COM-5 | 1119

MILITARY TECHNICAL COLLEGE
CAIRO - EGYPT

METHODOLOGY OF DEVELOPMENT OF AN ARABIC-BASED SPEECH PROCESSOR

Brig.Gen.Dr.Eng.N.M.El-Nadi *

ABSTRACT :The paper presents a methodology of developing a Text-to-Speech Arabic synthesizer based on the linear independent speech production model shown in Fig.(1) Relevent phonetical properties of Arabic are covered less the ambiguities ever associated in most references dealing with Arabic Phonology /1/,/2/,/3/,/4/,/5/&/6/.

Such ambiguities have always posed difficult problems that delayed the production of quality synthetic Arabic speech. To the best of the authors knowlege no Arabic voice processor of quality speech exists.

The paper describes how to apply state-of -the art techniques of digital signal processing for the estimation of the parameters of the speech synthesizer model for the production of Arabic Utterances.Numerous sonograms(time-frequency diagrams)for Arabic utterances are presented,which illustrate the relevence of the methodological preprocessing neede . The author presents a proposed model for text-to-speech synthesizer model favouring the concatenation of VCV (Vowel+Consonant+Vowel) segments rather than phonemes or allophones.This model will include the transitions from a vowel to a consonant and from a consonant to a vowel in a natural form and promises an improvement in performance.

* The author is curreatly the Chief of Chair of EW in the M.T.C.

I.Introduction

Since the introduction of the Texas Instruments Speak & Spell electronic spelling aid to the market in 1978,major semiconductor manufacturers have ever competed to produce "talking" products . Different types of voice processors are introduced , some share the basic configuration, and the others may utilize altogether different concept for synthesis. All types of speech processors are still besed on the linear independent speech production model shown in Fig (I).

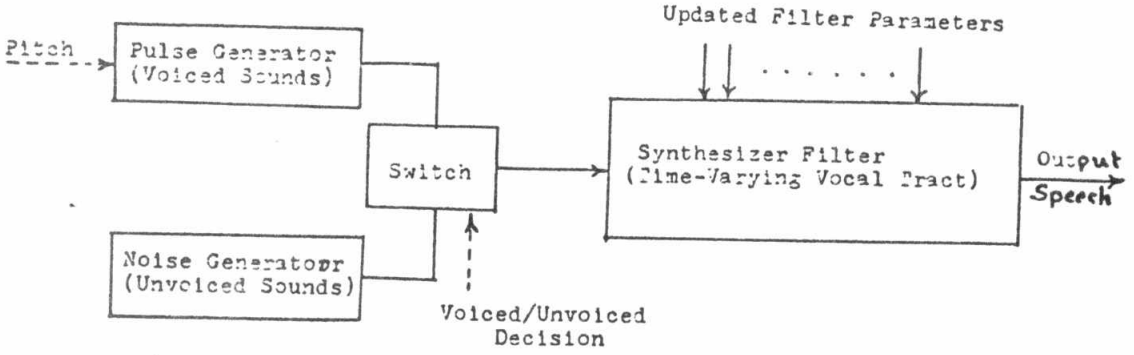


Fig.(1) Model for linear independent speech production mechanism

Techniques of digital speech processing are being adapted and algorithms developed to utilize the technical features offered by these different processors.

2.Arabic Phonology

Consonants,Vowels,Length,and Velarization:

Fig,(2) illustrates a diagram of the vocal apparatus and the points of articulation for the different Arabic consonants.Awareness of these points give correct pronunciation necessary for obtaining correct raw data for analysis . We all know how much time is needed to study the "letter outlets" as a premium before gaining proficiency in the articulation of the "Quraa'n"

Fig(2)&Fig(3)illustrate different Arabic consonants and vowels (together with their relevent phonemes and allophones).Aquisition of typical sound utterances has been performed in the lab. and Fig(4) illustrates the important remark that the relative duration (length)of each sound produces a variation in meaning. The sonograms developed show that in the case of a "stop";the explosion comes after a long withholding;in the case of a "flap"; the flaps are repeated ; in the case of a "nasal";the vibration of the vocal chords and the flow of breath through the nasal passage last longer.

One important note and feature of Arabic concerns the so called "Velarization".A velarized sound is one that is produced with the tongue flattened and grooved from the mid back,i.e.,at a point contiguous to the Velum.The velarized consonants of the Arabic Alphabet are :

(S): ص , (D): ض , (T) ط ,(B) : ظ ,(L) as in " الله "
&(R) as in " رَحْ "

COM-5 1121

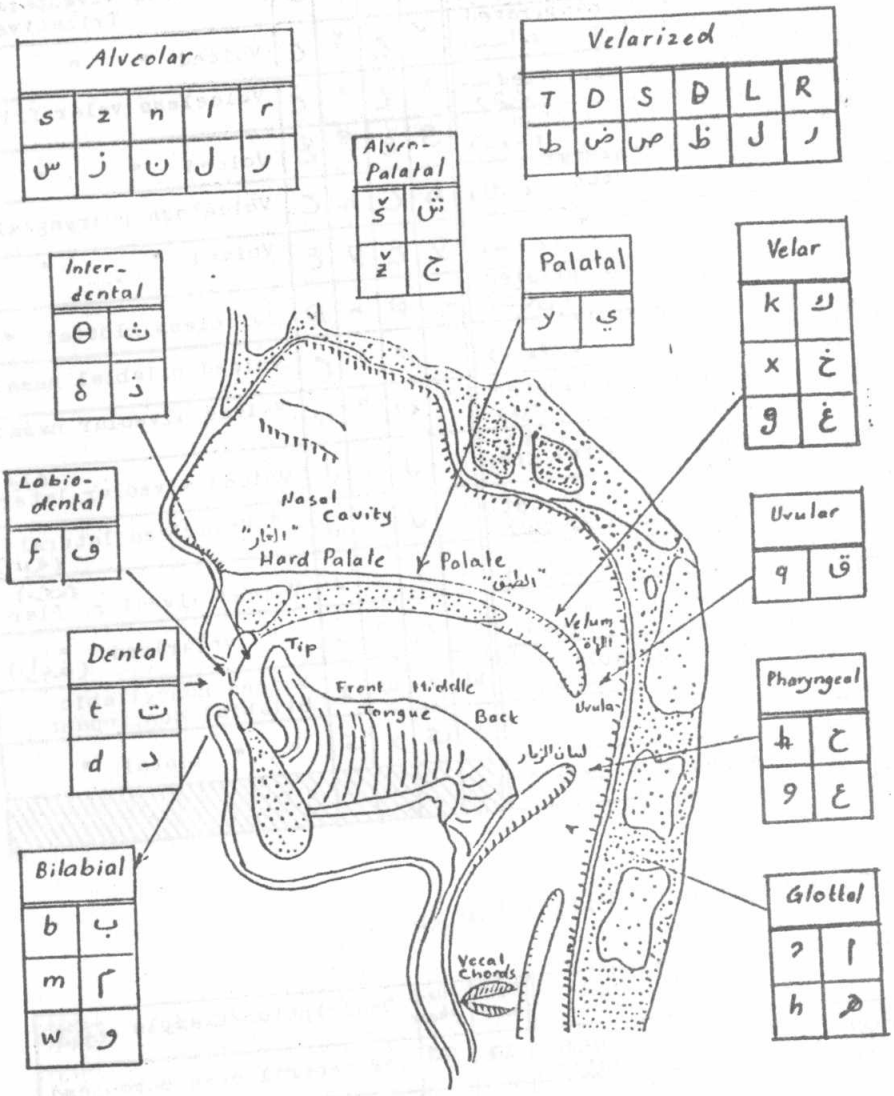


Fig.2 Points of Articulation of Arabic Consonants

COM-5 | 1122

a-Consonants

Phoneme		Allophone		Description
b		p	ب	Voiceless bilabial stop (حبس)
		b	ب	Voiced bilabial stop (باب)
t		t	ت	Voiceless unaspirated dental stop (شتاء)
		t^h	ت	voiceless aspirated " " (وقت)
d	د	d	د	Voiced " " (دود)
T	ط	T	ط	Voiceless dental velarized stop (طار)
D	ض	D	ض	Voiced " " " (مضى)
k	ك	k	ك	Voiceless unaspirated velar stop (شكى)
		k^h	ك	Voiceless aspirated velar stop (سمك)
q	ق	q	ق	Voiceless unaspirated velar stop (قال)
?	ء	?	ء	Voiceless glottal stop (أنا)
f	ف	f	ف	" labio-dental fricat (فلسوف)
θ	ث	θ	ف	" inter-dental " (ثلث)
ð	ذ	ð	ذ	Voiced " " " (ذنب)
s	س	s	س	Voiceless alveolar " (سوس)
z	ز	z	س	Voiced " "
S	ص	S	ص	Voiceless velarized "

Phoneme		Allophone		Description
ð	ظ	ð	ظ	Voiced velarized inter-dental fricative (ظن)
š	ش	š	ش	Voiceless alveopalatal fricative.
ž	ج	ž	ج	Voiced " "
x	خ	x	خ	Voiceless velar fricat
ɣ	غ	ɣ	غ	Voiced " "
ħ	ح	ħ	ح	Voiceless pharyngeal "
ʕ	ع	ʕ	ع	Voiced " "
h	ه	h	ه	Voiceless glottal ؟
m	م	m	م	Voiced bilabial nasal
n	ن	n	ن	Voiced alveolar nasal
l	ل	l	ل	Voiced alveolar latera (ليل)
ɫ	ل	ɫ	ل	" Velarized lateral (الله)
ɾ	ر	ɾ	ر	Voiced alveolar flap (ربح)
r	ر	r	ر	" velarized " (راحة)
w	و	w	و	Voiced monosyllabic bilabial continuant
y	ى	y	ى	" " platal "

b-Vowels, semivowels & diphthongs

Phoneme	Allophone	Description/Example
ii	ii	High front close unrounded (قيل)
	ɨɨ	" back " " (قميص)
i	i	" front open " (بين)
	ɨ	" back " " (ضرس)
	ɛ	Mid front " " بدل i في نهايات الكلمات
	ɛ̈	" back " " بدل ɨ في نهايات الكلمات
a	æ	Low front close " (شاب)
	ə	Mid central " " (لهيب)
aa	ææ	Low front " " (لُبّ)

Phoneme	Allophone	Description/Example
ɑɑ	ɑɑ	low central open unrounded (طلب)
uu	uu	High back close unrounded (سوق)
ʌʌ	ʌ	" " open " (كوخ)
	o	Mid " close " بدل u في نهايات الكلمات
ay	æi	Combination of æ and i (كي)
	əi	" " ə " i (ضيف)
aw	æu	" " æ " u (لو)
	əu	" " ə " u (صكوك)

Fig.(3)- Consonantal&Vowel Phonemes of Arabic with their major Allophonic submembers

One important note concerning Velarization, is that a vellarized consonant tends to velarize the rest of the word in which it occurs.For example , استطاع is actually pronounced اصطاع (with velarized vowels) .

Supra-Segmental Features :

These prosadic features of speech,namely;(Stress,Intonation, Juncture,Pause & Rhythm)must be very carefully analyzed and encoded in the algorithm for concatenating allophones and words. The following notes serves the purpose description and illustration (via developed sonograms and DFT waveforms)these important features of Arabic Phonetics :

Stress:

Stress is the force (relative force) used in pronouncing sounds . The occurence of different stress levels in Arabic is predictable according to the rules illstrated in Fig(6).The rules assigns different levels of stress to different syllables in words according to the syllabic structure of the word.Abundance of stresses in Arabic gives it its unique "stuccato" quality,and actually reduces the possibility of syllable jamming noticeable in English utterances .

Intonation:

Intonation means the changes in the pitch of the voice during speech.In English,for example,there are four relative pitch (intonational)evels.The combination of different pitches and their formation as glides or contours do chang the meaning of the word .

In Arabic there are four relative phonemic levels. They are combined in glides whenever the speaker changes from one level to the other.One important rule here, is that the pitch glide is always accompanied by a premary stress and is called a primary contour . More primary contours are found in Arabic utterances than in English utterances of the same duration , because of the greater abundance of primary stresses in Arabic.

Examples of intonational levels in Arabic arevillstrated in Fig(5).

Juncture:

There are two juncture phonemes in Arabic :

1. a plus juncture ,symbolized by (+), and
2. a minimal juncture ,symbolized by (-).

Sentences with only(+)junctures are ambiguous.without any juncture,the syllabic structure for the placement of the stress not clear, and the utterance will be ambiguous . Adding minimal junctures instead of plus junctures will make the syllabic structure clear, but the utterance will still be ambiguous . In very slow speech one would expect many more plus junctures than in repid or even normal speech.Note also the effect of velarization before and after plus and minimal junctures.The following examples illustrate this remark:

1. across minimal juncture boundaries :

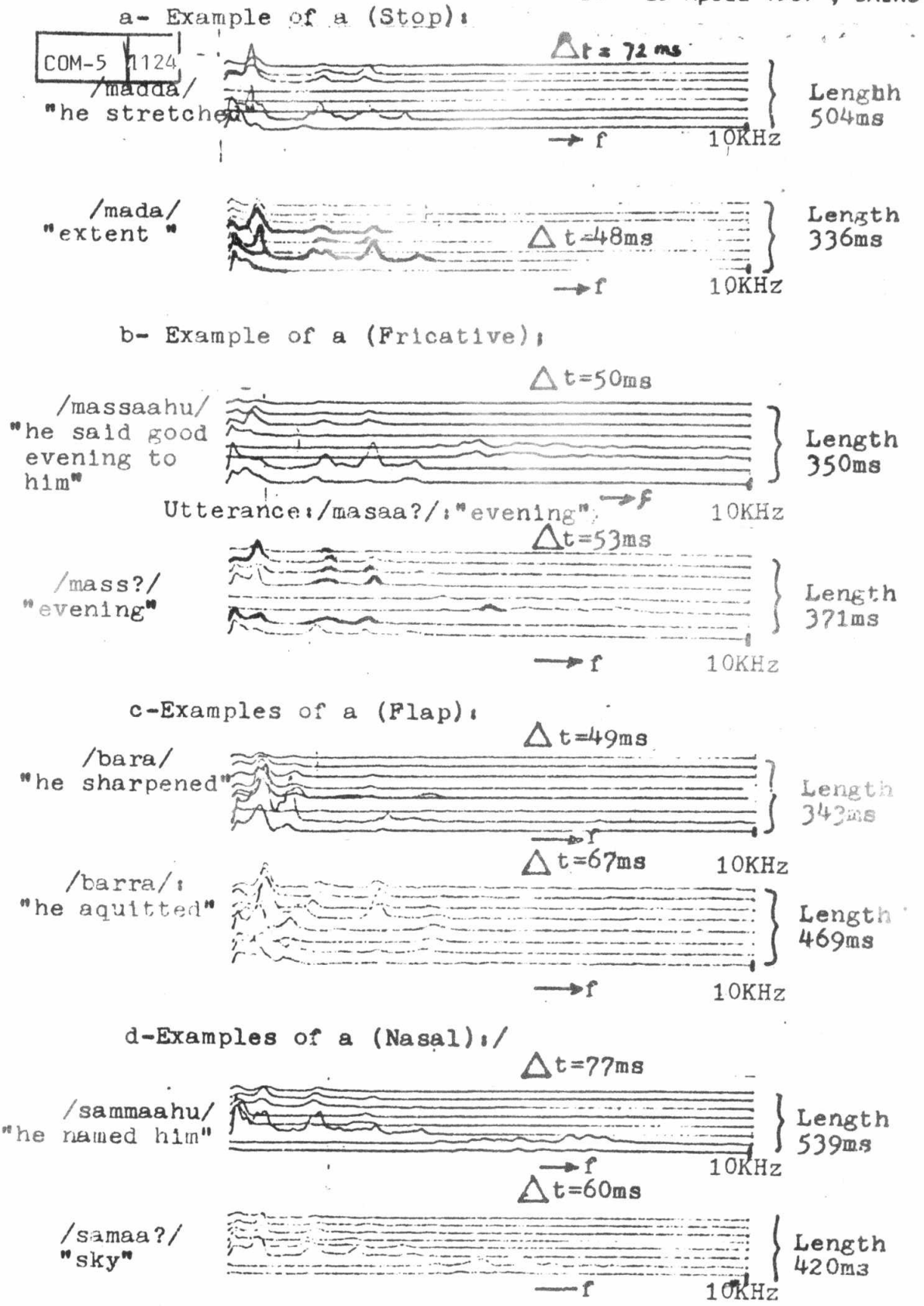


Fig.(4). Phonemic Attribution of Length Duration of typical Utterances (Raster&FFT waveforms:Transform size 512pts,Range10KHz &resolution 50Hz).

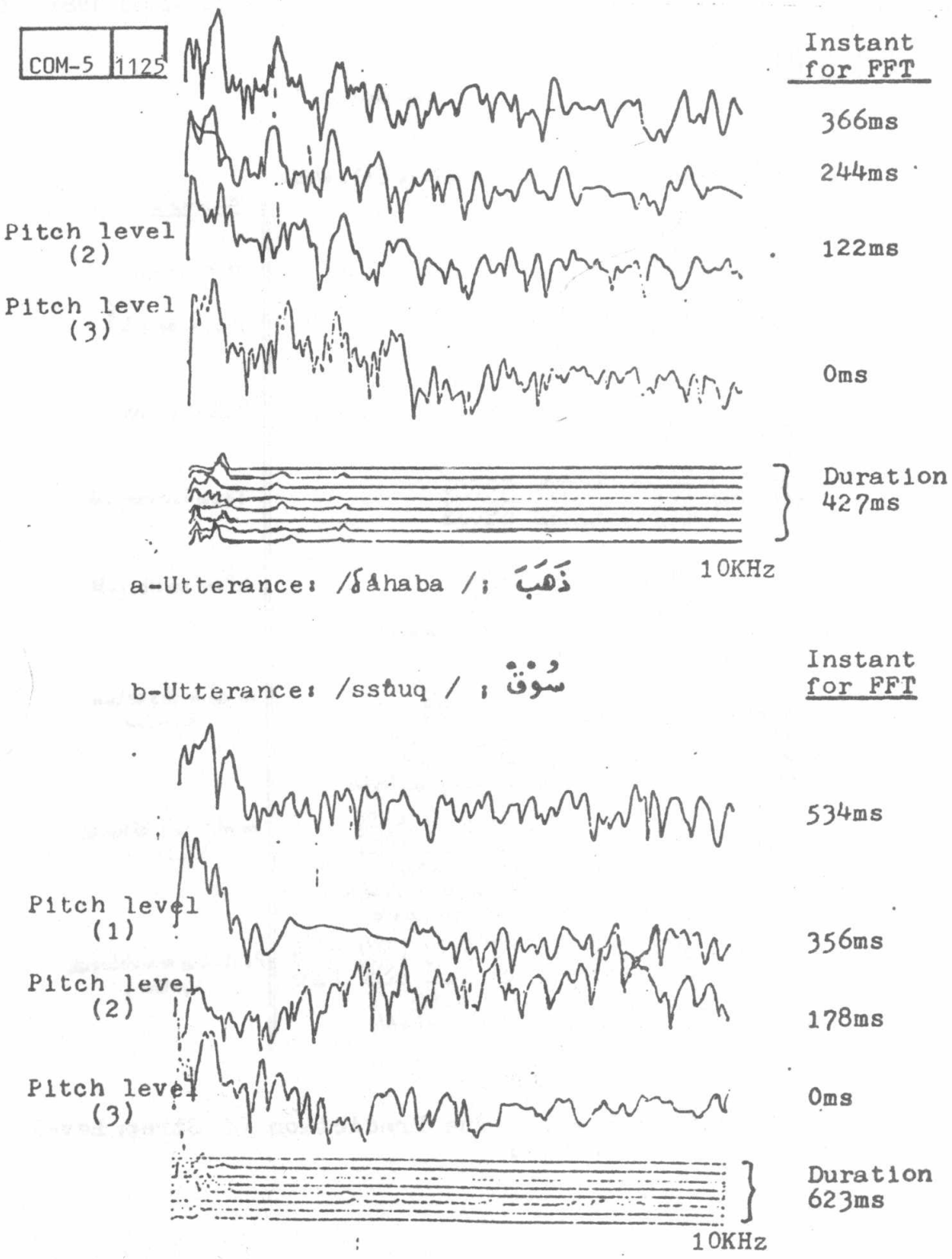
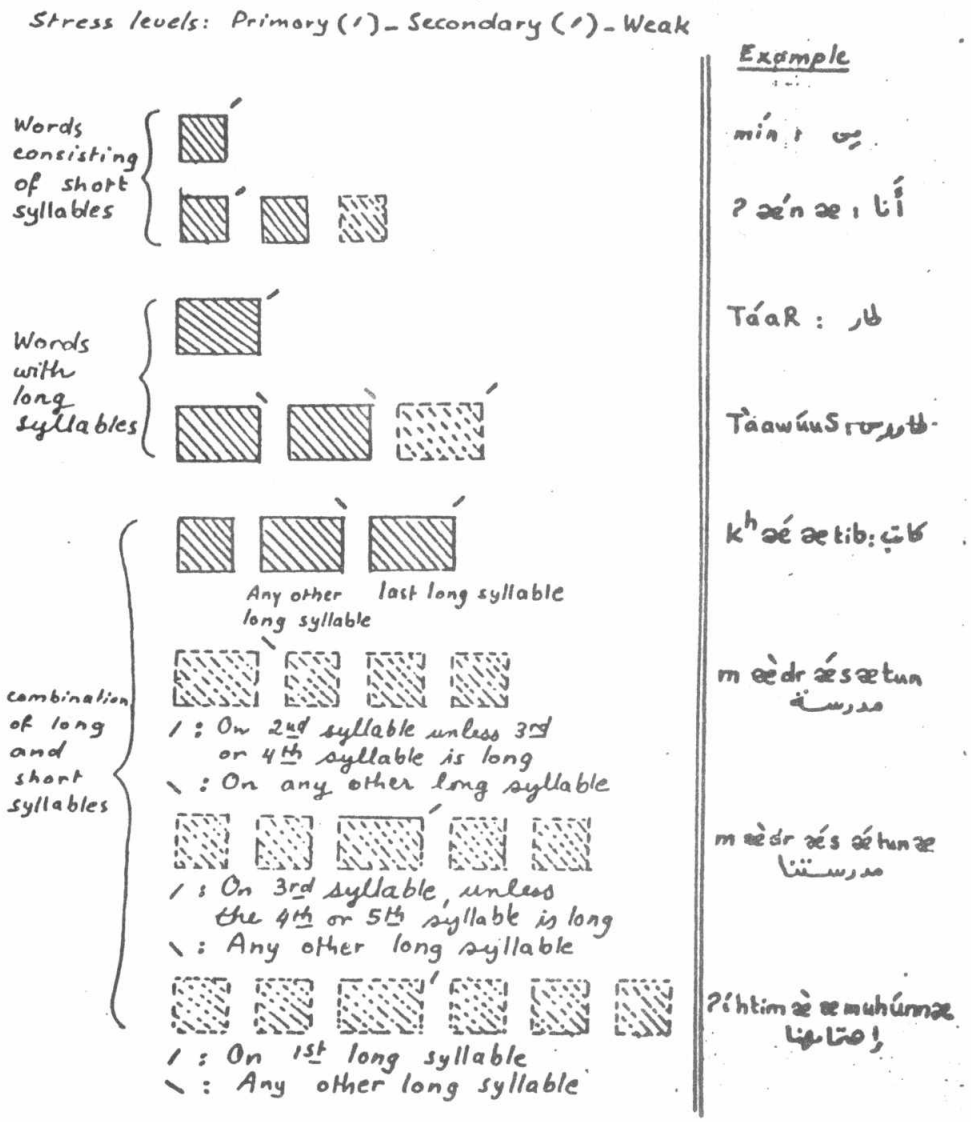


Fig.(5). Raster & FFT waveforms(Transform size 512pts, Range10KHz &Resolution 50Hz) to illustrate Intonational Pitch Variations during typical word utterances
(Pitch level variations1...4,4is the highest level)

COM-5 | 1126



Fig(6)-Rules For The Prediction Of Stress Levels In Arabic

a.(D): (máDa-Lwáladu):"the boy signed " ,
(d): (rádda-lwálada):"he returned the boy "
b.(R): (?alźáaRu-Lqaríibu):"the close neighbor",
(r): (?albí?ru-lqaríibu):"the close well "

2. across plus juncture boundaries :

a.(s): (l æ æ +s æ líim+waLáa + Saaḥíbuhu)
b.(R): (yéa + Rabb)"Oh,Lord",
(r): (yáé æ + r æ mzi) "Oh,Ramsay".

Pause:

In Arabic there are two pause phonemes- a short one longer than the plus juncture discussed above, and a long one which occurs after complete or independent utterances .

Rhythm:

Rhythm in Arabic is stress-timed.This means that just the same time elapses between two primary stresses.

3. Speech Processors

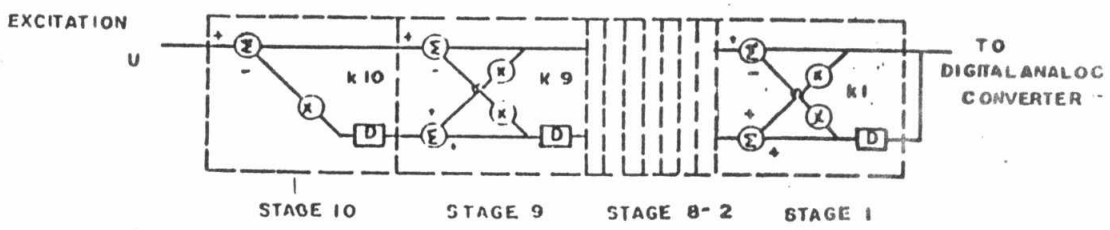
A legendary speech processing chip has been the TITMS5100. Fig(7) illustrates the LPC filter,(10) sections,present within the architecture of this chip.LPC (Linear Predictive Coding) is very successfuly applied in compression of speech .

As long as the speech waveform is stationary (for time intervals not exceeding 15ms),an LPC model of the vocal tract can be constructed using liear prediction.The time varying parameters of the model can be stored in a ROM and are accessed by a microprocessor to transfer these parameters to the filter periodically. The Architecture and function of this chip are illustrated in Fig (7) & Fig(8). The following is a brief description warranted by the lack of references concerning it .

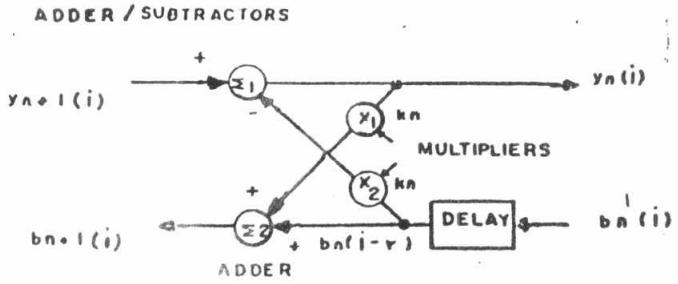
The first 9 stages of the LPC filter carry out two multiplications and two additions on their two digital inputs before passing the results backwards and forwards to their neighboring sections. Fig(7) illustrates the filter's nth stage during the i-th time cycle in response to the data(Y) derived from the (U) and the (B) data (feedback data).In this representation ,the subscript of the (Y) and (B) data refers to the stage in which the data is generated. In all of the calculations performed by the filter , the Y-and B-data,as well as the coefficients K1-K10 are multibit numbers.The coefficients K1-K10 may vary between +1 and-1 and are periodically updated .

A block diagram of the actual digital 10-stage lattice filter is shown in Fig(8). The filter includes a pipeline multiplier, an adder/subtracter circuit,adelay circuit,a shift register and a latch memory . The pipeline multiplier output is one input to the adder/subtracter.

The pipeline multiplier performs all twenty multiplications required by the lattice filter.It receives either (Y)or (B) via the data bus,and the coefficients K1-K10 from the K-stack of the shift register by the data lines.It initiates a different



a-The 10section Lattice Digital Filter



b- Detail of one section of the filter

Fig(7)- LPC Filter of the TI-TMS 5100 Chip

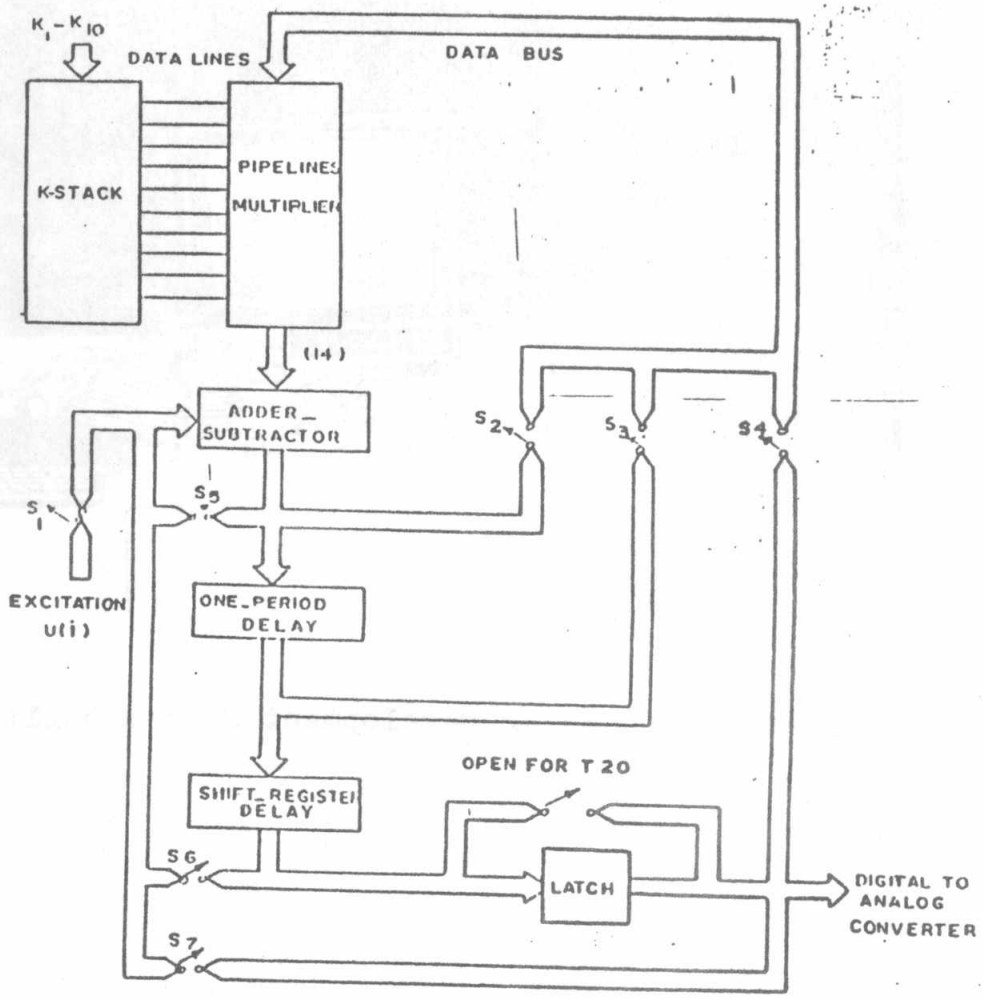
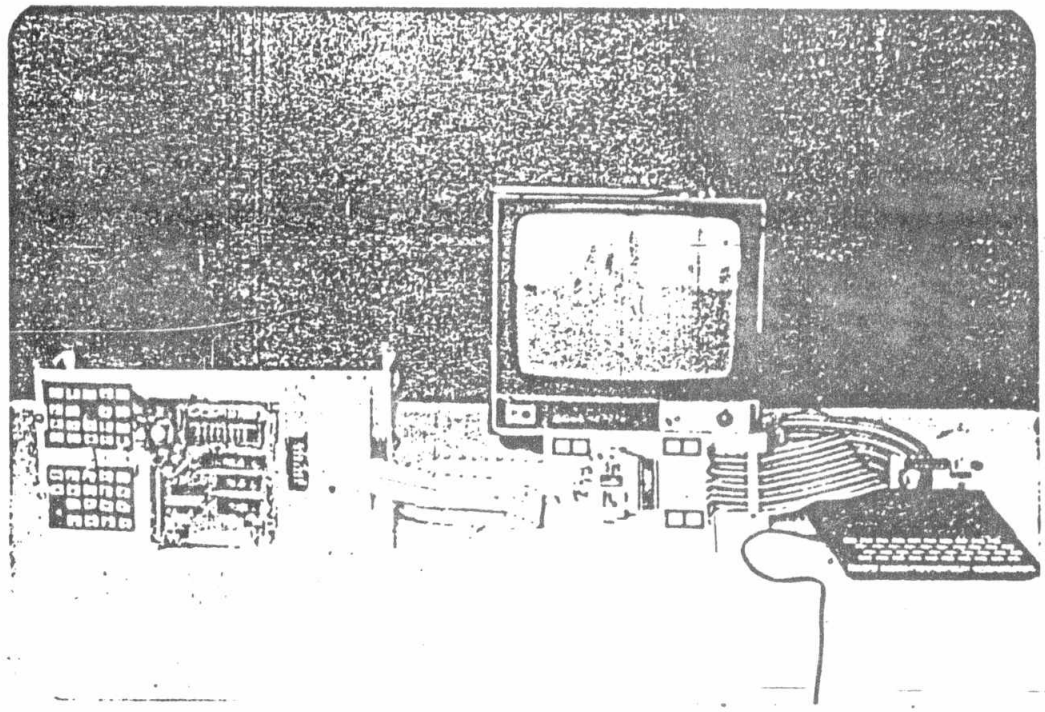


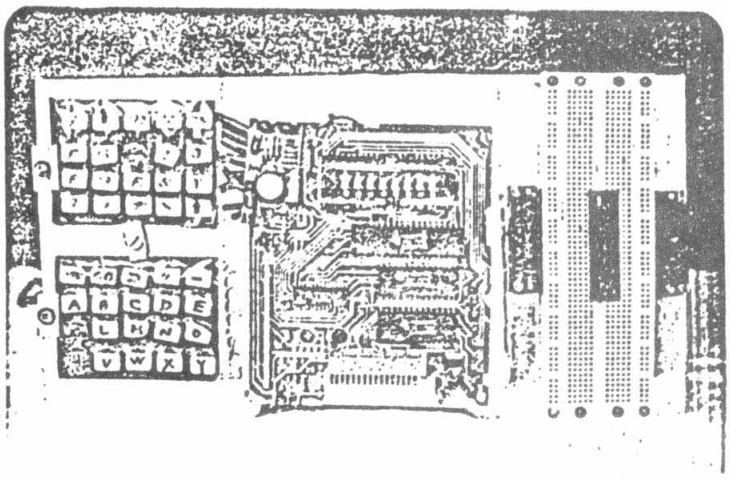
Fig (8)-Block Scheme of theDigital 10-section lattice Fiter of the TMS5100 Chip.

multiplication operation every 5-microsecond interval.One multiplication operation requires 8 time periods, so the pipeline multiplier has 8 stages.Therefore there are 8 multiplications coded data are applied to the digital filter every 20 milliseconds . A 10KHz output speech sample rate requires a time cycle of 100 microsecond.This allows the basic operation of the adder/subtracter,the muliplier and the shift registers to be accomplished in 5 microsecond time periods.A max. of 49 bits are neede to update 13 parameters of the algorithm every 20ms.They can be split as follows :

Energy (amplitude information)	:	4 bits
Repeat	:	1 bit
Pitch (frequency)	:	5 bits
The 10 reflection coefficients		
K1&K2	:	5 bits
K3&K7	:	4 bits
K8&K10	:	3 bits



Fig(9)- Speech Processing Development System built by the author

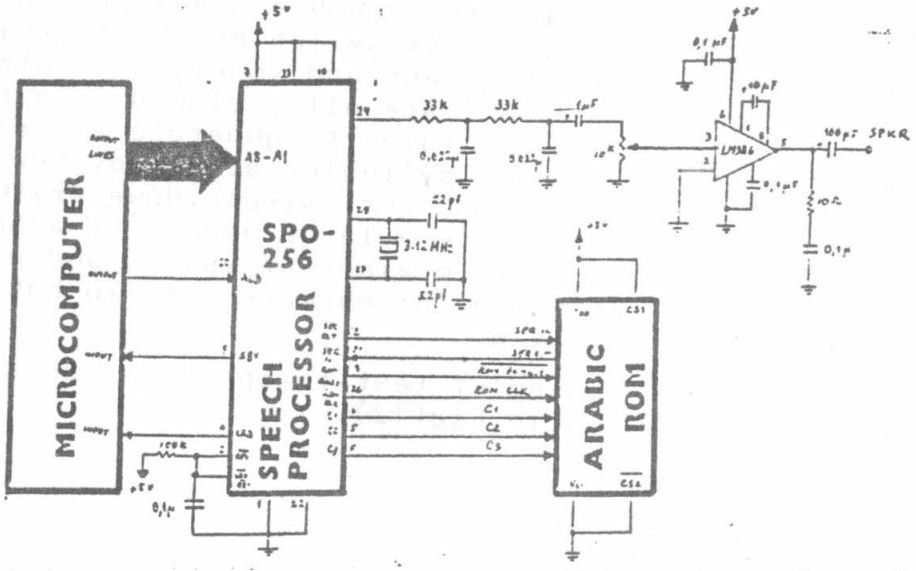


Fig(10)- Typical Speech Processor Emulation Board

Another modern voice processing chip is the General Instruments SPO-256 chip.It is analogous to the TMS5100,utilizes14-section digital filter representation of the human vocal tract.

Fig(11) illustrates a proposed utilization of this chip in a circuit configuration using a specially prepared Arabic Speech ROM for syhthesis of Arabic.

Fig(9) ,Fig(10) show a development system built by the author for sound evaluation using modern voice synthesizer chips.



Fig(11)-Circuit Configuration of the General Instruments SPO-256 speech processor for speech synthesis.

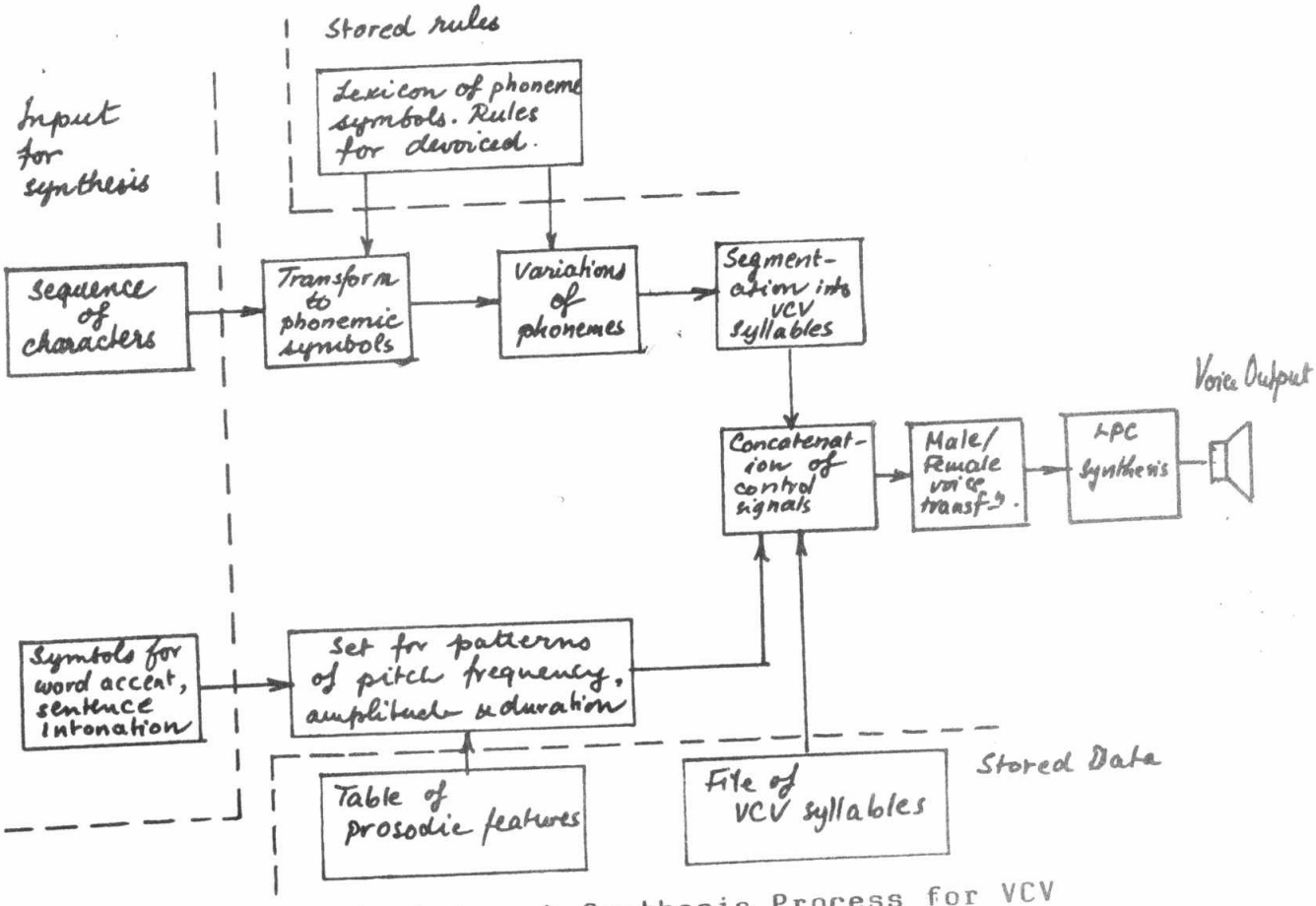
4.A Proposed Model For a Text-To-Speech Arabic-Based Synthesizer

Text-to-speech means speech synthesis from written messages in books and papers or from messages generated from a typewiter keyboard . Its ideal form can simulate the human action of reading books or papers aloud,therefore ,the following processes must be realized:(1)word boundary detection and segmentation of a sentence into phrases and words; (2) transformation of words into phonemic symbol sequences;(3)addition or derivation of word accent and sentence intonation;and (3),from the results of (2) and (3) generation of control signals for a speech synthesizer by stored data and rules.

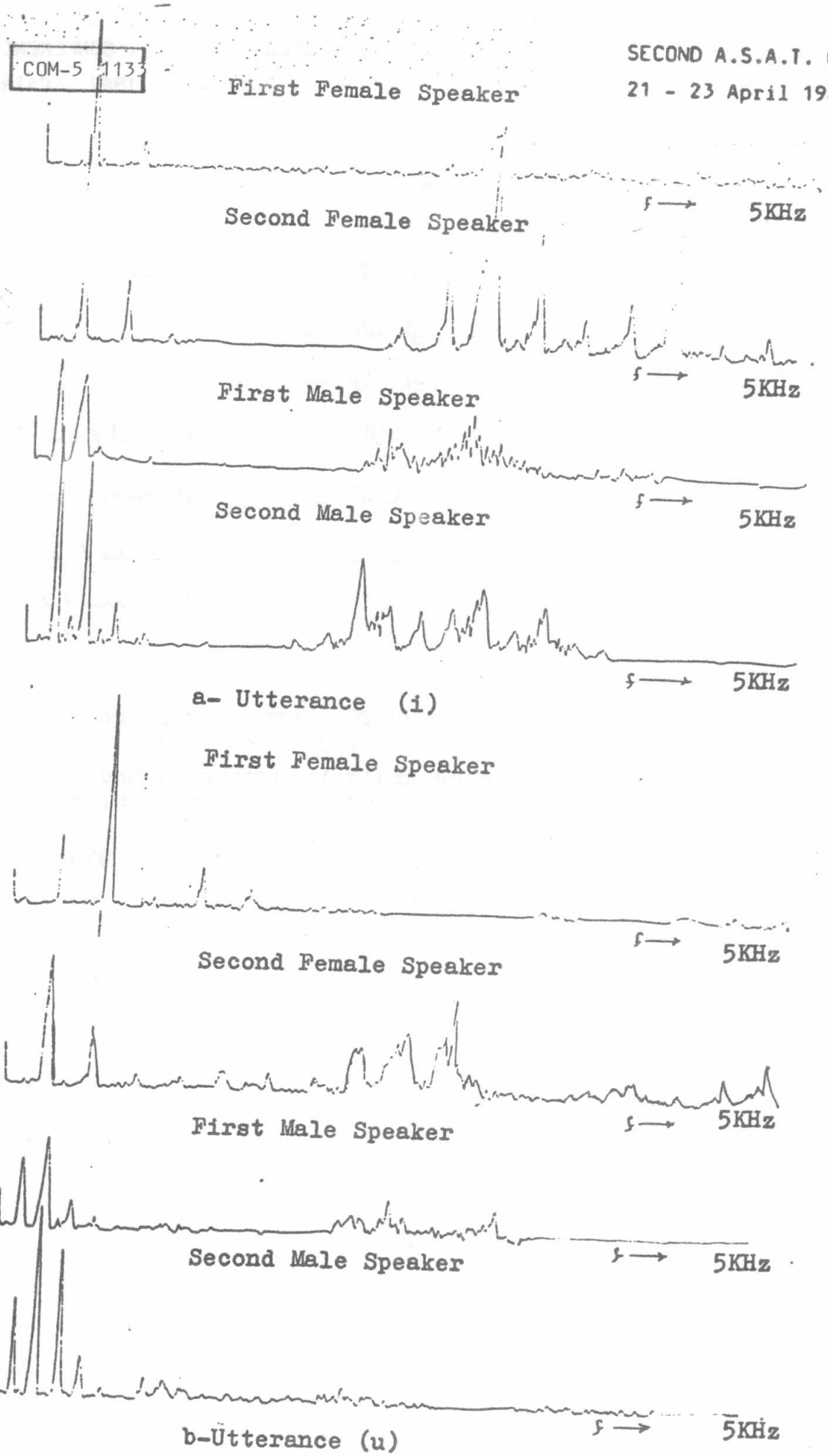
An example for typical text-to-speech synthesizer are those working on Phoneme principles.Prosodic rules,rules for the control of pitch period,duration,and amplitude are hard problems to overcome.The fundamental information of prosodic features of a sentence includes punctuation and stress marks on word syllables . The duration and pitch period for each phoneme are adjusted according to this prosodic information.Rules for the control of prosodic features are stored in the form of a table,wherein variations are sorted by acoustical features of each phoneme jodified by intonation,stress,and pause marks in a sentence.

The advantage of the phoneme-based synthesizer is that only about 50 phonemes are necessary for the synthesis of words .However,in the generation of the transitional part from phoneme to phoneme , consideration must be given to the rank order of the phoneme , duration of the transitional part,and the upper and lower limit of forment frequencies at a phoneme boundary.The process of transitional part generation becomes very complex but these rules still cause some degradation of voice quality.To improve the results of synthesis by phonemes,the synthesis by VCV(vowel+consonant+vowel) is proposed.By using VCV syllables,the transition between vowel and consonant,consonant and vowel,and coatculation betweer the first and second vowel are all included in a natural form. So there is no need to have rules to generate transitional parts.This method can concatenate syllables at steady vowel parts and thus concatenation rules become very simple.However,the number of units increases greatly and a very large amount of acoustical data must be stored in advance.If m kinds of vowels and n kinds of consonants are used, the number of phomemes is m+n but the number of VCV syllables is $m^2 \times n$.

A block diagram of the system of text-to-speech synthesis using VCV syllables is shown in Fig(12).



Fig(12)-Speech Synthesis Process for VCV



(13)- Differences in the Formant structure in Male and Female utterances of example phonemes

References

I) Arabic References:

١ ـ د / رمضان عبد التواب " التطور اللغوى ،مظاهره وعلله وقوانينه "
مكتبة الخانكى ـ القاهرة ـ ص ١٧ ـ ٤٦

٢ ـ " " " : المدخل الى علم اللغة " مكتبة الخانكى ـ القاهرة .

٣ ـ د / ابراهيم انيس : "الاصوات اللغوية" مكتبة الخانكى ـ القاهرة ١٩٤٧

٤ ـ د / محمود السعران : "علم اللغة " ص ١٩ ـ ٢٢٠

٥ ـ د / عبد الرحمن أيوب : "اصوات اللغة ١٩٦٨

٦ ـ د / سعد عبد العزيز مصلوح : "دراسة السمع والكلام ـ القاهرة ١٩٨٠ .

II)English References:

(1)- Technical Data of the SP0256 Narrator Speech Processor,Radio Shack Catalog No.276-1784

(2)- J.L.Flanagan :"Speech Analysis,Synthesis and Perception",Springer Verlag,Berlin&New York,1972

(3)-J.D.Markel &A.H.Gray(Jr):"Linear Prediction of Speech",Springer-Verlag,Berlin&New York,1976