However, the relationship between speech quality and acoustic parameters is not very well understood [Flan,72a]. Speech perception is a multistage process. Initially, it is concerned with the processing of acoustic information by humans and its decoding into linguistic units. The acoustic signal is only the starting point for a complex dynamic decoding process [Jako,63]. Since the mechanics of the decoding process is not well understood, speech perception is a complex one [Stev,75].

It is known [Stev,75; Lieb,67] that an individual experiences more difficulty in identifying isolated words than in recognizing the same words in the context of a sentence. Extending this further, it is seen that isolated words are more easily identified than isolated phonemes or syllables. These facts support the hypothesis that, in the speech perception process, decoding of syntactic (and larger) units takes place in parallel with the detection (based on acoustic information) of phonetic features and segments. This also suggests that much of the phonetic information is not extracted from the speech wave, but must be inferred from semantic context and derived from the listener's experience [Stev,75].

In the implementation of voice response systems one of the main constraints is the quality of the speech output. By quality of the speech output we mean speech output which is intelligible and natural sounding. It is possible to obtain synthetic speech which sounds natural but it is not intelligible. Also, synthetic speech can be intelligible but sound unnatural to the human ear. For voice response systems both characteristics must be presented if the system is to be useful.

The evaluation of speech output quality is subjective by nature, because it is the end user who decides whether

or not the response is intelligible and natural sounding. Besides, there are no well-established methods which can be used to quantitatively evaluate the speech quality [Flan,72a]. Most of the methods used to evaluate speech rely on the conventional articulation test. In this test, an individual is asked to write down the equivalent of the utterances he listens to. Very often, the utterances are composed of isolated vowels or isolated syllables. The analysis of the listener's responses is done using statistical methods such as Multi-dimensional Scaling [Krus,64].

In [Shes,79] a comparison of natural and synthetic speech based on a pattern recognition approach was presented. For this experiment, a set of utterances were recorded by a male speaker. The same utterances were synthesized in a VOTRAX synthesizer and recorded. Both speech waves forms were filtered at 5KHz and sampled at 10KHz in order to obtain a digital representation. The digital speech wave was classified into voice and unvoiced speech-sounds. The outcome of the comparison was that the synthetic speech contained more unvoiced sounds than the natural speech.

A schematic representation of the approach for testing the quality of speech is given in Figure 4.2. The term

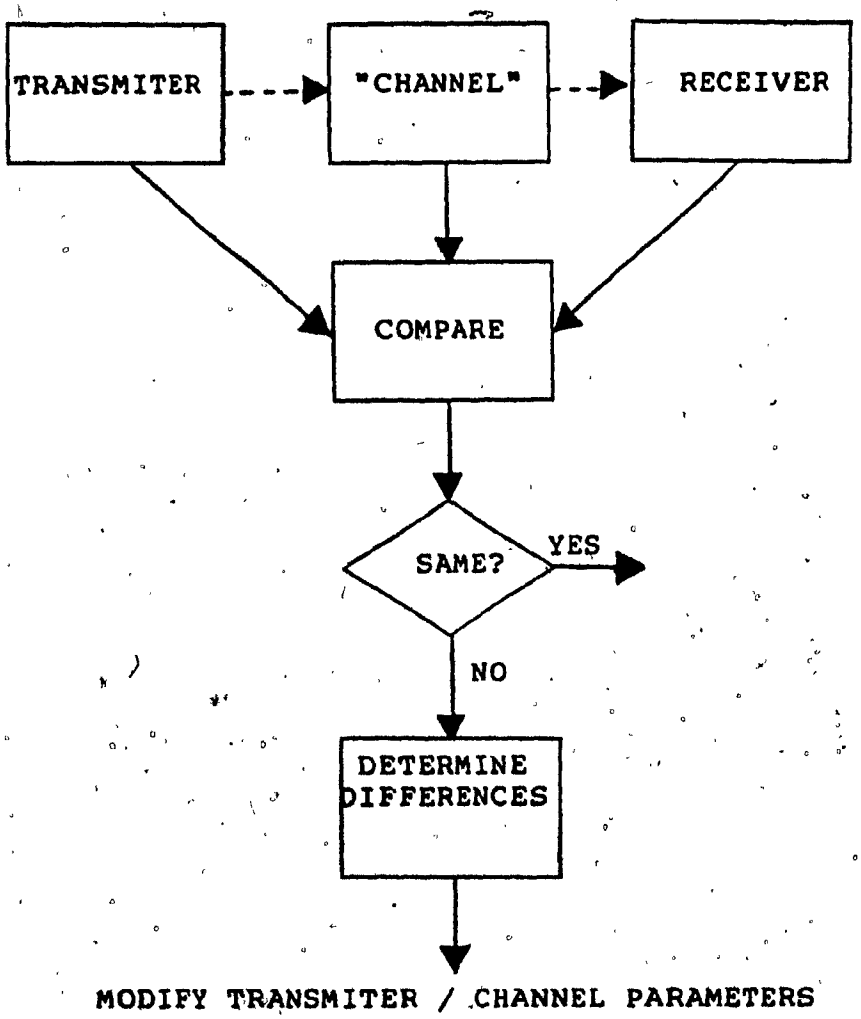


Figure 4.2 Schematic Representation of the Approach for Testing Quality of Speech

channel in this figure might refer either to a physical communication channel or to a transformation process used for speech coding to enable effective storage of the speech.

4.3 Suprasegmental Features of Speech:

The phoneme, the syllable and the breath group or sense group are very often referred to as "segmental features". In normal speech, utterances are constituted by segmental features modified by stress, rhythm, pitch, intonation, inflections, and so forth. That is why the term suprasegmental features is used to denote stress, rhythm, pitch, and intonation. In some cases, a change of pitch or stress implies a different syntactic unit. Some examples of this situation are the words: contract, export, frequent, present, progress, record, subject. In all these words, stress at the beginning of the word denotes a noun/adjective while the stress after the first vowel sound results in a verb [Jone,76]. In Spanish, the placement of stress within a word, contrasts between present and past tense, as in "camino" (I walk) and "caminó" (He walked), between noun and verb, as in "rio" (river) and "rió" (He laughed), or between adjective and verb, as in "corto" (short) and "cortó" (He cut).

In tone languages, like Mandarin, an utterance with a high pitch may have an altogether different meaning from that of the same utterance when uttered with a low pitch. In other languages, like English, variations in pitch do not change the meaning of a particular word. However, in English, a sentence can be a command or a request depending on the intonation of the utterance. Consider the following sentence in Spanish:

/bendras algun dia/ (You will come some day)

The intonation variations and the length of the pauses between utterances can give rise to a great number of different meanings. A rising pitch at the end of the utterance denotes a question. A falling pitch at the end of the utterance indicates a statement. The attributes of speech sounds such as stress, strees patterns, intonation, pause length, or pitch variations are commonly referred to as suprasegmental features.

The prominence of a syllable within a word is known as stress. Such prominence of a syllable can be interpreted as time duration or as a loudness with respect to neighbouring syllables. On the part of the speaker, stress is related to an increased vocal effort or muscular energy. Pitch variations are very often indicative of stressed syllables. In English, stress is used to differentiate between words that have the same spelling but different

meanings as in "desert vs. desert". Also, stress is employed to emphasize words within a sentence as in "He lent you two books". In some other situations, stress serves the purpose of conveying a feeling as in "How was your vacation? It was great!". Besides, stress is a primary element for the determination of the intonation contour for an utterance.

For the generation of natural-sounding synthetic speech, both segmental as well as suprasegmental features are necessary. In general, the use of the suprasegmental features in continous speech is subtly controlled by the speaker. In the written text the suprasegmental features are not evident. Therefore, in a text-to-speech synthesis procedure the suprasegmental features to be employed have to be inferred either from orthographic rules or from the "knowledge" of the discourse. It is our contention that by assuming a limited discourse for a text-to-speech synthesis system, the suprasegmental features can be derived mostly from orthographic rules.

Three intonation levels can be used, for non-emphatic Spanish utterances in the context of a system which deals with affirmative, negative, and interrogative sentences. This is only an approximation to what a native-speaker actually does while speaking. The desired intonation level

for the synthetic voice is specified by a level marker in the "phonetic transcription string" derived from step five in the procedure READ-SPEAK (Section 3.3). For this purpose, a level one marker is inserted in the phonetic transcription to denote the begining of an orthographic phrase. Then, a level two marker is introduced to indicate the ocurrence of the first stressed syllable in the orthographic phrase. The intonation level remains there until the occurrence of the final stressed syllable in the orthographic phrase. For interrogative sentences, a level three marker is inserted to indicate the last stressed syllable in the sentence. In any other type of sentences or orthographic phrase, a level one marker is used to denote the final stressed syllable.

4.4 Acoustical Correlates:

According to Denes [Dene,70], one of the central problems in the experimental research of language is to correlate parameters of linguistic descriptions and the measurement from physical speech events. Further, he notes that all the relevant acoustic features of the speech signal can be accurately measured by the instruments now available. Even more, there is a way to verify the perceptual effects of such measurements, through the use of speech synthesis techniques. However, the acoustical

features that correlate to linguistic features have not been unambiguosly established and it remains as an open problem. In Denes' opinion [Dene,70], a close interaction between linguists and physical scientists is a necessary step towards solving this problem.

The correlation of parameters of linguistic descriptions with the fundamental frequency (F_0) of the speech wave has been the subject of several studies. Intonation is one of the suprasegmental features conveyed through variations of the fundamental frequency. A detailed study of fundamental frequency patterns and their relationship to linguistic features was conducted by O'Shaughnessy [O'Sh,79]. He showed in several experiments that fundamental frequency contours follow particular patterns with respect to sentence type, syntactic construction, and other linguistic features. The F_0 pattern followed by an answer is noticeably influenced by the context on which a question is asked. One of the conclusions given by O'Shaughnessy [O'Sh,79] is that pitch, intonation, and stress correlate with fundamental frequency, amplitude, and duration.

Murillo, Berdichevsky, and Cutler [Muri,79] conducted an experiment for the analysis of the fundamental frequency of Spanish declarative statements. They found that the F_0

variation of each syllable could be approximated by a straight line. From their analyses, we observe that F_0 frequency rises steadly from the begining until the occurrence of the first stressed syllable in the sentence. Then F_0 falls to a lower level which is maintained steadily until the last stressed syllable in the sentence. For the declarative sentence given in [Muri,79], the F_0 contour drops to zero. It is possible that for interrogative sentences the F_0 contour rises instead of falling to zero. Further experimentation would be necessary to determine such factors.

Knowledge of acoustical correlates to linguistic features is also useful for speech recognition purposes. In the study conducted by Lea [Lea,80b], a set of 255 English sentences was spoken by one talker. Sentence type, phrase structure, stress patterns, and phonetic sequences were considered in the design of the sentences. The results of the study are summarized below.

Determination of the acoustical correlates to parameters of linguistic descriptions has been a difficult task. Perhaps as Denes [Dene,70] suggested, new definitions of linguistic descriptions may be necessary.

F_0 pattern	Cases
Rises steadly at the begining of a sentence, until the first stress, where it peaks.	99.0%
Falls, after a peak value in the last stress, to a low value at the end of each declarative, command, and WH-question.	99.5%
Falls, then peaks within the last stress of yes/no questions, and rises throughout subsequent unstresses.	95.0%
Its value on unstressed syllables is lower than on all preceding stresses. Also, the values of the preceding and the following stresses are at or above the value of the unstressed syllable.	91.0%

CHAPTER V

SENTENCE GENERATION AND ITS APLICATION TO DATABASES

5.1 Grammars for Sentence Generation:

The study of language has attracted the interest of man for thousands of years. A Sumerian grammar (2,000 B.C. aprox.) is one of the earliest linguistic works [Jako,72]. Until the nineteenth century the study of the language was oriented towards establishing a supposedly uniform parent language. Several techniques for language analysis have flourished since the end of the nineteenth century. Among the techniques originated in this century are those of Tagmemic Analysis [Pike,76], Stratificational Grammar [Makk,73], Case Grammar [Fill,68], and Transformational Generative Grammar [Chom,65].

Noam Chomsky adopted the view of language acquisition as an innate process [Chom,57]. He considered that in human beings language is "internalized knowledge" represented by a system of rules [Chom,66]. One outgrowth of the work of Chomsky and his colleagues was the Theory of Transformational Grammar, TTG for brevity [Chom,57; Chom,63; Chom,65; Chom,68]. The objective of TTG is to

characterize the abstract grammar which human beings possess.

In TTG, there are two basic dichotomies:

(i). Surface Structure vs. Deep Structure.

and

(ii). Performance vs. Competence.

Surface structure refers to what we actually hear or see in written form and deep structure refers to the representation of the meaning of an uttered or a written sentence. For example

a). Jim wrote the program

b). The program was written by Jim

Both sentences have the same deep structure (semantic interpretation) but different surface structure (syntactic structure). Now consider the sentence:

c). Jim sent the letter.

Comparing sentences (a) and (c), we find that they do not have the same deep structure. However, they have the same surface structure, that is:

noun-1 + verb (tense) + article + noun-2

Competence and performance refer to a subject who is skilled in the language under study. Performance is the way an individual speaks, or writes a language with all of its regularities and irregularities. On the other hand,

the concept of competence refers to a subject's ability to interpret and understand any sentence in the language. The claim of TTG is that all native speakers of a language have essentially the same competence but not necesarily the same performance.

For any grammar of a language, the TTG [Chom,65] has components dealing with:

(i).-syntax,

(ii).-morphophonemics, and

(iii).-semantics.

The syntactic component has two subcomponents which are:

-phrase structure (PS) rules and lexicon.

-transformational (T) rules.

Deep structures are generated by the PS-rules operating with the lexicon. The fundamental gramatical categories of the language and word order are introduced by the PS-rules.

The task of the transformational rules is to operate on the deep structure obtained from the previous subcomponent of the syntactic component. In doing so, the T-rules insert, delete, move and substitute elements in order to produce a surface structure without modifying the deep structure. Once a surface structure is obtained, it forms an input to the morphophonemics component. The phonological and morphological structures are considered by

the morphophonemic rules. The morphophonemic rules map the surface structure into phonetic transcriptions. The third component of TTG, that is the semantic component, maps deep structures into semantic interpretations. Augmented Transition Network (ATN) grammars are another tool for the study of languages that was developed by Woods [Wood,70]. The ATN is one of the widely used representations of deep structures in natural language understanding systems and in question-answering systems. An ATN can be visuallized as a collection of directed graphs with labeled states and labeled arcs, a start state, and a set of final states. In addition, a test and a sequence of "actions" are defined on each arc. For an arc to be traversed, the test must be satisfied and then the "actions" are executed. The label on an arc can be a nonterminal symbol, thus allowing recursion. Woods proved that ATN grammars have the generative power of a Turing machine [Wood,70].

5.2 Speech Output from Database Systems:

Natural language communication between man and computer is desirable in many database (DB) applications. It is often the case that a decision maker needs access to data managed by a database system (DBS). However, in order to interact with the DBS, he has to resort to the assistance of a person who is skilled in the details of the

computer system and the database system. This situation is undesirable, particularly when dealing with sensitive data. A natural language interface would enable the decision maker to interact directly with the DBS. A database system designed for non-expert users, as in banking systems for example, is accessed by casual users. In this situation, the user does not want to or can not learn a query language in order to make use of the database (such as to know his/her balance in his/her bank account).

In this general scope, a natural language interface between man and computer is a complex task. However, under the assumption of a particular application, a good degree of success can be achieved. For a particular application, the discourse will be limited, which reduces the vocabulary and the number of possible syntactic forms. Even more, some semantic interpretation might be feasible. From the point of view of the user there is no loss of generality and from the point of view of the implementor the complexity of the problem is reduced. Languages so designed are commonly referred to as "natural" language [Bens,79], or natural language interfaces.

"Natural" languages used in a computer system can be considered as a particular cases of formal languages [Bens,79]. In the design and implementation of "natural"

language systems, several aspects of the theory of formal languages are useful and relevant. "Natural" language systems can be broadly classified into two categories:

- "Natural" language understanding (NLU) systems.

- "Natural" language generating (NLG) systems.

The task of a NLU system is, for a given user's query, to derive the "semantics" of it and then to code it in a suitable form for the input to a DBS. The reverse process is achieved by a NLG system. The NLG produces a syntactic structure in the "natural" language, from a "semantic" representation. Generally, the interactive NLU systems include some kind of a response generation or a NLG system for the dialogue with a user.

The user's query and the system's response can be presented in several modes, as in textual or acoustic forms. As stated earlier, the NLU systems operate in a limited context. In TORUS [Mylo,75], the domain of the discourse is information about graduate students in the Computer Science Department of the University of Toronto. The user's query and the system's response are in textual form. In LADDER [Hend,78] (Language access to distributed data with error recovery), the domain of discourse is

restricted to a database meant for the manufacturing and maintenance of US navy ships. The user's query as well as the response from the system are in a textual form.

Our present interest is in sentence generation as per a natural language and its presentation in acoustic form. There are at least two ways to obtain the acoustic output from the NLG component: In Figure 5.2 the block might generate the response in the form of control parameters which could directly drive a synthesizer unit. In the other approach, the NLG component generates texts which are input to a text-to-speech component. In turn, the text-to-speech component produces the control parameters for a synthesizer unit. We consider the second approach suitable for our objectives because of the smaller binding among the diverse components. In addition, each component can be developed almost independently and we could benefit from the earlier researches in NLG.

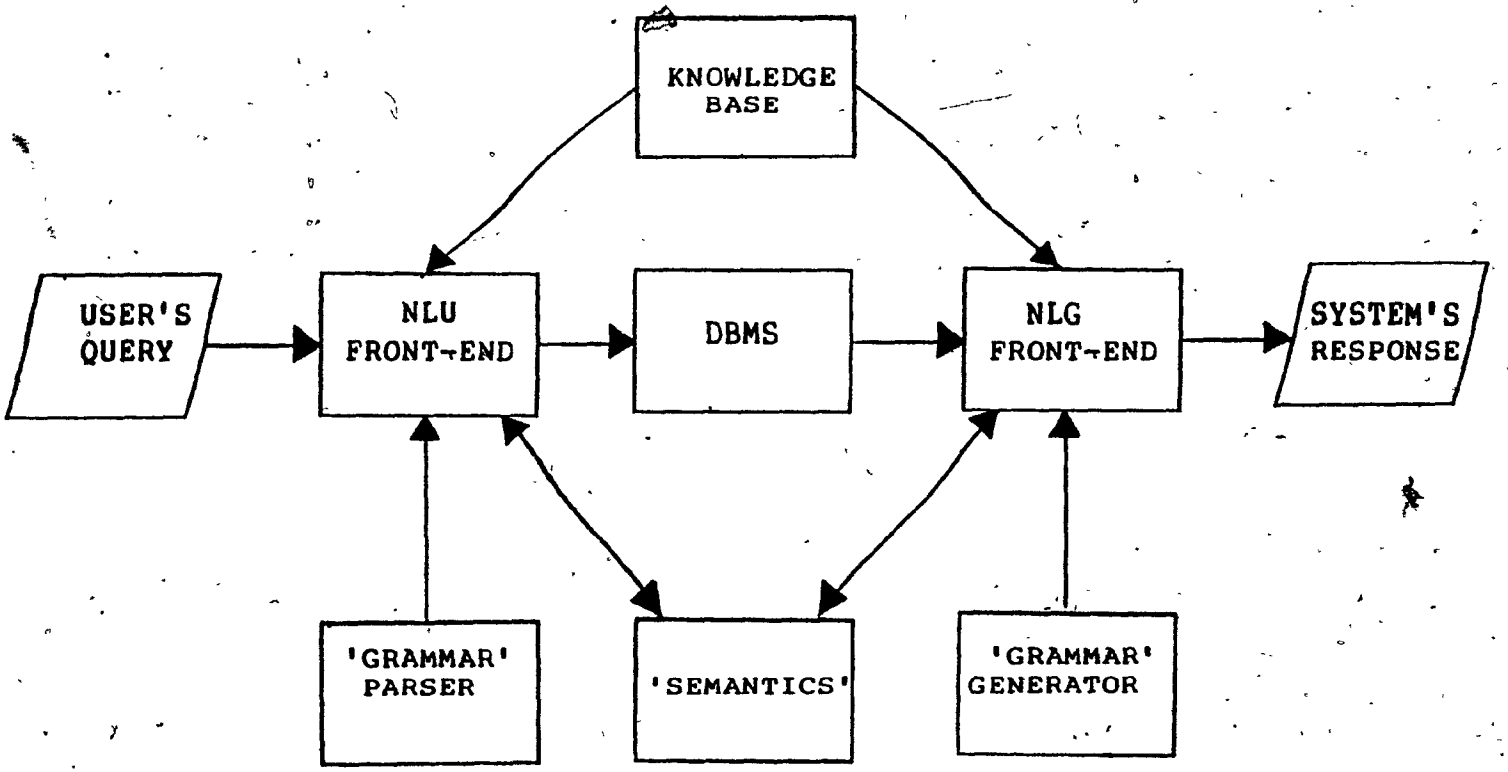


Figure 5.1 Natural Language as an Interface in DBMS

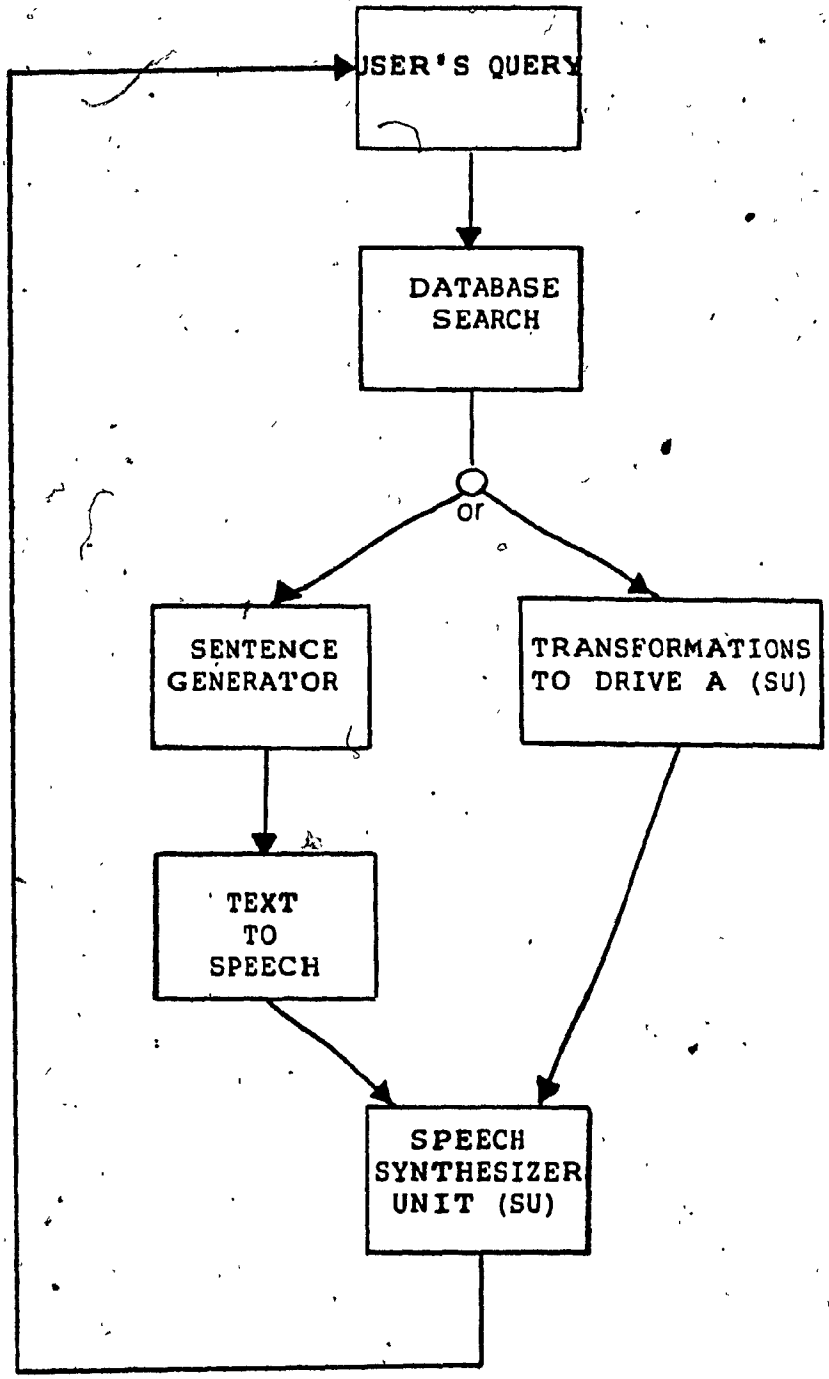


Figure 5.2 DBMS Speech response

5.3 A Case Study :

In this section we make use of an example database which provides information about student registration. The various stages of the case study are shown in Figure 5.3. The central interest is on the sentence generation module and the text-to-speech module. Therefore, suitable assumptions are made on the presentation of the user's query, and the selectors for the sentence generator to make the case study simple. The user's query could be expressed in a SEQUEL-like language. The problem of generating the values for selectors which control the sentences generated are discussed at length in [Mylo,75; Wong,75]. In TORUS, a question-answering component which creates a semantic network provides a set of arguments to the selector. Then, the selector copies part of the semantic network into a database and maps the resulting graph onto a "more surface representation" [Wong,75].

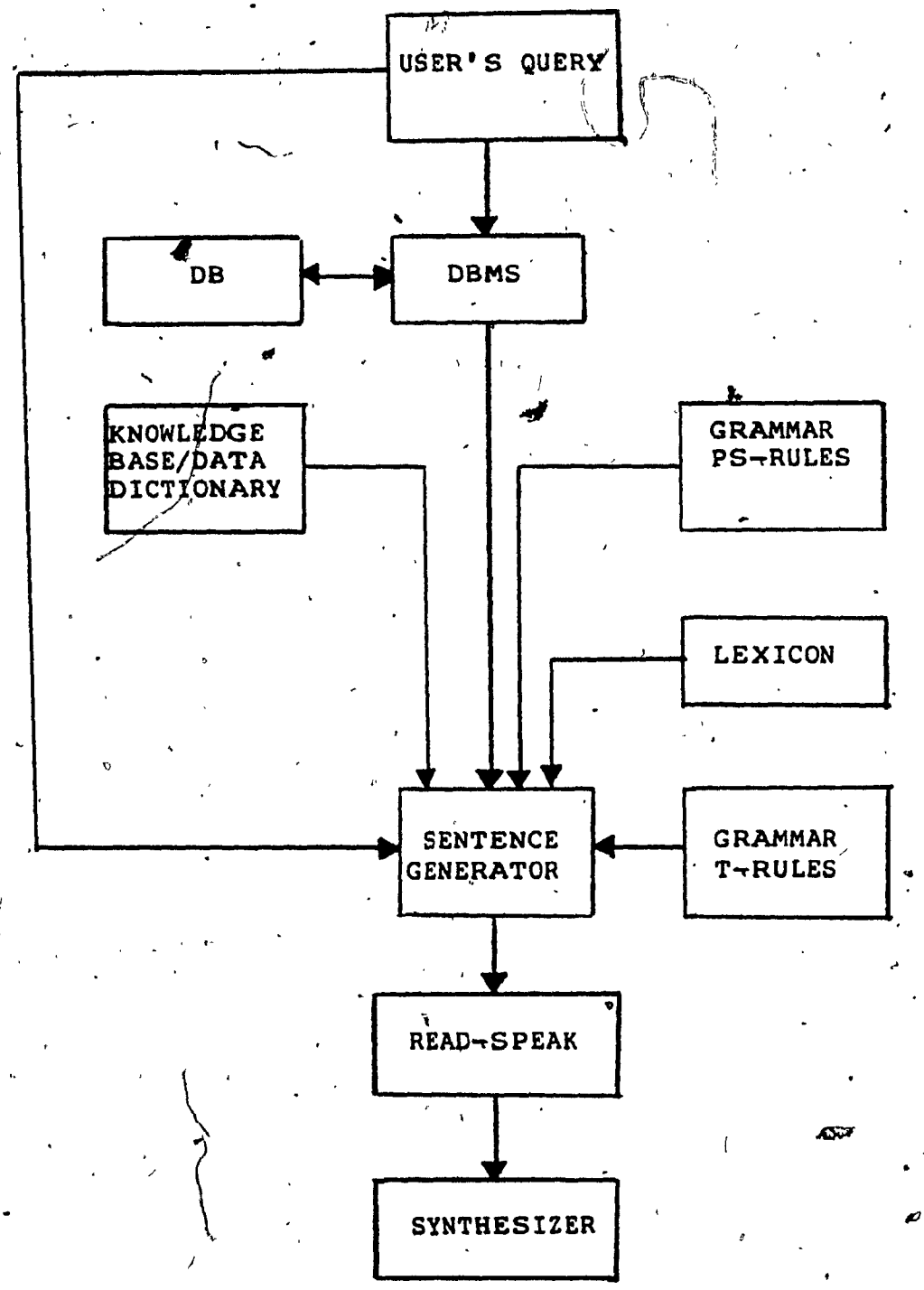


Figure 5.3 Prototype Database with Speech-Output.

The internal representation of a deep structure and the translation into surface structure are the primary characteristics of a sentence generation system. For this purpose, Bates [Bate,78], suggest the use of an ATN grammar along with a lexicon. The different stages of our case of study are as despicted in Fig.5.4, it involves the following:

- A set of parameters that determines the syntactic units needed in the ultimate sentence output.
- A phrase structure grammar which controls the production process.
- A set of T-rules or transformational rules for mapping the intermediate sentential forms into the surface structures or sentences.

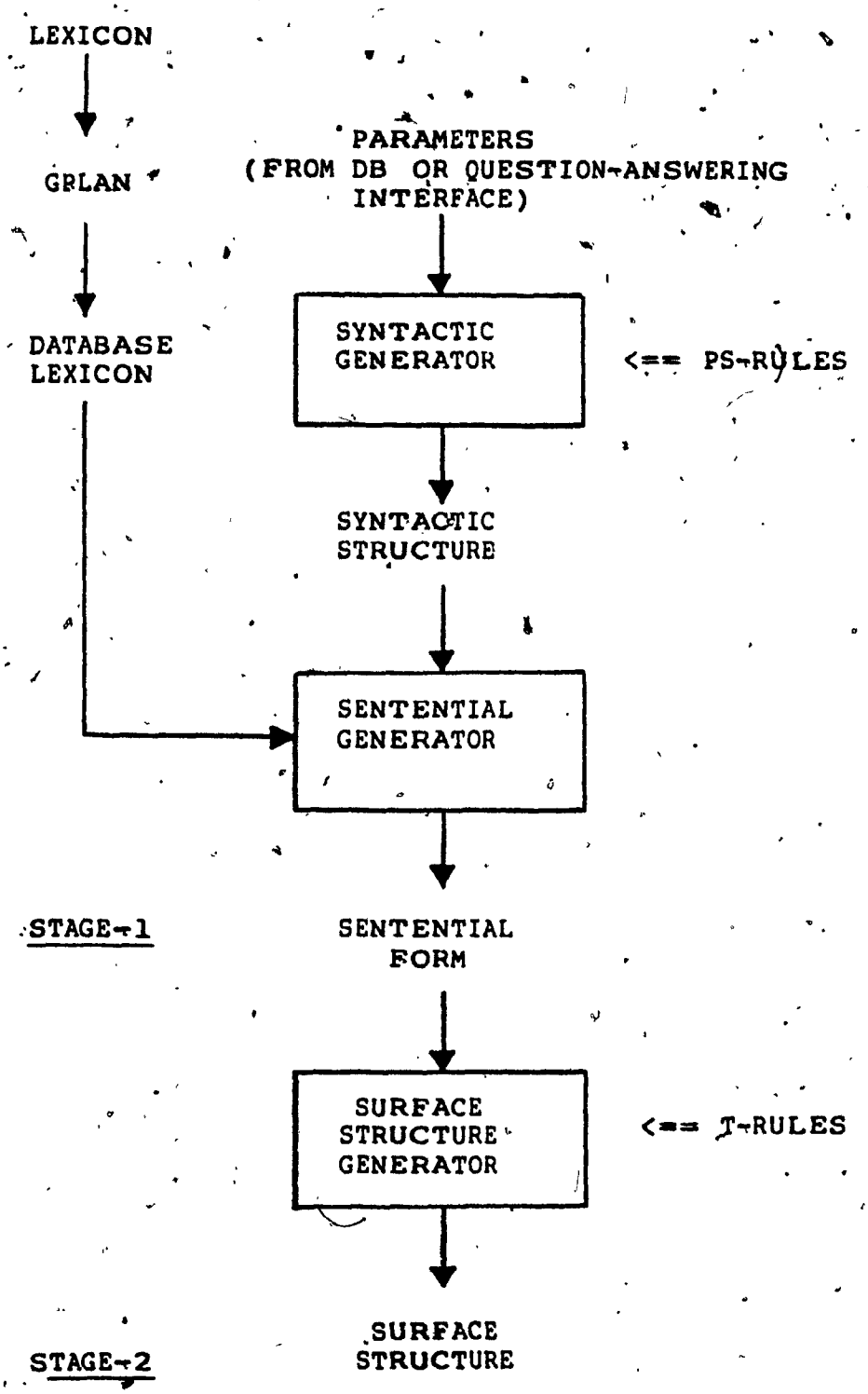


Figure 5.4 Sentence Generator Block Diagram

For the production of "sentential forms", the terminal symbols are specified by a set of parameters. A database lexicon is used to verify the compatibility among terminal symbols and in some cases to provide additional terminal symbols. In the database lexicon, the relations among terminal symbols are represented as a network model database [Date,77]. The GPLAN database management system is used for the generation and access of the database lexicon. GPLAN is a set of user-callable FORTRAN subroutines, that supports a network model of data. The database lexicon includes

- nouns

  curso (course), alumno (student), profesor (professor)

- verbs

  aprobar (to pass), impartir (to teach), reprobar (to fail), inscribir (to register), dar (to offer)

- auxiliary verbs

  haber, ser, estar

- prepositions

  por (by), en (in)

In this case study, the sentence generator has a close analogy to the syntactic component of a transformational grammar. In fact the phrase structure rules as well as the transformational rules here employed are based on the

subset of the Transformational Grammar proposed by R.L.Hadlich [Hadl,71] for Spanish. The PS-rules considered are:

```
        S ------> (neg) (Q) NP VP
       NP ------> (det) N (pl)
       VP ------> aux verbal
      aux ------> asp't
       't ------> [+past]
      asp ------> [+perfv] | [+subs]
   verbal ------> V (NP (passv))
    passv ------> SER passive | SE passive
```

Where:

S	= sentence	VP	= verb phrase
neg	= negative	aux	= auxiliary
Q	= question marker	asp	= aspect
NP	= noun phrase	t	= tense
det	= determiner	passv	= passive (two modes)
N	= noun	perfv	= perfective
pl	= plural	subs	= subsequent

Transformational rules are rewrite rules which map trees onto trees without changing the deep structure. Any T-rule has three basic parts, namely structural description (SD), condition (C), and structural change (SC). In order to apply a T-rule the condition must be satisfied and the

SD must match the deep structure of the sentence. The structural change (SD) specifies the changes to be made in the sentence. As an illustration on how a T-rule operates, the following example is given,

Consider the rules "SER passive" and "affix shift":

Rule "SER passive"

Structural description (SD):

NP_1 X V NP_2 passv

Condition (C) for the rule to be applied: none

Structural change (SC):

NP_1 X V NP_2 passv ==> NP_2 X ser -d- V por NP_1

Rule "affix shift"

Structural description (SD):

affix V

Condition (C) for the rule to be applied:

affix = -r, ndo, -do, -d-, or aspect-tense-person-number

Structural change (SC):

affix V ==> V affix

and the sentential form:

JIMENEZ [+perfv] [+past] IMPARTIR EL CURSO 352 passv

(Jimenez [+perfv] [+past] teach the course 352 passv)

The application of the rule SER PASSIVE would produce the the structural change:

EL CURSO 352 [+perfv] [+past] SER -D- IMPARTIR POR JIMENEZ

In turn, the application of the rule AFFIX SHIFT would give the following structure:

EL CURSO 352 [+perfv] [+past] SER IMPARTIR -D- POR JIMENEZ

At this stage, it is only necessary to obtain the past tense and the aspect verb modifications:

EL CURSO 352 FUE IMPARTIDO POR JIMENEZ

(The course was taught by Jimenez)

Finally this sentence will be the output from the sentence generator. The text-to-speech module transforms a surface structure into a phonetic transcription. The procedure READ-SPEAK (see sec. 2.3 and sec. 4.3) takes as its input the surface-structure sentences and generates a phonetic transcription by denoting phonemes, syllable boundaries, and stressed syllables. Intonation level markers are also inserted in the phonetic transcription. These markers can contribute to determine the appropriate intonation contour in the synthesized speech (Fig. 5.5). For the above example, the output from READ-SPEAK would be:

(1)EL# (2) KUR -SO #352# FUE # IM-PAR- TI (1)-DO # POR // JI- (2)- ME (1)-NEZ

Where we have used the following notation:

stressed syllable	‾
intonation level marker	(1), (2), or (3)
word boundary	#
syllable boundary	-
pause marker	//

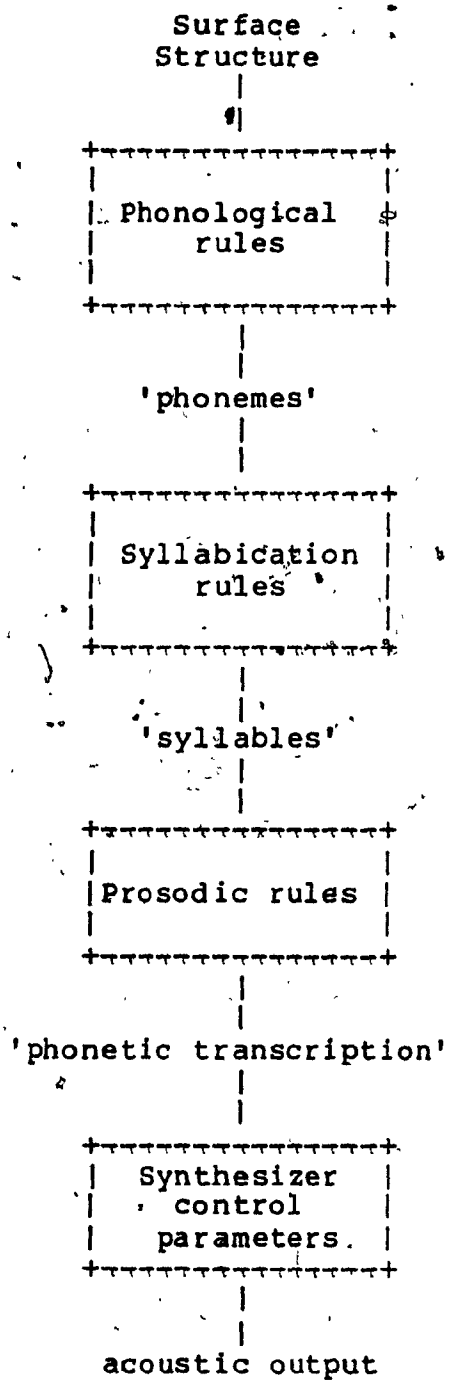


Figure 5.5 Text-to-Speech Generator

Relational schema of the sample database is given below.

STUD-BIO: <ID#,name,tel#,sex,age> Bio-data

STUD-FIN: <ID#,COURSE#,grade,year> Courses finished by a student

CORS-REQ: <ID#,COURSE#> Courses desired for registration

PRE-REQ: <COURSE#,PRE-REQ#> Prerequisites for a course

CORS-OFER: <COURSE#,SECTION#, time, prof#, available-places>

REGI-STU: <COURSE#,SECTION#,ID#> Registered students

PROF-NAME <PROF#,name> Professor name

COURSE-TITLE <COURSE#,title> Course title

Sample queries and responses from the sentence generator and the text-to-speech synthesis module are shown below.

Query: ¿Que curso aprobo AGUIRRE en 1980?.

(SEQUEL-type statement select course#
from STUD-FIN
where (grade<>'F'
and name=AGUIRRE
and year=80))

From query: curso aprobar AGUIRRE [+past]
N V N t
(course pass AGUIRRE past)

From database: 231 B student name

From data-dictionary: grade=B ==> aprobar
(pass).

PS-rules:

S ==> NP_1 [+perfv] [+past] V NP_2
AGUIRRE [+perfv] [+past] APROBAR CURSO 231

T-rules:

(verb agreement, tense and person)
AGUIRRE APROBO EL CURSO 231.

READ-SPEAK:

'syllables'

A- GI -RRE # A-PRO- BO # EL # CUR -SO # 231

'phonetic transcription'

(1)A- (2) GI -RRE # A-PRO- BO # EL # KUR (1)-SO # 231

Query: ¿Quien imparte el curso 241?

(SEQUEL-type statement: select prof-name
from CORS-OFER
where course#='241')

From database: prof-name = NAVARRO

From query: profesor impartir curso [-past]
N V N t
professor teach course past

PS-rules S ==> N [-perfv] [-past] det N
NAVARRO [-perfv] [-past] IMPARTIR EL CURSO 241

T-rules
(verb agreement, tense and person)
NAVARRO IMPARTE EL CURSO 241

READ-SPEAK:

'syllables'

NA- BA -RRO # IM- PAR -TE # EL # KUR -SÓ # 241

'phonetic transcription'

(1)NA- (2) BA -RRO # IM- PAR -TE # EL # KUR (1)-SO # 241

CHAPTER VI

CONCLUSIONS AND SUGGESTIONS FOR FUTURE RESEARCH

The goal of this thesis has been the speech output from data bases for machine to man communication. Two sequential stages in achieving this goal have been: (a) Couching the data base responses into sentences; (b) Transforming the sentences into speech by means of text-to-speech rules. In achieving the goal of this thesis, results from diverse fields of research are useful and necessary. Natural language generation, knowledge representation in the context of databases, text-to-speech translation, and development of speech synthesizer units are indispensable areas for our purposes. As many interdisciplinary fields are involved, the contributions of this thesis are centered in its breadth rather than its depth.

In our examples, we consider Spanish as the language for machine-to-man communication. From our early experiments with VOTRAX VS-6, a phoneme based synthesizer for synthesizing Spanish speech, we found the limitations in this approach. Keeping these limitations and the characteristics of Spanish in view, it was proposed to use

syllables as the main concatenative unit for speech synthesis. For this purpose, a syllabication algorithm is evolved that accepts Spanish texts and divides them into syllables. This algorithm has been tested on the 2000 most frequent Spanish words. Since no formant synthesizer is available to us at this time, neither in hardware nor in software, we have not generated the acoustic output from the syllabicated words. However, digital representations of the most common syllables are generated and stored on magnetic disks for future research.

In order to generate non-mechanical and natural sounding speech, we need to use prosodics. In this thesis we have considered stress, intonation countours, and pauses for this purpose. The Spanish orthography is helpful in determinining which syllables should be stressed. We have used a notational scheme for writing the phonetic transcription of a text that uses stress markers, intonation level markers, and pause markers. Further research is required in correlating the prosodic features to acoustical parameters of speech such as the fundamental frequency, signal amplitude, pause, duration, etc. Work in this direction for English has been done by O'Shaugnessy [O'Sh,79].

For the sentence generation part of the thesis, we

have used the results of other researchers. In particular, heavy use was made of the Transformational Grammar for Spanish proposed by Hadlich [Hadl,71]. An example database on student registration is used to describe the mechanics of sentence generation. Using a small vocabulary and a subset of the Transformational Grammar for Spanish, we are able to generate sample sentences. To simplify the programming task, which otherwise would require many man-years, we assumed the selectors to be available. Such selectors control the selection of terminals and non-terminals for the generation of sentences. The chosen subset of the Transformational Grammar of Spanish is represented as a network and the selectors are used for navigation in the network. The outcome of such a navigation is an intermediate sentence which is then transformed by a set of transformational rules to produce grammatically correct Spanish sentences. For the purposes of creating the network database and for navigation, we have used the GPLAN database management system which supports a network model of data.

The example sentences produced in the above manner are syllabicated by means of the algorithm described in the earlier part of the thesis and the phonetic trancriptions are also given. This was done essentially for the sake of completeness.

Based on the experience gained from the work done for this thesis, further research and developmental work are recomended in the following three major areas:

a). Development of a formant synthesizer, perhaps a combined hardware and software system, that would facilitate syllable-based synthesis of speech. Such systems are available in software in some of the speech research centers such as those at MIT and BNR/INRS at Montreal. The LSI chips meant for signal processing, as the Intel 2902, are useful for this purpose. If flexible speech synthesizers were available, syllables could be examined as a concatenative unit for speech synthesis, and smoothing of inter-syllable boundaries could be studied.

b). Development of a software system is recommended to produce responses from a database in the form of natural language sentences. The dialogue generation part of the natural language understanding systems, such as the LIFER system developed at Stanford Research Institute, is capable of achieving this goal. However, devising of the algorithms for the selectors that control the sentence generation is an interesting research problem.

c). As pointed out by the well-known researchers in this field, such as J. Allen, considerable research is required in understanding the speech process and

speech perception, finding the acoustical correlates of the prosodics, and examining the linguistic aspects and acoustical cues of speech communication.

R E F E R E N C E S

[Ains,74] Ainsworth,W.A. "Performance of a speech synthesis program". Int. J. Man-Machine Studies, Vol.6, pp.493-511, 1974.

[Ains,73] Ainsworth,W.A. "A system for converting English text into speech". IEEE Trans. Audio and Electronics. Vol.AU-21, No.3, pp.288-290, June 1973.

[Alle,77] Allen,J. "Synthesis of speech from unrestricted text". in Linguistic Structures Processing. Zampolli,A.(Ed.), pp.1-30, Amsterdam: North Holland, 1977.

[Alle,76] Allen,J. "Synthesis of speech from unrestricted text". Proc.IEEE. Vol.64, No.4, pp.433-442, April 1976.

[Alle,68] Allen, J. "Machine-to-man communication by speech. Part II: synthesis of prosodic features of speech by rule". Proc. AFIPS Spring J. Computer Conf. Vol.32, pp.339-344, 1968.

[Ball,80] Ball,A.J.S., Bochmann,G.V., and Gecsi,J. "Videotext networks". Computer. Vol.13, No.12, pp.8-14, October 1980.

[Bate,78] Bates,M. "The theory and practice of augmented transition network grammars". in Natural Language Communication with Computers. Bolc,L.(Ed.), pp.191-260, Heildelberg: Springer, 1978.

[Bell,78] Bell,A. and Hooper,J.B. (Eds.) Syllables and Segments. New York: North Holland, 1978.

[Bens,79] Benson,D.B. "Formal languages vis-a-vis 'natural' languages". in Computers in Language Research by Sedelow,W.A. and Sedelow,S.Y.(Eds.). pp.98-164, The Hague: Mouton, 1979.

[Bert,77] Bertinetto,P.M., Miotti,C., Sandri,S., and Vivalda,E. "An interactive synthesis system for the detection of Italian prosodic rules". CSELT Rapporti tecnici, Vol.v, No.5, pp.325-331, Dec.1977.

[Bros,70] Brosnaham,L.F. and Malmberg B. Introduction to Phonetics. Cambridge: W.Heffer, 1970.

[Chaf,76] Chafcouloff,M. Vingt Cinq Annes de Recherches en Synthese de la Parole. Paris: CNRS, 1976.

[Chom,68] Chomsky,N. and Halle,M. The Sound Pattern of English. New York: Harper and Row, 1968.

[Chom,66] Chomsky,N. Cartesian Linguistics: A Chapter in the History of Rationalist Thought. New York: Harper, 1966.

[Chom,65] Chomsky,N. Aspects of the Theory of Syntax. Cambridge: The MIT Press, 1965.

[Chom,63] Chomsky,N. and Miller,G.A. "Introduction to the formal analysis of natural languages". in Handbook of Mathematical Psychology, by Bush,R.R., Galanter,E.H., and Luce,R.D. Vol.2, pp.269-321. New York: Wiley, 1963.

[Chom,57] Chomsky,N. Syntactic Structures. The Hague: Mouton, 1957.

[Cohe,52] Cohen,A. The Phonemes of English. The Hague: Mouton, 1952.

[Coke,76] Coker,C.H. "A model of articulatory dynamics and control". Proc. IEEE, Vol.64, No.4, pp.452-459, April 1976.

[Coke,73] Coker,C.H., Umeda,N., Browman,C.P. "Automatic synthesis from ordinary English text", IEEE Trans. on Audio and Electroacoustics, Vol.AU-21, No.3, pp.293-297, June 1973.

[Coke,67] Coker,C.H. "Synthesis by rule from articulatory parameters". Proc. Conf. Speech Communication and Processing, Paper A9, pp.52-53, 1967.

[d'Ag,79] d'Agapeyeff,A. "Developments in microelectronics and the potential for business". Paper to Special Seminar Series for Directors of Anglo - Overseas Transport Ltd. March, 1979.

[Date,77] Date,C.J. An Introduction to Database Systems. Addison-Wesley, 1977.

[Dela,65] Delattre,P. Comparing the Phonetic Features of English, French, German, and Spanish. Heidelberg: Julius Groos Verlag, 1965.

[Dene,70] Denes,P.B. "The use of speech analysis and synthesis in speech training". in Prosodic Feature Analysis. Leon,P.R., Faure,G. and Rigault,A.(Eds.), pp.192-201, Ottawa: Libraire Didier, 1970.

[Dewe,71] Dewey,G. English Spelling: Roadblock to reading. New York: Columbia University, 1971.

[Dudl,39] Dudley,H., Riesz,R.R., and Watkins,S.S.A., "A synthetic speaker". J. Franklin Institute, Vol.227, pp.739-764, June 1939. reprinted in Speech Synthesis, Flanagan,J.L., and Rabiner,L.R.(Eds.), pp.190-215, Pennsylvania: Dowden, 1973.

[Elov,76] Elovitz,H.S., Johnson,R.W., McHugh,A., and Shore,J.E. "Automatic Translation of English text to Phonetics by means of letter-to-sound rules". Naval Research Laboratory Report 7948, Washington, January 1976.

[Fall,78] Fallside,F. and Young,S. "Speech output from a computer-controlled water-supply network". Proc. IEE, Vol.125, No.2, pp.157-161, February 1978.

[Fant,73] Fant,G. Speech Sounds and Features. Cambridge: MIT Press, 1973.

[Fant,68] Fant,G. "Analysis and synthesis of speech processes". in Manual of Phonetics, (Ed.) B. Malmberg, pp.173-277, Amsterdam: North Holland, 1968.

[Fant,60] Fant,G. Acoustic Theory of Speech Production. The Hague: Mouton, 1960.

[Fant,59] Fant,G. "The acoustics of speech". Proc. Third Int. Congr. Acoustics pp.188-201, 1959.

[Fill,68] Fillmore,C.J. "The case for case". in Universal in Linguistic Theory, Bach,E., and Harms,R.T.(Eds.), pp.1-88, New York: Holt, 1968.

[Flan,76] Flanagan,J.L. "Computers that talk and listens: man machine communication by voice". Proc. IEEE, Vol.64, No.4, pp.405-415, April 1976.

[Flan,72a] Flanagan,J.L. Speech Analysis Synthesis and Perception. 2nd ed. New York: Springer Verlag, 1972.

[Flan,72b] Flanagan,J.L. "Voices of Men and Machine". J.Acoustical Soc. Am., Vol.51, No.5, Part I, pp.1375-1387, 1972.

[Flan,70] Flanagan,J.L., Coker,C.H., Rabiner,L.R. Schafer,R.W., and Umeda,N. "Synthetic voices for computers". IEEE Spectrum, Vol.7, No.10, pp.22-45, October 1970.

[Frie,69] Friedman,J. "A computer system for transformational grammar". Comm. of the ACM, Vol.12, No.6, pp.341-348, 1969.

[Fry ,79] Fry,D.B. The Physics of Speech. Cambridge: University Press 1979.

[Fry ,76] Fry,D.B. Acoustic Phonetics: A Course of Basic Readings. Cambridge: University Press, 1976.

[Fuji,78] Fujimura,O., and Lovins,J. "Syllables as concatenative phonetic units". in Syllables and Segments, Bell,A., and Hooper,J.B.(Eds.), pp.107-140, Amsterdam: North-Holland, 1978.

[Gagn,78] Gagnon,R.T. "VOTRAX real-time hardware for phoneme synthesis of speech". IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. pp.175-176, 1978.

[Gili,71] Gili-Gaya,S., Elementos de Fonética General. Madrid: Editorial Gredos, 1971.

[Gold,75] Goldman,N. "Conceptual Generation". in Conceptual Information Processing. Schank,R.C.(Ed.), pp.289-358, Amsterdam: North-Holland, 1975.

[Gree,66] Greenberg,J. (Ed.) Universals of Language. Cambridge: MIT Press, 1966.

[Hadl,71] Hadlich,R.L. A Transformational Grammar of Spanish. New Jersey: Prentice-Hall, 1971.

[Hase,77] Haseman,W.D., and Whinston,A.B. Introduction to Data Management. Illinois: Irvin, pp.392-410, 1977.

[Hend,78] Hendrix,G.G., Sacerdoti,E.D., Segalowicz,D., and Slocum,J. "Developing a natural language interface to complex data". ACM Trans. Database Systems, Vol.3, No.2, pp.105-147, June 1978.

[Hill,79] Hill,D.R. "Using speech to communicate with machines". Man/Computer Communication, Vol.2, pp.193-221, England: Infotech International, 1979.

[Holm,64] Holmes,J.N., Mattingly,G.I., and Shearme,J.N. "Speech synthesis by rule". Language and Speech. Vol.7, pp.127-143, 1964.

[Hunt,80] Hunt,M.J., Lennig,M., and Mermelstein,P. "Experiments in syllable-based recognition of continuous speech". IEEE Int. Conf. Acoustics, Speech, and Signal Processing. pp.880-883, 1980.

[Itur,74] Iturriaga,R. "Demostración formal de algunas propiedades de un algoritmo para silabear palabras en Español". Comunicaciones Técnicas. CIMAS-UNAM, México. Serie B, Vol.5, No.77, 1974.

[Jako,72] Jakobson,R. "Verbal Communication". Communication. pp.39-44, San Francisco: Freeman, 1972

[Jako,63] Jakobson,R., Fant,G., and Halle,M. Preliminaries to Speech Analysis: The Distinctive Feature and Their Correlates. Cambridge: MIT Press, 1963.

[Jone,76] Jones,D. The Phoneme, its Nature and Use. Cambridge: University Press, 1976.

[Juil,74] Juilland,A. and Chang-Rodriguez,E. Frequency Dictionary of Spanish Words. The Hague: Mouton, 1964.

[Kapl,81] Kaplan,G. "If they could talk(and some do), chips might ask: Is there a market?". The Institute. Vol.5, No.5, pp.6, May, 1981.

[Kell,61] Kelly,J.L. and Gerstman,L.J. "An artificial talker driven from a phonetic input". Journal of The Acoustical Society of America. Vol.33, pp.835, 1961.

[Kers,76] Kerschberg,L., Ozkarahan,E.A., and Pacheco, J.E.S. "A synthetic English query language for a relational associative processor". U. of Toronto, Tech.Report CSRG-68, April, 1976.

[Kiel,78] Kielczenski,G. "Digital synthesis of speech and its prosodic features by means of a microphonemic method". in Speech Communication with Computers. L.Bolc.(Ed.) London: Macmillan, pp.183-206, 1978.

[Klei,70] Klein,S., and Kuppin,M.A. "An interactive heuristic program for learning transformational grammars". Computer Studies in the Humanities and Verbal Behavior. Vol.3, No.3, pp.144-162, 1970.

[Krus,64] Kruskal,J. "Nonmetric multidimensional scaling". Psychometrika. Vol.29, pp.115-129, 1964.

[Lara,74] Lara,F.L. and Garcia Hidalgo,I. El Uso de la Computadora Electrónica en la Elaboración del Diccionario Español de México(DEM). Cuadernos de Trabajo del INAM. México, 1974.

[Lea,80a] Lea,W.A.(Ed.) Trends in Speech Recognition. New Jersey: Prentice-Hall, 1980.

[Lea,80b] Lea,W.A. "Prosodic aids to speech recognition". in Trends in Speech Recognition. Lea,W.A.(Ed.), pp.166-205, New Jersey: Prentice-Hall, 1980.

[Lema,79] Lemaitre,C.,Archundia,E., and Outon,A. "Un sistema de consulta en Español de una base de datos". Departamento de Matemáticas. Fac. Ciencias, UNAM, México, 1979.

[Lenn,67] Lennenberg,E. The Biological Foundations of Language. New York: Wiley, 1967.

[Lieb,67] Lieberman,P. Intonation, Perception and Language. Cambridge: MIT Press, 1967.

[Lilj,68] Liljencrants,J. "The OVE-III speech synthesizer". IEEE Trans. Audio-Electronics. Vol.AU-16, No.1, pp.137-140, 1968.

[Lind,79] Lindblon,B. and Ohman,S. Frontiers of Speech Communication Research. London: Academic Press, 1979.

[Makk,73] Makkai,A. and Lockwood,D.G. Reading in Stratificational Linguistics. U. of Alabama Press, 1973.

[Malm,68] Malmberg,B. "The linguistic basis of phonetics". in Manual of Phonetics. (Ed.) Malmberg,B., pp.1-16, Amsterdam: North-Holland, 1968.

[Malm,65] Malmberg,B. Estudios de Fonética Hispánica. Madrid: CSIC, 1965.

[Malm,63] Malmberg,B. Phonetics. New York: Dover, 1963.

[Malm,55] Malmberg,B. "The phonetic basis for syllable division". Studia Linguistica 9, pp.80-87, 1955.

[Mang,78] Mangold,H. and Stall,D.S. "Principles of text controlled speech synthesis with special application to German". in Speech Communication with Computers. (Ed.) Bolc,L. pp.139-182, London: Macmillan, 1978.

[Mark,76] Markel,J.D. and Gray,A.H. Linear Prediction of Speech. New York: Springer-Verlag, 1976.

[Matt,77] Mattingly,I.G. "Syllable synthesis". Haskins Laboratories. Status Report on Speech Research, SR-49, 1977.

[McDo,75] McDonald,D. "Preliminary report on a program for generating natural language". Advance Papers of the 4th Joint Conference on Artificial Intelligence. Tbilisi, USSR, pp.401-405, September 1975.

[McIl,74] McIlroy,M.D. "Synthetic English speech by rule". Bell Telephone Laboratories. March, 1978.

[Mezz,74] Mezzalama,M., and Rusconi,E. "A general system for synthesizing speech". in Speech Communication.Seminar,Stockholm.Vol.2, Speech Production and Synthesis by Rule. Fant,G.(Ed.) New York: Wiley, 1974.

[Muri,79] Murillo,G.,Bedichevsky,F., and Culter,C. "Analysis of formant and pitch information for Spanish phonemes". IEEE Int. Conf. Acoustics, Speech and Signal Processing. pp.914-916, 1979.

[Mylo,75] Mylopoulos,J., Borgida,A., Cohen,P., Rossopoulus,N., Tsotsos,J., and Wong,H. "TORUS.A natural language understanding system for data management". Advance Papers of The 4th. Joint Conf. on Artificial Intelligence. Tbilis USSR, pp.414-421, Septembre, 1975

[Nava,68] Navarro,T. Studies in Spanish Phonology. Florida: U of Miami Press, 1968.

[Nava,63] Navarro,T. Manual de Pronunciación Española. Madrid: Publicaciones de la Revista de Filología Española, 1963.

[Ochs,74] Ochsman,R.B., and Chapanis, A. "The effects of 10 communication modes on the behaviour of teams during co-operative problems solving". Int. J. Man-Machine Studies, Vol.6, pp.579-619, 1974.

[Oliv,74] Olive,J.P. "Speech synthesis by rule". Speech Communication. Seminar, Stockholm, Vol.2 Speech Production and Synthesis by Rule. Fant, G.(Ed.). New York: Wiley,1974.

[Oppe,78] Oppenheim,A.V. "Digital processing of speech". in Application of Digital Signal Processing". Oppenheim,A.V.(Ed.), Prentice-Hall, 1978.

[O'Sh,79] O'Shaughnessy,D. "Linguistic features in fundamental frequency patterns". Journal of Phonetics, Vol.7, pp.119-145, 1979.

[Pike,76] Pike,K.L., and Brend,R.M. Tagmemics. The Hague: Mouton, 1976.

[Pike,65] Pike,K.L. The Intonation of American English. Ann Arbor: U. of Michigan Press, 1965.

[Pinn,79] Pinnell,J.P. "Speech synthesis in real-time by microprocessor control". Master Thesis, Dept. Electrical Engineering. Mc.Gill University, 1979.

[Pott,50] Potter,R.K. and Steinberg,J.C. "Towards the specification of speech". JASA, Vol.22, No.6, pp.807-820, November 1950.

[Pulg,70] Pulgram,E. Syllable, Word, Nexus, Cursus. The Hague: Mouton, 1970.

[Quil,69] Quillian,M.R. "The teachable language comprehender". Comm. ACM Vol.12, No.8, pp.459-476, August, 1969.

[Rabi,78] Rabiner,R.L., and Schafer,R.W. Digital Processing of Speech Signals. New Jersey: Prentice-Hall, 1978.

[Rabi,76] Rabiner,R.L.,Schafer,R.W. "Digital techniques for computer voice response. Implementations and applications". Proc.IEEE. Vol.64, No.4, pp.416-433, April, 1976.

[Rabi,68] Rabiner,R.L. "Digital-formant synthesizer for speech synthesis". J.A.S.A., Vol.43, pp.822-828, 1968.

[Real,74] Real Academia Española. Esbozo de una Nueva Gramática de la Lengua Española. Madrid: Espasa-Calpe, 1974.

[Rode,79] Rodet,X. and Delatre,J-L. "Time-domain speech synthesis-by-rule using a flexible and fast signal management system". IEEE Int. Conf. Acoustics, Speech, and Signal Processing. pp.895-898, 1979.

[Ruhl,76] Ruhlen,M. "A guide to the languages of the world". Language Universals Project. Stanford University, 1976.

[Sapo,62] Saporta,S. and Contreras,H. A Phonological Grammar of Spanish. Seattle: U. of Washington Press, 1962.

[Seco,76] Seco,R. Manual de Gramática Española. Madrid: Aguilar, 1976.

[Sher,78] Sherwood,B.A. "Fast text-to-speech algorithms for Esperanto, Spanish, Italian, Russian, and English". Int. J. Man-Machine Studies, Vol.10, pp.669-692, 1978.

[Shes,79] Sheshadri,S. and Waldron,M.B. "A pattern recognition approach to compare natural and synthesized speech". IEEE Int. Conf Acoustics, Speech, and Signal Processing. pp.777-780, 1979.

[Simm,72] Simmon,R.F. and Slocum,J. "Generating English discourse from semantic networks". Comm. ACM, Vol.15, No.10, pp.891-905, 1972.

[Simp,20] Simplified Spelling Board. Handbook of Simplified Spelling, 1920.

[Sone,68] Sonesson,B. "The functional anatomy of the speech organs". in Manual of Phonetics, (Ed.) Malmberg,B., pp.45-75, Amsterdam: North Holland, 1978.

[Stev,75] Stevens,K.N. "Speech perception". in The Nervous System. Vol.3:Human Communication and Its Disorders. Tower,D.B.(Ed.). pp.163-171, New York: Raven, 1975.

[Suen,76] Suen,C.Y. "Computer synthesis of Mandarin". IEEE Int. Conf. Acoustics, Speech and Signal Processing. pp.698-700, 1976.

[Supp,79] Suppes,P. "Current trends in computer-assisted instruction". Advances in Computers. Vol.18 (Ed.) Yovits,M.C., pp.173-230, 1979.

[Umed,76] Umeda,N. "Linguistic rules for text-to-speech synthesis". Proc. IEEE. Vol.64, No.4, pp.443-451, April 1976.

[Viva,79] Vivalda,E., Sandri,S., and Miotti,C. "Real-time text processing for Italian speech processing". IEEE Int. Conf. on Acoustics, Speech and Signal Processing. pp.880-883, April 1979.

[Walt,78] Waltz,D.L. "An English language question answering system". Comm. ACM, Vol.21, No.7, pp.526-534, July 1978.

[Wigg,78] Wiggins,R. and Brantingham,L. "Three-chips system synthesizes human speech". Electronics. Vol.51, pp.109-116, August 31, 1978.

[Wijx,66] Wijx,A. Rules of Pronunciation for the English Language. London: Oxford Univ. Press, 1966.

[Wino,73] Winograd,T. "A procedural model of language understanding". Computers Models of Thought and Language. Schank,R.C., and Colby, K.M.(Eds.) pp.153-186. San Francisco: Freeman, 1973

[Witt,77] Witten,I.H., and Madams,P.H.C. "The telephone enquire service a man-machine system using synthetic speech". Int. J. Man-Machine Studies. Vol.9, No.4, pp.449-464, July, 1977.

[Wong,75] Wong,H.K.T. "Generating English sentences from semantic structures". U. of Toronto, Dept. Computer Science, Technical Report No.84. August, 1975.

[Wood,77] Woods,W.A. "Lunar rocks in natural English: explorations in natural language question answering". in Linguistic Structures Processing. Zampolli,A.(ed.) Amsterdam: North-Holland, 1977.

[Wood,70] Woods,W.A. "Transition network grammars for natural language analysis". Comm. ACM. Vol.13, No.10, pp.591-606, 1970.

[Zamp,77] Zampolli,A.(Ed.) Linguistic Structures Processing. Amsterdam: North-Holland, 1977.

A P P E N D I X I

TOTAL NUMBER OF SYLLABLES 596

/DE/	/EL/	/A/	/LA/
/KE/	/I/	/EN/	/ES/
/NO/	/E/	/TO/	/TE/
/DO/	/KO/	/RA/	/O/
/NA/	/MO/	/TA/	/PA/
/KON/	/YO/	/SU/	/SI/
/POR/	/SO/	/U/	/KA/
/SE/	/UN/	/MA/	/LO/
/RO/	/TU/	/PO/	/SA/
/TI/	/DA/	/LLA/	/PE/
/TRO/	/ZION/	/MI/	/BA/
/BI/	/MAS/	/BRE/	/LLO/
/DI/	/NI/	/LI/	/ME/
/RRE/	/GO/	/NE/	/DAD/
/BE/	/ZI/	/AN/	/TAN/
/ZIA/	/BER/	/CHO/	/I/
/MU/	/JO/	/AL/	/ZE/
/ñO/	/BO/	/MEN/	/YA/
/JE/	/TED/	/US/	/IN/
/RI/	/PRE/	/BIEN/	/PUES/
/TRA/	/LLE/	/DES/	/PRO/
/KUAN/	/DON/	/FI/	/TIE/

/RIO/	/RE/	/SIN/	/LE/
/RRA/	/ZA/	/GUN/	/MIS/
/ZIEN/	/OM/	/DOS/	/KU/
/KI/	/SER/	/FUE/	/BRA/
/ZIO/	/NUES/	/AS/	/AY/
/MOS/	/TRE/	/PRI/	/GRAN/
/BLE/	/MER/	/SON/	/MUY/
/MIEN/	/DIO/	/KEL/	/LU/
/TAM/	/TAL/	/PI/	/[illegible]/
/PAR/	/BEZ/	/SEN/	/TOR/
/IS/	/EKS/	/GU/	/PU/
/SION/	/TEN/	/PER/	/JER/
/JA/	/TER/	/NUE/	/AR/
/ZER/	/ñA/	/KIE/	/BUE/
/ñOR/	/RAL/	/GA/	/GRA/
/TAR/	/BEN/	/RIA/	/AUN/
/SIEM/	/KIEN/	/KUAL/	/DU/
/AK/	/DEN/	/ñOL/	/MUN/
/KLA/	/JEN/	/TIEM/	/MIL/
/DRE/	/FOR/	/TRES/	/DIS/
/BLI/	/BIO/	/PIO/	/OR/
/KRE/	/KRI/	/ER/	/NER/
/ZES/	/BIE/	/MAR/	/LOR/
/CHA/	/BLO/	/RRI/	/GAR/
/NU/	/FA/	/CHE/	/BRO/
/LAR/	/KUL/	/ZON/	/LEN/

/KUEN/	/PUE/	/GLO/	/RRO/
/IM/	/NIN/	/BIS/	/AU/
/ZO/	/RAN/	/KOM/	/ZIE/
/TES/	/SIS/	/DIOS/	/ZIU/
/SAR/	/ZIN/	/NAL/	/JI/
/ZIAL/	/FUER/	/KUA/	/RES/
/LEK/	/RIOR/	/TAD/	/BES/
/BIR/	/NOR/	/OY/	/BAR/
/NOS/	/PIE/	/KUER/	/PUN/
/UL/	/OB/	/LUZ/	/KAM/
/JOR/	/DAR/	/PEN/	/ZIER/
/TON/	/MOR/	/KAR/	/INS/
/PRIN/	/NUN/	/DIA/	/GE/
/FEK/	/DIE/	/BIER/	/TRAS/
/TUD/	/SAN/	/BAN/	/PLE/
/NOM/	/ZEN/	/YOR/	/LUE/
/GUA/	/AM/	/PEK/	/FIN/
/NIO/	/SUE/	/TOY/	/FUN/
/EM/	/SEIS/	/TUAL/	/NEN/
/GRU/	/PLA/	/LLI/	/RRES/
/BLA/	/JAR/	/TIN/	/BAS/
/RIR/	/MAL/	/PUER/	/BIA/
/GUS/	/REN/	/NIS/	/KUR/
/TIS/	/MAN/	/FRAN/	/FAL/
/MUER/	/JUN/	/SOY/	/RAK/
/REK/	/PAL/	/LIA/	/ZIL/

/GIEN/	/BLAN/	/BOL/	/JEM/
/DOK/	/GUO/	/JION/	/IR/
/BOR/	/SIA/	/SOL/	/MES/
/TRAN/	/TRAR/	/GRO/	/GRE/
/GUAL/	/LEY/	/KOR/	/BEIN/
/KONS/	/SIE/	/DOR/	/NION/
/BOY/	/AI/	/PRON/	/RIEN/
/KAN/	/PLO/	/GRI/	/TUA/
/KUAR/	/DUK/	/KOS/	/LIS/
/TEK/	/FO/	/NES/	/BLAR/
/PAZ/	/TRUK/	/BOZ/	/FON/
/KUE/	/SIM/	/DAN/	/FER/
/FREN/	/LAN/	/RREY/	/DRA/
/JU/	/KAU/	/BIL/	/BOS/
/RIE/	/SIEN/	/TAS/	/BRI/
/KIER/	/SOM/	/KRIP/	/KRIS/
/ZEP/	/ZIR/	/TEL/	/OK/
/RON/	/SUL/	/TREIN/	/GLES/
/PLAN/	/FRA/	/JOS/	/JUS/
/PEL/	/JIO/	/ON/	/RREI/
/ZUL/	/DER/	/RRIEN/	/YER/
/TRIA/	/CHI/	/TRI/	/LON/
/BIN/	/FLOR/	/LIR/	/TIA/
/DIEZ/	/KLU/	/KUES/	/TION/
/LLON/	/TIO/	/KIN/	/MON/
/PIN/	/KAL/	/TEM/	/LIZ/

/GLE/	/TIR/	/TRIS/	/PRAK/
/LIO/	/SUN/	/ET/	/AB/
/GI/	/IZ/	/BIK/	/NAR/
/FAN/	/KES/	/KRIA/	/KROS/
/SUA/	/RER/	/GIR/	/SOR/
/BUS/	/KAS/	/LUN/	/SAL/
/NAS/	/ZIS/	/PRUE/	/SUER/
/TRU/	/TUM/	/FIES/	/PLI/
/MIE/	/TEKS/	/GRIE/	/OS/
/SUR/	/KUI/	/RIAL/	/ñE/
/BAL/	/DIN/	/FLUEN/	/CHEN/
/FES/	/AD/	/DIG/	/FIR/
/GUAR/	/SAK/	/DIEN/	/RROR/
/BU/	/DRO/	/FU/	/JUI/
/TIL/	/ñIA/	/FIE/	/DUL/
/GAS/	/MIA/	/RRUI/	/JUE/
/MAG/	/RED/	/TRAL/	/YEN/
/DIAR/	/BUES/	/JES/	/MIR/
/RAR/	/RROS/	/TREN/	/EU/
/FEN/	/FRE/	/JEL/	/PUL/
/PLIO/	/PREN/	/SUBS/	/KRO/
/PIEL/	/BUEL/	/GOL/	/PIEN/
/DUO/	/LUM/	/NUA/	/PAN/
/PLU/	/MAI/	/DUS/	/FUEN/
/MIN/	/NIR/	/TRIUN/	/GRIS/
/JAN/	/KUNS/	/LUS/	/FRON/

/LLAN/	/FAK/	/FRI/	/GAN/
/BAI/	/PAI/	/PLAR/	/BON/
/FIAN/	/GLA/	/SEK/	/TAK/
/YEK/	/BLIO/	/FRU/	/LIM/
/LUD/	/PIER/	/BIU/	/DAS/
/PRESN/	/ñAN/	/DIAN/	/FAS/
/NIA/	/NIE/	/DUAL/	/IE/
/ZEL/	/LAS/	/LLER/	/GES/
/JUEZ/	/KIS/	/KLI/	/LLU/
/RREU/	/DRAL/	/GRAL/	/NUS/
/ROI/	/RRIO/	/SIO/	/SUAL/
/TIZ/	/UE/	/DIK/	/LLAZ/
/MUE/	/PON/	/DRI/	/FIER/
/LLAR/	/LOJ/	/POS/	/BUL/
/FRES/	/RRAS/	/DEU/	/DRON/
/FIEL/	/MIEM/	/RRED/	/RRIN/
/RROZ/	/SEKS/	/YUN/	/ZED/
/AIS/	/DIAL/	/FRAI/	/GOS/
/KRUZ/	/NIEN/	/NUO/	/PAS/
/PLEN/	/POL/	/SEP/	/YU/
/BEL/	/DUE/	/DUM/	/FRIO/
/MIO/	/NEL/	/BRIL/	/FIL/
/LIE/	/MUL/	/PRAR/	/ZAR/