„O viață liniștită și modestă aduce mai multă bucurie decât urmărirea succesului asociat cu anxietatea constantă”, Einstein a scris aceste cuvinte în germană pe antetul hotelului Imperial Tokyo.

ASRSV – curs 7
Sinteza Semnalului Vocal

- **Introducere**
- **Sisteme de sinteza clasice**
- **Tehnici de sinteza**

http://research.spa.aalto.fi/publications/theses/lemmetty_mst/chap2.html
Development of speech synthesizers

https://www.cs.cmu.edu/~srallaba/Learn_Synthesis/intro.html

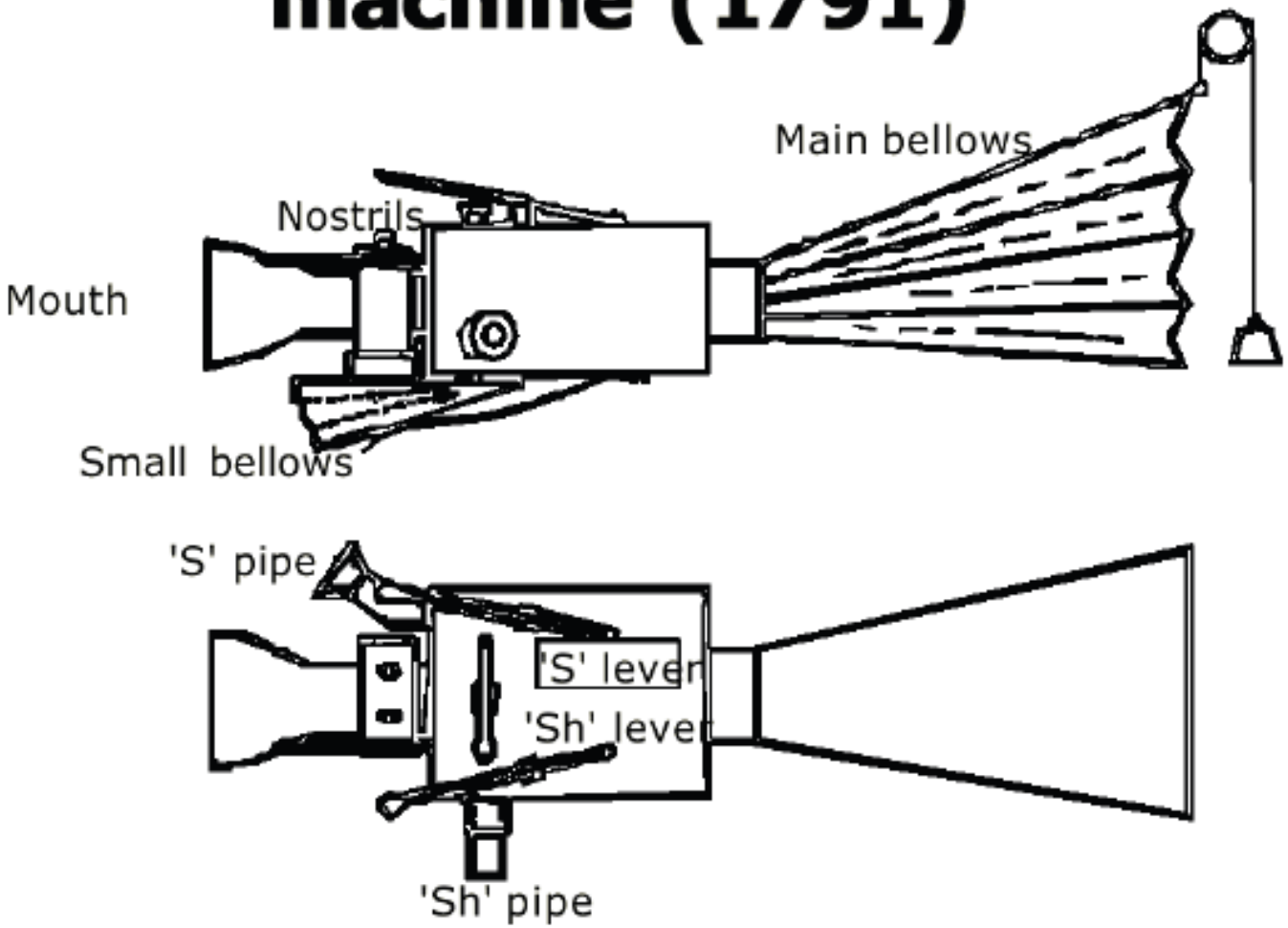
Sinteza vocală este generarea semnalului vocal sintetic utilizând intrare text (TTS) și modele care capturează informații fonetice, prosodice și acustice pentru a reproduce vorbirea umană într-un mod cat mai natural și inteligibil.

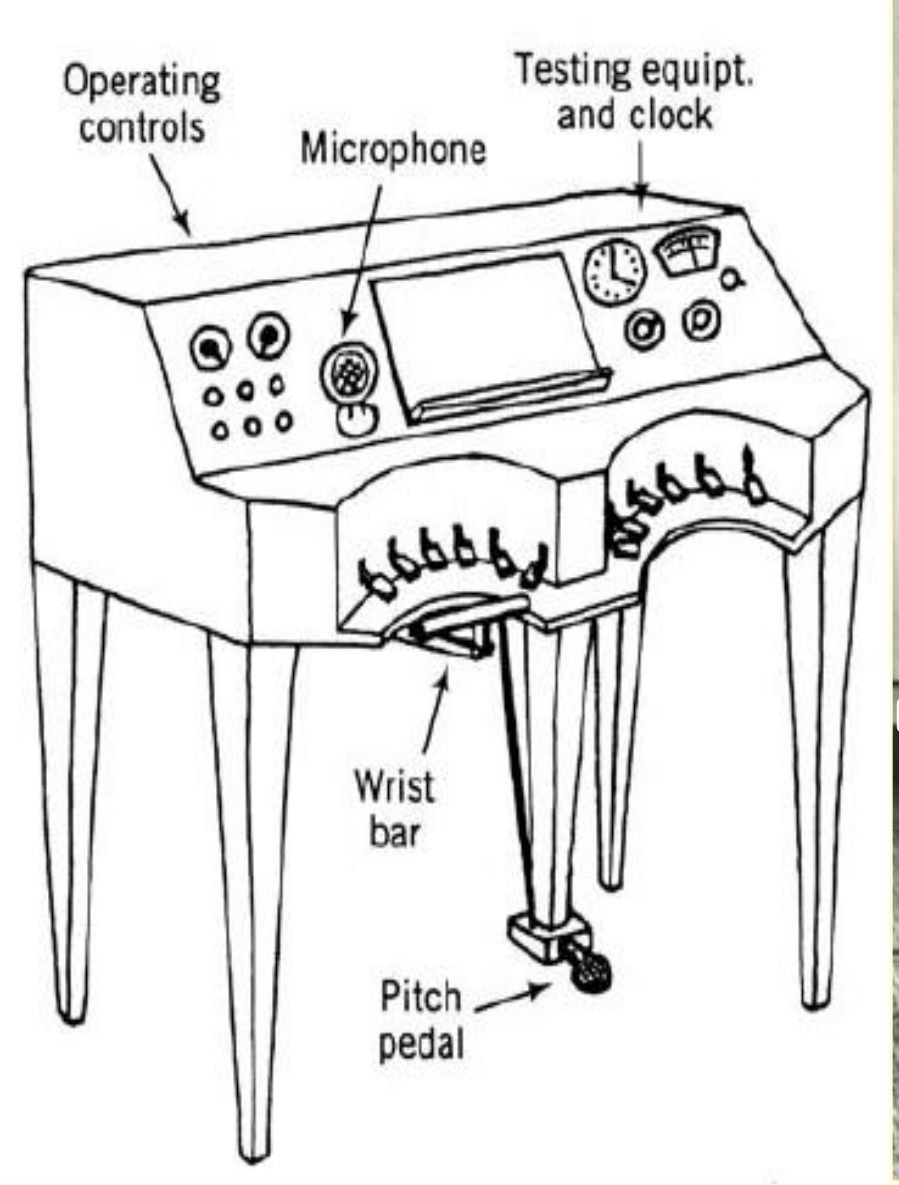
Obiective tehnice: se urmărește atât fidelitatea perceptuală (naturalitate, timbru, prosodie), cât și robustețea în diferite limbi, accente și voci, cu resurse de calcul compatibile cu aplicațiile în timp real.

Arhitecturi majore:

- pre-electronic (sinteza mecanică),
- electronică (formant-based),concatenativă (unit selection), parametrică și cele neuronale end-to-end actuale (Tacotron, WaveNet, transformere audio).

Von Kempelen's talking machine (1791)

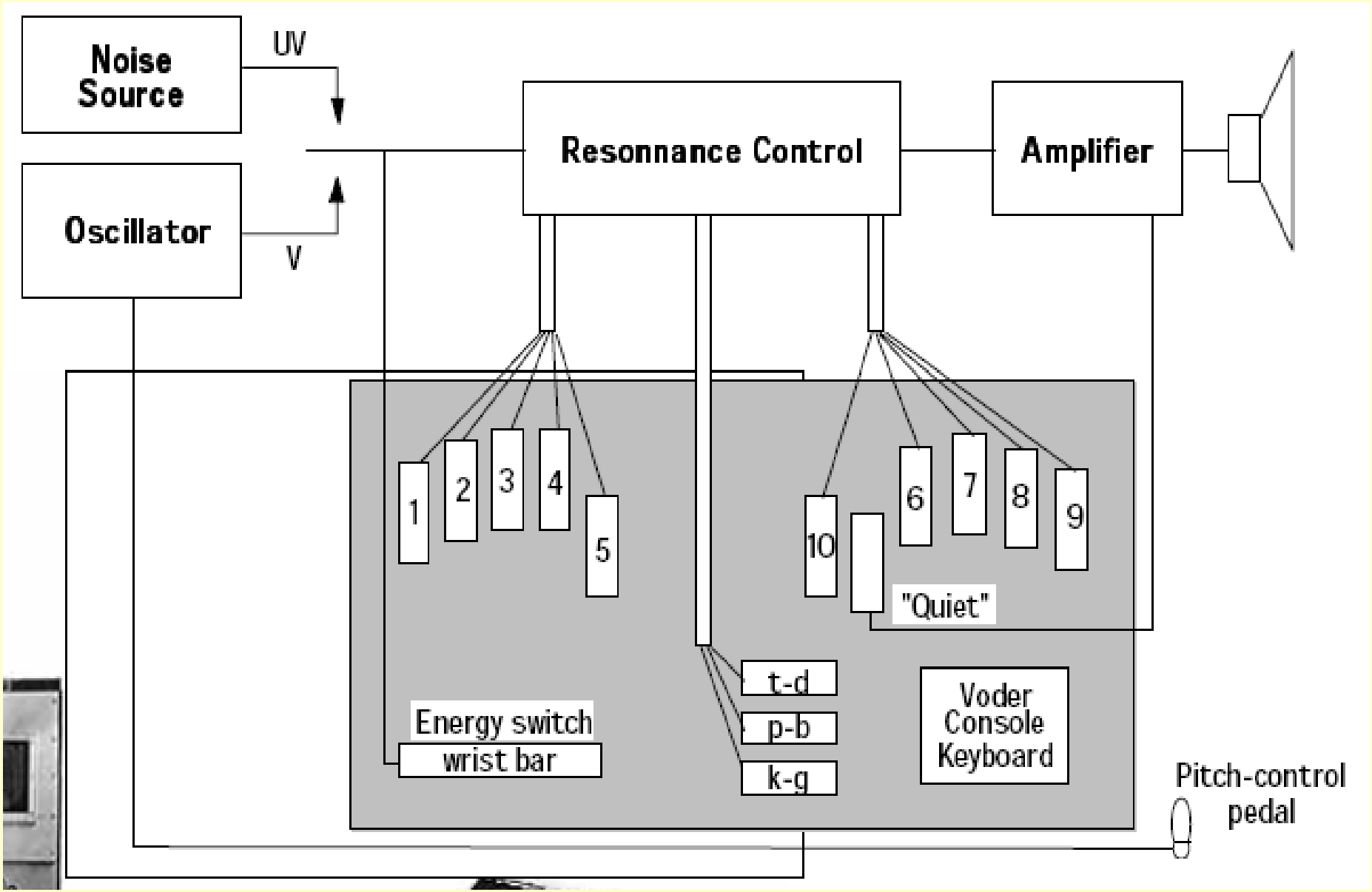




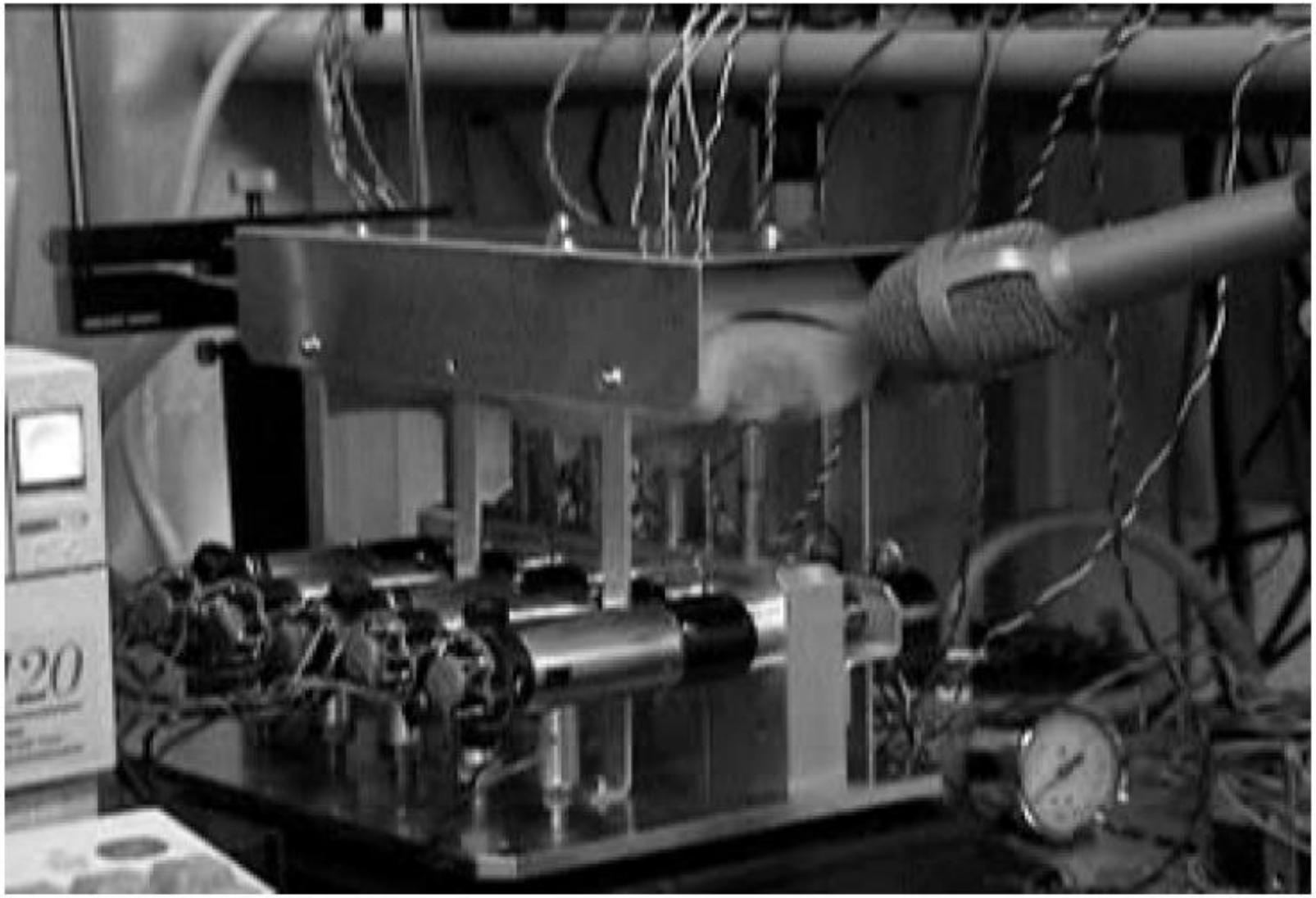
1939 –VODER- Homer Dudley (Bell Labs.)
(Voice Operated DEmonstratoR)

Wave Sound

(.wav)


Noise Source
UV
Oscillator
V
Resonnance Control
Amplifier
1
2
3
4
5
6
7
8
9
10
"Quiet"
t-d
p-b
k-g
Energy switch
wrist bar
Voder Console Keyboard
Pitch-control pedal

Cercetari in sinteza electro-mecanica a vorbirii



http://www.eng.kagawa-u.ac.jp/~sawada/

Tehnologia sintezei vorbirii bazată pe învățare profundă

- Odată cu dezvoltarea științei și tehnologiei informatice, inteligibilitatea și naturaleța vorbirii sintetizate au fost mult îmbunătățite datorită progresului continuu al tehnicilor de sinteza TTS:
  - de la metodele bazate pe formanți
  - la metodele de concatenare a formelor de undă
  - metode bazate pe selecția unităților de sinteza
  - la metodele de sinteză statistică parametrică a vorbirii (SPSS) bazate pe modele Markov ascunse (HMM)

- La metodele bazate pe 'deep-learning' care :
  - Learning from Audio: Wave Forms
  - Learning from Audio: Time Domain Features
  - Learning from Audio: Fourier Transformation
  - Learning from Audio: Spectrograms
  - Learning from Audio: Pitch and Chromagrams

THE BEST TEXT-TO-SPEECH SOFTWARE

Click the links below to go to the provider's website:

1. Amazon Polly
2. Linguatec Voice Reader
3. Capti Personal
4. NaturalReader
5. Voice Dream Reader

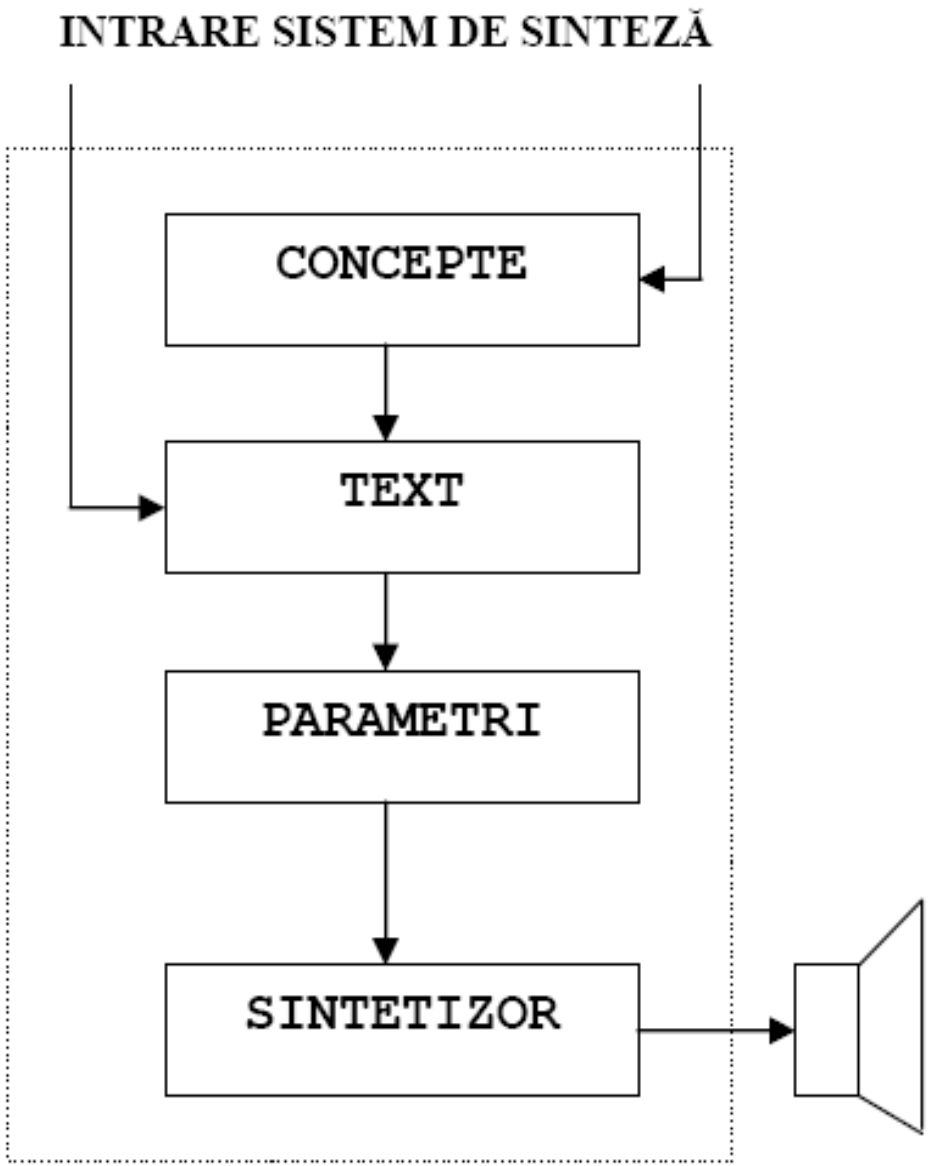
Or, jump to:
Best free text-to-speech apps.

1. Amazon Polly
2. Linguatec Voice Reader
3. Capti Personal
4. NaturalReader
5. Voice Dream Reader

jump to:
Best free text-to-speech apps.

- **Alexa** - serviciul de voce cloud-based al companiei Amazon, disponibil pe zeci de milioane de dispozitive de la producătorii de dispozitive Amazon și de la terți.
- Cu Alexa, putem experimenta un dialog natural care oferă clienților o modalitate mai intuitivă de a interacționa cu tehnologia pe care o folosesc în fiecare zi.

Sisteme de sinteza vocala clasice



Sistem de sinteză

Moduri de sinteza:

- **mod grosier** *(mem.mare, vocabular limitat)*
- **prin fragmentarea vorbirii** *(difoneme, demisilabe, reguli)*

Sintetizoarele se pot clasifica după *domeniul în care se realizează prelucrările* (timp sau frecvenţă)

- sintetizoare pe bază de eşantioane (amplitudine-timp)
- sintetizoare utilizând modulaţia delta
- sintetizoare utilizând generatoare de frecvenţă
- sintetizoare de canal (vocoder de canal)
- sintetizoare pe bază de predicţie liniară (cu excitaţie F0, reziduală sau multiimpuls)
- sintetizoare formantice
- sintetizoare complexe (bazate pe forme de undă) cu eliminarea redundanţei vorbirii

După **tehnica** folosită:

1 **generare directă** (codarea formei de undă) – include tehnici de prelucrare a formei de unda a SV inregistrat (direct/codat) => generarea mesajelor vocale

2 generarea vocii conform unui **model (analiză-sinteză)** - realizate prin modelarea producerii SV, la care vocea => parametri =>sintetizorul => mesajul dorit. (sintetizoarele formantice, LPC,...)

3 **simularea tractului vocal**

- *metoda analogiei cu tractul vocal* care simulează propagarea undei acustice prin tractul vocal;
- *metoda analogiei terminale* care simulează structura spectrului cu caracteristicile de rezonanţă şi antirezonanţă reproducând procesul articulator.

4 **generare pe bază de reguli** – SV => pe baza reg. fonetice, lingvistice pornind de la litere/grupuri fonemice + caracteristici prozodice

5 **metode neuronale end-to-end** actuale (Tacotron, WaveNet, transformere audio).

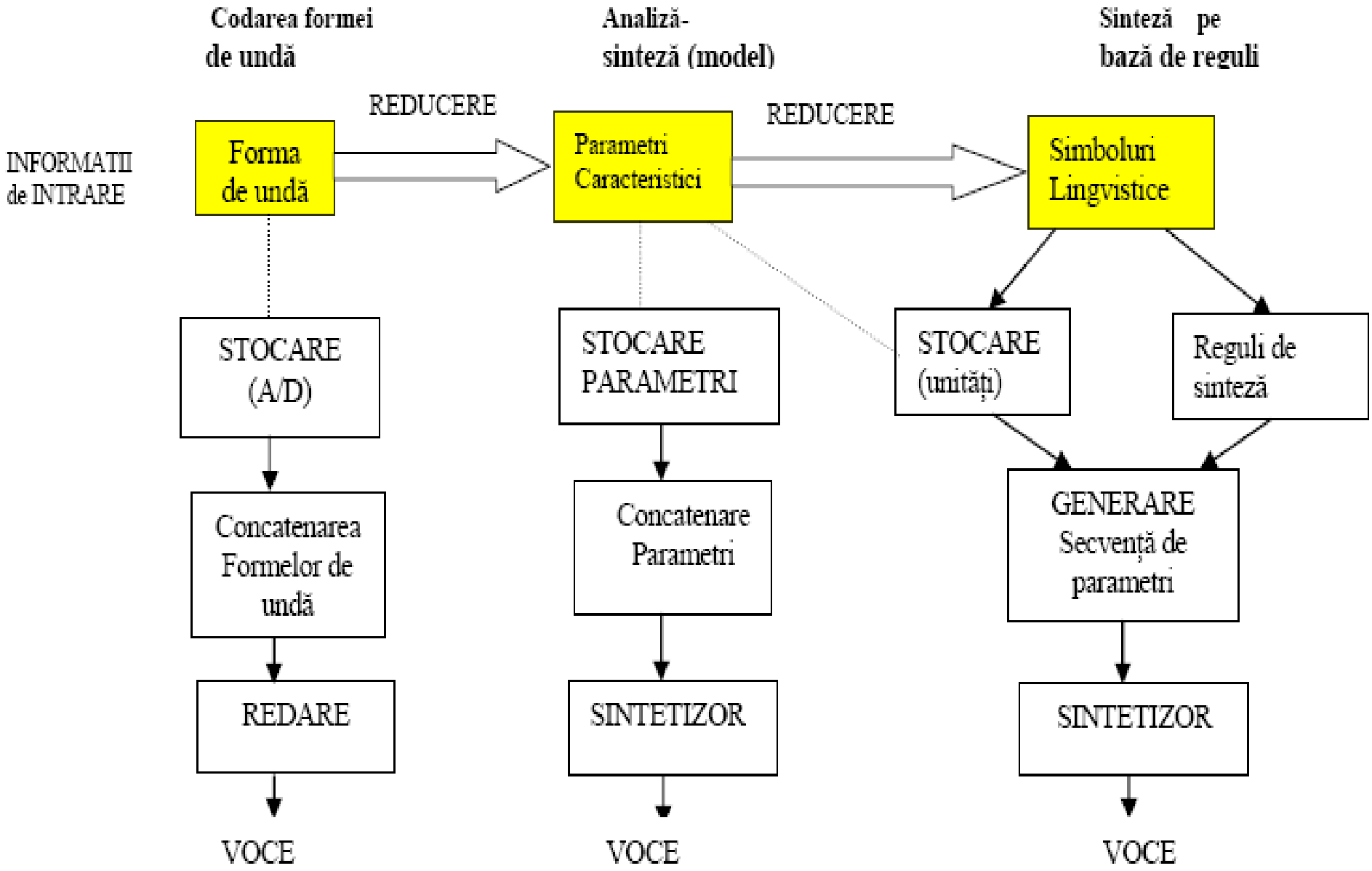


Fig2. Principiile de bază ale metodelor de sinteză şi corelațiile între ele

CARACTERISTICI	CODAREA FORMEI	ANALIZĂ-SINTEZĂ	SINTEZĂ PRIN REGULI
INTELIGIBILITATE	Mare	Mare	Medie
NATURALEȚE	Mare	Medie	Medie
MARIME VOCABULAR	Mică(<500)	Mare >(1000)	Fără restricții
DEBIT BINAR	24-64Kbps	2.4-9.6Kbps	50-75bps
SEMNAL STOCAT ÎNTR-UN MBIT	15-40s	100s-7min.	Fără restricții
UNITĂȚI STOCATE	Cuvinte,silabe, propoziții	Cuvinte,silabe, propoziții	Foneme, silabe (CV,VCV,CVC, morfeme,..)
COMPLEXITATE	Redusă	Medie	Mare
RESURSE	Memorie	Procesor+Memorie	Procesor+Memorie
EXEMPLE TIPURI	PCM,ADPCM	Vocoder canal,LPC,..	Concatenare formă de undă, Vocodere (canal/cepstral/LPC)

Tabel 1. Caracteristici şi performanțe ale tehnicilor de sinteză

Criterii de calitate pentru evaluarea sintezei

- **Inteligibilitatea**
  (cuvinte, propoziţii)
- **Naturaleţea**
  (calitatea modelarii unităţilor ; prozodia, accentuarea, ezitările etc)

Fluiditatea
(cursivitatea)

- **Potrivirea prozodiei**
  (Expresii, continuări, întrebări,exclamatii)

- **Măsurători perceptuale:** MOS (mean opinion score), SIM (similaritatea percepută), preference tests; evaluări în diferite limbi pentru inteligibilitate și naturalete.

- **Măsurători obiective:** distanțe între spectre (MSE pe spectrogramă), dinamica F0, articularea (corelații cu seturi fonetice), statistici de intelligibilitate bazate pe n-gram –

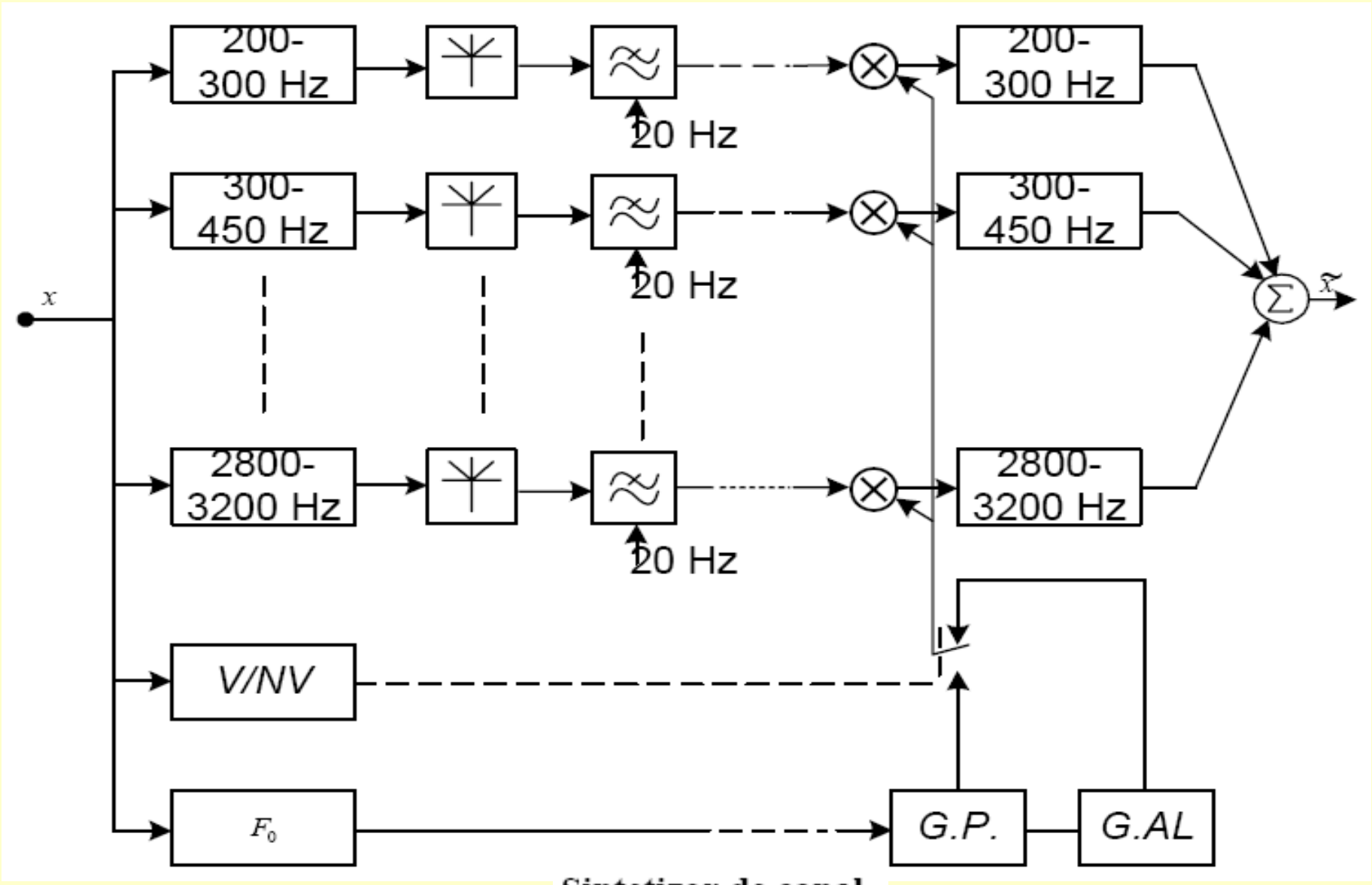
- **Evaluare în timp real:** latențe, consum energetic, flexibilitate în personalizarea vocii, scalabilitate pe mai multe voci.

- **Criminalistică și fiabilitate:** detectabilitatea sintezei, metode de watermarking audio, reduceri ale riscurilor de impersonare/clonare.

https://towardsdatascience.com/state-of-the-art-of-speech-synthesis-at-the-end-of-may-2021-6ace4fd512f2

1. SINTETIZOARE CU GENERARE DIRECTĂ (CODAREA FORMEI DE UNDĂ)

- Genereaza direct forme de undă sonoră pe baza unor măsurători efectuate asupra unor semnale reale



Sintetizor de canal

- Sinteza constă în concatenarea directă a unui număr oarecare de segmente SV
- Segmentele sunt anterior memorate, apoi sunt concatenate câteva dintre acestea, formând fonemele, pe urmă din acestea se formează silabe, morpheme ,… cuvintele.

Sintetizorul de canal – *consta in parametrizarea anvelopei spectrului SV pe termen scurt*
- se foloseşte un set de filtre (FTB 10-20) (sintetizor de canal / Channel Vocoder);
- estimarea sonor/nesonor => CODATE
- ***F0***
- **RECEPTIE** => reconstituirea cu un set FTB ≡ , excitat cu F0/zg, redresat, FTJ
Banda ~20Hz – anvelopa sp.=> 300Hz

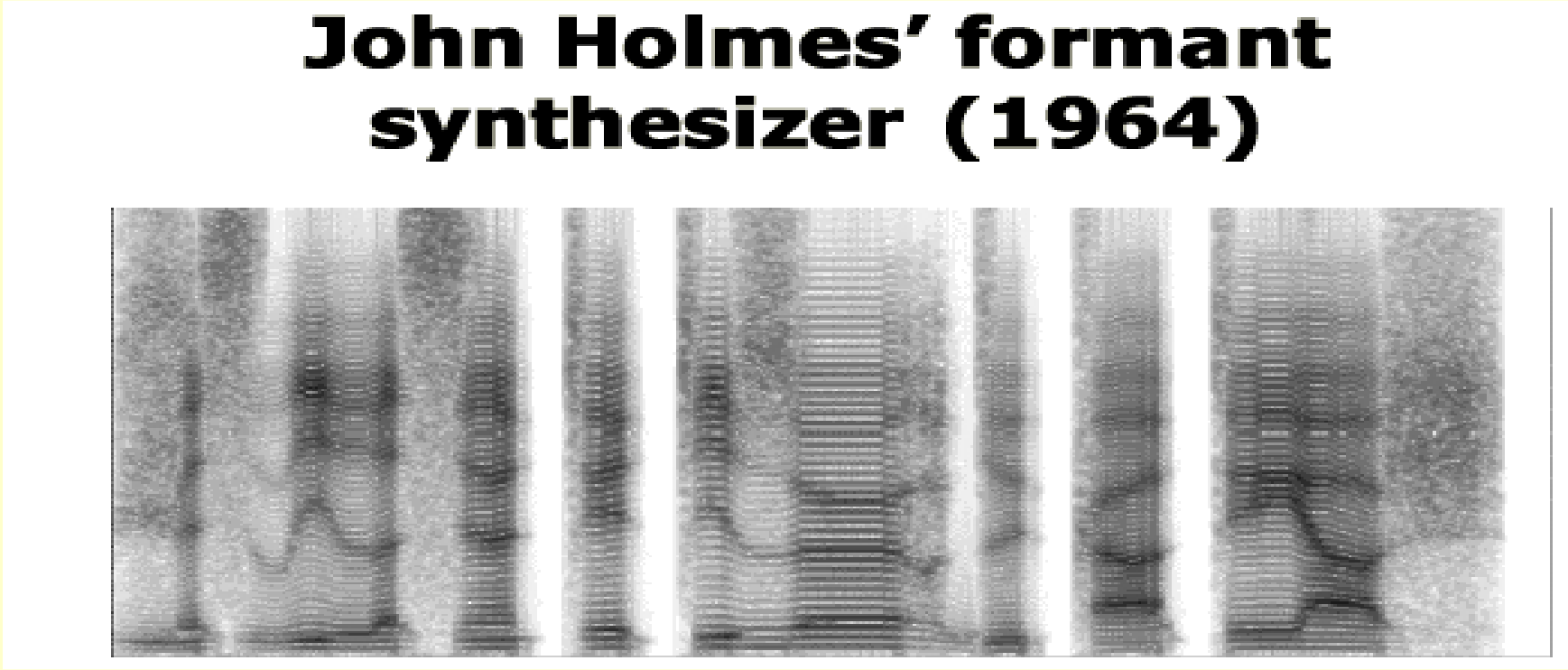
Avantaje:
- simplitatea şi memoria relativ redusă necesară stocării datelor
-calitatea şi inteligibilitatea este acceptabilă în anumite aplicaţii fără pretenţii

Dezavantaje
- descrierea spectrului este o aproximaţie mai mult sau mai puţin grosieră în funcţie de numărul de canale utilizate, de lăţimea lor de bandă

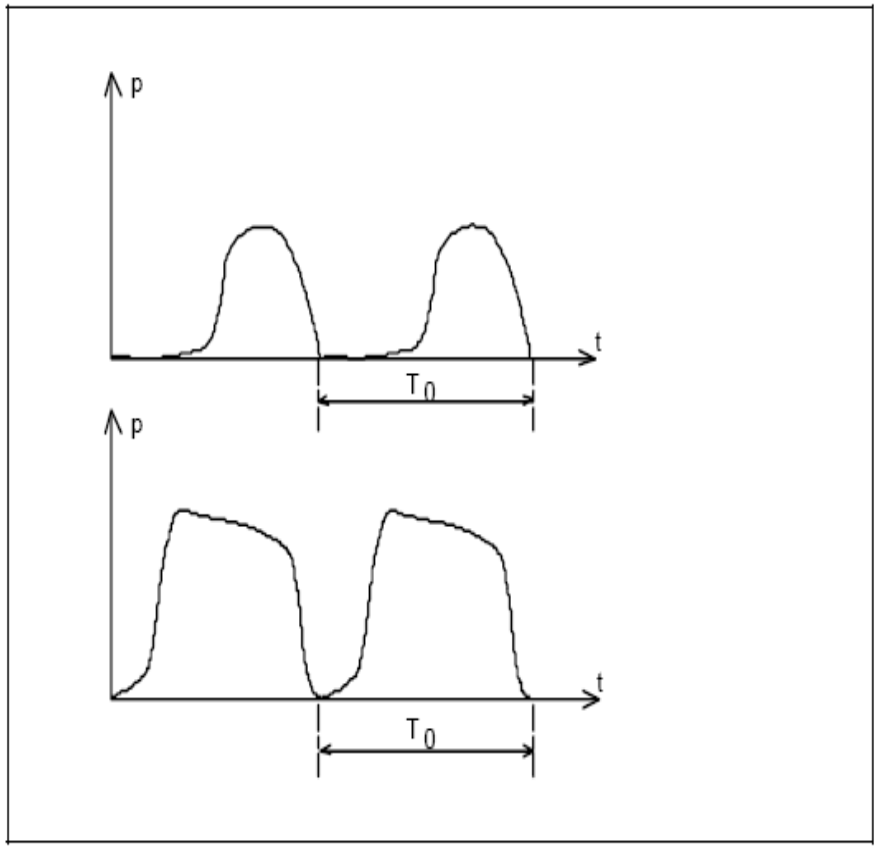
2. SISTEME DE SINTEZĂ A VOCII PE BAZA UNUI MODEL

SINTEZA FORMANTICĂ – 1950, Lawrence şi Fant.

- In timp ce SV parcurge distanţa de la glotă până la buze, spectrul larg al excitaţiei este modulat datorită selectivităţii în frecvenţă a tractului vocal, rezultând formanţii
- semnalul este generat printr-un sistem, ***care simulează transmitanţa tractului vocal, definită prin frecvenţele de rezonanţă (formanţi-rez. Ord.2)-*** caracterizat prin frecvenţa de rezonanţa Fk şi banda de trecere Bk.



REALIZAREA EXCITATIEI



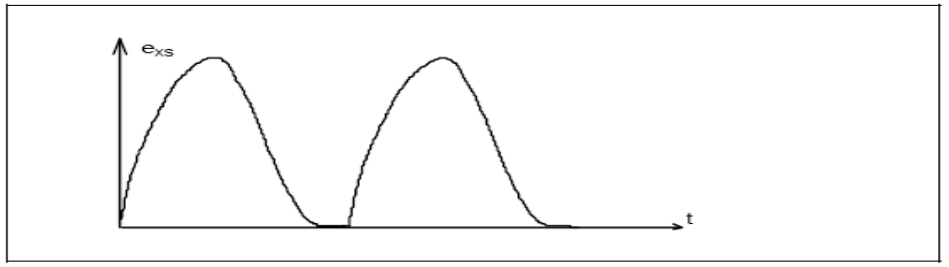
. Forme de undă glotice (Flanagan)

Reproducerea spectrului glotal se realizeaza cu diferite aproximaţii :

- aproximaţia triunghiulară

- aproximaţia polinomială

- excitarea unui filtru trece jos de ordinul doi, printr-un tren de impulsuri



Undă glotică folosită de Fant

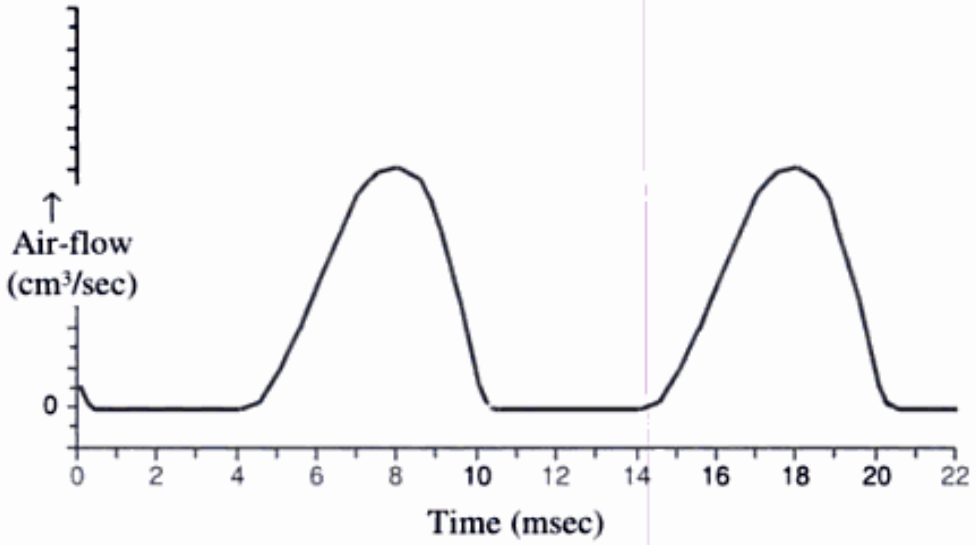


Figure 8.8 Two periods of an idealised glottal volume velocity wave with an F0 of 100 Hz

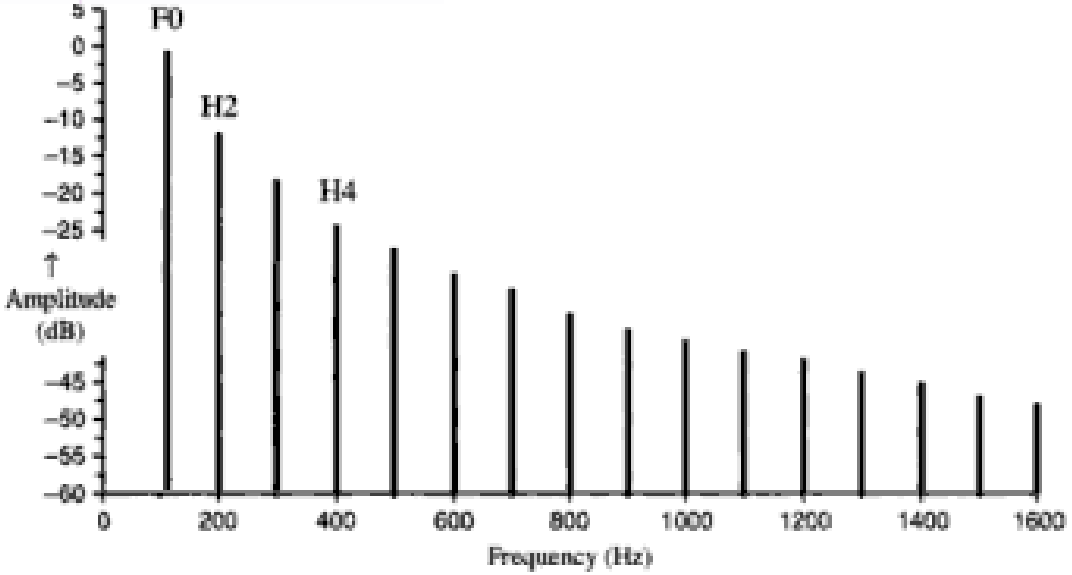
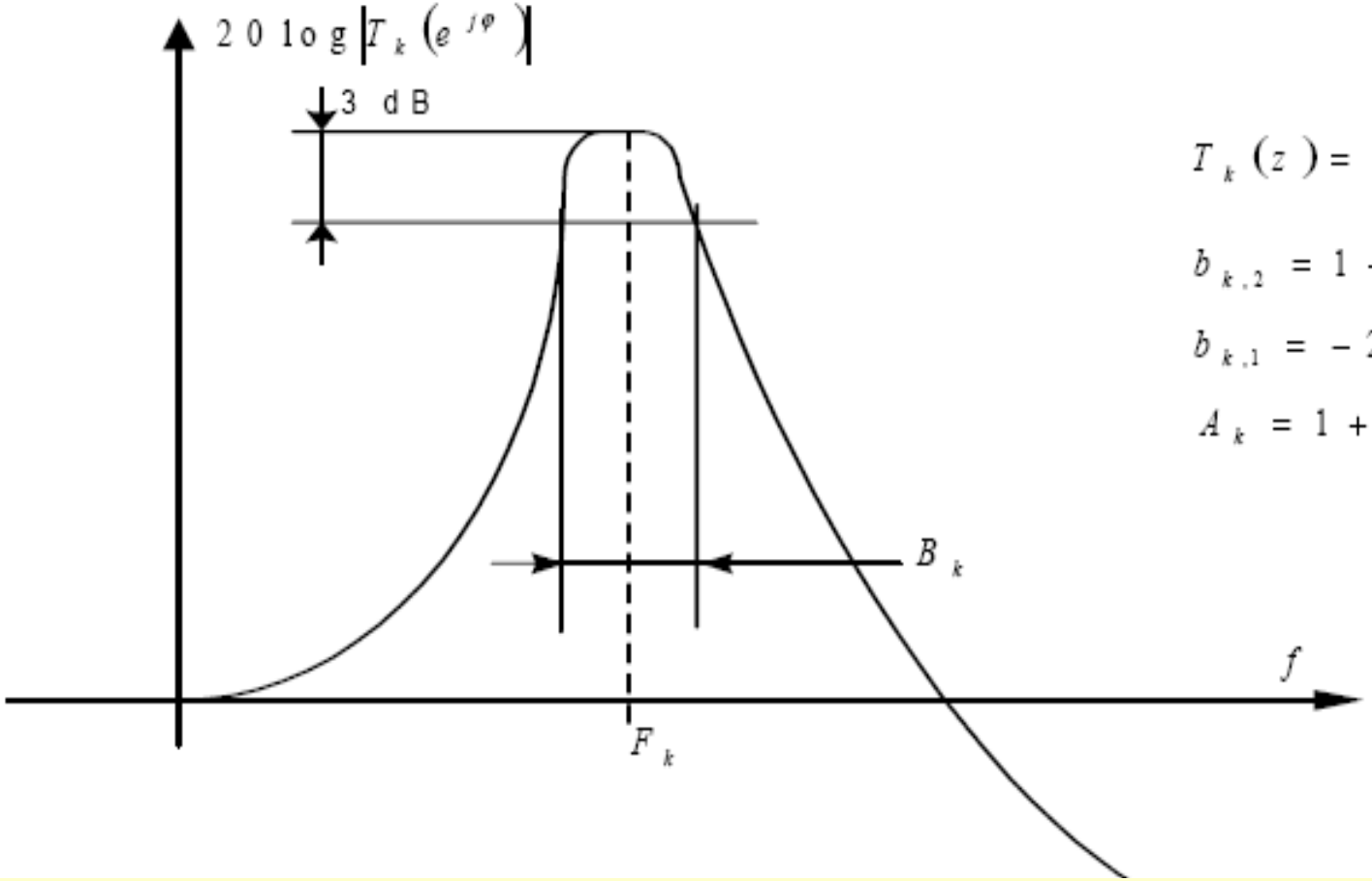


Figure 8.9 Idealised spectrum (up to 1.6 kHz) of the glottal volume velocity wave shown in Figure 8.8. The fundamental (F0) and second and fourth harmonics (H2, H4) are indicated

2.1. SINTEZA FORMANTICĂ - SINTEZA ÎN CASCADĂ

- Acest tip de sinteza constă în înserierea a 3-4 rezonatori de ordinul 2, care realizează fiecare, câte un formant de frecvenţa F_k şi banda de trecere B_k
- Transmitanţa se normalizeaza pentru a obţine câştig nul la frecvenţă zero.



$$T_k(z) = \frac{A_k}{1 + b_{k,1} z^{-1} + b_{k,2} z^{-2}}$$

$$b_{k,2} = 1 - 2\pi B_k / f_e$$

$$b_{k,1} = -2 b_{k,2} \cos 2\pi F_k / f_e$$

$$A_k = 1 + b_{k1} + b_{k2}$$

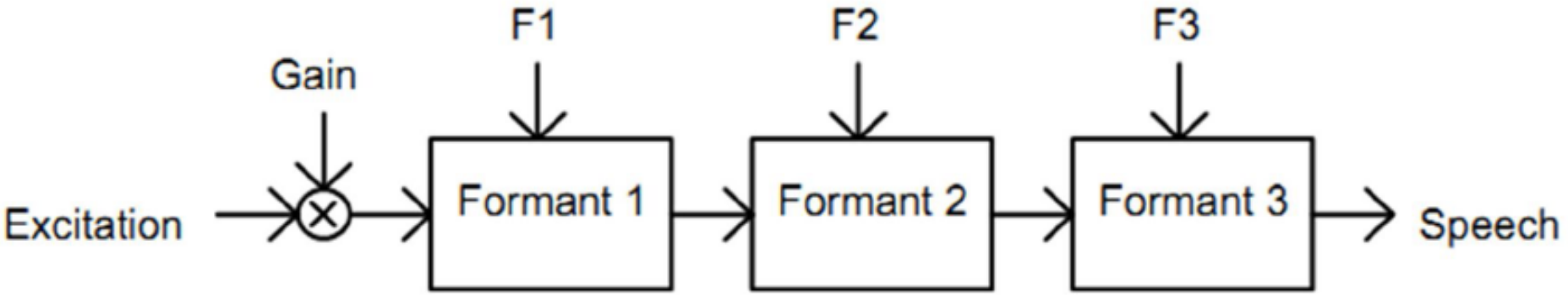
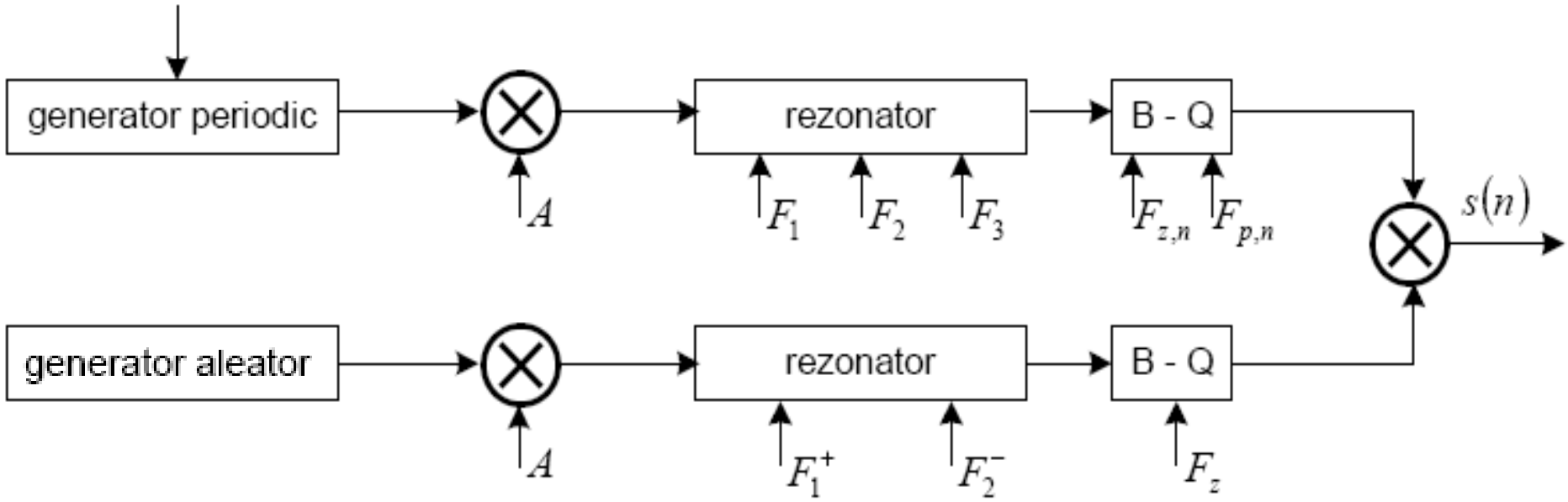


Fig. 3. Basic structure of cascade formant synthesizer [2]

Configuraţia serie - se bazează pe ***funcţia de transfer numai poli a tractului vocal***, ceea ce este ***insuficient*** pentru reprezentarea tuturor sunetelor vocii
Amplitudinea formanţilor superiori este funcţie de a formanţilor inferiori, deoarece aceştia din urmă introduc în spectru o pantă de -12 dB/octavă. Acest fenomen nu se poate simula prin înserierea filtrelor



$$V_k(z) = \frac{B_k}{1 + b_{1_k} z^{-1} + b_{2_k} z^{-2}}$$

$$b_{2_k} = 1 - 2\pi B_k / f_{es}$$

$$b_{1_k} = -2b_{2_k} \cos(2\pi F_k / f_{es})$$

$$B_k = 1 + b_{1_k} + b_{2_k}$$

$$T(z) = \prod_{k=1}^{K} \frac{B_k}{1 + b_{1_k} z^{-1} + b_{2_k} z^{-2}}$$

	F (Hz)	B (Hz)
Sunete sonore		
Primul formant (F_1)	100 – 1100	45 – 130
Al doilea formant (F_2)	500 – 2500	5 – 190
Al treilea formant (F_3)	1500 – 3500	70 – 260
Rezonanță nazală ($F_{p,n}$)	200 – 1000	100
Antirezonanța ($F_{z,n}$)	200 – 1000	
Sunete nesonore		
Primul formant	200 – 500	60 – 300
Al doilea formant	1500 – 3500	60 – 200
Antiformant	800 – 2000	

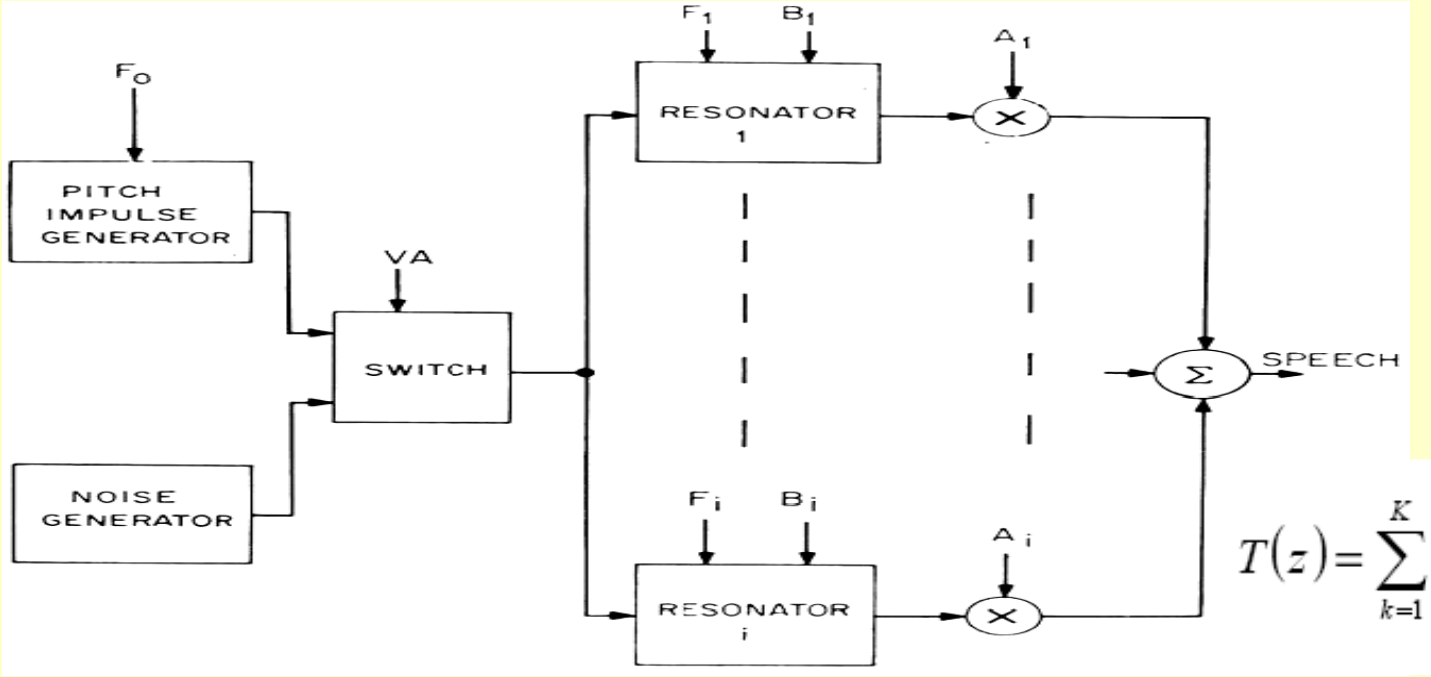
- In tabel se indică domeniile de frecvenţe ale parametrilor F_k şi B_k (sunete sonore), $F_{z,n}$ şi $F_{p,n}$ (sunete nazale) respectiv, şi F_z (sunete fricative), valori determinate statistic

2.2. SINTEZA FORMANTICĂ - SINTEZA ÎN PARALEL

- Este mai delicată decât în cazul sintezei în cascadă, deoarece câştigul asociat fiecărui formant trebuie cunoscut

-Aceste câştiguri corespund rezidurilor polilor rezultaţi, din descompunerea transmitanţei T(z) în fracţii simple

-Calcularea şi realizarea lor trebuie să se facă cu mare precizie, fără a introduce zerouri în transmitanţă



$$T(z)=\sum_{k=1}^{K}\frac{h_k}{1+b_{1_k}z^{-1}+b_{2_k}z^{-2}}$$

Configuraţia paralelă - ***filtre cu doi poli complex conjugaţi*** este mai generală

- Se poate obţine o funcţie de transfer cu polii şi zerourile dorite motiv pentru care această metodă este convenabilă pentru generarea sunetelor care conţin zerouri (de ex. sunetele nazale).

Avantaje

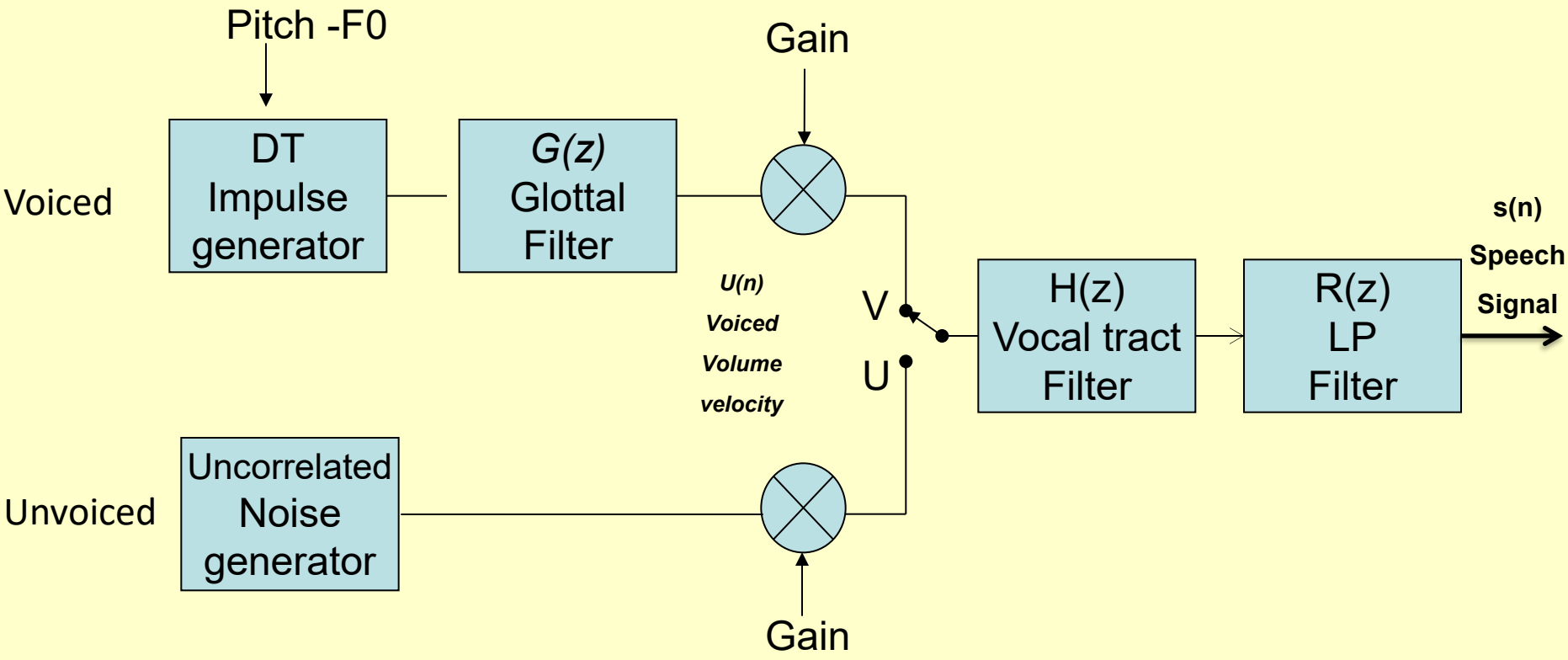
- Parametrii de sinteză sunt în ***relaţie strânsă cu producerea şi propagarea sunetului prin tractul vocal***
- Dacă se respectă condiţia continuităţii în evoluţia parametrilor, sintetizoarele formantice pot genera sunete sintetice cu ***sonoritate plăcută***
- Pe lânga tipurile de excitaţie uzuale (sonor, nesonor) sintetizoarele formantice permit utilizarea ***excitaţiei mixte***

Dezavantaje

- Problema majoră a sintetizoarelor formantice este **obţinerea datelor** cu care se va opera
- *Urmărirea traiectoriei formanţilor (formant tracking)* este o sarcină foarte **dificilă** datorită faptului că această analiză nu poate fi automatizată complet
- Determinarea *lăţimilor de bandă* aferente este şi mai *problematică* (formanţii sunt apropiaţi cu benzi suprapuse)

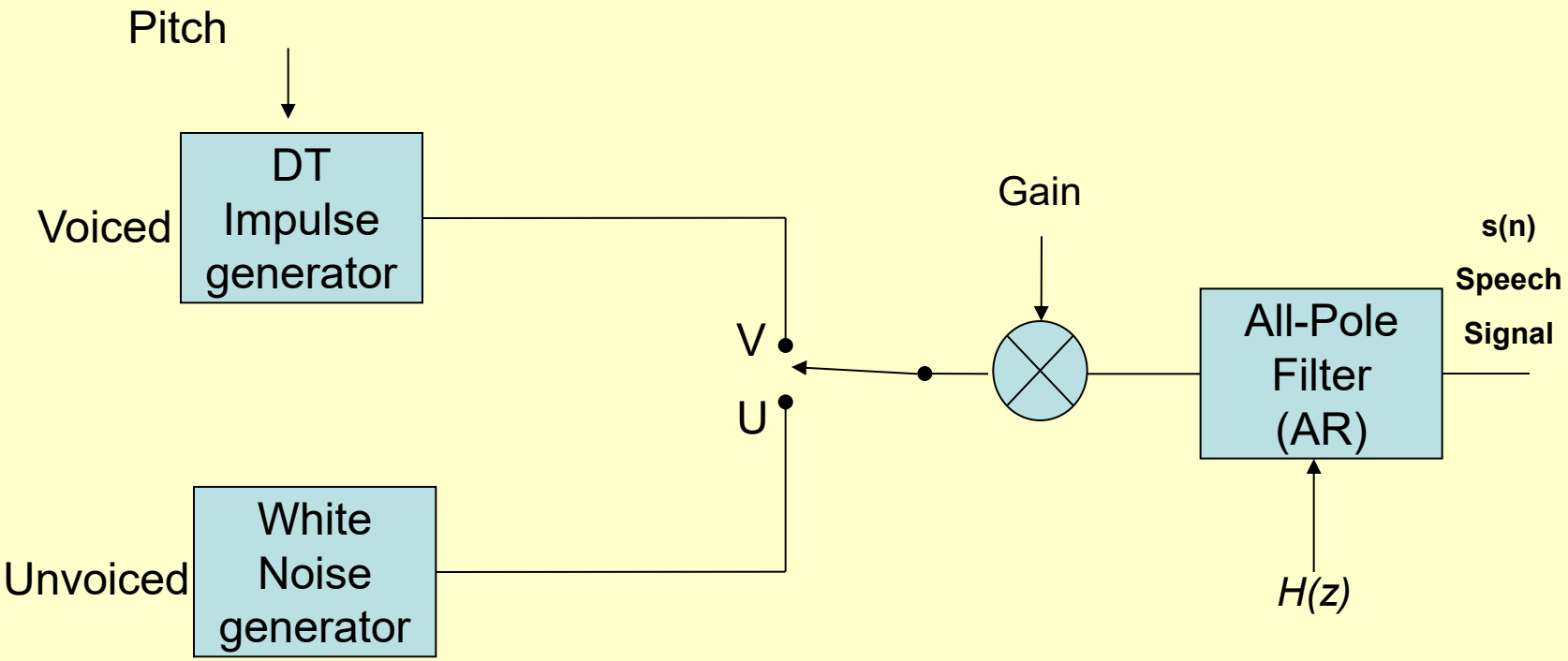
2.3. SINTEZA BAZATĂ PE PREDICŢIA LINIARĂ

- Modelul Real:



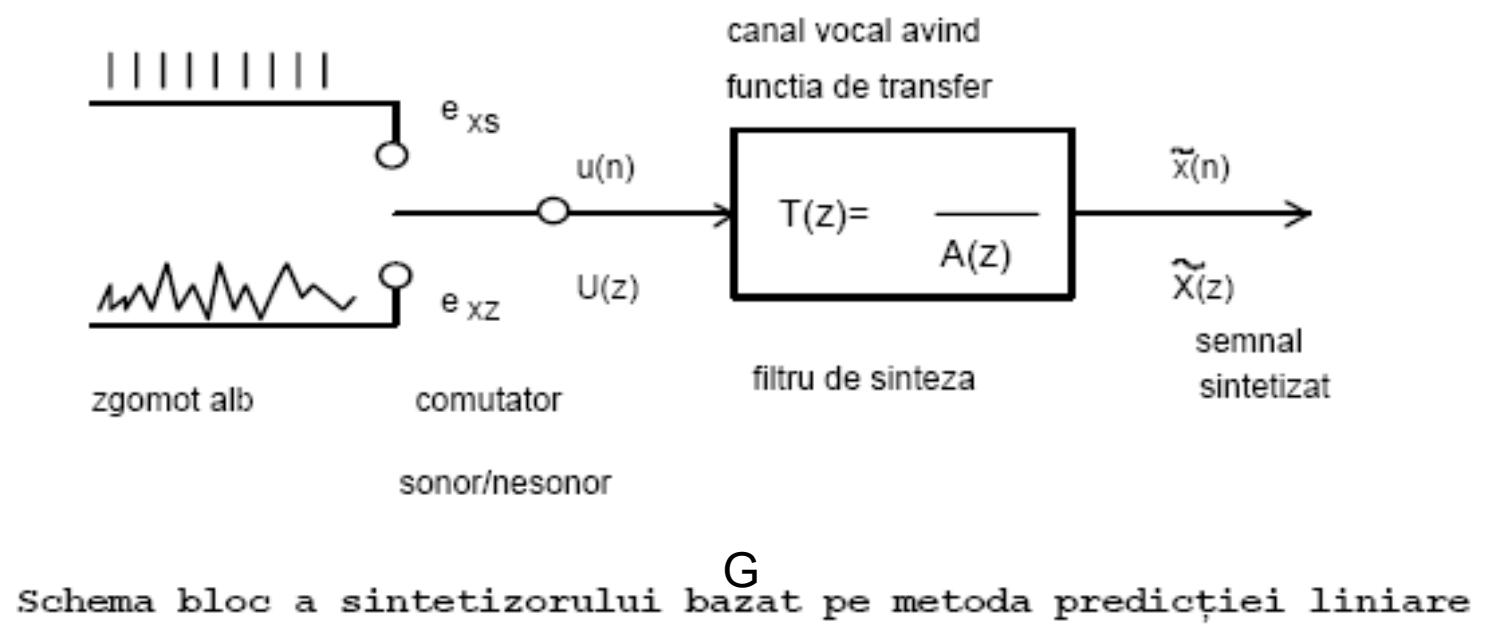
Modelul LPC real de producere a SV

- Modelul LPC simplificat utilizat in analiza



Modelul LPC

- are ca obiect estimarea parametrilor sistemului susceptibil de a genera un semnal artificial cât mai apropiat de semnalul original x.
- excitaţia se realizează fie cu un tren periodic de impulsuri, fie un zgomot alb
- pe lângă transmitanţa canalului vocal se ia în considerare şi forma reală de excitaţie (unda glotala)



Schema bloc a sintetizorului bazat pe metoda predicției liniare

Parametri modelului LPC:

- coeficienţii filtrului numeric A(z), p
- decizia sonor/nesonor
- F0
- Câştigul, G

Decizia sonor/nesonor este un parametru necesar pentru care s-au elaborat diferite metode altele fata de cele uzuale:

1.metoda bazată pe transformata Fourier - Se ştie că energia cadrelor sonore este concentrata în partea inferioară spectrului (0-Fes/4), iar energia cadrelor nesonore este concentrat în partea superioara a spectrului (Fes/4-Fes/2).

2. metoda bazată pe compararea energiei SV original cu a celui derivat. Experimental s-a constatat că *energia semnalului sonor original este mai mare decat a celui derivat,* iar în cazul semnalului nesonor acest raport este inversat.

3. metoda bazată pe NTZ a SV derivat. NTZ a semnalului derivat în cadrele nesonore a semnalului vocal este mai mare decât în cadrele sonore.

4. metoda bazată pe NTZ al derivatei funcţiei de autocorelaţie a SV original. NTZ al derivatei funcţiei de autocorelaţie a SV este considerabil mai mare în cazul semnalelor nesonore, decat în cazul semnalelor sonore.

Avantaje:

- Predicţia liniară a devenit o metodă simplu de utilizat pentru simulare pe calculator
- Algoritmii utilizaţi pot fi uşor implementaţi hard
- Procesul de analiză este complet automatizat
- Calitatea vocii sintetice este destul de bună, în special din punctul de vedere al timbrului

Dezavantaje:

- decizia între excitaţia sonoră/nesonoră nu este întotdeauna satisfăcatoare
- Metodele noi (RELP, Modelul sursei mixte etc.) care şi-au propus înlăturarea acestui neajuns, dar au tendinţa de a schimba timbrul vocii, deoarece excitaţia are un spectru plat mai intins

3. SISTEME DE SINTEZĂ A VOCII PRIN SIMULAREA TRACTULUI VOCAL

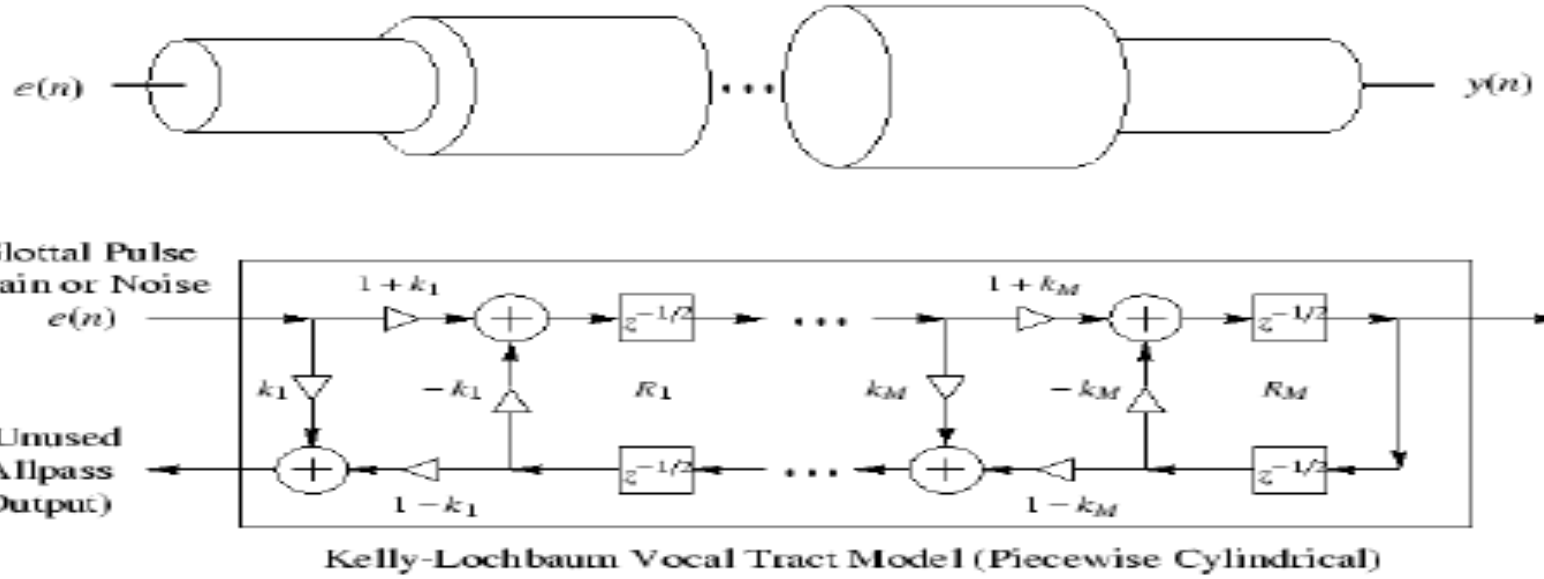
- **Principalul neajuns** al metodelor prezentate anterior este faptul că acestea ***neglijează interacţiunea dintre sursă şi tractul vocal***

- s-a constatat că diferite configuraţii ale tractului vocal pot produce sunete cu spectru identic

- altă abordare pentru descrierea acustică a tractului vocal este ***simularea directă a generării şi propagării undelor sonore*** în interiorul sistemului fonator

- simularea acusticii tractului vocal poate fi combinată cu un model articulator
- sunt 2 metode capabile sa produca SV prin replicarea electro-acustica a mecanismului de producere a vorbirii:

- ***METODA ANALOGIEI TRACTULUI VOCAL*** - care simuleaza propagarea undei acustice prin tractul vocal
- ***METODA ANALOGIEI TERMINALE*** - care simuleaza structura spectrului de frecventa (caracteristicile rezonantelor si antirezonantelor care reproduc procesul de articulare)



Kelly-Lochbaum Vocal Tract Model (Piecewise Cylindrical)

3.1. METODA ANALOGIEI TRACTULUI VOCAL

- datele referitoare la forma tractului vocal sunt furnizate de *modelul articulator,* respectiv de *modelul corzilor vocale*

- Modelele propuse pentru simularea vibraţiilor corzilor vocale:
  - **one-mass** de J. L. Flanagan şi J. L. Landgraf;
  - **two-mass** de K. Ishizaka şi J. L. Flanagan;
  - **multi-mass** de I. R. Tietze;
  - **two-beam** de R. Descont, J. Y. Aulage şi B. Guerin

- *încearcă să calculeze cantitatea de aer glotal rezultată prin parametri ca: presiunea subglotală, tensiunea corziilor vocale şi forma tractului vocal*
- simularea propagării sunetului în interiorul tractului vocal *în domeniul timp şi în domeniul frecvenţă*

Punctul de plecare îl constituie **ecuaţiile Webster:**

$$\frac{\delta^2 p}{\delta x^2} + \frac{1}{A}\frac{\delta p}{\delta x}\frac{\delta A}{\delta x} = \frac{1}{C^2}\frac{\delta^2 p}{\delta t^2} \qquad A\frac{\delta}{\delta x}\left(\frac{1}{A}\frac{\delta U}{\delta x}\right) = \frac{1}{C^2}\frac{\delta^2 U}{\delta t^2}$$

unde **p(x,t) si U(x,t)** sunt **presiunea,** respectiv **viteza volumică** în momentul **t** şi poziţia **x**, iar **A(x)** este aria secţiunii transversale a tractului vocal. ****

3.2. METODA ANALOGIEI TERMINALE

- simuleaza producerea SV printr-un circuit electric constand din cascadarea sau punerea in paralel a unor circuite care descriu rezonantele (formantii) si antirezonantele (antiformantii)

-F_k si B_k ale fiecarui circuit sunt variabile (metoda Formantica)

- conexiunea in cascada este avantajoasa in stabilirea automata a raportului intre amplitudinile formant/antiformant (vocale-structura spectrala)

- conexiunea paralela e avantajoasa ptr ca forma spectrala finala poate fi mai precis simulata (nazale, fricative)

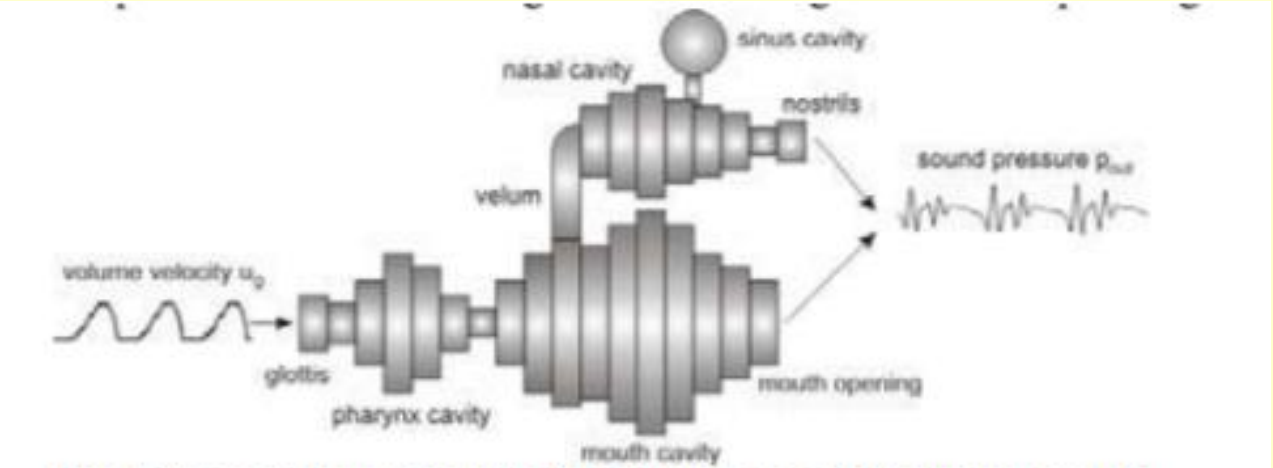


Fig. 2: An articulatory model of the human speech production system

4. SINTEZA PE BAZĂ DE REGULI

- se construiește vocea prin reguli fonetice, fonologice și fonetico-prosodice derivând direct din descrieri lingvistice ale limbii țintă. (pe baza unor secvenţe de simboluri *fonetice / silabice / text)*
-descrierea fonetică a textului (foneme, reguli fonetice), reguli de pronunție, reguli de transformare din text în formă canonică (etalon)
- modele de prosodie: reguli pentru intonație, creșterea/reducerea intensității (accentuare), durata fonemelor, pauzele.
- interpretabilitate ridicată: parametrii sunt în mod explicit controlați, facilitând studierea relațiilor între text și vorbire.
-adecvate pentru vocabular limitat sau când este nevoie de control strict asupra caracteristicilor vocale.
-naturaletea și eficienta la vocabulare mari sunt adesea limitate; adaptarea la dialecte/limbi diferite necesită efort semnificativ în definirea regulilor.

- caracteristicile *unităţilor fundamentale* de vorbire (silabe, foneme sau vorbirea pe T0) sunt stocate şi se concatenează pe baza unor reguli
- Caracteristicile prozodice (F0, A,..) sunt de asemenea controlate prin reguli
- *Calitatea unităţilor fundamentale* pentru sinteză ca şi *regulile de control* joacă un rol esenţial în această metodă şi trebuie să se bazeze pe caracteristicile fonetice şi lingvistice ale vorbirii naturale specific limbii respective
-In cazul sistemelor *bazate pe forma de undă:* daca unităţile fundamentale pentru sinteză sunt *fonemele* (extrase din vorbirea naturală sau artificială) *memoria* necesară stocării este *foarte mică* (de obicei sunt 30-50 de foneme), *dar regulile de conectare ale fonemelor sunt foarte complicate, iar calitatea sintezei este redusa*
-se utilizează frecvent unităţi mai mari decât *fonii sau alofonii* (foneme dependente de context) - pentru o sinteză de calitate sunt necesare alte unităţi fundamentale ~ 10^3-10^4

Table 16.4 Unit types in English assuming a phone set of 42 phonemes. Longer units produce higher quality at the expense of more storage. The number of units is generally below the absolute maximum in theory: i.e., out of the $42^3 = 74{,}088$ possible triphones, only about 30,000 occur in practice.

Unit length	Unit type	#Units	Quality
Short	Phoneme	42	Low
↓	Diphone	~1500	↓
	Triphone	~30K	
	Demisyllable	~2000	
	Syllable	~11K	
	Word	100K–1.5M	
	Phrase	∞	
Long	Sentence	∞	High

Alegerea unitatilor

Figure 16.2 Block diagram of a synthesis-by-rule system. Pitch and formants are listed as the only parameters of the synthesizer for convenience. In practice, such system has about 40 parameters.

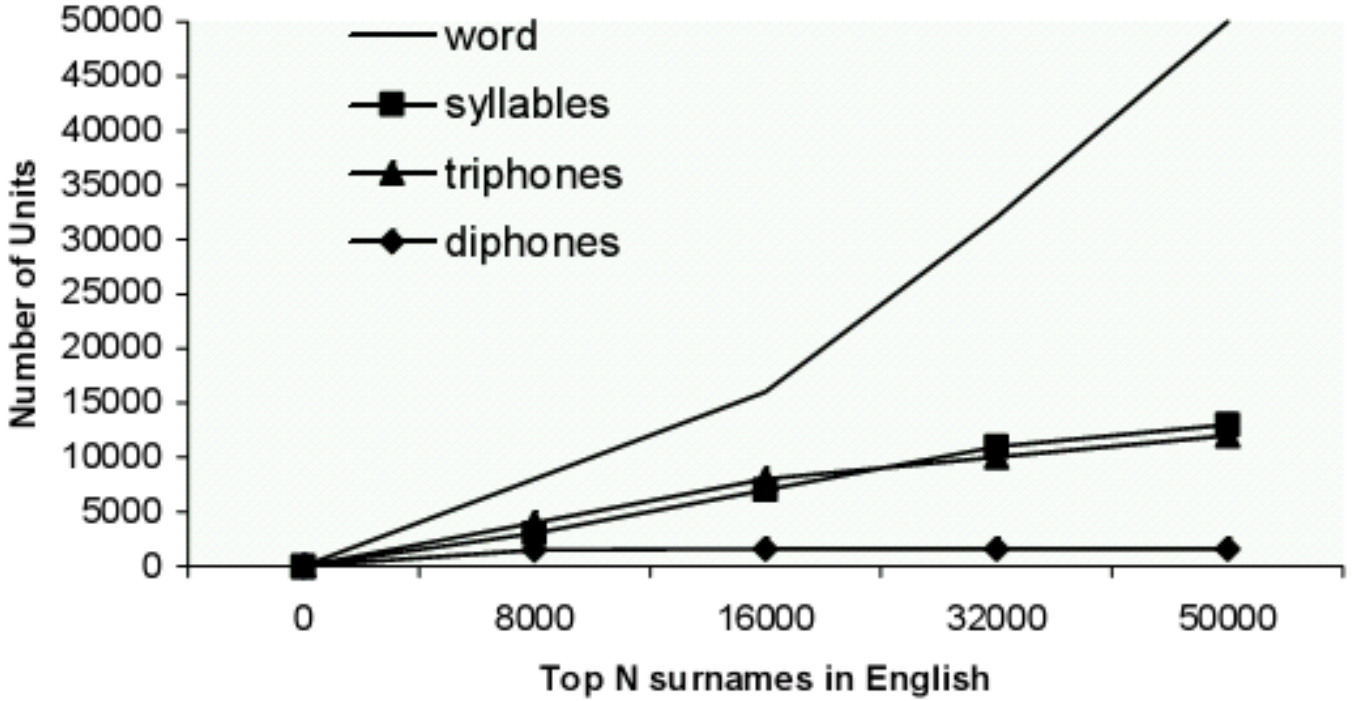


Figure 16.5 Coverage with different number of units displays the number of units of different types required to generate the top N surnames in the United States [34].

CONCATENAREA UNITATILOR fundamentale

- CUVINTE

-Nu are acoperire completă pentru domenii largi

-Capacitatea limitată de a modifica F0, amplitudinea şi durata fără a pierde din naturaleţe şi inteligibilitatea

- Necesitatea unei baze de date imense pentru a extrage mai multe versiuni ale fiecărui cuvânt

- SUB-UNITATI (silabe, demisilabe, difoni,…)

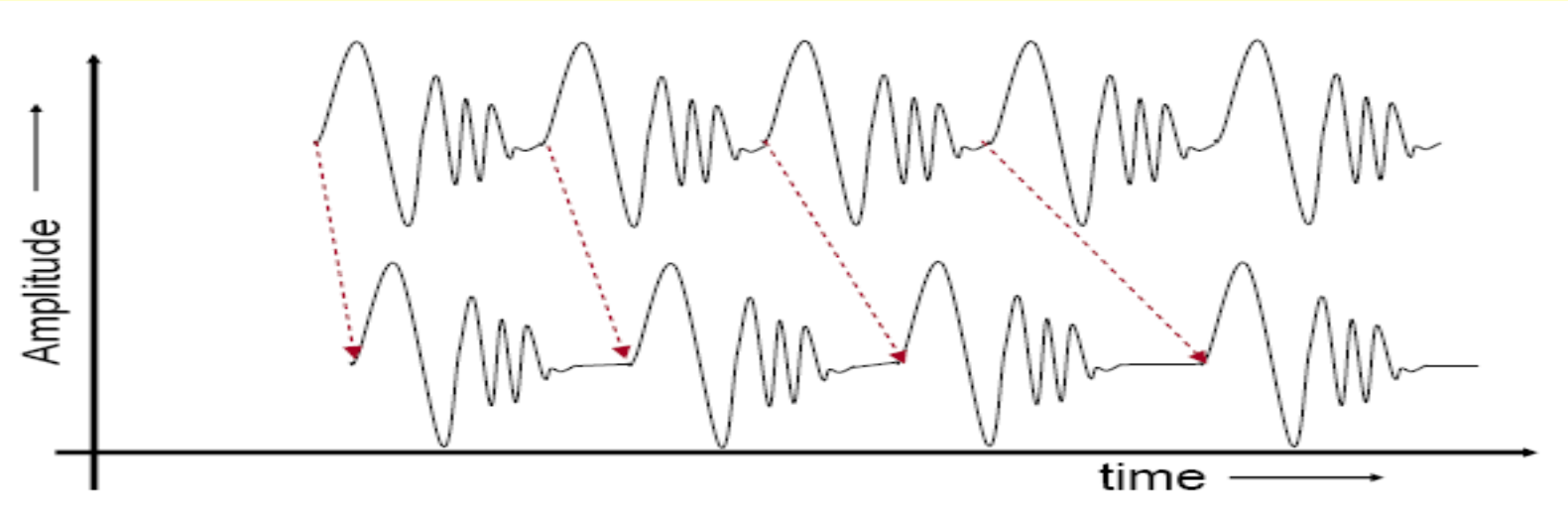
- *greu de izolat în context din cauza coarticularii*
- necesita variante alofonice pentru a caracteriza subunităţile în toate contextele in care apar
- se cere efort mare de prelucrare a semnalului pentru a netezi conexiunile cu unitatile adiacente

CONCATENAREA UNITATILOR REPREZENTATE

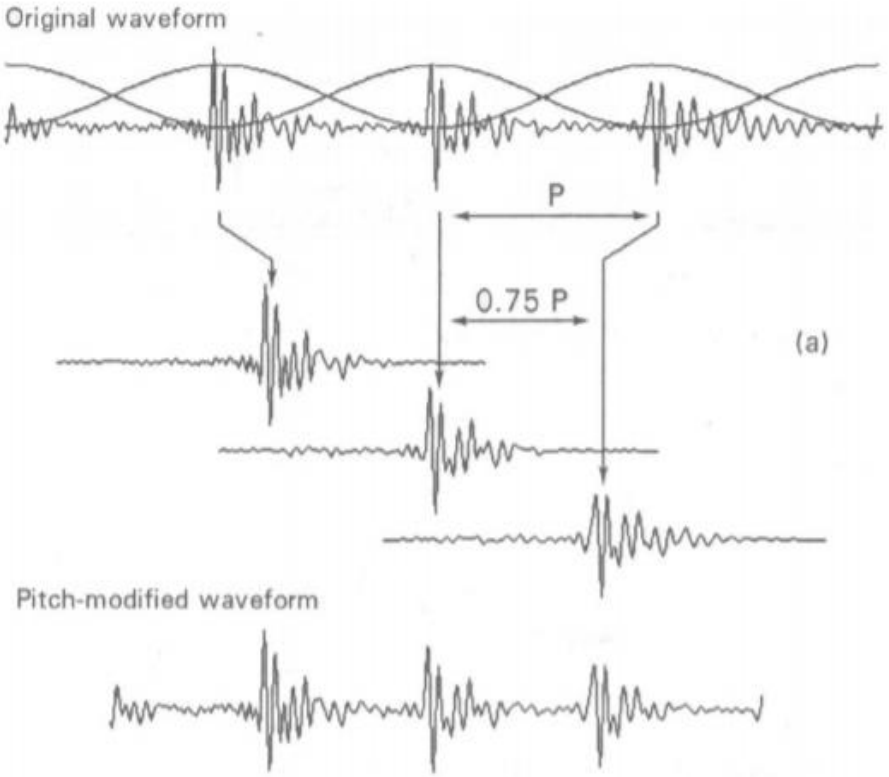
- LPC

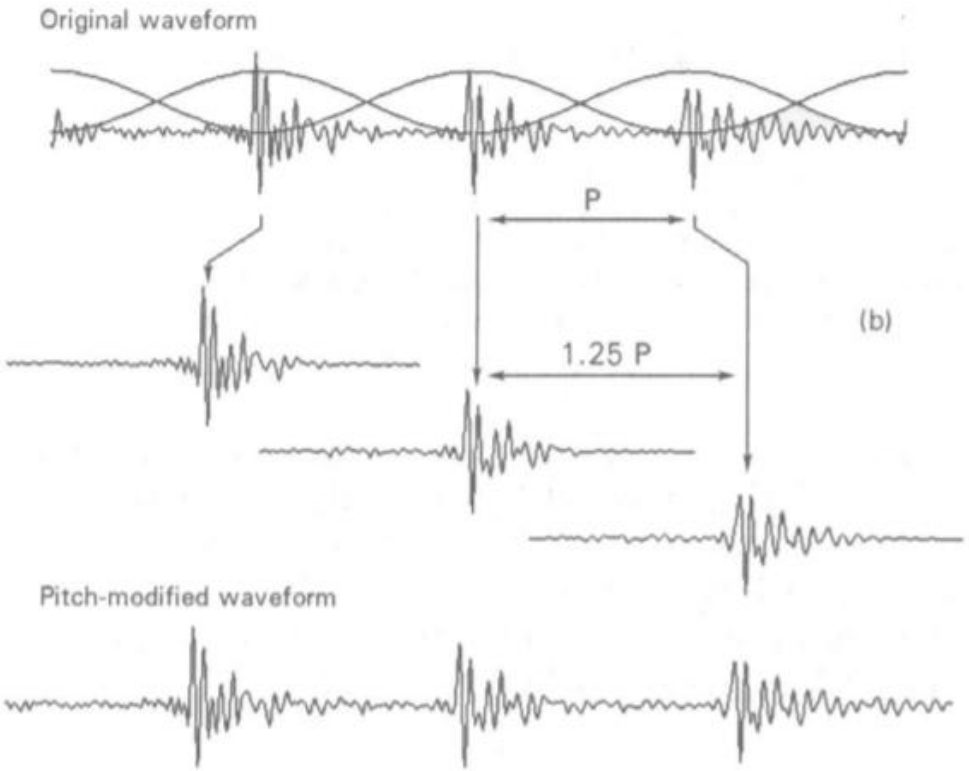
-Simplu, uşor de concatentat unităţile, eficient pentru modificarea F0
- Nu funcţionează bine pentru sunete nazale (lipsa zerouri nazale)
- Excitaţie glotala nu este corecta (excitatia asumata ca impulsuri)
- LPC - nu funcţionează bine pentru excitaţii mixte

- TD-PSOLA—Time Domain, Pitch Synchronous Overlap Add Synthesis - Modificarea eficienta a prozodiei (sincron cu F0)
  - Nu cere o uniformizare la punctele de jonctiune

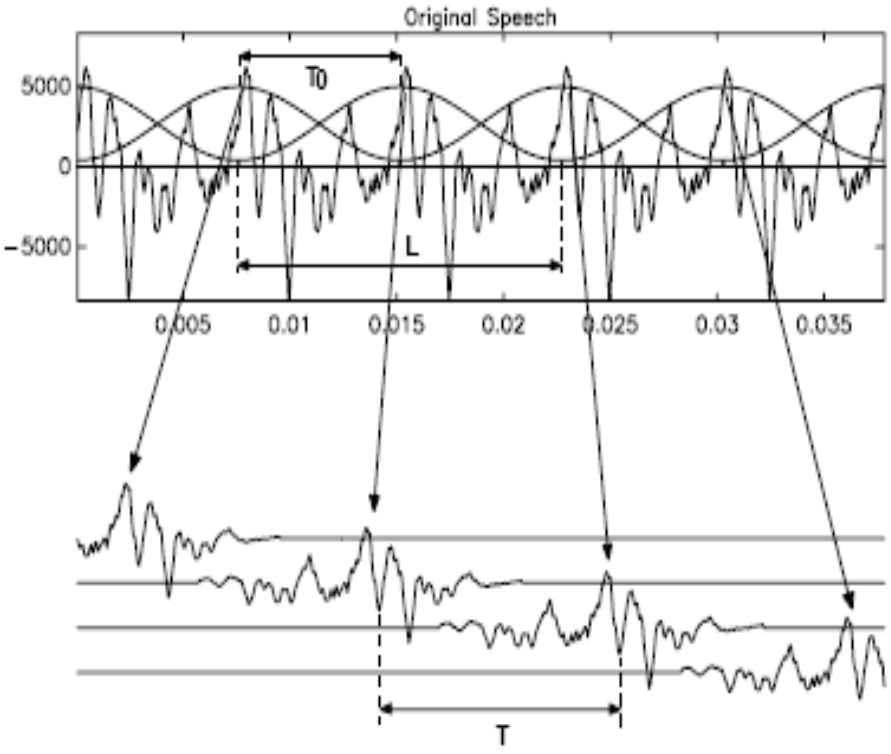


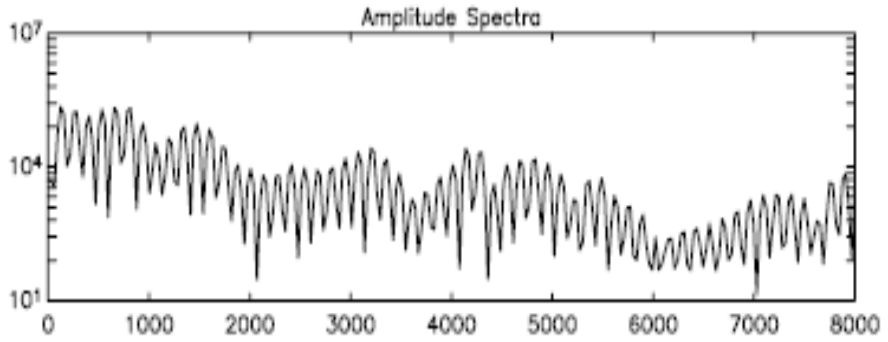
Pitch-scaling

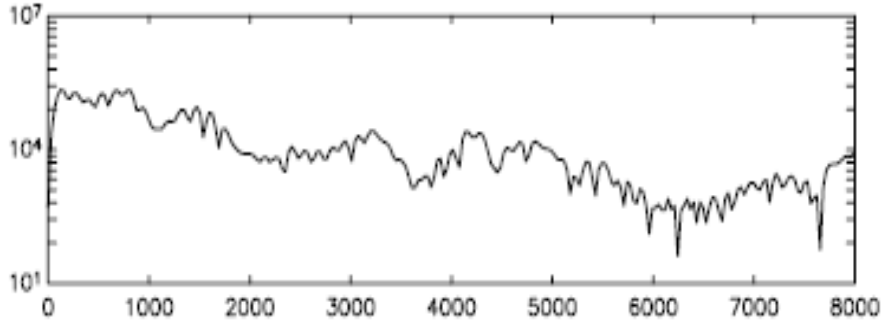


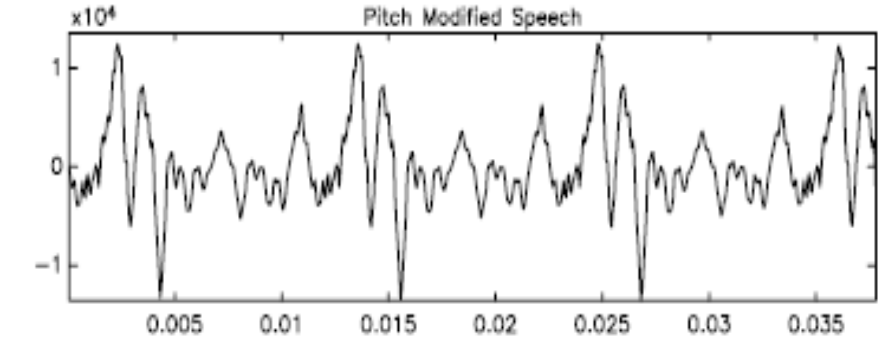


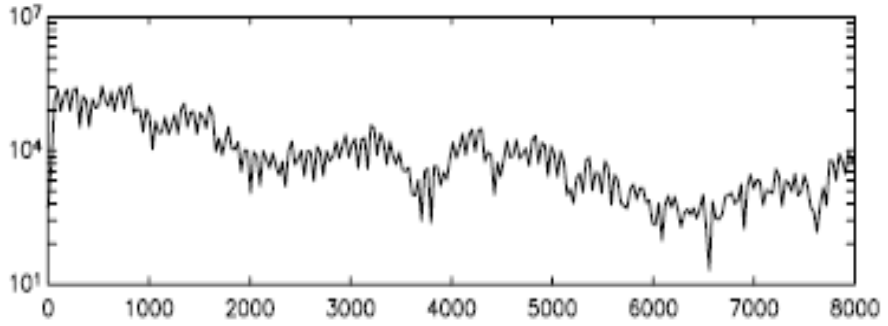
SINTEZA TD-PSOLA











MODIFICARILE TEMPORALE DUC LA DISTORSIUNI SPECTRALE

(MBROLA- The *M*ulti-*B*and *R*e-synthesis *O*ver*L*ap-*A*dd)

- sinteza in domeniu timp care combina avantajele PSOLA (calcul redus, modificarea prozodiei), cu un algoritm hibrid timp–frecvenţă, off-line pentru uniformizarea tranziţiilor.

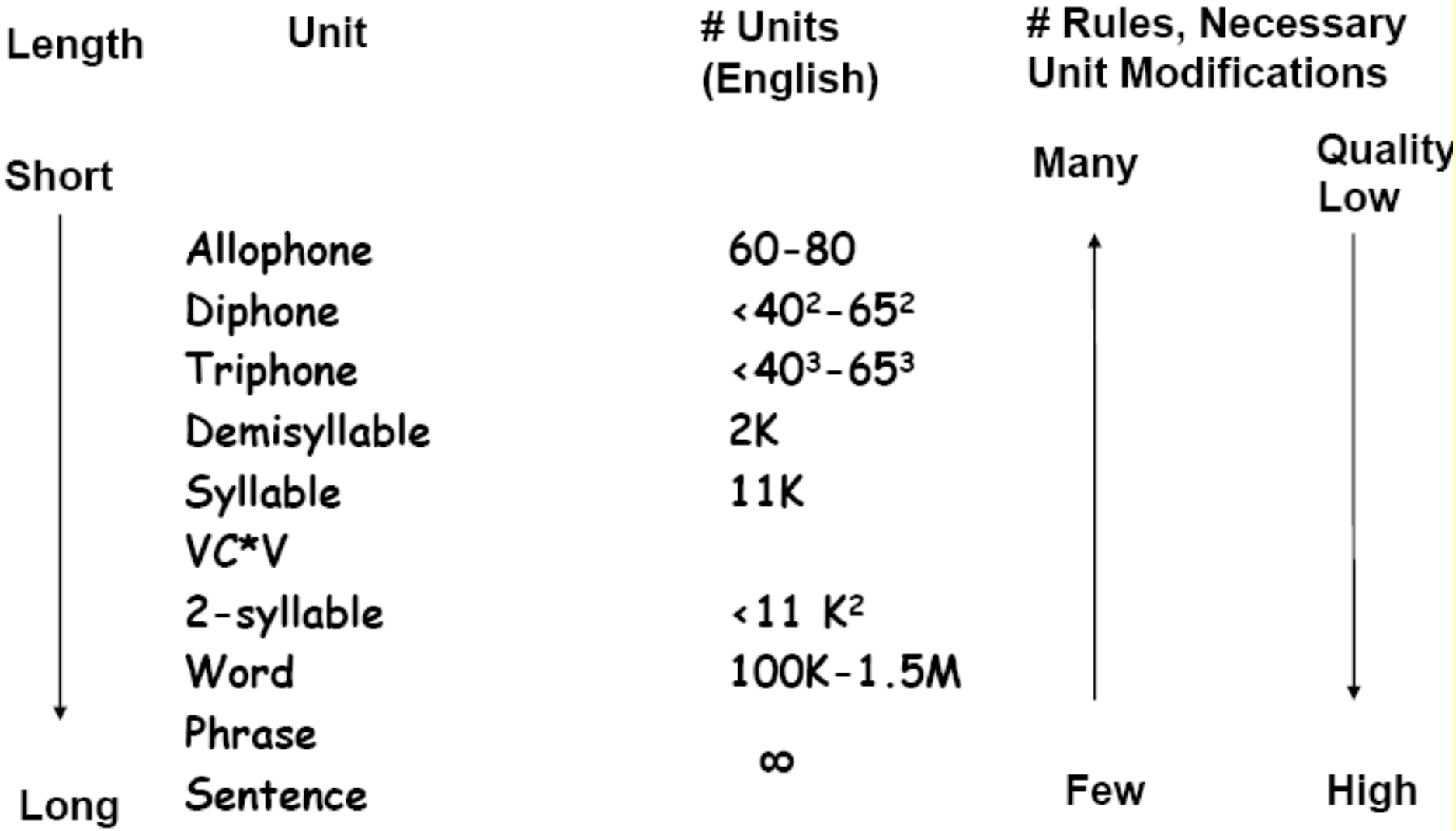
Exemplu.

Pentru limba japoneză este nevoie de ~200 de silabe de tip CV/VC (consoană–vocală) sau 5000-6000 de unităţi de tip CVC (putând fi reduse la ~1000) sau între 700-800 de unităţi de tip VCV.
De exemplu cuvântul SAKURA = cireş poate fi reprezentat cu unităţi :

CV : SA+KU+RA
CVC: SAK+KUR+RA
VCV: SA+AKU+URA.

Unităţile CVC sunt conectate pe consoană, iar cele VCV pe vocală fiecare prezentând anumite avantaje în uşurinţa de conectare.

Pentru limba engleză sunt > de 3500 de silabe care se extind la 10.000 dacă se consideră şi alofonii.
-silabele se descompun în diade/difoni (400-1000) sau demisilabe (~1000) care au reguli mai simple de concatenare.


Length
Unit
Units (English)
Rules, Necessary Unit Modifications
Short
Many
Quality
Low
Allophone
60-80
Diphone
<40²-65²
Triphone
<40³-65³
Demisyllable
2K
Syllable
11K
VC*V
2-syllable
<11 K²
Word
100K-1.5M
Phrase
∞
Sentence
Long
Few
High

5. Metode neuronale end-to-end actuale

Iată cateva din cercetările si abordările populare și actuale ale sintezei vocale:

- WaveNet: A Generative Model for Raw Audio
- Tacotron: Towards End-toEnd Speech Synthesis
- Deep Voice 1: Real-time Neural Text-to-Speech
- Deep Voice 2: Multi-Speaker Neural Text-to-Speech
- Deep Voice 3: Scaling Text-to-speech With Convolutional Sequence Learning
- Parallel WaveNet: Fast High-Fidelity Speech Synthesis
- Neural Voice Cloning with a Few Samples
- VoiceLoop: Voice Fitting and Synthesis via A Phonological Loop
- Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions

Tema-studiu individual.

- Noile abordari ale sintezei SV:

- (Wavenet) utilizarea retelelor neuronale, modelul folosit este probabilistic și autoregresiv

- (Tacotron) este un model generativ text-vorbire, care sintetizează vorbirea direct din perechi de text și audio. Tacotron generează vorbire la nivel de cadru și,prin urmare, este mai rapid decât metodele autoregresive la nivel de eșantion.

- (Deep Voice 1-3) este un sistem text-to-speech dezvoltat folosind deep neural networks, DNN.

- (Parallel WaveNet) Fast High-Fidelity Speech Synthesis - o metodă cunoscută sub denumirea de Probability Density Distillation, care antrenează o rețea paralelă feed-forward alimentata de la un WaveNet antrenat.

https://www.deepmind.com/blog/

https://heartbeat.fritz.ai/a-2019-guide-to-speech-synthesis-with-deep-learning-630afcafb9dd

Studiu.

https://www.techradar.com/best/best-text-to-speech-software

https://heartbeat.fritz.ai/a-2019-guide-to-speech-synthesis-with-deep-learning-630afcafb9dd