„Dacă există dorință, va exista o cale”.

ASRSV – curs 8

Sinteza Semnalului Vocal din Text (TTS synthesis)

- Introducere
- Analiza textului şi analiza lingvistică
- Sinteza vorbirii
- Sinteză bazată pe HMM (HTS)
- Noi abordari DNN

https://en.wikipedia.org/wiki/File:Fidelity_Chess_Challenger_Voice_speech_output.flac
Fidelity Voice Chess Challenger

- Producerea artificială a vorbirii umane este cunoscută sub numele de sinteză a vorbirii.
- Există câteva strategii tradiționale specifice pentru sinteza vorbirii: *concatenativa și parametrica.*
- În *abordarea concatenativă,* rostirile dintr-o bază mare de date sunt utilizate pentru a genera rostiri noi, audibile.
- În cazul în care este necesar un stil diferit de vorbire, este utilizată o altă bază de date cu rostiri. Aceasta limitează scalabilitatea acestei abordări.

- *Abordarea parametrică* folosește o voce înregistrată și o funcție cu un set de parametri care pot fi modificați pentru a schimba vocea.
- Aceste două abordări reprezintă modul clasic de a face sinteza vorbirii.

- Tehnologia bazată pe « învățarea profunda » este aplicabilă în sinteza text-to-speech, generare de muzică, generare de vorbire, dispozitive compatibile cu vorbirea, sisteme de navigație și accesibilitate pentru persoanele cu deficiențe de vedere.

MOTIVATIA dezvoltarii sistemelor TTS

- Creşterea popularitatii **sistemelor interactive** cu răspuns vocal, fac sistemele text-to-speech (TTS) tot mai atractive;
- Sistemele de mesagerie unificată **(UMS)** fac uz la accesul pe cale vocală la orice informaţie scrisă: fax, e-mail, baze de date text;
- Creşterea cererii de **sisteme de dialog,** inclusiv roboţi şi agenţi; (sistemele de dialog utilizeaza vocea ca intrare şi aşteaptă răspunsul în mod natural pe cale sonora de asemenea;)
- Acces vocal la **bazele de date** (liste de preţ, evenimente, telefon, banca…);
- **Sisteme de citire vocala** pentru persoanele cu deficienţe de vedere

Evaluarea calitatii sintezei

- De regulă, calitatea sintetizatoarelor sistemului TTS este evaluată din diferite aspecte, inclusiv inteligibilitatea, naturalețea și preferința vorbirii sintetice, precum și factorii de percepție umană, cum ar fi comprehensibilitatea.

Inteligibilitate: calitatea sunetului generat a fiecărui cuvânt produs într-o propoziție.
Naturaletea: calitatea vorbirii generate în ceea ce privește structura sa temporală, pronunția și redarea emoțiilor.
Preferință: alegerea de către ascultători a celui mai bun TTS; preferințele și naturalețea sunt influențate de sistemul TTS, calitatea semnalului și vocea, izolate și în combinație.
Comprehensibilitatea: gradul de înțelegere a mesajelor primite.

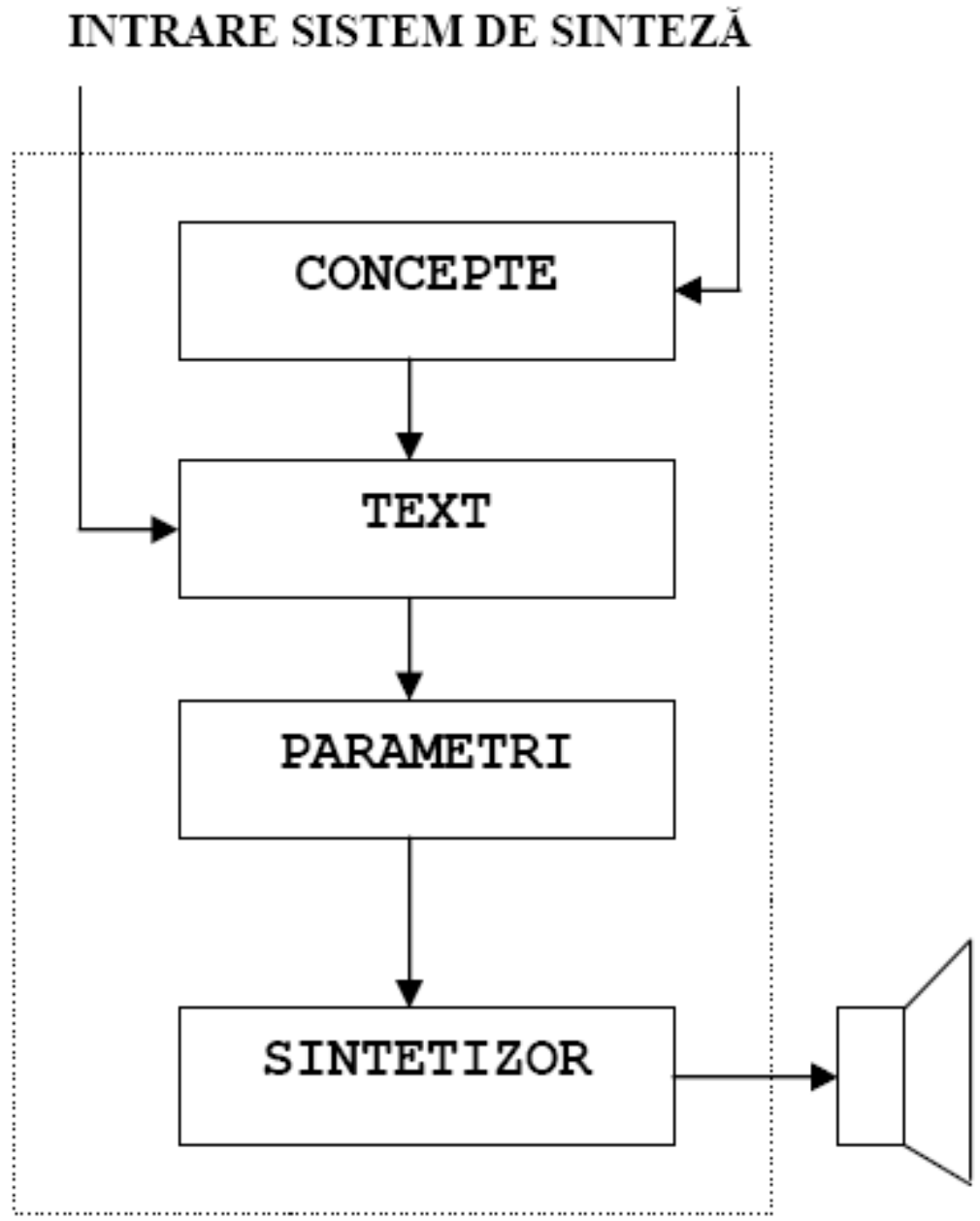


Fig. 1 Sistem de sinteză

SINTEZA TTS

- Conversia mesajelor text în voce sintetică cu caracteristicile dorite (pitch, viteza, amplitudine, mod de articulare, intonatie etc)
- Sistemele TTS ~ revers al sistemelor de recunoastere a vorbirii (Speech To Text)
- Dudley 1939 – VODER

- Problemele sintezei TTS:

- Pe langa pronunţarea corectă a cuvintelor
- un sistem TTS performant mai trebuie să mai rezolve:
  - *accentuarea sau nu a unor cuvinte;*
  - *divizarea propoziţiei în fraze intonaţionale (sens);*
  - *alegerea unui contur adecvat pentru F0 ;*
  - *durata unor cuvinte funcţie de poziţia ocupată în frază...*

Citirea e dificilă, de ce?

- lipsa in sistemele de scriere a unor specificatii importante pentru vorbire ~

Forma scrisă :

- cuvintele
- informatiile intonationale partial (!,?,.,)
- informatiile de separare pot lipsi (chineza, japoneza, thai)
  - (ideogramele chinezești indică numai sensul, nu și rostirea)
  - Ebraica (numai consoane)

Sarcina sistemelor TTS este complexă :

- încearcă imitarea cititorului uman
- având cunoştinţe gramaticale limitate
- neavând capacitatea de a înţelege ceea ce pronunţă (feedback)

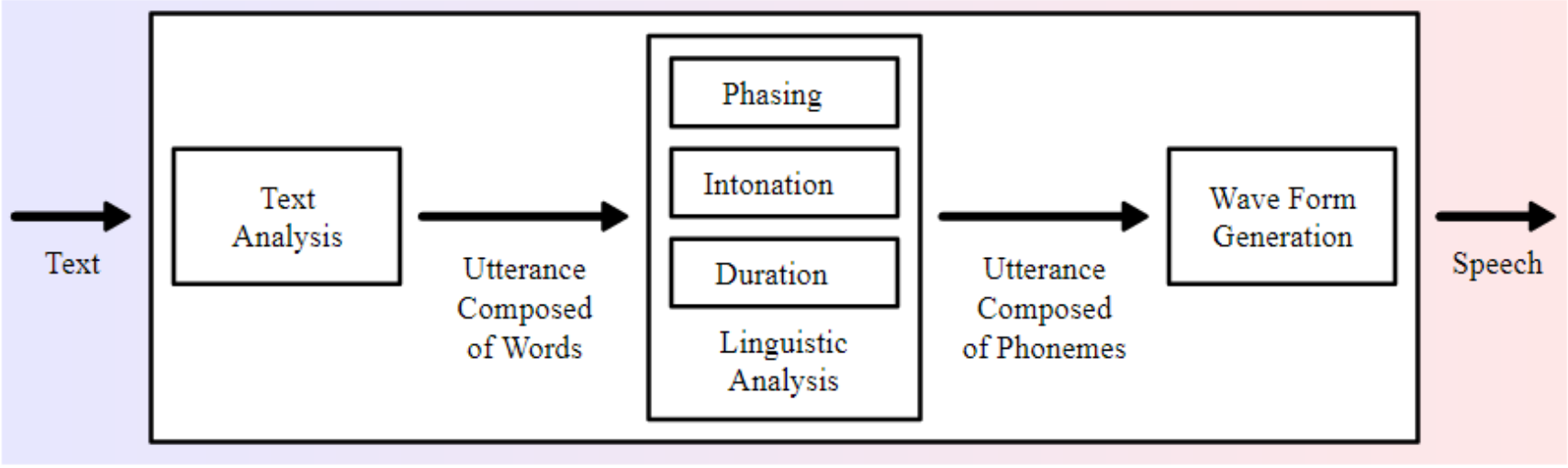
- **Problema sintezei TTS se divide în mod natural în două sub-probleme:**

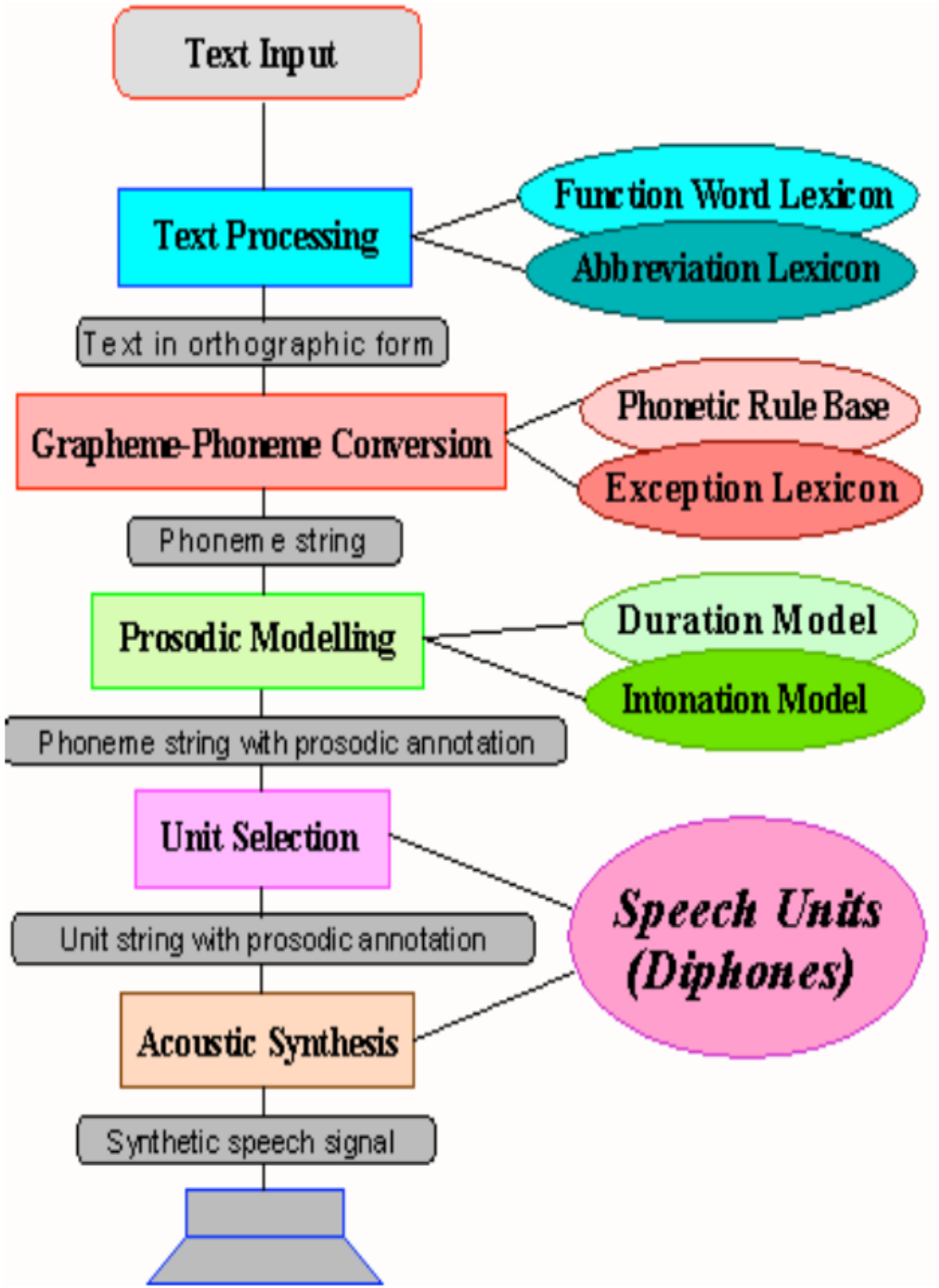
1. Conversia textului şi analiza lingvistică

-textul este o reprezentare imperfectă a limbajului în forme lingvistice de reprezentare, care includ informaţii legate de *fonemele* (sunetele/fonii) ce trebuiesc rostite*, duratele* lor, *localizarea pauzelor şi conturul* folosit pentru *F0.*

2. Sinteza vorbirii

- preia informaţiile anterioare şi le transformă în semnal vocal.
- Fiecare dintre cele două probleme se subdivid si ele în mai multe sarcini:





Textul primar (intrare)

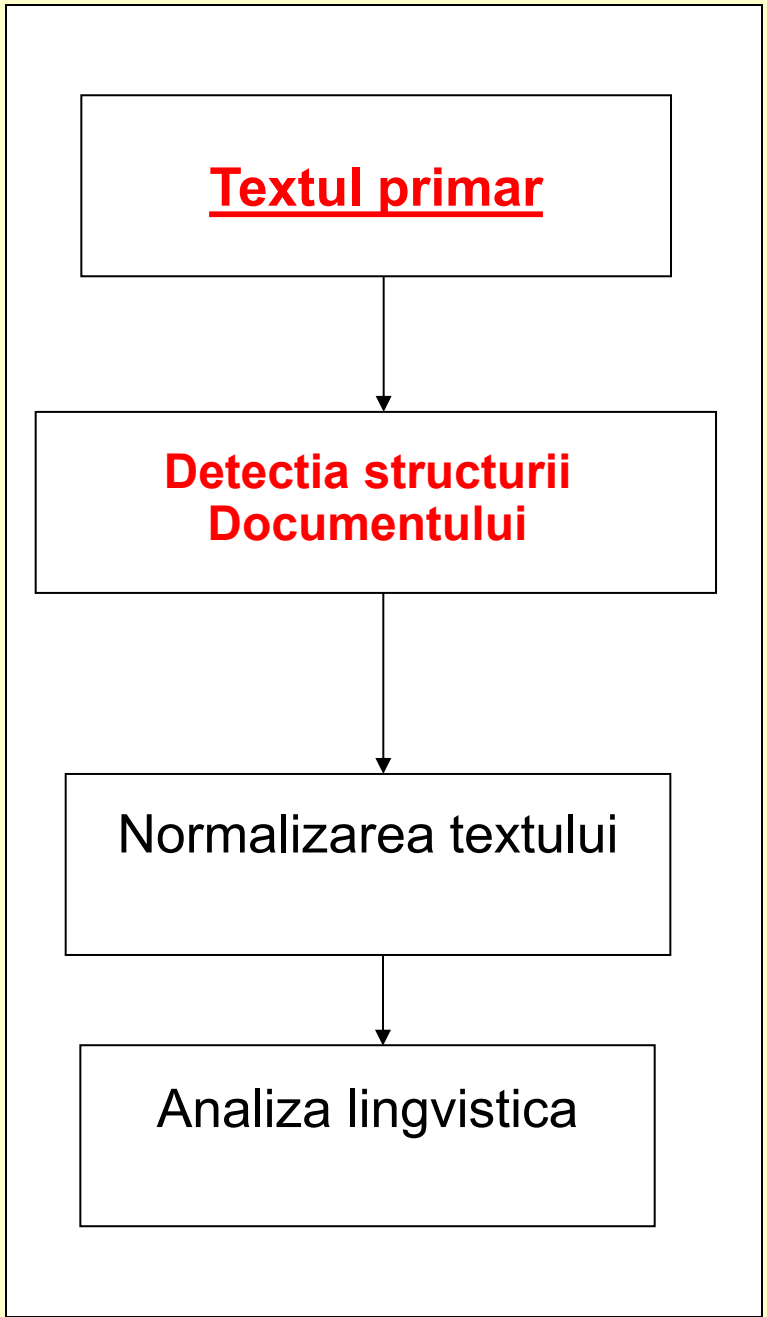
Analiza Textului
Detectia structurii Documentului
Normalizarea Textului
Analiza Linguistica

Analiza Fonetica
Conversia Grafeme-Foneme

Analiza Prozodica
Pitch si Atribuirea Duratelor

Sinteza Vorbirii (iesirea)
Selectia unitatilor
Generarea formei de unda

1. Analiza textului şi analiza lingvistică



1.1 Detectia structurii documentului

→ Furnizeaza informatii de context pentru modulele urmatoare

→ Textul în cod ASCII

→ divizarea textului în porţiuni rezonabile – propoziţiile (chineza) pe baza semnelor de punctuatie (ambiguitati)

→ Stabilirea tipului propozitiei

Jones lives at the end of St. James St.

→ Textul este divizat în cuvinte, pe baza spaţiilor si a unor reguli textuale simple

(I) (_) (know) (_) (1) (,) (000) (_) (words).

→ Segmentarea paragrafelor

→ detectia structurii de email, articol…

→ detectia structurii de pag. Web etc

1.2 Normalizarea Textului

- Normalizare textului - ***transformarea formelor non-ortografice*** (cum ar fi numere, simboluri) in ***forme ortografice***

- Modul de normalizare text: mai intai identifică ***tipul formei***, apoi o converteşte la reprezentarea ortografica

- Extensia abrevierilor/ (ex. Mr. = mister, D-na, Mrs.) - nu este banală (ex. St. = Street/Saint, m, m^2, kg);

- Extensia acronimelor: NATO, UTCN, ΦΠΑ, ΔΕΗ,

- Conversia/expandarea numerelor în cuvinte

(150 = o sută cincizeci ;

+40-264-432311 - dacă este un număr de telefon;

15.25lei - pret ;

23.09.2009- data;

11:56:12 - ora…

π =3.14159..simbol matematic

K – simbol chimic / kilo

1.3 Analiza morfologica

- atribuirea/etichetarea părţilor de vorbire cuvintelor propoziţiei
-(POS tagging =etichetarea partilor vorbirii)

ex. Ei au realizat proiectul. / *Pronume plural, verb trecut, substantiv articulat*)

- limbile vorbite (ex. engleza) diferite cuvinte din propoziţii sunt ***accentuate***, ceea ce constă în *deplasări în sus sau jos ale frecvenţei fundamentale*

- decizia de accentuare sau nu a unui cuvânt – (ne/accentuat)
- asignarea accentelor cuvintelor care în general sunt accentuate (substantive, verbe, adjective) neaccentuate cuvinte functionale (prepozitii, verbe auxiliare)
- Cuvintele functionale scurte tind sa devina neaccentuate ("cliticized") ex. l'arme fr.)

Ex. Am cerut prăjitură cu cireşe nu prăjitură cu mere. – dificila, contrast

1.4 Analiza fonetica

→ Alegerea pronunţiilor pentru cuvinte, prin reprezentarea ortografiei lor, adica *transcrierea fonetica automata a textului* – dependenta de limba

→ Crearea unui *set de reguli de conversie literă-sunet,* care crează o corespondenţă între *o secvenţă de grafeme şi una de foneme* cu posibile precizari diacritice (difera functie de ortografie)

→ Regula *"o litera-un fonem"* nu se respecta intotdeauna (elan, el, x- cs mixer, gz – examen, - regionalisme)

→ Pronuntarea izolata – dictionare fonetice, derivatele – liste de reguli

→ dicţionare de pronunţii care stau la baza conversiilor (engleza, franceza)

→ ambiguităţile de pronunţie a omografelor (ex. bass muzical sau peşte au pronunţii diferite [bejs] sau [bæs] – context (umbrele, mobilă, veselă, zări...)

→ în exprimarea intonaţională (o frază lungă este divizată în unităţi de sine stătătoare din punct de vedere intonaţional) – cuvinte functionale, sau :/;/

ex1. *El a coborât din maşină **şi** s-a plimbat o vreme prin parc.*

ex.2 *De câteva ore era între agonie şi extaz **în** aşteptarea rezultatului care-i putea schimba viaţa*.

1.5 Analiza prozodica

Prozodia –> caracteristica complexa rezultanta a mai multor factori:

- Pauzele dintre fraze
- Evolutia pitch/F0 in timp
- Durata fonemelor
- Taria/amplitudinea SV

- Elementele prozodiei sunt :

  - accentul (influenteaza durata si amplit. fonemei)
  - intonatia (evolutia F0)
  - ritmul (durata si viteza de sinteza)

- Naturaletea sintezei TTS ~ de tezaurul de contururi intonationale si sabloane ritmice

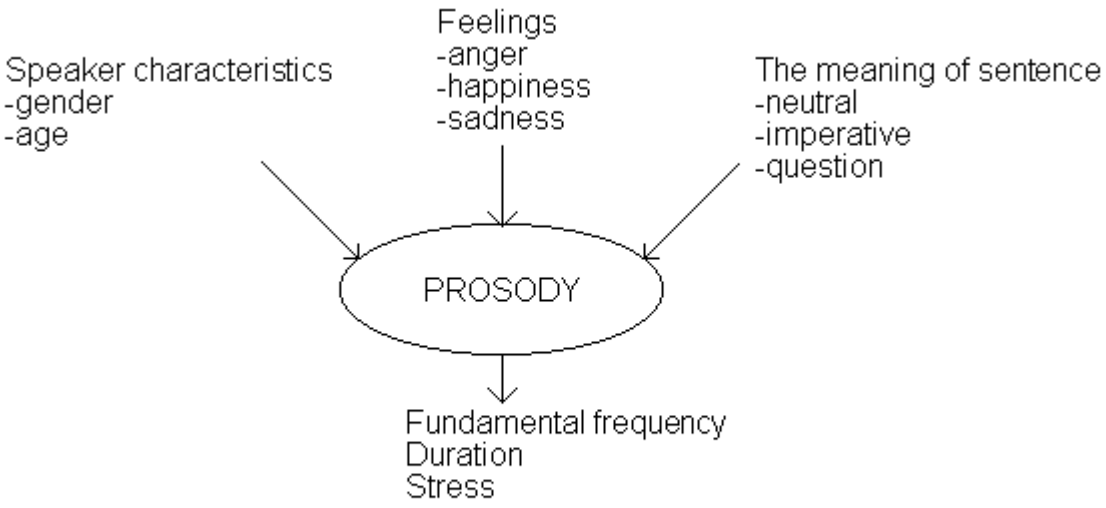


Fig. 4.1. Prosodic dependencies.

- Dupa stabilirea *şirului de foneme* ce trebuie sintetizat --> stabilirea *duratelor*

*F*actori care influenteaza durata :

- *tipul segmentului* în discuţie (de ex. În multe dialecte ale lb. engleze vocala /æ/ are o durată implicit mai lungă decât a vocalei /i/.
- *Accentul* - pe silaba din care face parte segmentul (durata, amplitudine fonem)
- *Calitatea* segmentelor învecinate (al, ac..)
- *Poziţia segmentului* în frază (elementele terminale sunt mai lungi)
- Modelarea duratei : reguli – formalizata D=X+n

Conturul intonational ~ metodele folosite din teoria prozodiei

Intrarile ptr modulul intonational:

- *ce silabe sunt accentuate în frază*
- *ce tipuri de accente se folosesc (joase/înalte)*
- *duratele tuturor segmentelor din rostire sunt calculate în blocul duratelor segmentelor*

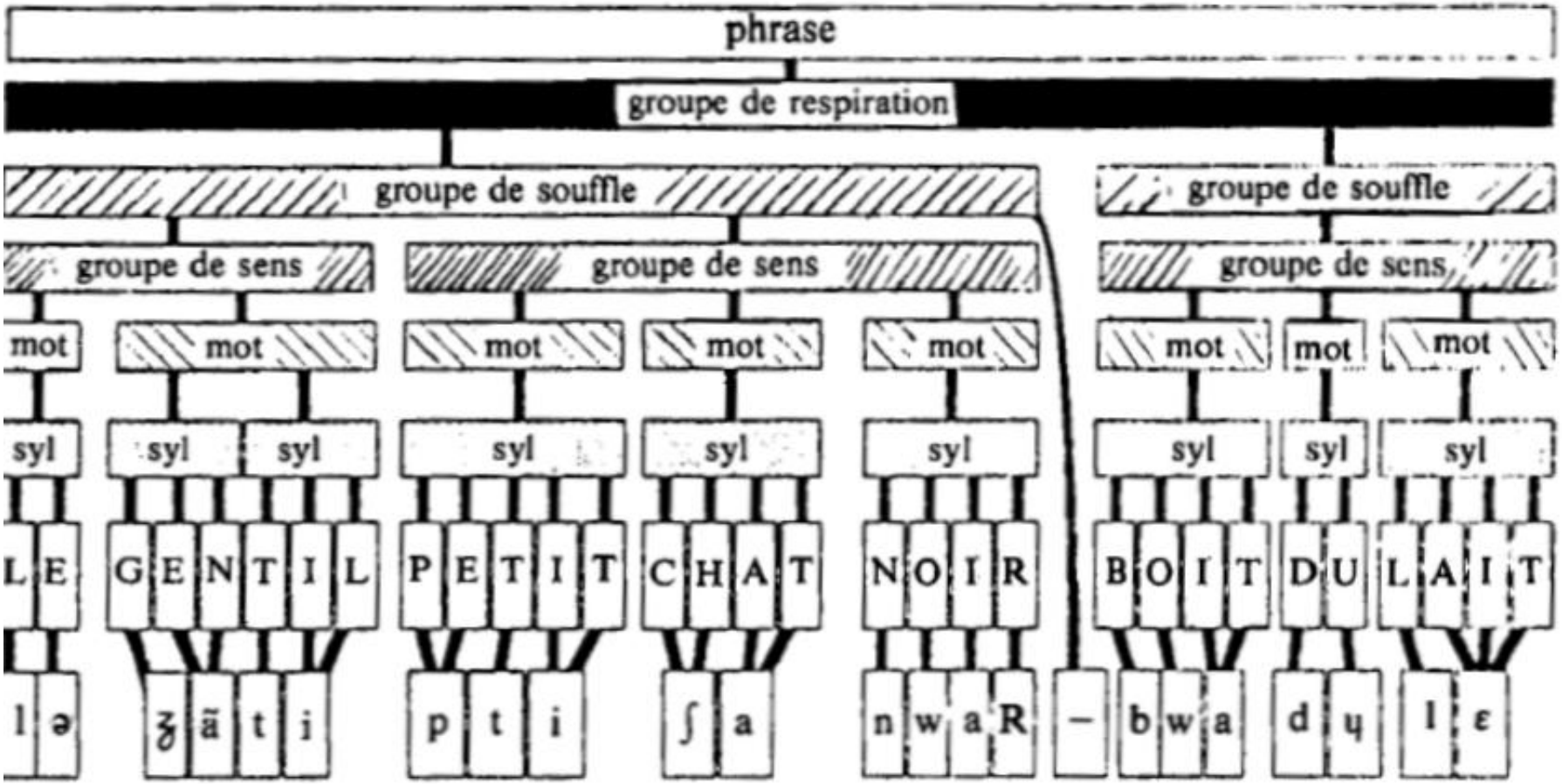


FIG. XV.20. — *Niveaux classiques de la structure prosodique sur un exemple.*

Un prim pas la stabilirea conturului F0 :

- ***setarea unor valori*** **pentru funcţia ton/frecvenţă** la momentele (timp) corespunzătoare fiecărui accent.

- ***Limitele iniţiale ale tonurilor*** sunt aliniate cu porţiunile de linişte plasate la începutul fiecărei fraze minore, în timp ce limitele finale ale tonurilor sunt aliniate la ultima vocală din frază.

- ***Momentele de accentuare*** sunt convertite în perechi F0/timp, valorile pentru F0 fiind calculate ţinând cont de proeminenţa accentelor şi de diverşi parametri ce caracterizează fraza

- *In final, conturul* pentru F0 se va calcula prin ***interpolarea perechilor F0/timp*** şi netezirea valorilor prin convoluţia cu o fereastră dreptungiulară

	H*												L*		L- H%			
do		you		really				want				to		see		all		of		it	
d	uw	y	uw	r	ih	l	iy	w	aa	n	t	t	ax	s	iy	ao	l	ah	v	ih	t
110	110	50	50	75	64	57	82	57	50	72	41	43	47	54	130	76	90	44	62	46	220

Figure 4.25 Output of the FESTIVAL (Black et al., 1999) generator for the sentence *Do you really want to see all of it?* The exact intonation contour is shown in Figure 4.26. Thanks to Paul Taylor for this figure.

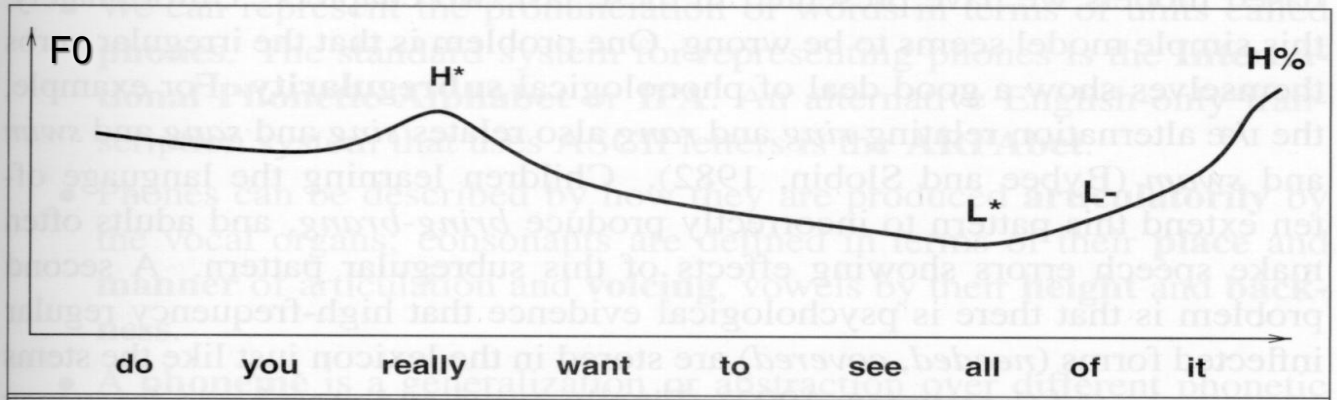
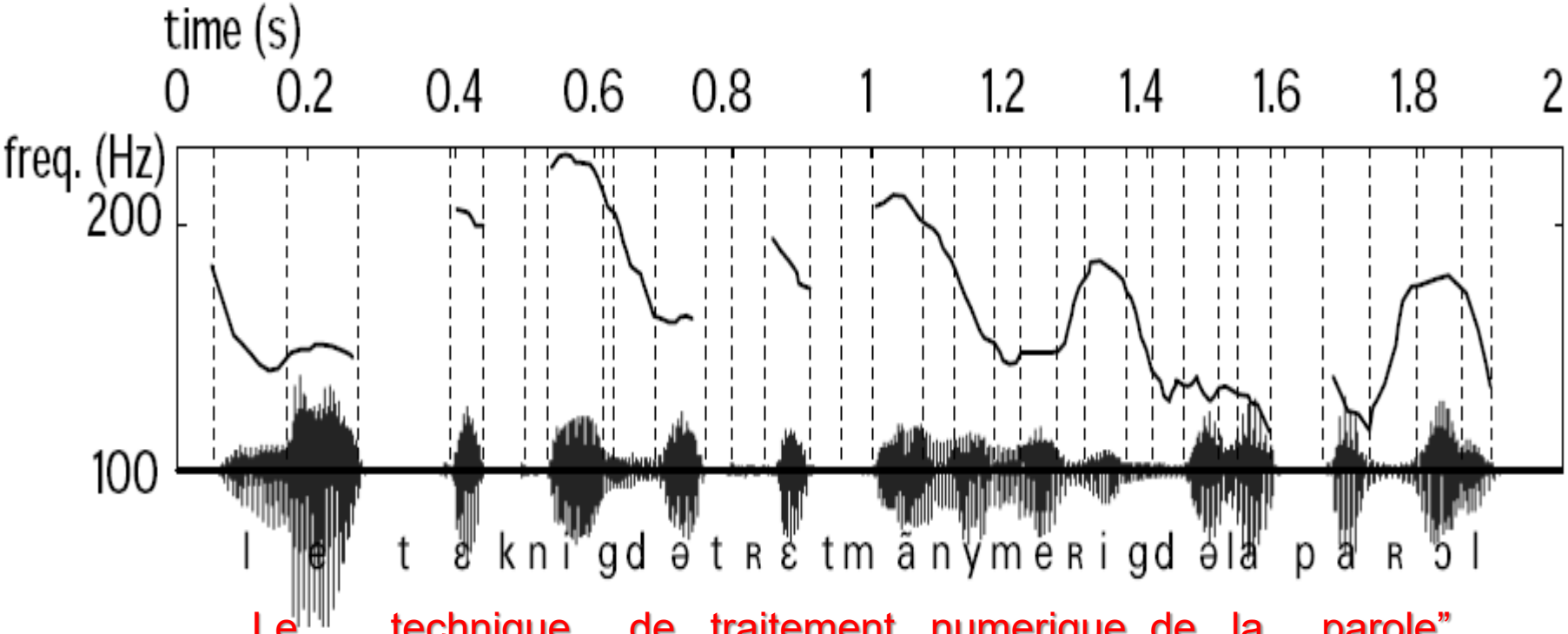


Figure 4.26 The F0 contour for the sample sentence generated by the FESTIVAL synthesis system in Figure 4.25, thanks to Paul Taylor.

H*	*Peak accent*	
L*	*Low accent*	
L*+H	*Scooped accent*	
L*+!H	*Scooped downstep accent*	
L+H*	*Rising peak accent*	
!H*	*Downstep high tone*	


time (s)
0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6 1.8 2
freq. (Hz)
200
100
l e t ɛ k n i g d ə t ʀ ɛ t m ã n y m e ʀ i g d ə l a p a ʀ ɔ l
„Le technique de traitement numerique de la parole”

2. Sinteza vorbirii

• Textul convertit în foneme, durate stabilite şi conturul intonaţiei determinat => sistemul poate calcula parametrii pentru sinteza vorbirii;

• Există două grade de libertate pentru calculul parametrilor unui sistem TTS clasic:

A1. sinteza pe bază de reguli sau
A2. sinteza prin concatenarea segmentelor anterior rostite şi selectate

B. tipul de parametri utilizaţi (LPC, formanţi, par. spectrali sau temporali)

Obs. *La schema de concatenare* orice tip de parametri se folosesc permit controlul independent *al amplitudinii, F0, vocalizării, temporizării şi posibilitatea de prelucrare spectrală adecvată.*

2.1. Sistemele bazate pe reguli

- sunt mai restrictive în alegerea parametrilor folosiţi şi în posibilitatea de control a dinamicii lor
- sunt mai economice din punct de vedere al memoriei

Ex. La un sistem AT&T sunt ~ 2900 de segmente/elemente de vorbire numite „diade”

- Alegerea adecvată a elementelor din baza acustică, se face în două etape :

- convertirea reprezentării fonemice în reprezentarea dată de elementele de dicţionar ale bazei acustice
- conectarea şi interpolarea parametrilor.

Ex. We are in the test. (text)

- care se transcrie fonetic (IPA) : /*wi ar ɪn ðə tɛst*/
- intrare în modulul de selecţie al diadelor:

/*/ /*-w/ /w-i/ /i-a-r/ /r- ɪ/ /ɪ -n/ /n-ð/ /ð-ə/ /ə-t/ /*/ /t-ɛ/ /ɛ-s/ /s/ /s-*/ /*/ /t-*/ /*/

http://festvox.org/festtut/notes/festtut_toc.html

Limba română

Limba română are o ortografie ***preponderent fonemică***, dar prezintă și numeroase excepții.
Printre acestea se pot menționa următoarele:

• litera *c* are valori fonetice diferite în grupurile *ce* și *ci* față de celelalte situații. Aceeași observație este valabilă pentru litera *g*;

• Grupul de litere *ch* din *che* și *chi* reprezintă o consoană oclusivă palatală surdă, care în limba română este echivalentă fonologic cu versiunea sa velară, notată în scris prin litera *c* (cu excepția grupurilor *ce* și *ci*). Aceeași observație este valabilă pentru litera *g*.

• nu există simboluri separate pentru semivocale; literele *i*, *u*, *e* și *o* putand reprezenta fie vocale, fie semivocale, în funcție de context, de exemplu litera *i* din cuvintele *știam* și *fiară* are valori fonetice distincte;

• nu se indică poziția accentului, de exemplu verbul *"intră"* poate fi la timpul prezent sau la perfectul simplu în funcție de plasarea accentului;

• nu se deosebește în scris vocala /i/ de palatalizare, de ex. litera *i* din cuvintele *(tu) umbli* și *(două) boli* se pronunță diferit;

• nu se notează în scris pronunția sonoră sau surdă a literei *x*, de exemplu în cuvintele *extrem* și *exemplu* ea se citește /ks/, respectiv /gz/;

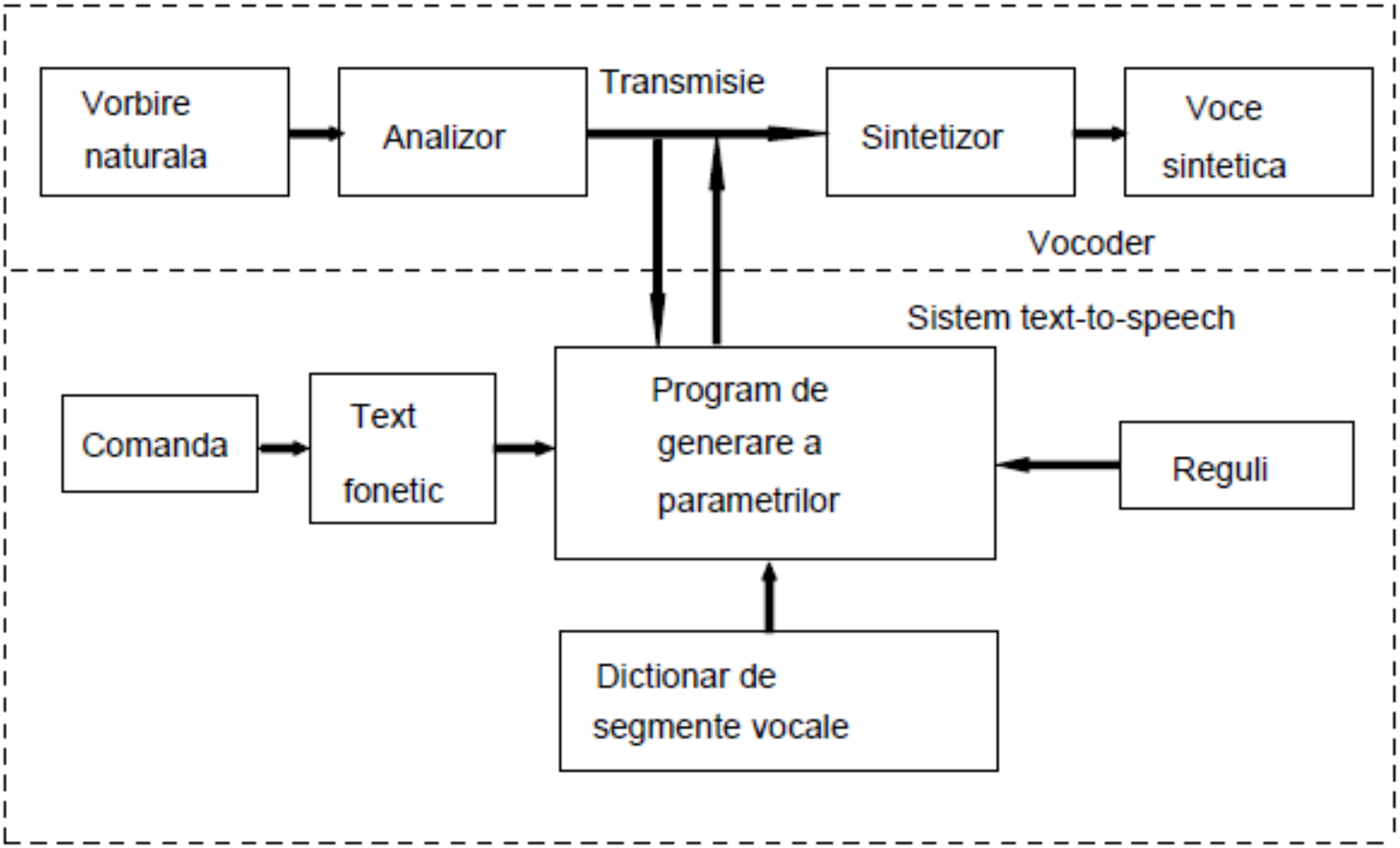
• Grupul de consoane /ks/ se scrie fie ca *x*, fie ca *cs*, de exemplu în *axă* și *ticsit*;

• cuvintele noi de origine străină sunt adesea păstrate în forma originală, de exemplu *watt*, *yoga*, *computer* sau parțial în *technețiu* (pronunțat /tehˈne.ʦju/);

• literele *î* și *â* corespund aceluiași fonem, vocala închisă centrală nerotunjită /ɨ/);

• literele *k*, *w*, *y* și *q* nu notează sunete distincte, ci se suprapun fonetic cu litere deja existente.

Exemplu de sistem pe bază de reguli



Schema bloc a unui vocoder legat de un sistem text-to-speech

2.2. Sinteza prin concatenare

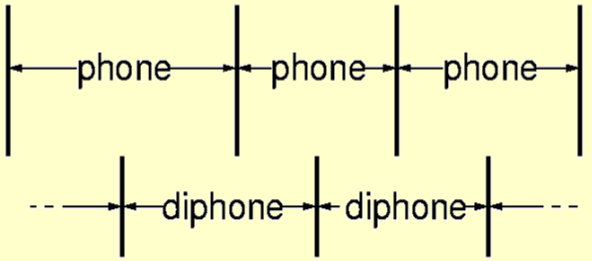
Se aleg subunităţi adecvate de vorbire (*difoni, demi-silabe sau trifoni*)

- înregistrarea şi prelucrarea rostirilor
- Segmentarea semnalului şi extragerea *unităţilor de vorbire*
- stocarea formelor de undă a segmentui (împreună cu contextul) şi informaţii extinse în baza de date
- extragerea parametrilor şi crearea bazei de date parametrice de segmente
  - utile pentru compresia de date
  - alegerea/modificarea mai usoara a prozodiei

-Exista o mare varietate de combinaţii de foneme şi contexte prozodice

- În engleză: 43 foni, 79.507 trifoni, numai ~70.000 utilizati

-Care dintre ei ar trebui să fie păstrate?



- Selectarea unitatilor ptr. sinteza concatenativa
  - Se înregistreaza un corpus mare de vorbire
  - la selectarea unitatii, corpus-ul este divizat în unităţi fonetice, indexate
- *La sinteza concatenativa, selecţia se face off-line şi manual !*

Metoda PSOLA - Pitch Synchronous Overlap and Add

O fereastră (ex. 2-perioade F0) este înmulţită cu semnalul

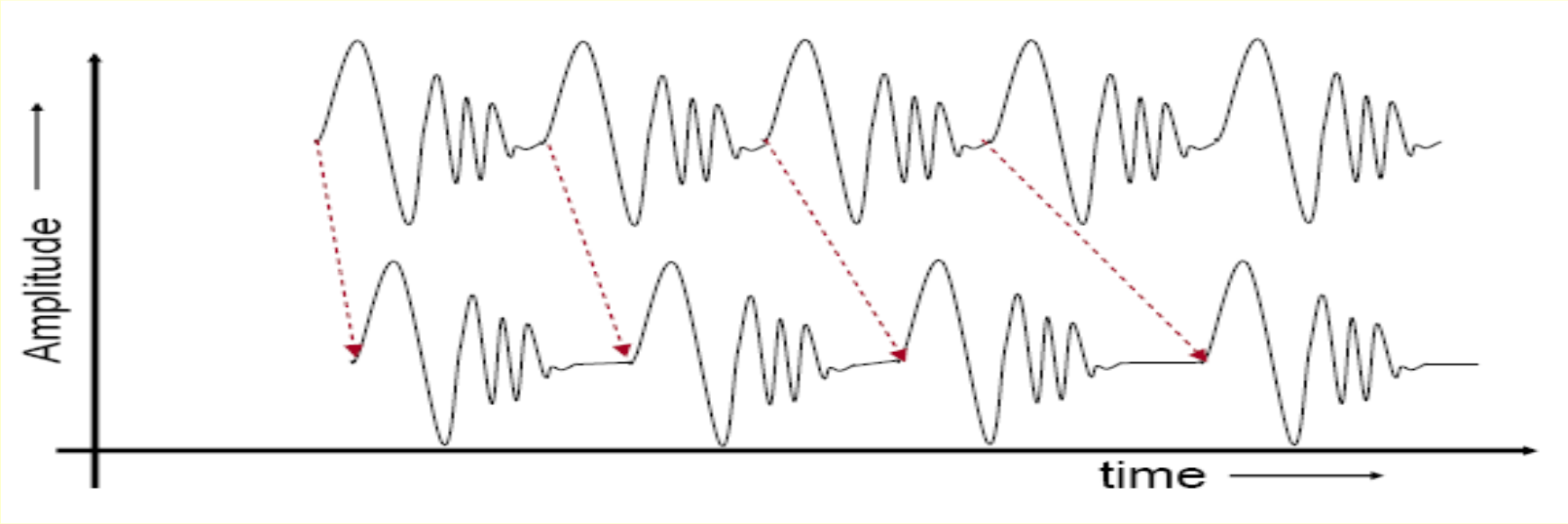
- Semnalul este divizat într-un set de semnale locale (diferit de zero doar pe intervalul fereastrei)

Modificarea Pitch

- deplasarea relativă a semnalelor locale
- Spaţierea reflectă durata pitch
- rezultate bune pentru β factor de modificare = [0.5 - 1.5]

Durată

- Semnalele locale se adaugă sau se elimină de la ieşire



Rezumat

- **Analiza textului** - Generarea de caracteristici lingvistice din text
- **Modele acustice** - Generează caracteristici acustice din caracteristici lingvistice
- **Vocodere** - Generarea formei de undă din caracteristici acustice

Analiza textului

- Segmentarea cuvintelor
  - Necesară pentru limbile cu cuvinte bazate pe caractere (de exemplu, chineza)

- Partea de vorbire (POST)
  - Etichetarea POS a fiecărui cuvânt permite o mai bună predicție a prozodiei și acuratețe G2P

- Predicția prozodiei
  - Prosodia înglobează intonația și alte aspecte importante ale vorbirii
  - Permite o sinteză mai realistă/naturala a vorbirii

- Conversia de la grafem la fonem (G2P)
  - Grafemele sunt diferitele moduri în care un fonem poate fi reprezentat în text
    - Ex: /k/ poate fi scris ca „c”, „k”, „ck”, „qu”, „ch”
  - Generarea fonemelor din grafeme de text
  - Folosit de obicei ca o plasă de siguranță pentru cuvintele care nu fac parte din vocabularul unui model

Tipuri de modele folosite

- **Metode clasice de procesare a semnalului**
  - Articulator
  - Formantic
  - Concatenativ

- **Sinteză parametrică statistică a vorbirii (Statistical Parametric Speech Synthesis)**
  - HMM

- **Metode neuronale**
  - Bazate pe RNN
  - Bazate pe CNN
  - GAN-uri

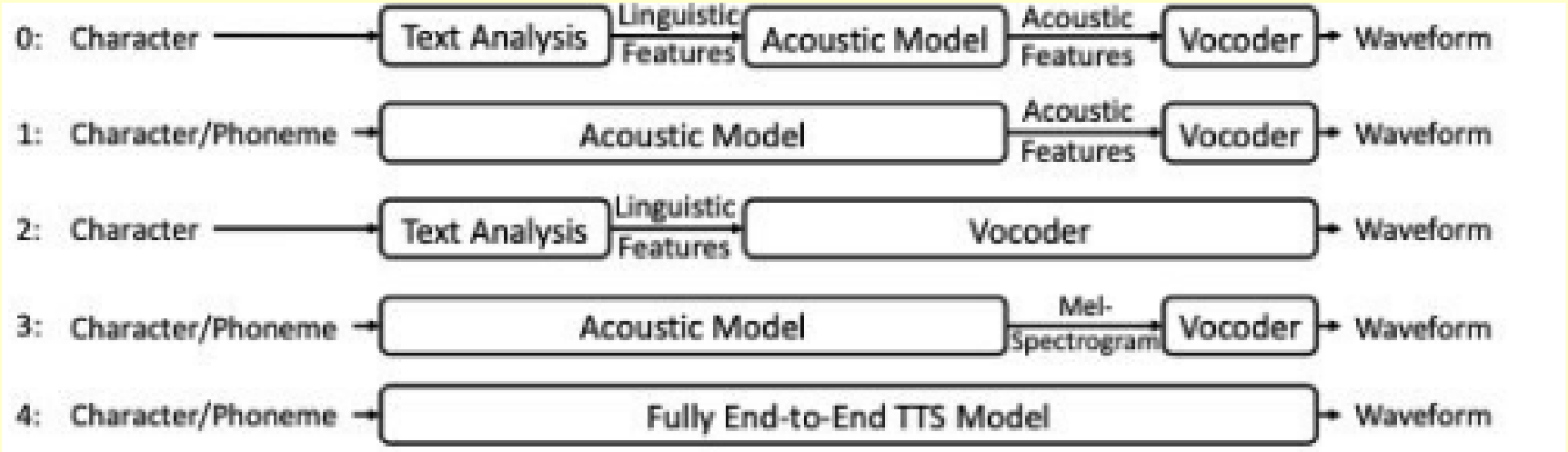
- **Metode probabilistice mai noi**
  - Modele de flux
  - Modele de difuzie

Caracteristici acustice

- Coeficienți Mel Cepstral (MCC)
- Spectrograme Mel
- Frecvență fundamentală (F0)
- Vocal/nevocal
- Coeficienți Cepstrali pe scara Bark ….

Vocodere

- Analiză/Sinteză
  - Utilizează caracteristici acustice, cum ar fi MFCC, aperiodicitatea benzii și (F0)
  - Creează forma de undă din caracteristicile acustice
- Vocodere autoregresive
- Vocodere bazate pe flux
- Vocodere bazate pe GAN
- Vocodere bazate pe difuzie



Stage	Models
0	SPSS [416, 356, 415, 425, 357]
1	ARST [375]
2	WaveNet [254], DeepVoice 1/2 [8, 87], Par. WaveNet [255], WaveRNN [150], HiFi-GAN [23]
3	DeepVoice 3 [270], Tacotron 2 [303], FastSpeech 1/2 [290, 292], WaveGlow [279], FloWaveNet [163]
4	Char2Wav [315], ClariNet [269], FastSpeech 2s [292], EATS [69], Wave-Tacotron [385], VITS [160]

Statistical Parametric Speech Synthesis

Procesul progresiv end-to-end pentru modelele TTS.

Sinteza parametrică statistică utilizează trei module de bază, în care modulul de analiză a textului convertește caracterele în caracteristici lingvistice, iar modelele acustice generează caracteristici acustice din caracteristicile lingvistice (unde caracteristicile acustice țintă sunt obținute prin analiza vocoderului), iar apoi vocoderele sintetizează forma de undă a vorbirii din caracteristicile acustice prin calcul parametric.

Etapa 1. In sinteza parametrică statistică explorează combinarea analizei textului și a modelului acustic într-un model acustic end-to-end care generează direct caracteristici acustice din secvența fonemică, apoi utilizează un vocoder în SPSS pentru a genera forma de undă.

Etapa 2. WaveNet este propus pentru prima dată pentru a genera direct forma de undă a vorbirii din caracteristicile lingvistice, care poate fi considerată o combinație între un model acustic și un vocoder. Acest tip de modele necesită în continuare un modul de analiză a textului pentru a genera caracteristici lingvistice.

Etapa 3. Tacotron este propus în continuare pentru a simplifica caracteristicile lingvistice și acustice, care prezice direct spectrogramele liniare din caractere/foneme cu un model encoder-atenție-decoder și convertește spectrogramele liniare în forme de undă cu Griffin-Lim [95]. Următoarele lucrări, precum DeepVoice 3 [270], Tacotron 2 [303], TransformerTTS [192] și FastSpeech 1/2 [290, 292], prezic spectrogramele mel din caractere/foneme și utilizează în continuare un vocoder neuronal, cum ar fi WaveNet [254], WaveRNN [150], WaveGlow [279], FloWaveNet [163] și Parallel WaveGAN[402], pentru a genera forma de undă.

Etapa 4. Unele modele TTS complet end-to-end sunt dezvoltate pentru sinteza directă a textului în formă de undă. Char2Wav utilizează un model codificator-atenție-decodificator bazat pe RNN pentru a genera caracteristici acustice din caractere, apoi utilizează SampleRNN pentru a genera forma de undă.

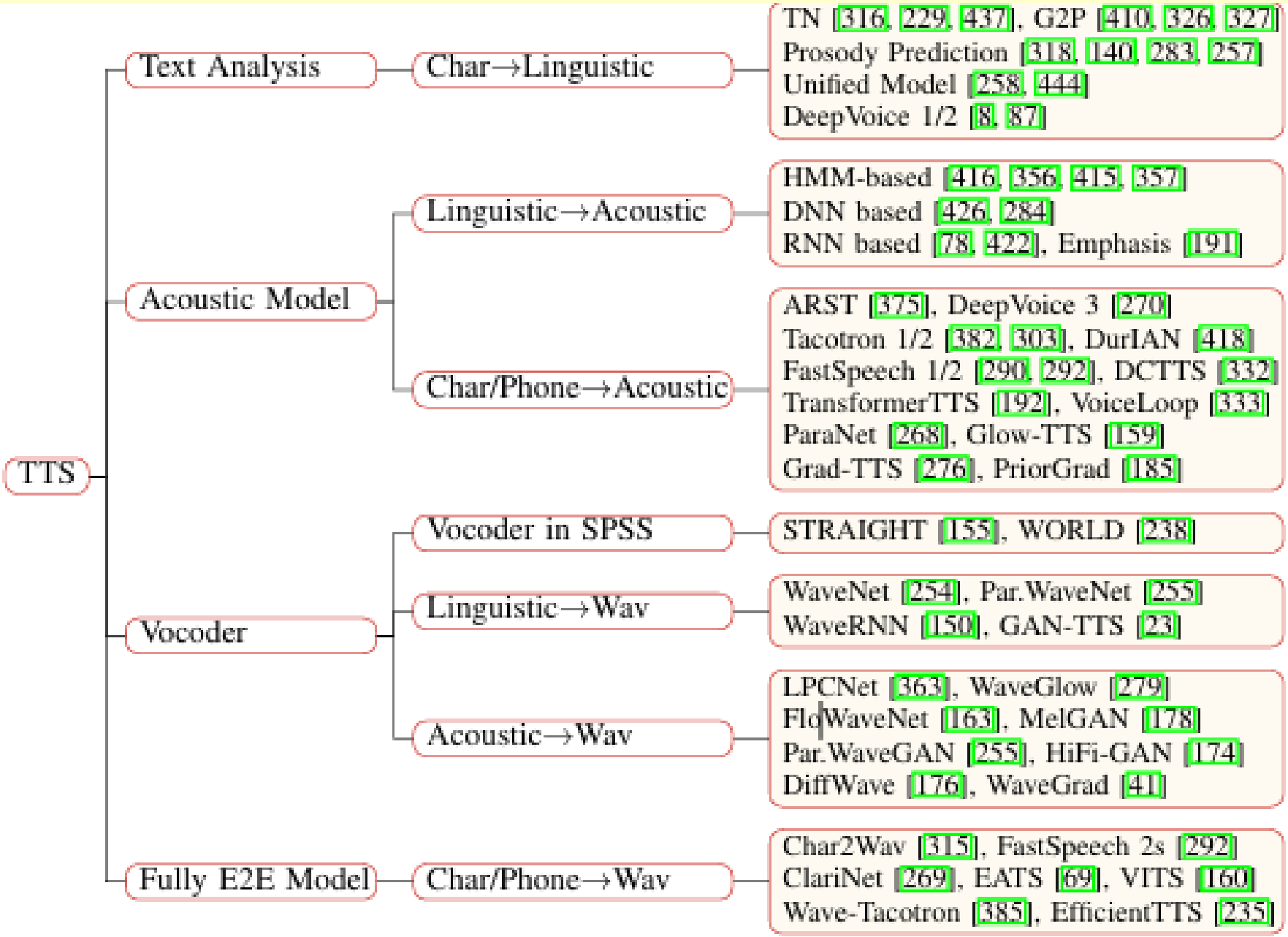
Cele două modele sunt reglate împreună pentru sinteza directă a vorbirii. În mod similar, **ClariNet** reglează împreună un model acustic autoregresiv și un vocoder nonautoregresiv pentru generarea directă a formei de undă.

FastSpeech generează direct vorbirea din text cu o structură complet paralelă, ceea ce poate accelera considerabil inferența.

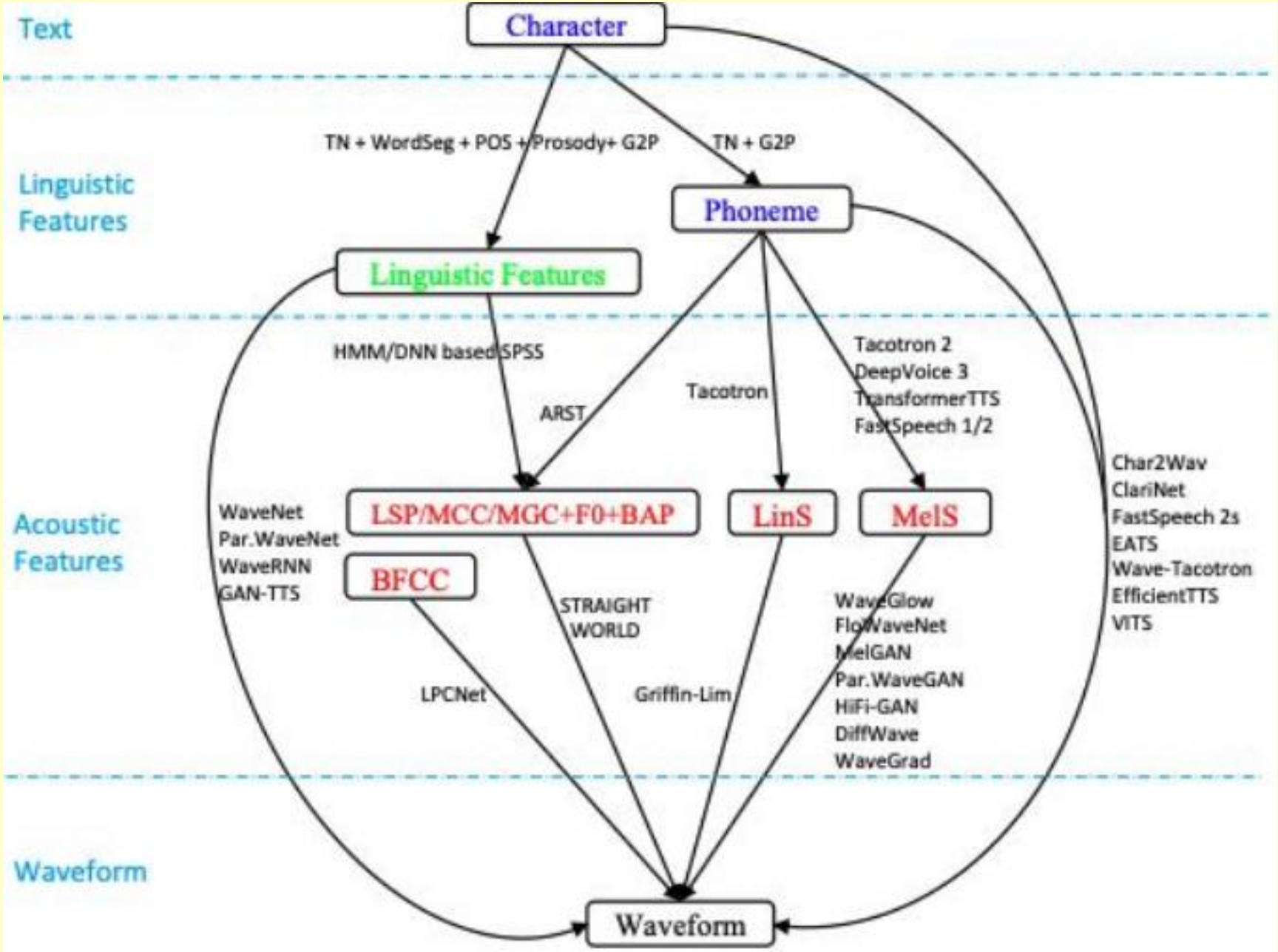
Pentru a atenua dificultatea antrenării comune text-formă de undă, acesta utilizează un decodor mel-spectrogramă auxiliar pentru a ajuta la învățarea reprezentărilor contextuale ale secvenței fonemelor.

EATS generează, de asemenea, direct forma de undă din caractere/foneme, care utilizează interpolarea duratei și pierderea dinamică lenta de înfășurare a timpului pentru învățarea alinierii end-to-end.

Wave-Tacotron construiește un decodor bazat pe pentru a genera direct forme de undă, care utilizează generarea paralelă de forme de undă în partea de flux, dar totuși generarea autoregresivă în partea Tacotron.



(a) A taxonomy of neural TTS.



Fluxul de date de la text la formă de undă.

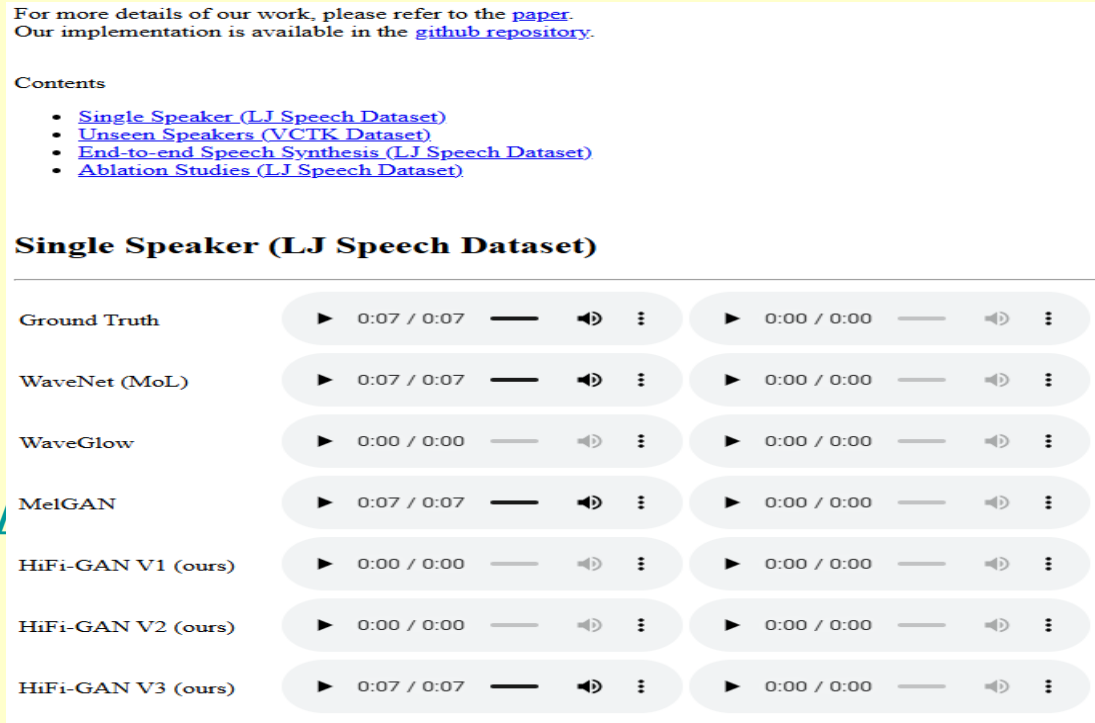
“Audio samples from "HiFi-GAN: Generative Adversarial Networks for Efficient and High-Fidelity Speech Synthesis“

“Mai multe lucrări recente privind sinteza vorbirii au utilizat *rețele adversariale generative (GAN)* pentru a produce forme de undă brute. Deși astfel de metode îmbunătățesc eficiența eșantionării și utilizarea memoriei, calitatea eșantionului lor nu a atins-o încă pe cea a modelelor generative autoregresive și bazate pe flux. HiFi-GAN realizează o sinteză vocală eficientă și de înaltă fidelitate. Deoarece sunetul vorbirii este format din semnale sinusoidale cu diferite perioade, demonstrăm că modelarea periodica al unui sunet este esențială pentru îmbunătățirea calității eșantionului. O evaluare umană subiectivă (MOS) a unui set de date cu un singur vorbitor indică faptul că metoda propusă demonstrează similitudinea cu calitatea umană, generând în același timp un sunet de înaltă fidelitate (la 22 kHz) de 167,9 ori mai rapid decât în timp real pe un singur GPU V100. Mai mult, arătăm generalitatea HiFi-GAN pentru inversarea spectrogramei-mel vorbitorilor noi și sinteza vorbirii end-to-end.

În cele din urmă, o versiune cu amprentă redusă a HiFi-GAN generează eșantioane de 13,4 ori mai rapid decât în timp real pe CPU, cu o calitate comparabilă cu o contrapartidă autoregresivă.”

https://wandb.ai/messlav/Hi-Fi-GAN/reports/

https://jik876.github.io/hifi-gan-demo/


For more details of our work, please refer to the paper.
Our implementation is available in the github repository.

Contents

- Single Speaker (LJ Speech Dataset)
- Unseen Speakers (VCTK Dataset)
- End-to-end Speech Synthesis (LJ Speech Dataset)
- Ablation Studies (LJ Speech Dataset)

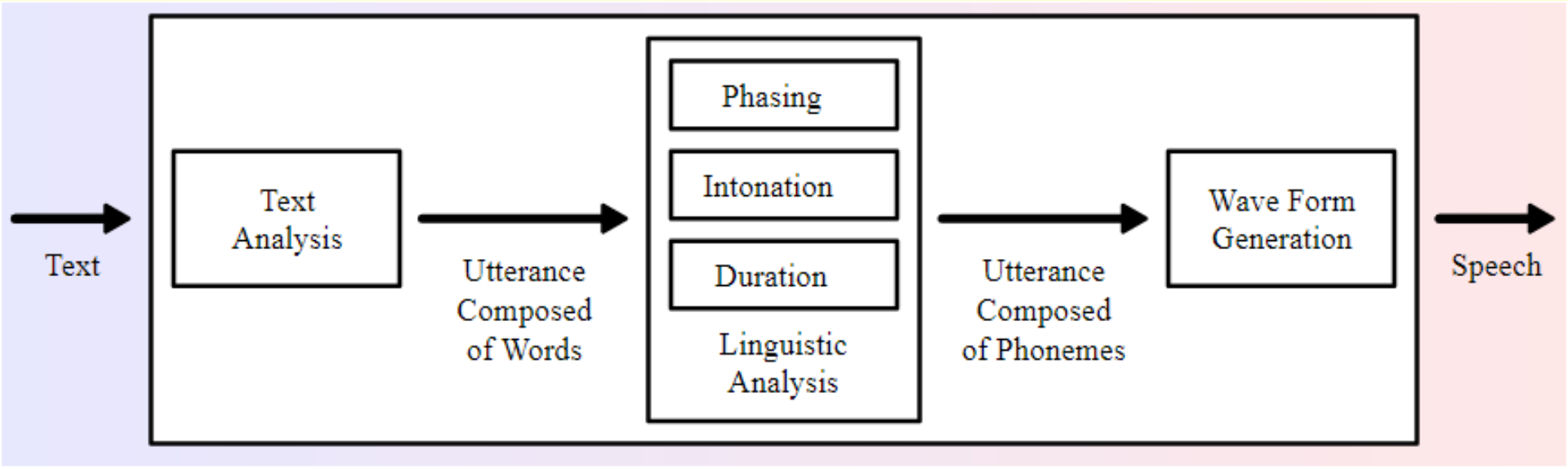
Single Speaker (LJ Speech Dataset)

Ground Truth	0:07 / 0:07	0:00 / 0:00
WaveNet (MoL)	0:07 / 0:07	0:00 / 0:00
WaveGlow	0:00 / 0:00	0:00 / 0:00
MelGAN	0:07 / 0:07	0:00 / 0:00
HiFi-GAN V1 (ours)	0:00 / 0:00	0:00 / 0:00
HiFi-GAN V2 (ours)	0:00 / 0:00	0:00 / 0:00
HiFi-GAN V3 (ours)	0:07 / 0:07	0:00 / 0:00

- *Noi modalități de a face sinteza TTS sunt cele bazate pe deep-learning.*
- cercetările din domeniu propun cateva abordări raspandite și actuale ale sintezei vorbirii:

- 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

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



https://en.wikipedia.org/wiki/Speech_synthesis
https://medium.com/sciforce/text-to-speech-synthesis-an-overview-641c18fcd35f
https://developer.nvidia.com/blog/generate-natural-sounding-speech-from-text-in-real-time/
https://paperswithcode.com/sota

https://arxiv.org/pdf/2106.15561.pdf A Survey on Neural Speech Synthesis
https://theaisummer.com/text-to-speech/
https://arxiv.org/pdf/2104.09995.pdf Review of end-to-end speech synthesis technology based on DL

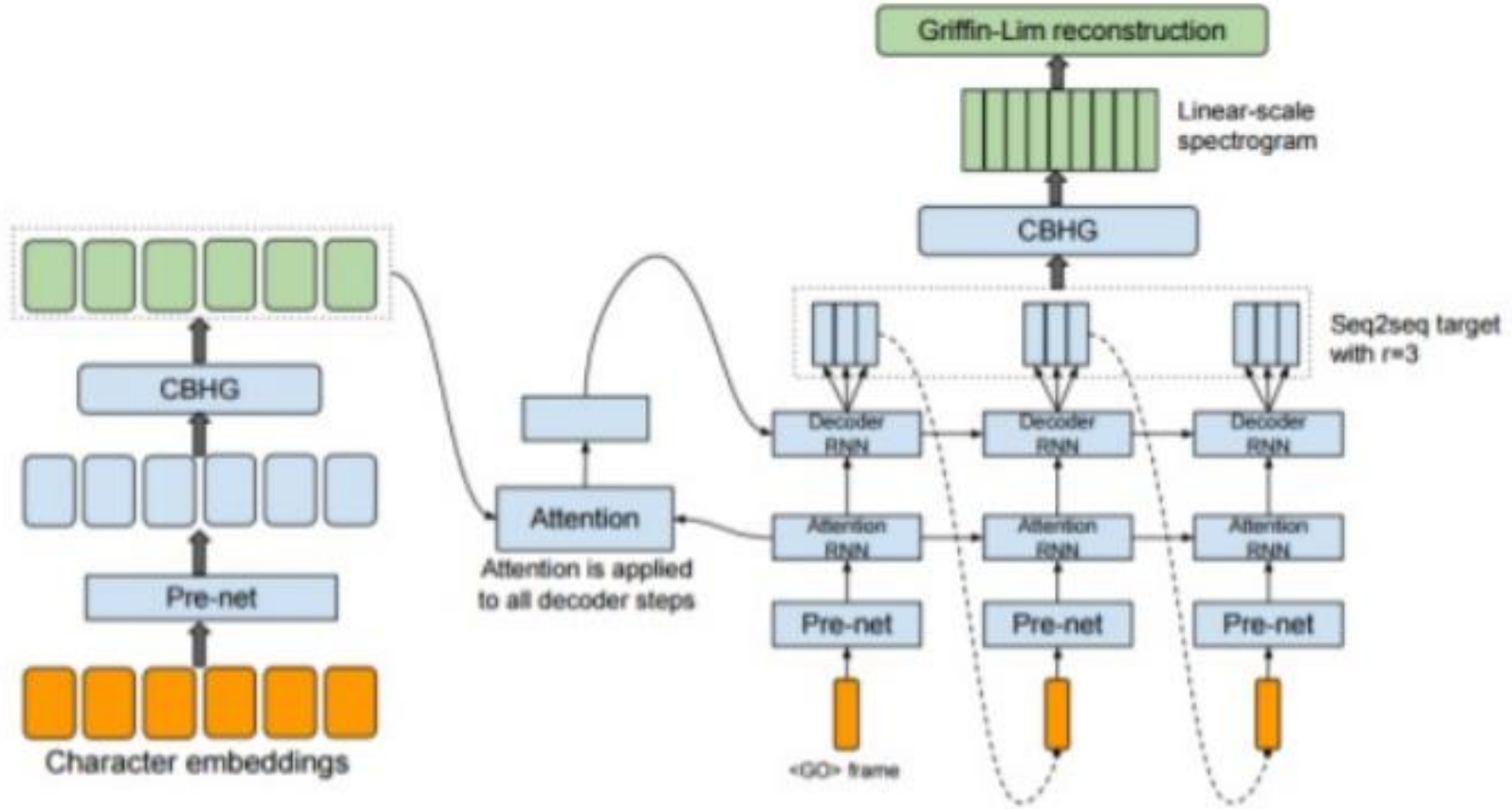
https://murf.ai/alternatives/ibm-watson-text-to-speech

Tacotron

- A TTS synthesis system typically consists of multiple stages, such as a text analysis frontend, an acoustic model and an audio synthesis module.
  - Building these components often requires extensive domain expertise and may contain brittle design choices.
- Tacotron, an e2e generative TTS model that synthesizes speech directly from characters.
- Given pairs, the model can be trained completely from scratch with random initialization.
- Some techniques to make the sequence-to-sequence framework perform well for this challenging task.

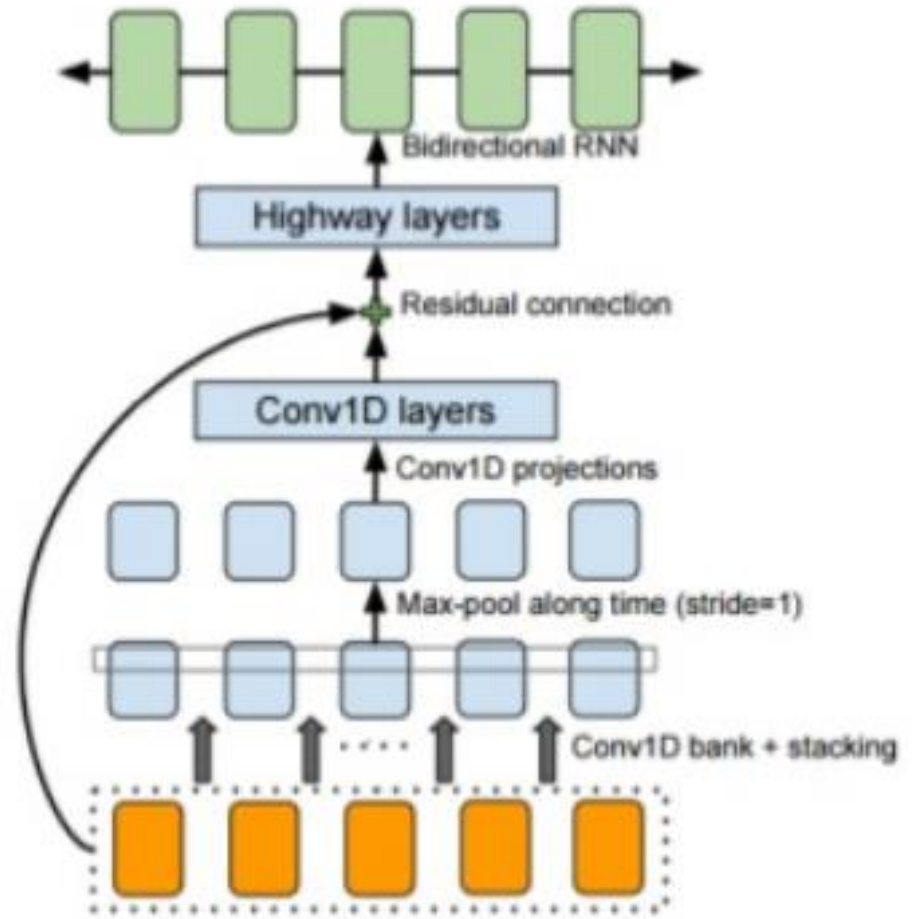
https://google.github.io/tacotron/publications/wave-tacotron/index.html
https://www.cellstrat.com/2020/01/15/text-to-speech-tts-using-tacotron/

Tacotron



Model architecture. The model takes characters as input and outputs the corresponding raw spectrogram, then fed to the Griffin-Lim reconstruction algorithm to synthesize speech.

Tacotron



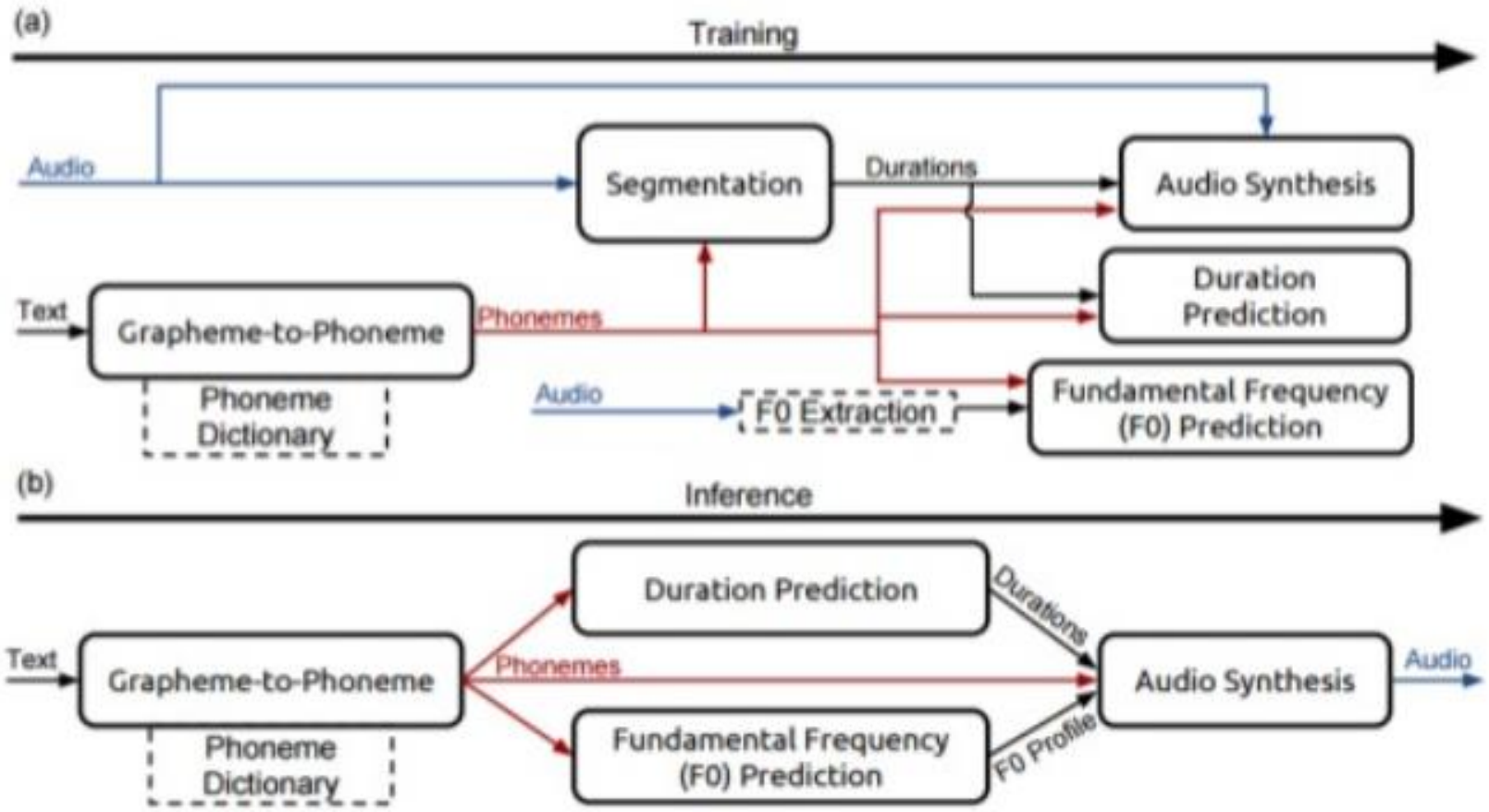
The CBHG (1-D convolution bank + highway network + bidirectional GRU) module

Deep Voice: Real-time Neural Text-to-Speech

- A production-quality TTS system constructed entirely from deep neural networks.
- Deep Voice is truly end-to-end neural speech synthesis.
- There are 5 major building blocks:
  - a segmentation model for locating phoneme boundaries
  - a grapheme-tophoneme conversion model
  - a phoneme duration prediction model
  - a fundamental frequency prediction model
  - an audio synthesis model
- For the segmentation model, performing phoneme boundary detection with deep neural networks using connectionist temporal classification (CTC) loss.
- For the audio synthesis model, a variant of WaveNet that requires fewer parameters and trains faster than the original.
- Simpler and more flexible than traditional text-tospeech systems, where each component requires laborious feature engineering and extensive domain expertise.
- Inference can be performed faster than real time and describe optimized WaveNet inference kernels on both CPU and GPU that achieve up to 400x speedups over existing implementations.

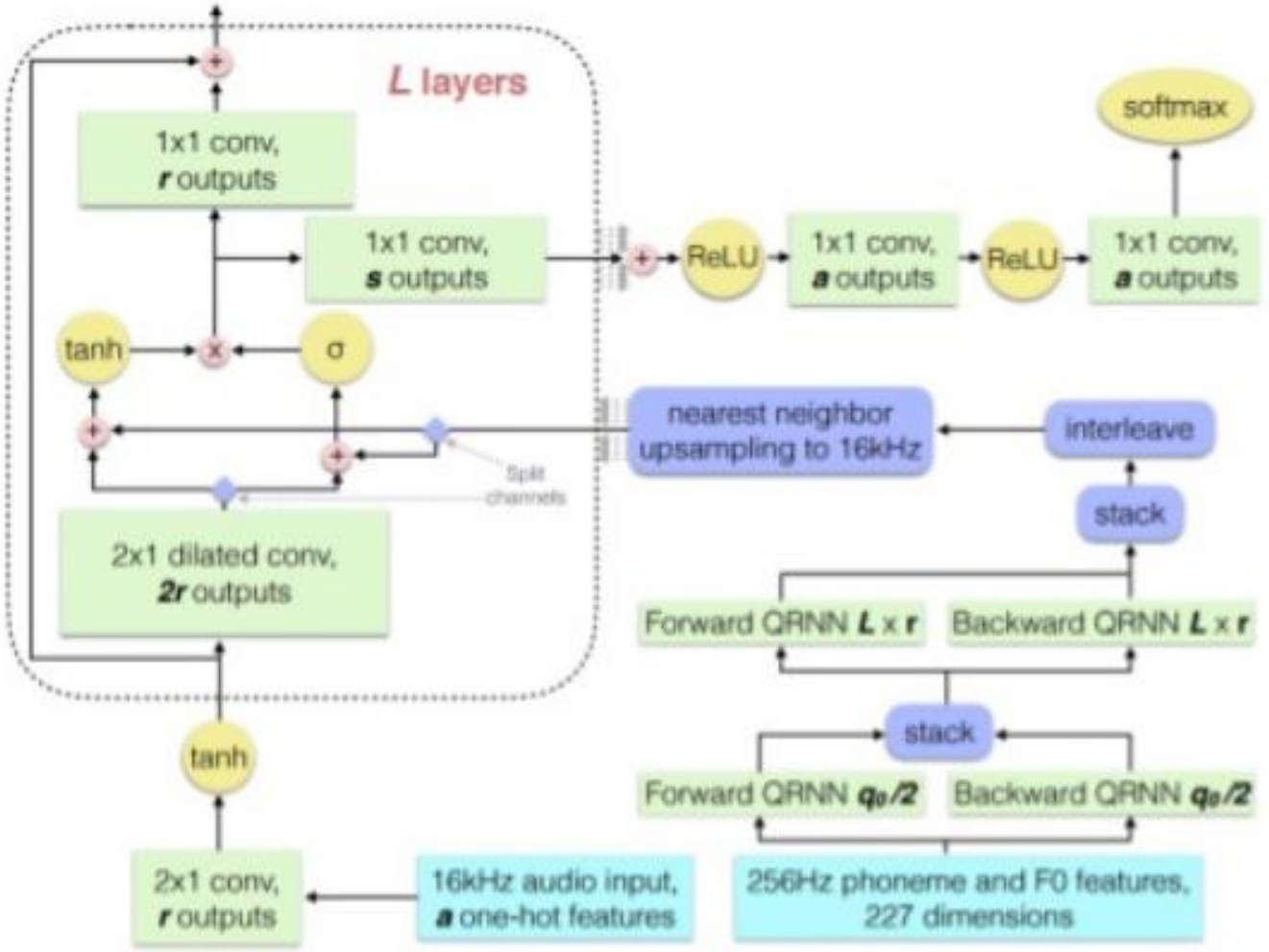
https://www.slideshare.net/yuhuang/voice-speech-synthesis-by-deep-learning

Deep Voice: Real-time Neural Text-to-Speech



Deep Voice: (a) training procedure and (b) inference procedure. In the system, the duration prediction model and the F0 prediction model are performed by a single neural network trained with a joint loss. The grapheme-to-phoneme model is used as a fallback for words that are not present in a phoneme dictionary, such as CMUDict.

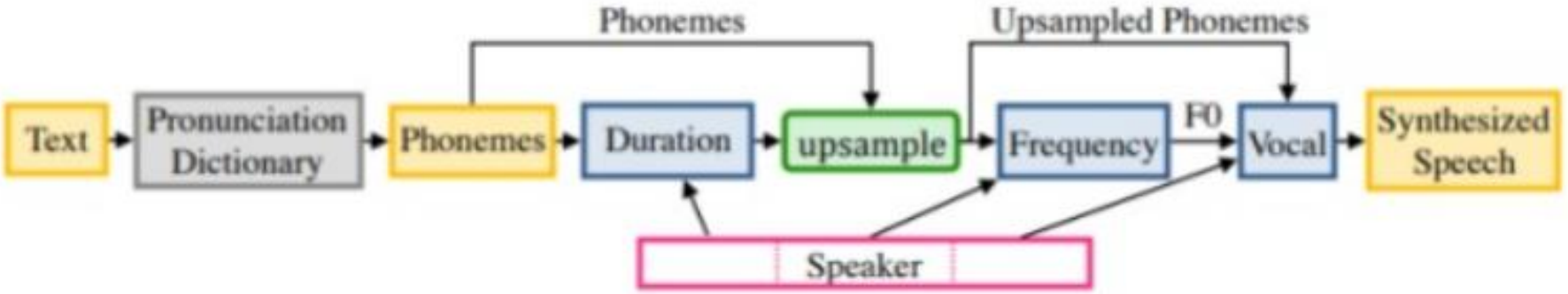
Deep Voice: Real-time Neural Text-to-Speech



The **modified WaveNet** architecture: teal inputs, green convolutions and QRNNs, yellow unary operations and softmax, pink binary operations, and indigo reshapes, transposes, and slices.

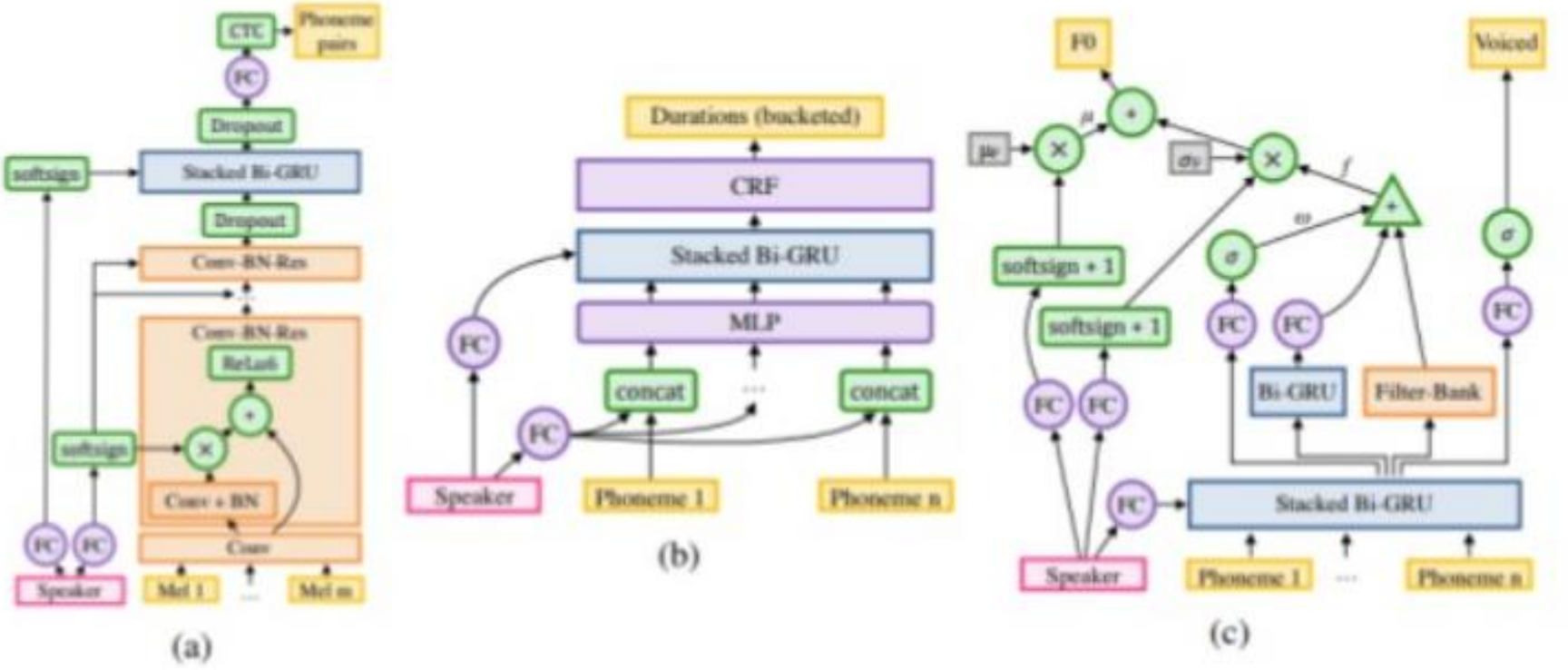
Deep Voice 2: Multi-speaker neural TTS

- A technique for augmenting neural TTS with low dimensional trainable speaker embeddings to generate different voices from a single model.
- Deep Voice 2, based on a similar pipeline with Deep Voice 1.
- Improve Tacotron by introducing a post-processing neural vocoder.



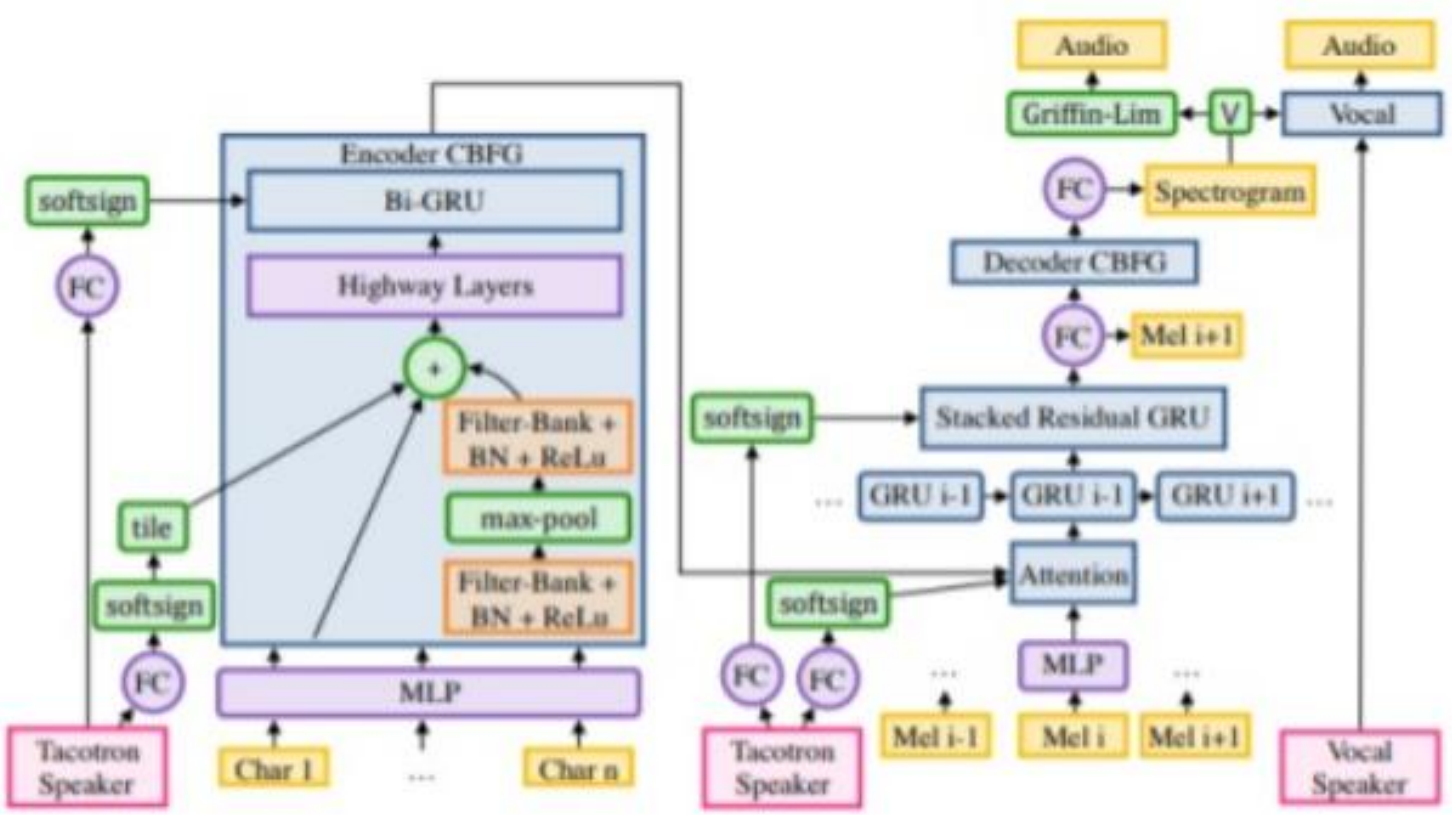
Inference system diagram

Deep Voice 2: Multi-speaker neural TTS



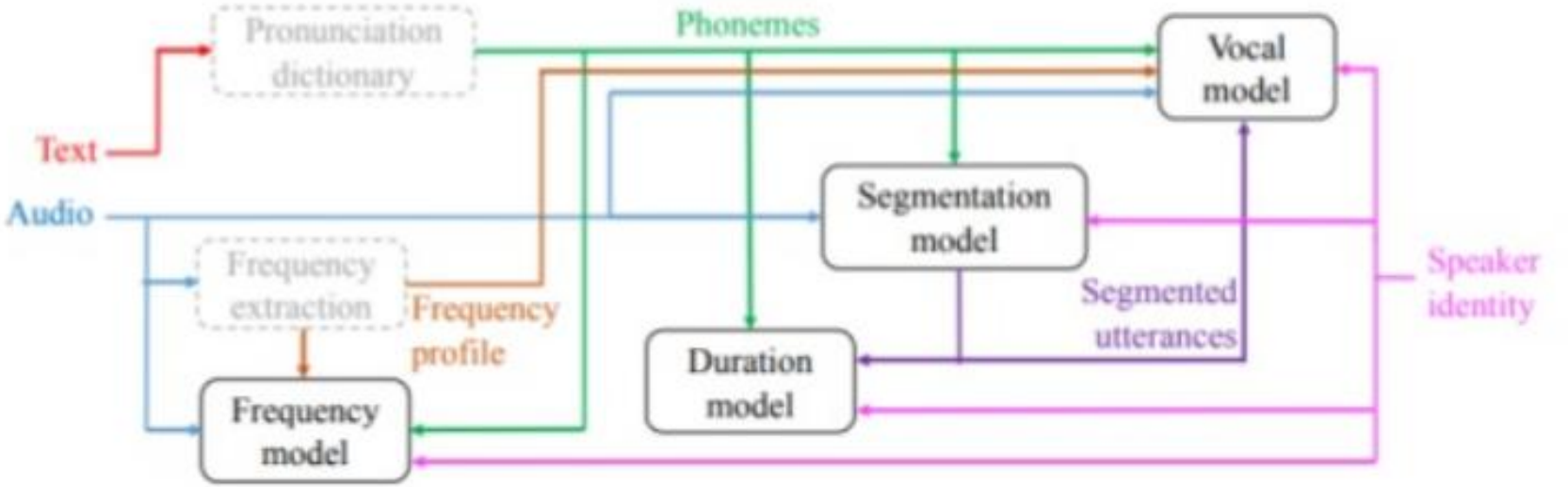
Architecture for multi-speaker (a) segmentation, (b) duration, and (c) frequency model.

Deep Voice 2: Multi-speaker neural TTS



Tacotron with speaker conditioning in the Encoder CBHG module and decoder with two Phonemes Upsampled Phonemes ways to convert spectrogram to audio: Griffin-Lim or our speaker-conditioned Vocal model.

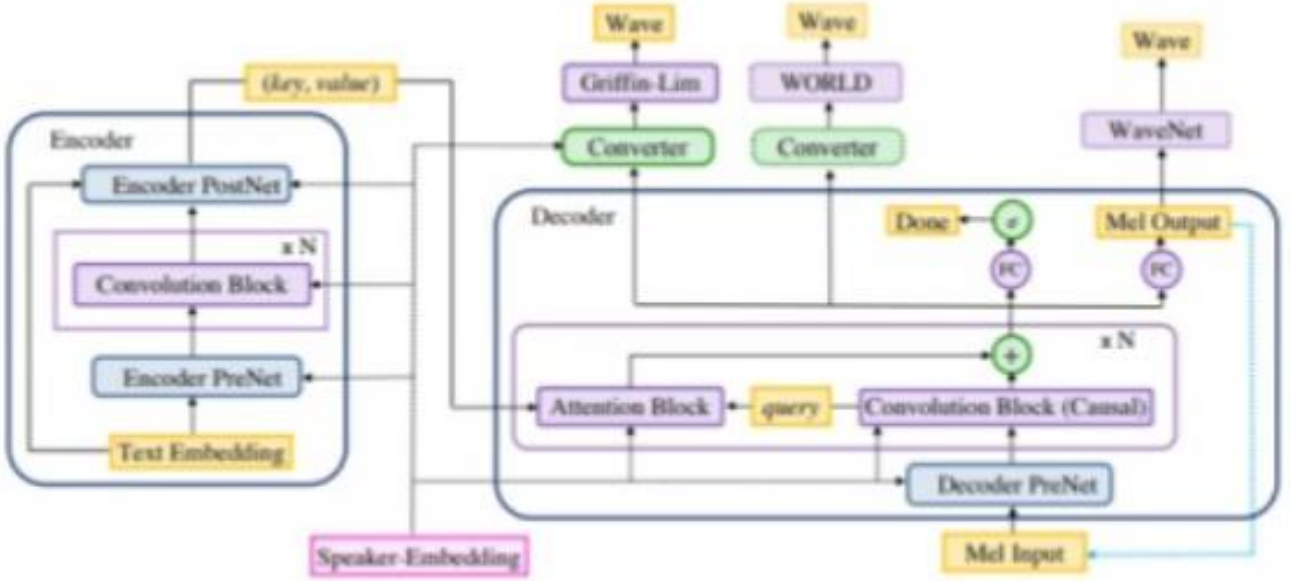
Deep Voice 2: Multi-speaker neural TTS



System diagram for training procedure for Deep Voice 2

Deep Voice 3: Scaling TTS with Convolutional Sequence Learning

- Deep Voice 3, a fully-convolutional attention-based neural TTS system.



Deep Voice 3 uses residual convolutional layers to encode text into per-timestep key and value vectors for an attention-based decoder. The decoder uses these to predict the mel-scale log magnitude spectrograms that correspond to the output audio. The hidden states of the decoder are then fed to a converter network to predict the vocoder parameters for waveform synthesis.

Sisteme de sinteză text-vorbire pentru limba română

- **Ivona** – Carmen, http://www.ivona.com/en/
- **Nuance** – Simona, www.nuance.com
- **Loquendo** – Ioana, www.loquendo.com
- **Baum** – Ancutza, Simona, www.baum.ro
- **Romanian Formant Synthesis** [Jitcă et al., 2002]
- **RomVox+** [Ferencz, 2000]
- **BRVox** [Bodo, 2009]
- **LIGHTVOX** [Buza, 2010]
- **RSS** [Stan 2011] www.romaniantts.com/new/rssdb/rssdb.html

- **AT&T Bell Labs**
- **MBROLA** http://tcts.fpms.ac.be/synthesis/mbrola.html, [Dutoit et al., 1996]
- **Phobos TTS** http://www.phobos.ro/demos/tts/index.html
- **eSpeak** http://espeak.sourceforge.net/
- **LingvoSoft Talking Dictionary** http://www.lingvosoft.com/
- https://www.slideshare.net/CynthiaKing30/a-short-introduction-to-texttospeech-synthesis

https://www.cepstral.com/en/demos

http://www.fromtexttospeech.com/

http://www.oddcast.com/home/demos/tts/tts_example.php

https://www.linguatec.de/en/demo/

https://www.naturalreaders.com/

http://www.voicedream.com/reader/

https://encyclopedia.pub/entry/2279

https://www.mathworks.com/help/dsp/ug/lpc-analysis-and-synthesis-of-speech.html

https://www.slideshare.net/CynthiaKing30/a-short-introduction-to-texttospeech-synthesis

IPA Symbol	ARPAbet Symbol	Word	IPA Transcription	ARPAbet Transcription
[p]	[p]	parsley	['parsli]	[p aa r s l iy]
[t]	[t]	tarragon	['tærəgɑn]	[t ae r ax g aa n]
[k]	[k]	catnip	['kætnip]	[k ae t n ix p]
[b]	[b]	bay	[beɪ]	[b ey]
[d]	[d]	dill	[dɪl]	[d ih l]
[g]	[g]	garlic	['gɑrlik]	[g aa r l ix k]
[m]	[m]	mint	[mɪnt]	[m ih n t]
[n]	[n]	nutmeg	['nʌtmɛg]	[n ah t m eh g
[ŋ]	[ng]	ginseng	['dʒɪnsɨŋ]	[jh ih n s ix ng]
[f]	[f]	fennel	['fɛnl̩]	[f eh n el]
[v]	[v]	clove	[kloʊv]	[k l ow v]
[θ]	[th]	thistle	['θɪsl̩]	[th ih s el]
[ð]	[dh]	heather	['hɛðɚ]	[h eh dh axr]
[s]	[s]	sage	[seɪdʒ]	[s ey jh]
[z]	[z]	hazelnut	['heɪzl̩nʌt]	[h ey z el n ah t]
[ʃ]	[sh]	squash	[skwɑʃ]	[s k w a sh]
[ʒ]	[zh]	ambrosia	[æm'broʊʒə]	[ae m b r ow zh ax]
[tʃ]	[ch]	chicory	['tʃɪkɚi]	[ch ih k axr iy]
[dʒ]	[jh]	sage	[seɪdʒ]	[s ey jh]
[l]	[l]	licorice	['lɪkɚɨʃ]	[l ih k axr ix sh]
[w]	[w]	kiwi	['kiwi]	[k iy w iy]
[r]	[r]	parsley	['pɑrsli]	[p aa r s l iy]
[j]	[y]	yew	[yu]	[y uw]
[h]	[h]	horseradish	['hɔrsrædɪʃ]	[h ao r s r ae d ih sh]
[ʔ]	[q]	uh-oh	[ʔʌʔoʊ]	[q ah q ow]
[ɾ]	[dx]	butter	['bʌɾɚ]	[b ah dx axr]
[ɾ̃]	[nx]	wintergreen	[wɪɾ̃ɚgrin]	[w ih nx axr g r i n]
[l̩]	[el]	thistle	['θɪsl̩]	[th ih s el]

Figure 4.1 IPA and ARPAbet symbols for transcription of English consonants.

IPA Symbol	ARPAbet Symbol	Word	IPA Transcription	ARPAbet Transcription
[i]	[iy]	lily	['lɪli]	[l ih l iy]
[ɪ]	[ih]	lily	['lɪli]	[l ih l iy]
[eɪ]	[ey]	daisy	['deɪzi]	[d ey z i]
[ɛ]	[eh]	poinsettia	[pɔɪn'sɛriə]	[p oy n s eh dx iy ax]
[æ]	[ae]	aster	['æstɚ]	[ae s t axr]
[ɑ]	[aa]	poppy	['pɑpi]	[p aa p i]
[ɔ]	[ao]	orchid	['ɔrkɪd]	[ao r k ix d]
[ʊ]	[uh]	woodruff	['wʊdrʌf]	[w uh d r ah f]
[oʊ]	[ow]	lotus	['loʊɾəs]	[l ow dx ax s]
[u]	[uw]	tulip	['tulɪp]	[t uw l ix p]
[ʌ]	[uh]	buttercup	['bʌɾɚˌkʌp]	[b uh dx axr k uh p]
[ɝ]	[er]	bird	['bɝd]	[b er d]
[aɪ]	[ay]	iris	['aɪrɪs]	[ay r ix s]
[aʊ]	[aw]	sunflower	['sʌnflaʊɚ]	[s ah n f l aw axr]
[ɔɪ]	[oy]	poinsettia	[pɔɪn'sɛriə]	[p oy n s eh dx iy ax]
[ju]	[y uw]	feverfew	[fivɚfju]	[f iy v axr f y u]
[ə]	[ax]	woodruff	['wʊdrəf]	[w uh d r ax f]
[ɨ]	[ix]	tulip	['tulɪp]	[t uw l ix p]
[ɚ]	[axr]	heather	['hɛðɚ]	[h eh dh axr]
[ʉ]	[ux]	dude[2]	[dʉd]	[d ux d]

Figure 4.2 IPA and ARPAbet symbols for transcription of English vowels.

TABELUL DIFONEMELOR LIMBII ROMÂNE UTILIZATE ÎN IMPLEMENTAREA SISTEMULUI DE SINTEZĂ TEXT-VORBIRE

	al doilea fonem																											
	a	ă	b	k	č	d	e	f	g	ğ	h	i	i,	î	j	l	m	n	o	p	r	s	ş	t	ţ	u	v	z	-
a	*		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
ă			*	*	*	*		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
b	*	*		*	*	*	*	*	*			*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
k	*	*			*		*					*	*	*		*		*	*		*	*	*	*	*	*	*		*	*
č	*	*					*					*	*	*					*								*			*
d	*	*	*		*		*			*	*	*	*	*	*		*	*	*		*		*				*	*		*
e	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
f	*	*					*					*	*	*		*		*	*		*				*	*	*			*
g	*	*					*					*	*	*	*	*	*	*	*		*						*		*	*
ğ	*	*					*					*	*	*					*								*			*
h	*	*					*	*				*	*	*		*		*	*		*				*		*			*
i	*		*	*	*	*	*	*	*	*	*	*	*		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
i,	*	*	*	*	*	*	*	*	*	*	*	*			*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
î	*	*	*	*	*	*	*	*	*	*	*	*	*		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
j	*	*	*			*	*		*			*	*	*					*	*							*			*
l	*	*	*	*	*	*	*	*	*	*		*	*	*			*	*	*	*	*		*	*	*	*	*	*		*
m	*	*	*		*		*	*				*	*	*		*		*	*	*	*	*	*	*	*	*	*	*	*	*
n	*	*		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
o	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
p	*	*			*	*	*					*	*	*		*		*	*		*	*	*	*	*	*	*			*
r	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
s	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		*	*	*	*	*	*	*	*
ş	*	*		*			*	*				*	*	*	*	*	*	*	*	*	*	*			*		*	*		*
t	*	*	*	*	*	*	*	*	*			*	*	*	*	*	*	*	*	*	*	*	*	*	*		*	*		*
ţ	*	*		*			*					*	*	*					*								*			*
u	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
v	*	*					*					*	*	*		*		*	*		*						*		*	*
z	*	*	*			*	*		*			*	*	*	*	*	*	*	*		*				*		*	*		*
-	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	