Computer Audio

Speech Synthesis

Master's Degree in Electrical and Computer Engineering

Faculdade de Engenharia da Universidade do Porto

Introduction

Speech Synthesis (SS or TTS)

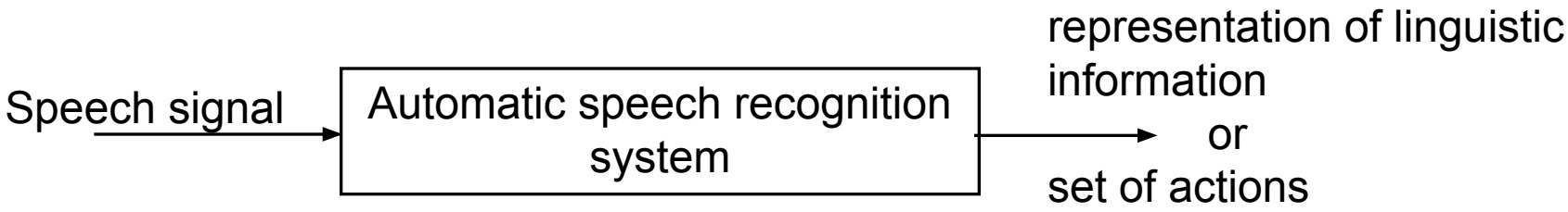
- Speech Synthesis (SS) or Text-to-speech synthesis (TTS) is based on the automatic production of the acoustic content of the speech signal from the respective written content,

  It is necessary to consider that the human informant or speaker, once chosen, is unique, with his or her audio profile, with the texts to be “read” or converted to synthetic speech being variable.

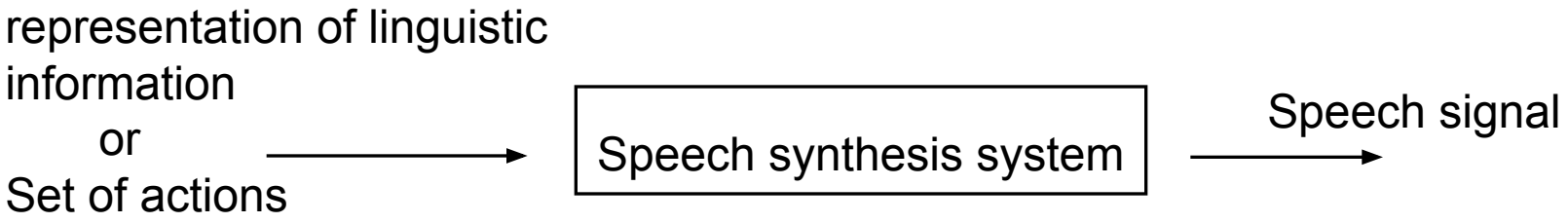
- TTS systems automatically recognize patterns, this time associated with the input text to convert into a speech signal:
-
  - classifying them into certain structural (grammatical) representations and associating the respective acoustic “readings”
  - or leading to the production of sound utterances (announcements, phrases, etc.) in response to a certain set of actions

Comparison with speech recognition

- Speech recognition function



- Speech synthesis function



TTS difficulty factors

- Pronunciation complexity:
  - extensive vocabularies (>10^5 words) – ambiguity (homographs: spoon (cutlery) and spoon (pick up)) between words is frequent
  - complex linguistic structures influence utterance – levels: lexical; syntactic; semantic; pragmatic
  - realistic oral production – the symbols of the text to be synthesized must be read in full, numerals, dates, etc.)
  - combination of linguistic information together with para-linguistic (prosody) and extra-linguistic (emotional states, etc.) information

- Continuous nature of the synthetic speech signal
  - dilution of boundaries between linguistic units
  - pronounced co-articulation effects

TTS difficulty factors *(cont.)*

- Variability of acoustic trajectories and their temporal evolution
  - strongly depends on the speaker (in general; accentuated by local or regional accents)
  - also varying (although to a lesser extent, in general) for each speaker, depending on the speed of speech, physical or mental state, etc..

- The problem of SF is multi-disciplinary (psycho-acoustics, acoustics, linguistics) - it is a vast and complex task to develop techniques capable of effectively integrating this knowledge to obtain sufficient naturalness of synthetic speech.

Examples of TTS Uses

- Telecommunications (telephone access systems to data and services; etc.)
- “Office” environment (reading documents; diaries; controlling computer systems; etc.)
- Support systems for specific professional activities (reading medical reports; memos, etc.)
- Factory environment (voice outputs; information outputs; etc.)
- Support systems for disabled people (screen reader, systems without looking: clock, compass, etc.)
- Messages from household utensils and appliances for control and information
- Leasure activities, etc.

Levels of speech analysis

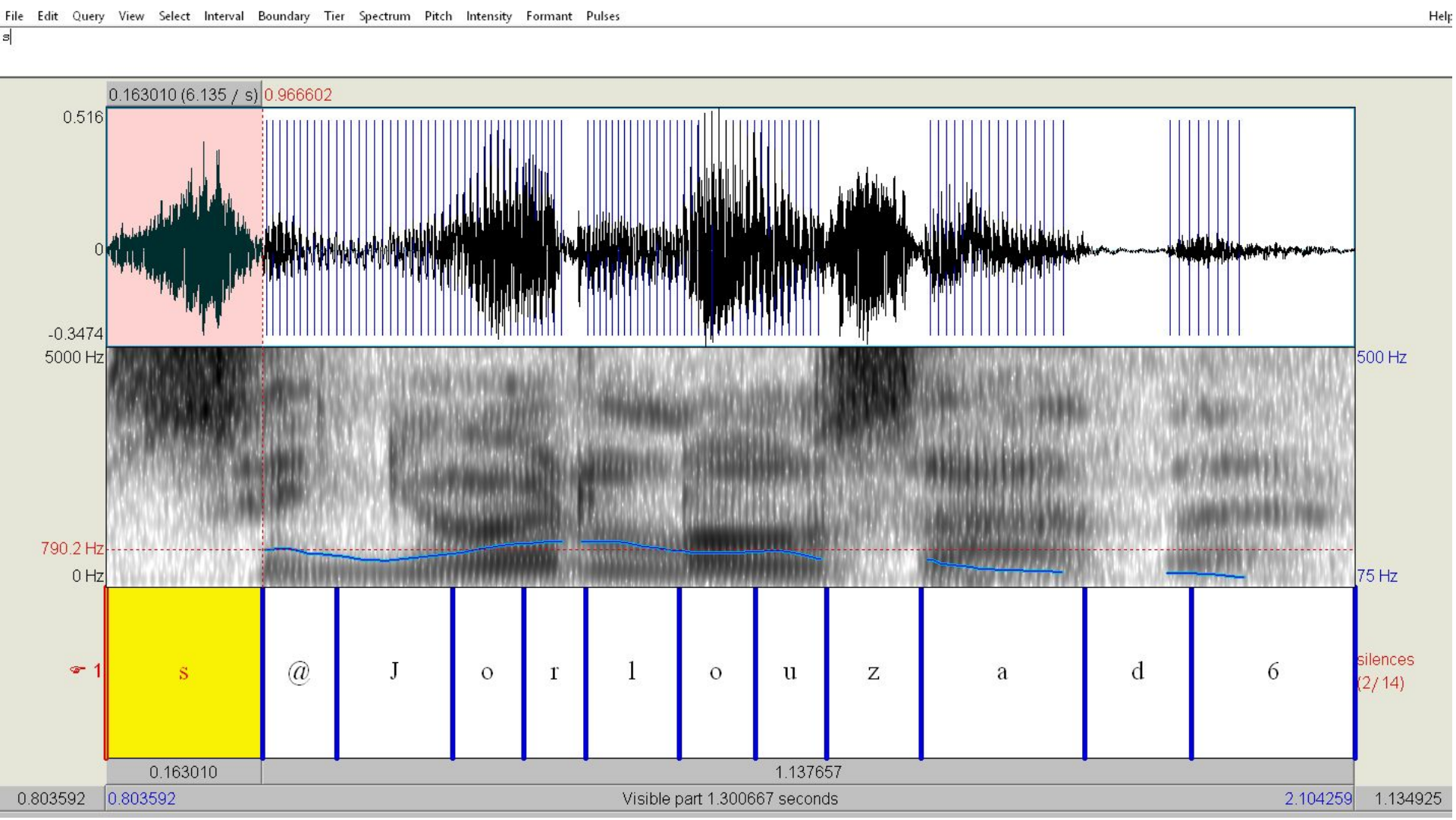
Level…	Speech is…
Application	The interaction between people and speech-based systems
Utterance	A message defined by speech
Segment	Sufficiently large segment to distinguish one word from the others
Signal	Analogue of pressure or velocity waves

ref[a]: Witten, 1986, in Chris Rowden, “Speech Processing”, McGraw-Hill, 1992.

Example of one utterance (via tables)

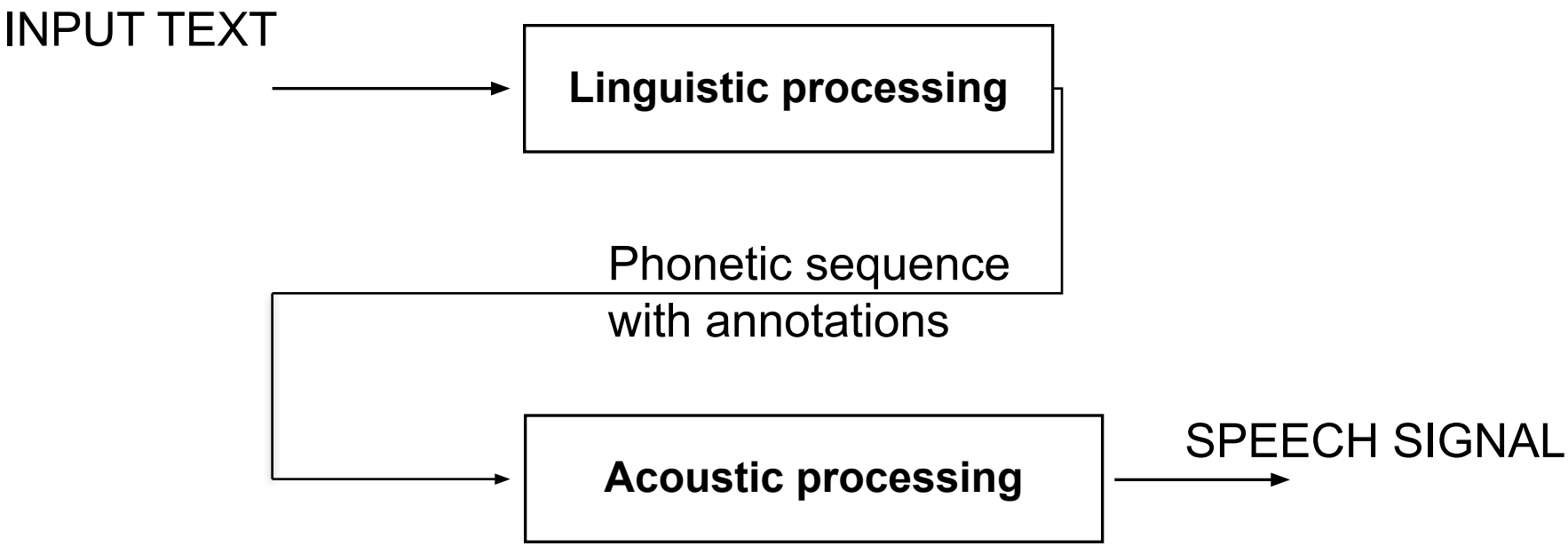
Phrase (texto)	Sr.					Lousada						
Pronounceable text	Senhor					lousada						
Phonemic segments (SAMPA)	/s@Jor/					/louzad6/						
Linguistic function	noun					noun						
Phrase function Phrase mark-up	Noun group Beginning of phrase											end of phrase
Segments list (alofones)	s	@	J	o	r	l	o	u	z	a	d	6
accentuation				1						1		
f0 (Hz) (x - undefined)	x	138	127	146	155	150	137	135	x	108	100	96
Duration (ms)	163	77	121	74	65	97	80	74	100	168	111	170
Control parameters (10 ms) or waveform												
Total word duration(ms)	500					800						

Example of a Phrase (cont.)



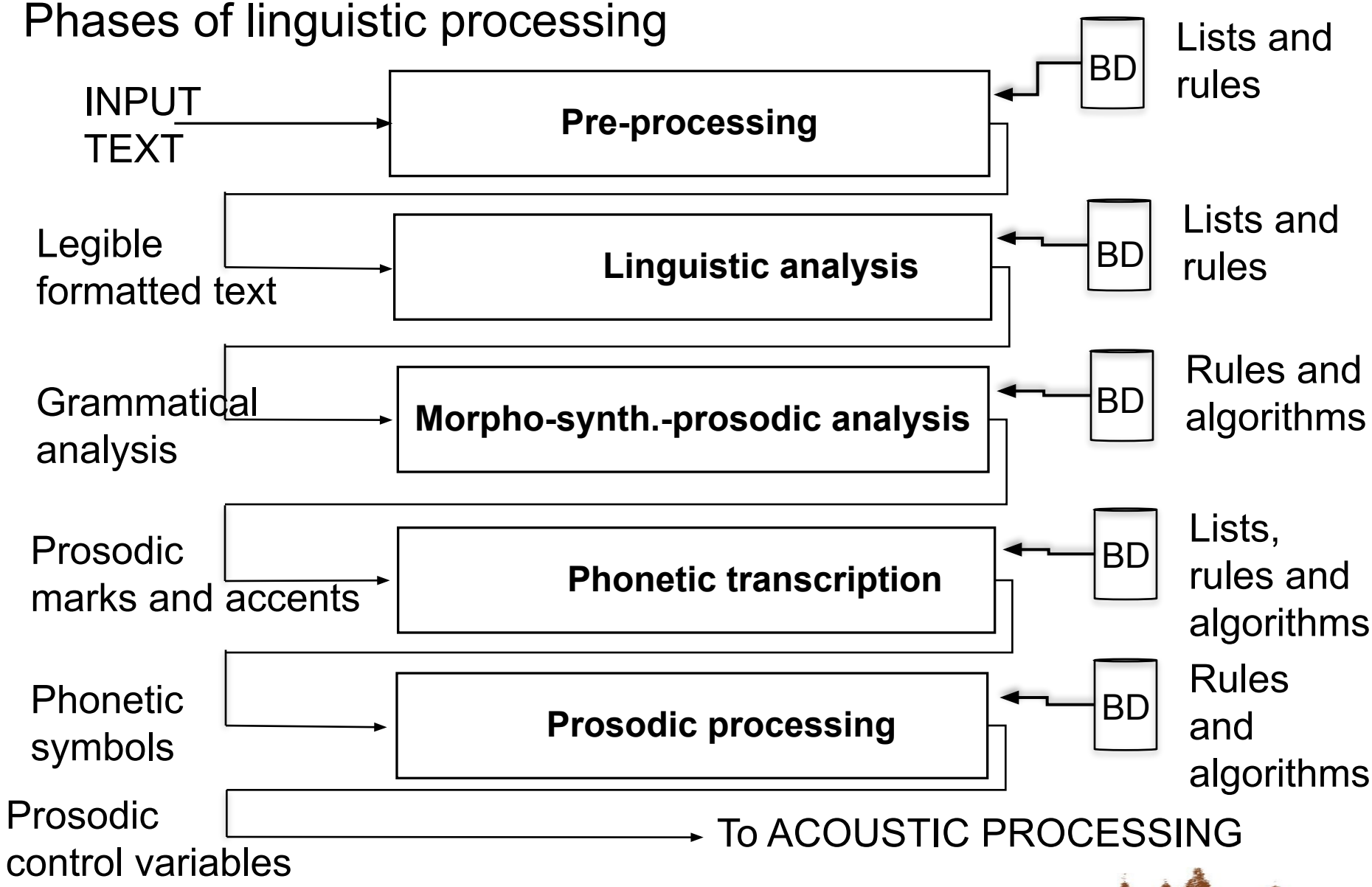
Text-to-speech conversion

- Text-to-speech (TTS) conversion phases



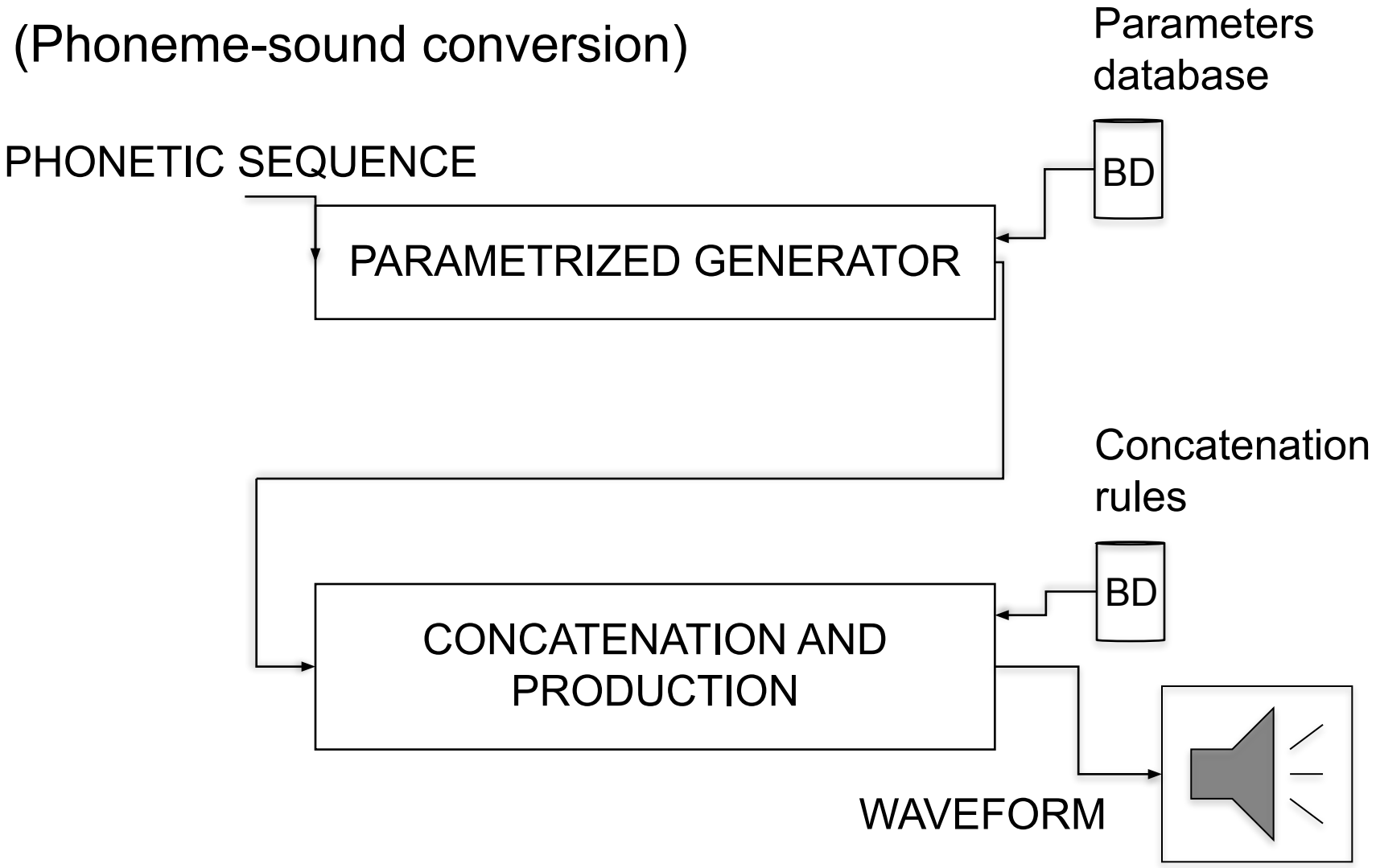
Linguistic Processing

- Phases of linguistic processing



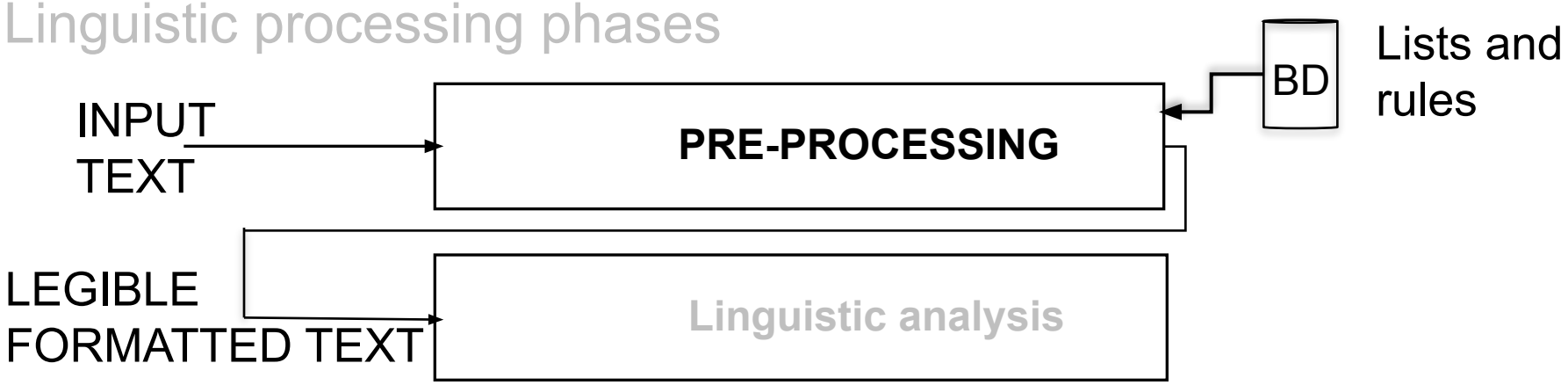
Acoustic Processing

- Acoustic processing phases
  (Phoneme-sound conversion)



Pre-processing

- Linguistic processing phases



Unrestricted conversion of text to readable text:

(Word or phoneme substitution)

- Characters conversion;
- Numerals
- Abbreviations
- Acronyms.

PRE-PROCESSING (CONT.)

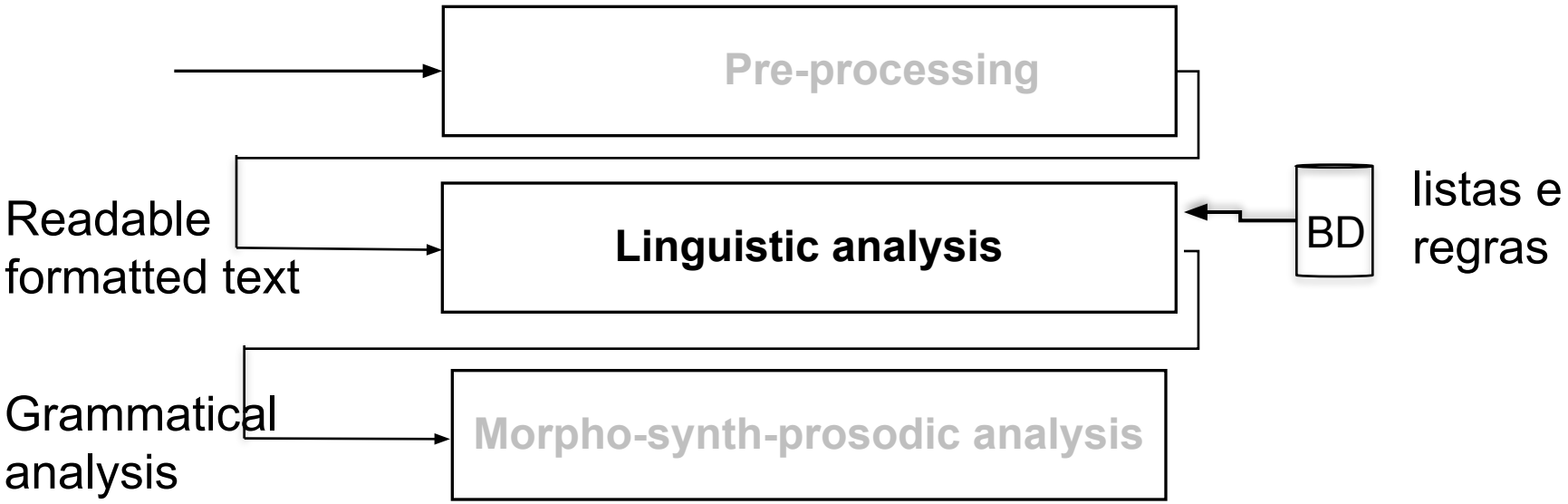
Unrestricted text into readable text conversion - examples:

(Replacement of word or phoneme substitution

- Characters conversion: ex:
  - units: ms -> milisecond; diverse: 4/8-> four in eight…
- Numerals: ex:
  - cardinals: 1 000 023 -> one million and twenty three;
  - ordinals: 346º -> three hundred and fourty-sixth;
  - romans: MMXXII -> two thousand; and twenty-two
  - decimals: 2,2 -> two comma two;
  - scientific: 1,5e-5 -> one comma five times tem elevated to 5;
  - hexadecimal: FFA3 -> éfe éfe ei three;
  - telephone; 225081837 -> two-two, five-zero-eight, one-eight-three-seven;
  - monetary: 12.500€50 -> twelve thousand and five-hundred euros and fifty cents;
  - time: 21:45 -> twenty one (hours) and fourty five (minutes); 7:37 pm – seven and thirty seven in the afternoon…
- Abbreviations:
  - Ms -> madam; engº -> engineer; sat. -> saturday
- Acronyms:
  - UN -> /iuen/; OCDE -> /Ocidii/; FEUP -> /feup/ (faculty of engineering of the university of porto)

Linguistic analysis

- Phases of linguistic processing



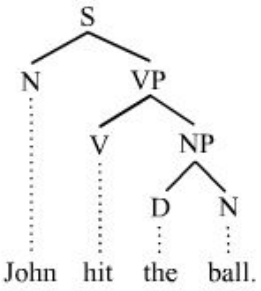
Analysis and classification of the grammatical functions of each word:

- lexical;
- morphologic
- synthatic
- semantic

Linguistic analysis (cont.)

Analysis and classification of the grammatical functions of each word (see “Vocabulário ortográfico Comum da Língua Portuguesa” at https://voc.cplp.org/index.php?action=von&csl=pt):

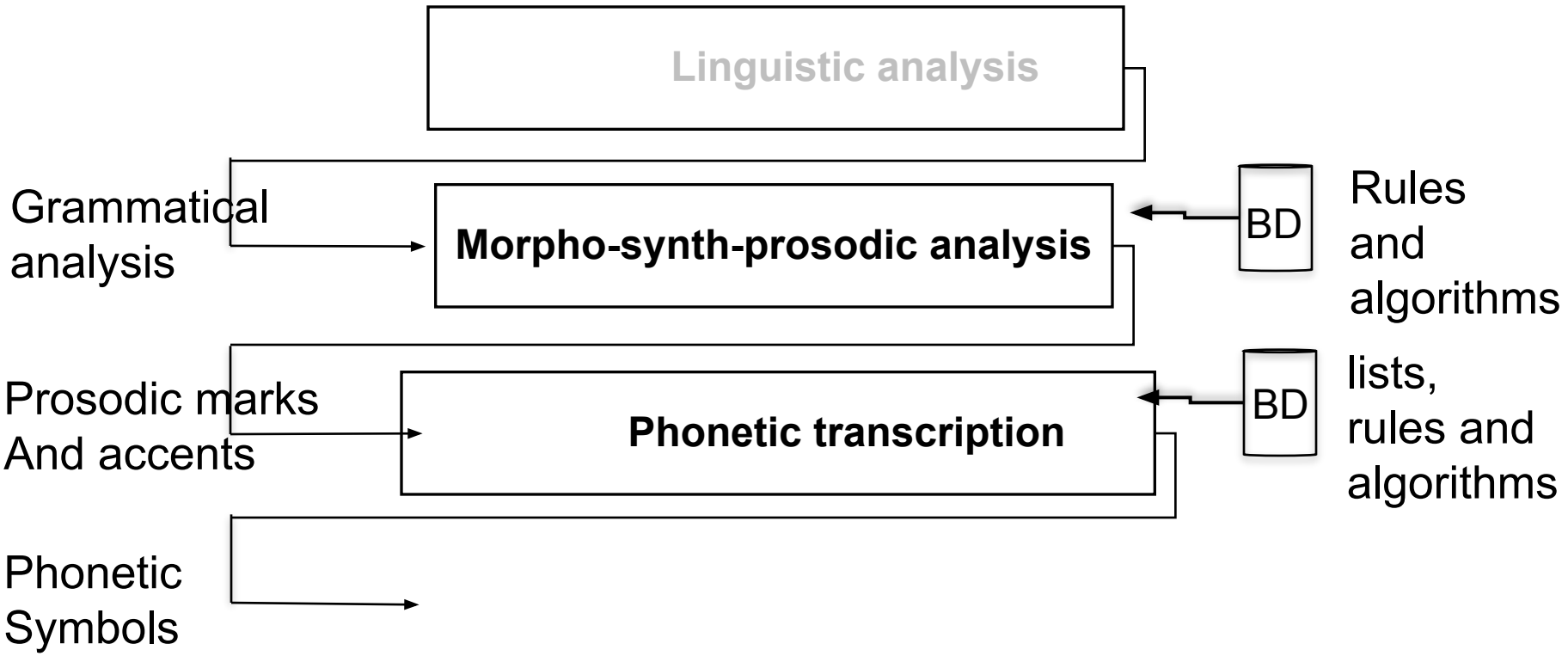
- Lexicologic: uses a dictionary with a relevant lexicon (verb forms, common expressions), also containing a table of prefixes and suffixes and grammatical content rules for entries that do not exist in the dictionary; determines symbols (tokens)
- Morphologic: analysis of words (*tokens*). Base is a *morpheme* named (*tronco)* stem, free *morpheme*, which is the basis of the meaning, to which other *morphemes* denominated *affixes*, allowing *preffixes* and/or *suffixes:*
  - *ex: in feliz mente*
- Synthatic: is the study of the logic meaning of specific frases or parts sentences, it consists in obtaining an analysis of the text in the a structural tree through the use of na analyser software (par



(see at: https://www.analyticssteps.com/blogs/syntactic-analysis-overview)

- Semantics: study of the exact meaning of the text in its ensemble to disambiguate meanings of words and establish relationships. It has impact on prosody.

- Phases of linguistic processing



To annotate syntactic-prosodic frontiers and define accents:

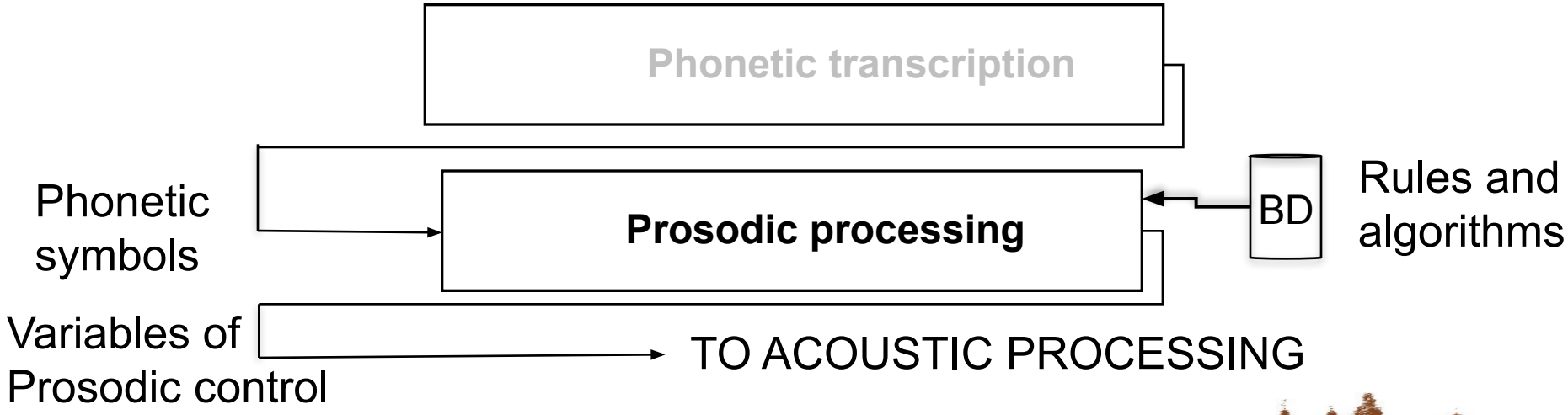
- Boundaries are defined by the logical relationship between consecutive structures
- Accents are defined by emphasis.

To do the phonetic transcription:

- Prepare the sequence of phonetic codes from the text with markings from previous tasks
- Context-dependent rules are used that also take into account pauses in the context, accentuation and co-articulation effects.
- In Portuguese, the conversion of vowels in the text to phonetic representation is quite complex. It depends on the position in the word, adjacent phonemes and accentuation.
- The process consists of 2 steps:
  - Tabular rules, using a decreasing priority listing of rules and a character increment in function of the conversions that are done
  - Rules in the form of a procedure, executed by an algorithm that is more efficient than a table because it is more generic, for example: [vog][s][vog] (text: casa) -> [vog][z][vog] (SAMPA: caza).

PROSODIC PROCESSING

- Phases of linguistic processing

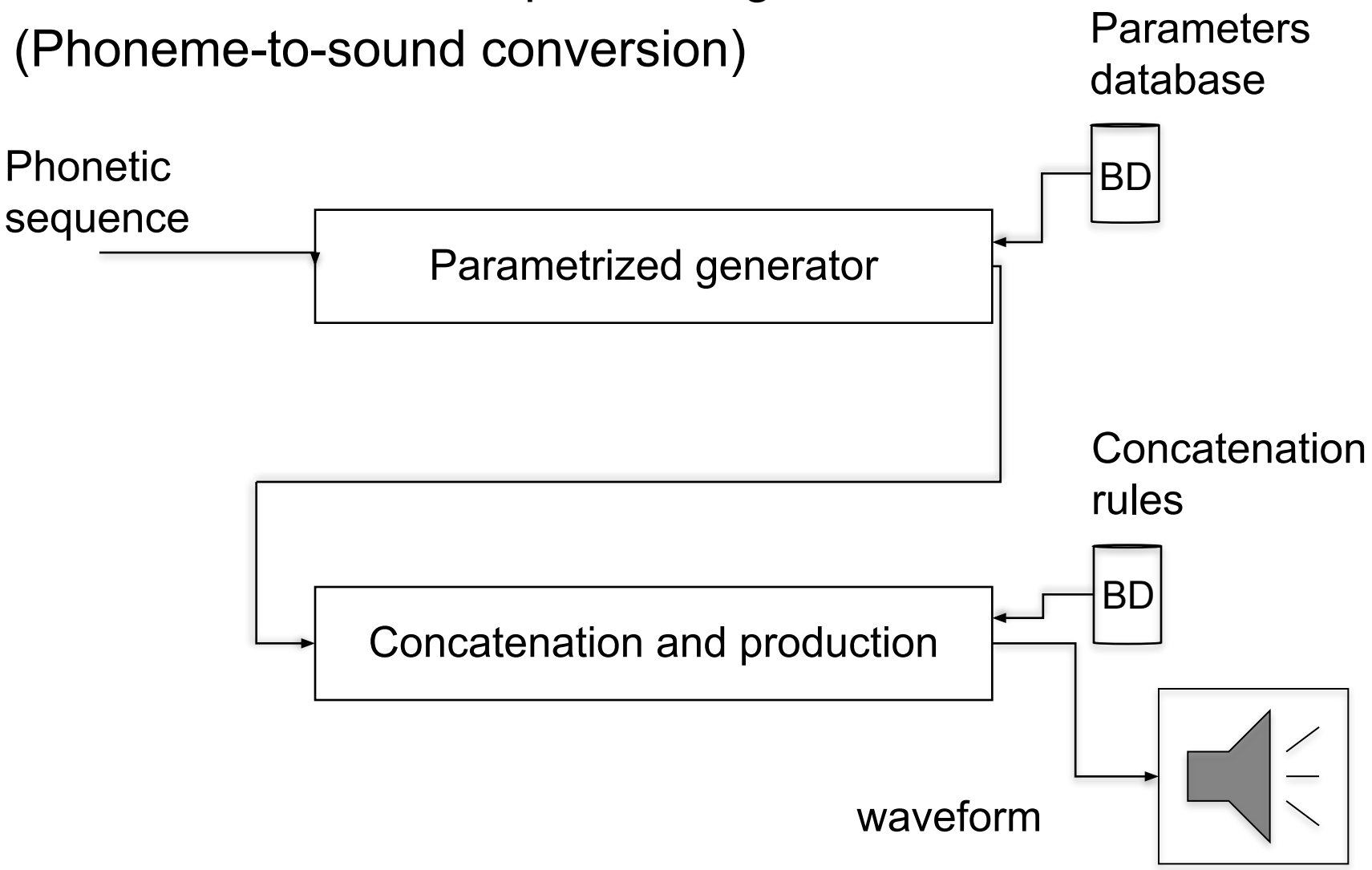


PROSODIC PROCESSING (CONT.)

- Adjustments to some phonetic codes, for example nasal vowels:
  - [e][m][cons] -> [e~][cons]
- Marking and adjustment of the tonic syllable:
  - Graphemes already marked produce the marking of the stressed syllable of the word; articles, conjunctions and prefixes are not accented
  - The absence of accent marks (accents) produces in words ending in {al, el, ol, ar, er, ir, az, ez, iz, oz, uz} an accent on the last syllable
  - In the absence of accents and in cases not yet covered, the general rule for accentuation is that of the penultimate syllable.
  - The marking of the stressed vowel is associated with a double increase in duration, an increase in f0 and amplitude, decreasing until the end of the word.
- Ajuste da entoação das orações:
  - Declarative case: decay of f0 before the end of the sentence, following the accentuation of the penultimate word. In the last word there is no variation in the stressed syllable
  - Interrogative case, 1- no interrogative word - rise in intonation on the last word with final descent unless it is accentuated on the last syllable; 2- with an interrogative word (how, where, which, etc.): a rise is also added at the beginning of the interrogative word and a sharp descent at the beginning of the next word
  - In the presence of a comma, in addition to inserting the pause before it, f0 is dropped from the penultimate syllable of the previous word
- Adjusting the rhythm programming:
  - In addition to the already changed durations, promote the joining of articles with the next word and connect some words.

Acoustic Processing

- Phases of the acoustic processing
  (Phoneme-to-sound conversion)



ACOUSTIC PROCESSING (CONT.1)

- The input to this module is the representation of phonetic codes with prosodic marks and indication of final durations.
- This is followed by the preparation of the parameter sequence or the preparation of pre-recorded waveform units corresponding to the phonetic sequence, and then finally the waveform is generated.

The two phases are clearly interconnected and the choice of sound production tactic strongly affects the parameter sequence preparation phase..

The end result is always a concatenation of acoustic units.

Two importante cases:

- Previously stored units (segment database) with somewhat difficult prosodic processing
- Units generated by analogy to the vocal tract (database of parameters and rules with easier prosodic processing

ACOUSTIC PROCESSING (CONT.2)

- Previously stored units (segments database):

Speech unit	quantity	Duration (ms)	Memory (LPC – 2 kbps)	Computational complexity	Segmental quality	Supra-segmental quality (prosodic)
Word	40000	350	28 Mbit	Low	High	Low
Part of word	12000	250	6 Mbit	n/a	Medium-high	Medium-Low
Syllable	5000	200	2 Mb	Medium	Medium	Medium
Phoneme	40	80	6,4 kbit	Medium	Medium-Low	Medium-high
Diphone	1600 *	80	192 kbit	Alta	Medium-high	High

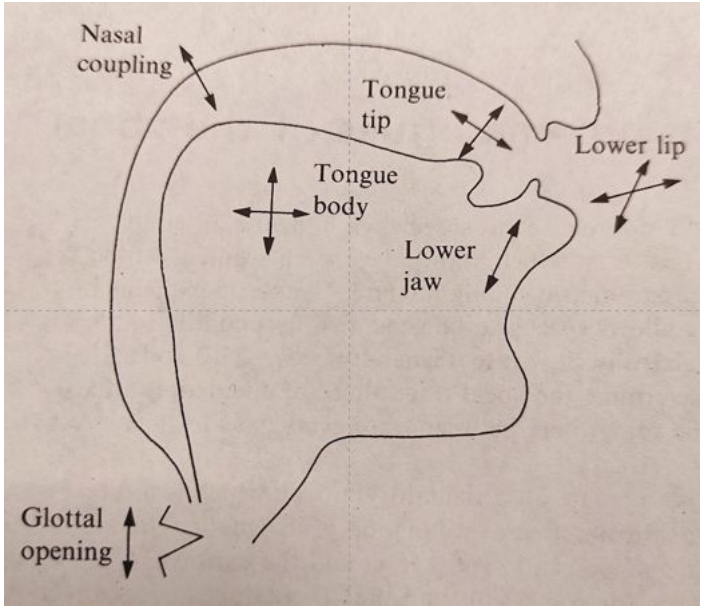
Table 6.3 in Rowden, Chris, Speech Processing, *400 only by co-articulation

ACOUSTIC PROCESSING (CONT.3)

- Units generated by analogy to the vocal tract (database of parameters and rules) with easier prosodic processing:
  - Articulatory models:

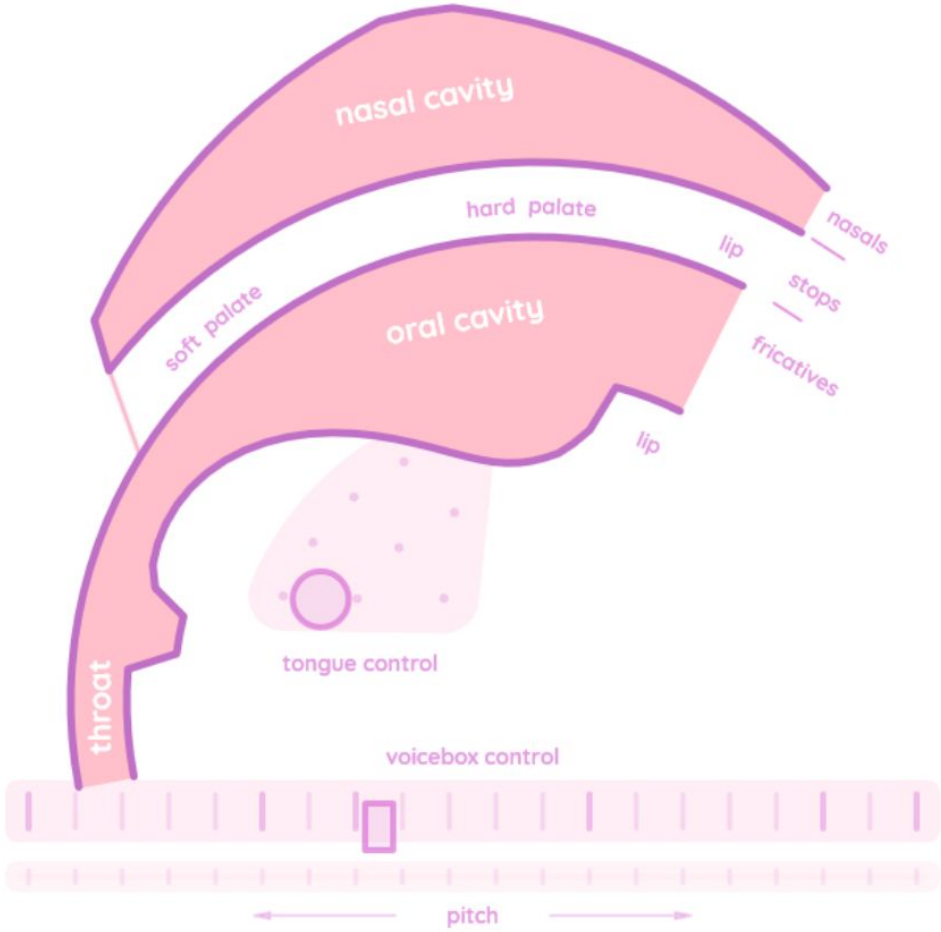
Units generated by analogy to the vocal tract (database of parameters and rules) with easier prosodic processing

  - Each configuration must have defined the very exact position and speed of the articulators – tongue, jaw, lips, velum or soft palate, glottic opening with easier prosodic processing



ACOUSTIC PROCESSING (CONT.3A)

- Pink Trombone (example):
- https://www.imaginary.org/program/pink-trombone



ACOUSTIC PROCESSING (CONT.4)

- Units generated by analogy to the vocal tract (database of parameters and rules) with easier prosodic processing:
  - Spectral models:

Based on the description of speech production through LPC or formants, with 4 resonant filters, glottal and noise pulse generator.

There are 3 configurations: cascade chain (series), parallel chain and mixed chain.

Cascade formant filters:

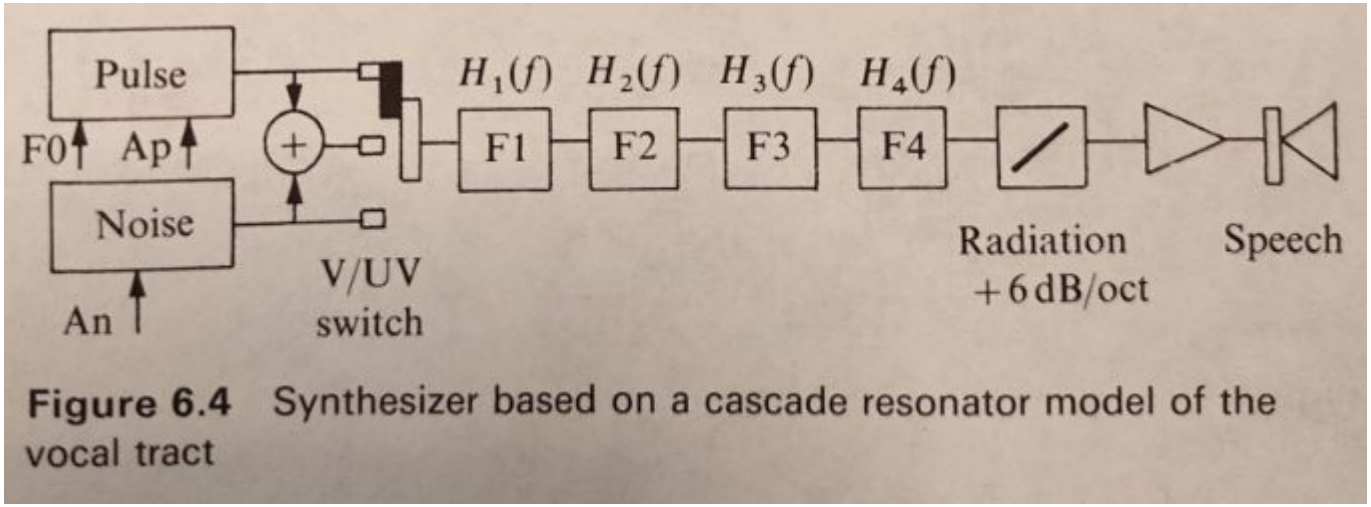


Figure 6.4 Synthesizer based on a cascade resonator model of the vocal tract

in Rowden, Chris, Speech Processing

ACOUSTIC PROCESSING (CONT.5)

Cascades formant filters:

Equations:

Complex conjugated poles zk=σk+j2πFkT e zk*

$$|z_k| = e^{-\sigma_k T}$$

$$\theta_k = 2\pi F_k T$$

Second-order filter:

$$V_k(z) = \frac{(1 - 2|z_k|\cos(2\pi F_k T) + |z_k|^2)}{(1 - 2|z_k|\cos(2\pi F_k T)z^{-1} + |z_k|^2 z^{-2})}$$

2M order filter:

$$V(z) = \prod_{k=1}^{M} V_k(z)$$

in Teixeira, João Paulo, Análise e Síntese da Fala

ACOUSTIC PROCESSING (CONT.6)

Parallel formant filters:

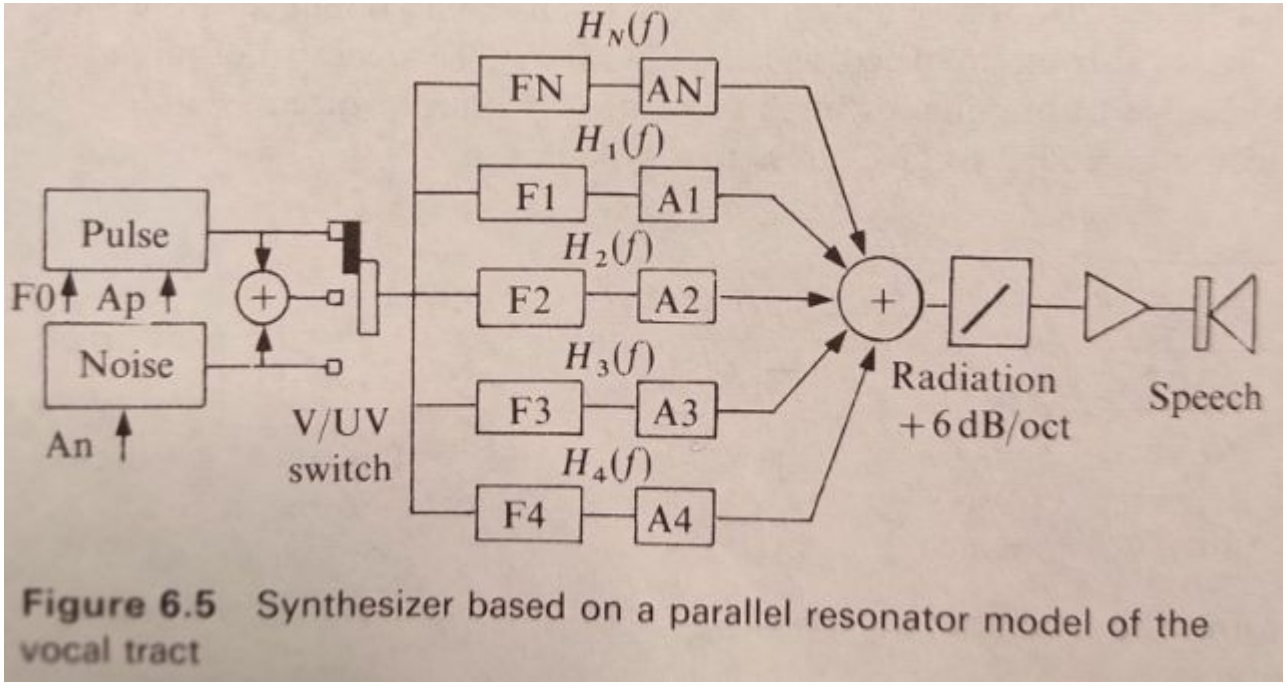


Figure 6.5 Synthesizer based on a parallel resonator model of the vocal tract

in Rowden, Chris, Speech Processing

ACOUSTIC PROCESSING (CONT.7)

Series and parallel formant filters (mixed structure):

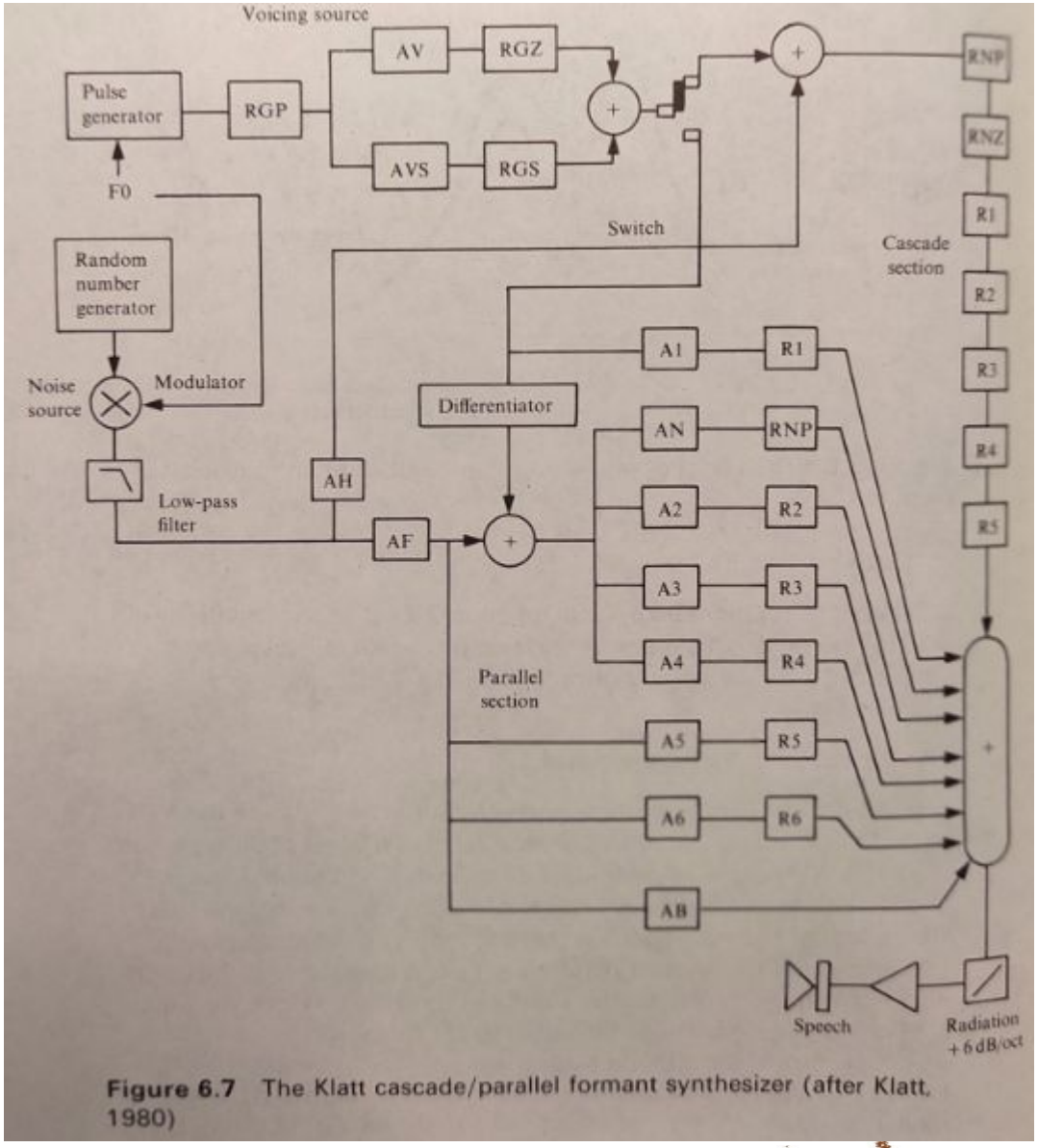
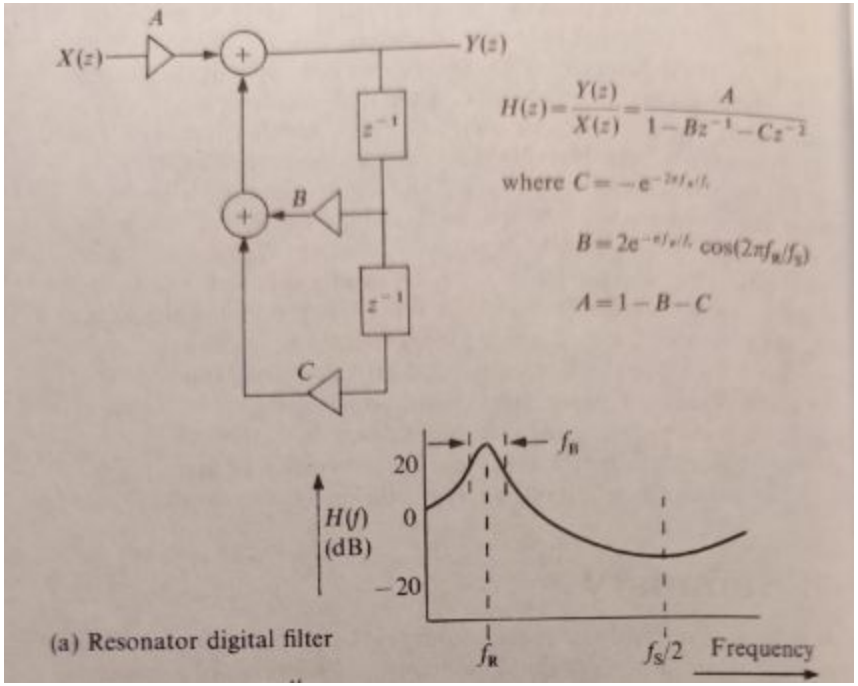


Figure 6.7 The Klatt cascade/parallel formant synthesizer (after Klatt, 1980)

in Rowden, Chris, Speech Processing

ACOUSTIC PROCESSING (CONT.8)

Ressonant and anti-ressonant formant filters:



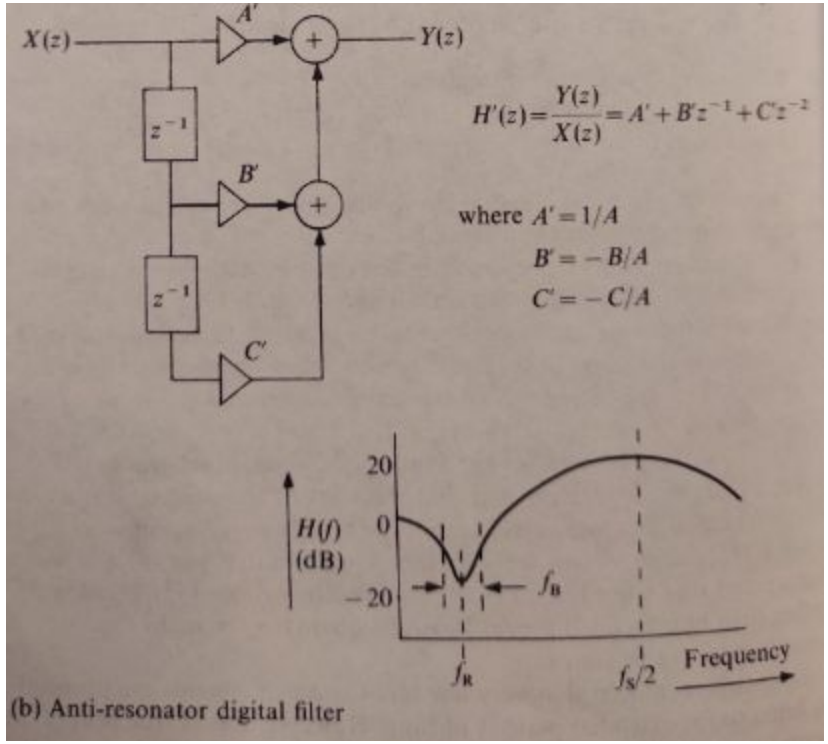


Figure 6.8 Digital filters for resonators and antiresonators

in Rowden, Chris, Speech Processing

Prosodic processing

- Prosodic changes in speech
- Pitch Synchronous Overlap and Add (PSOLA):

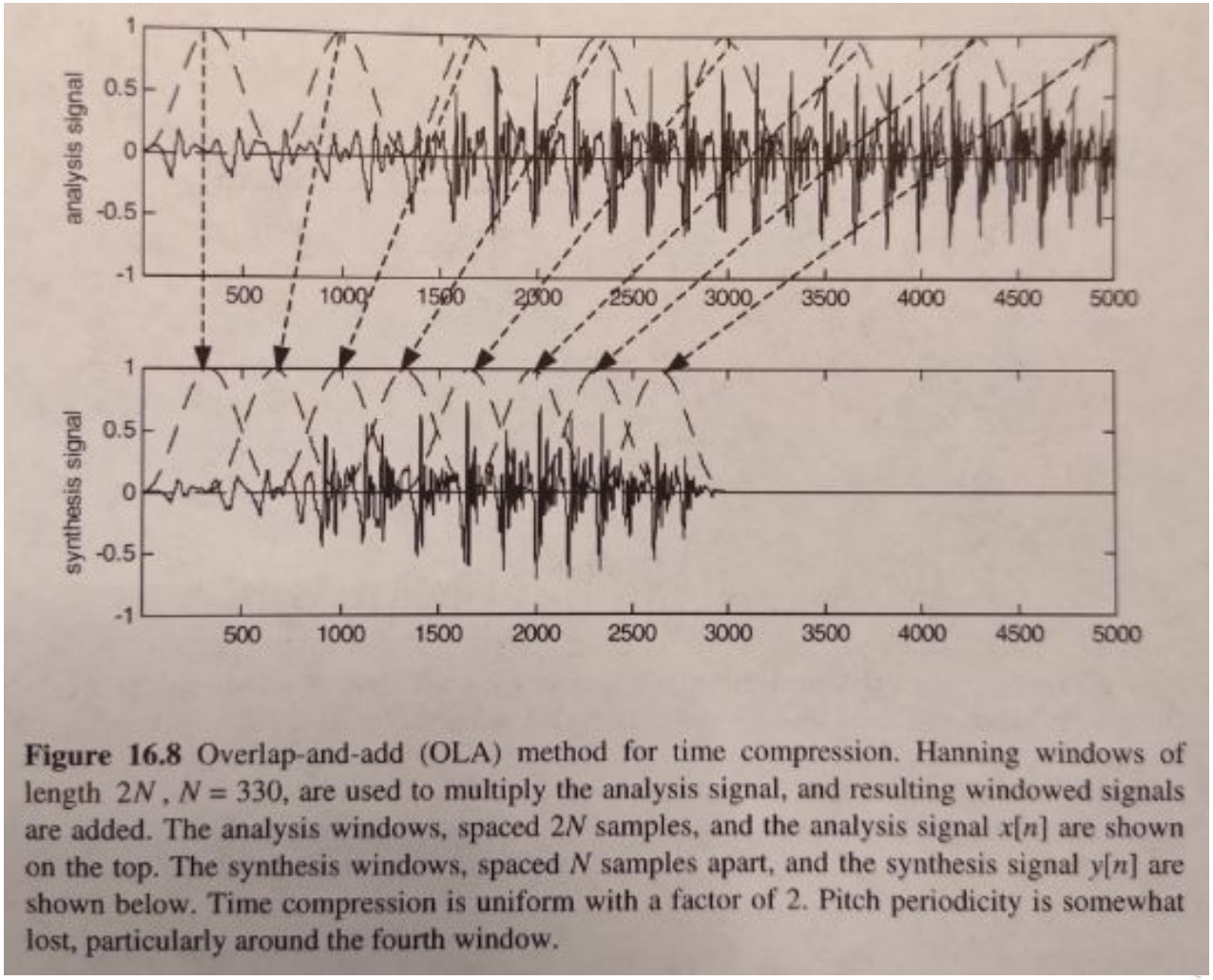


Figure 16.8 Overlap-and-add (OLA) method for time compression. Hanning windows of length $2N$, $N = 330$, are used to multiply the analysis signal, and resulting windowed signals are added. The analysis windows, spaced $2N$ samples, and the analysis signal $x[n]$ are shown on the top. The synthesis windows, spaced N samples apart, and the synthesis signal $y[n]$ are shown below. Time compression is uniform with a factor of 2. Pitch periodicity is somewhat lost, particularly around the fourth window.

Prosodic processing

- Pitch marks detection:

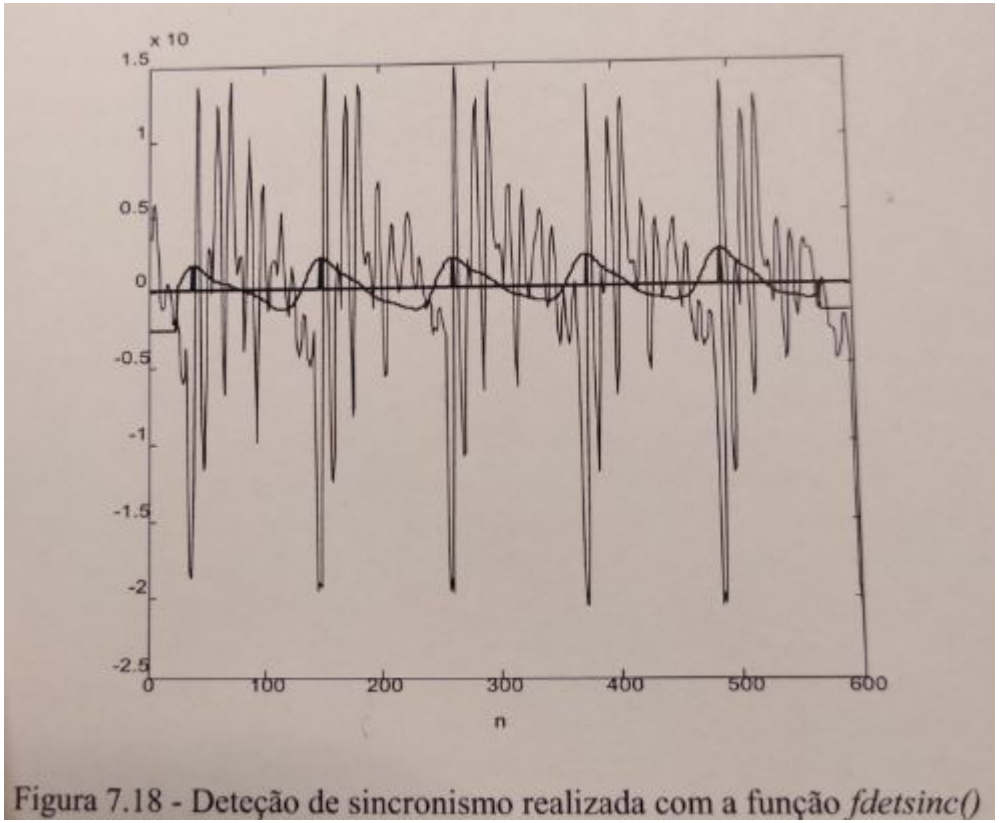


Figura 7.18 - Deteção de sincronismo realizada com a função *fdetsinc()*

Prosodic processing

- Pitch Synchronous Overlap and Add (PSOLA) (cont.):

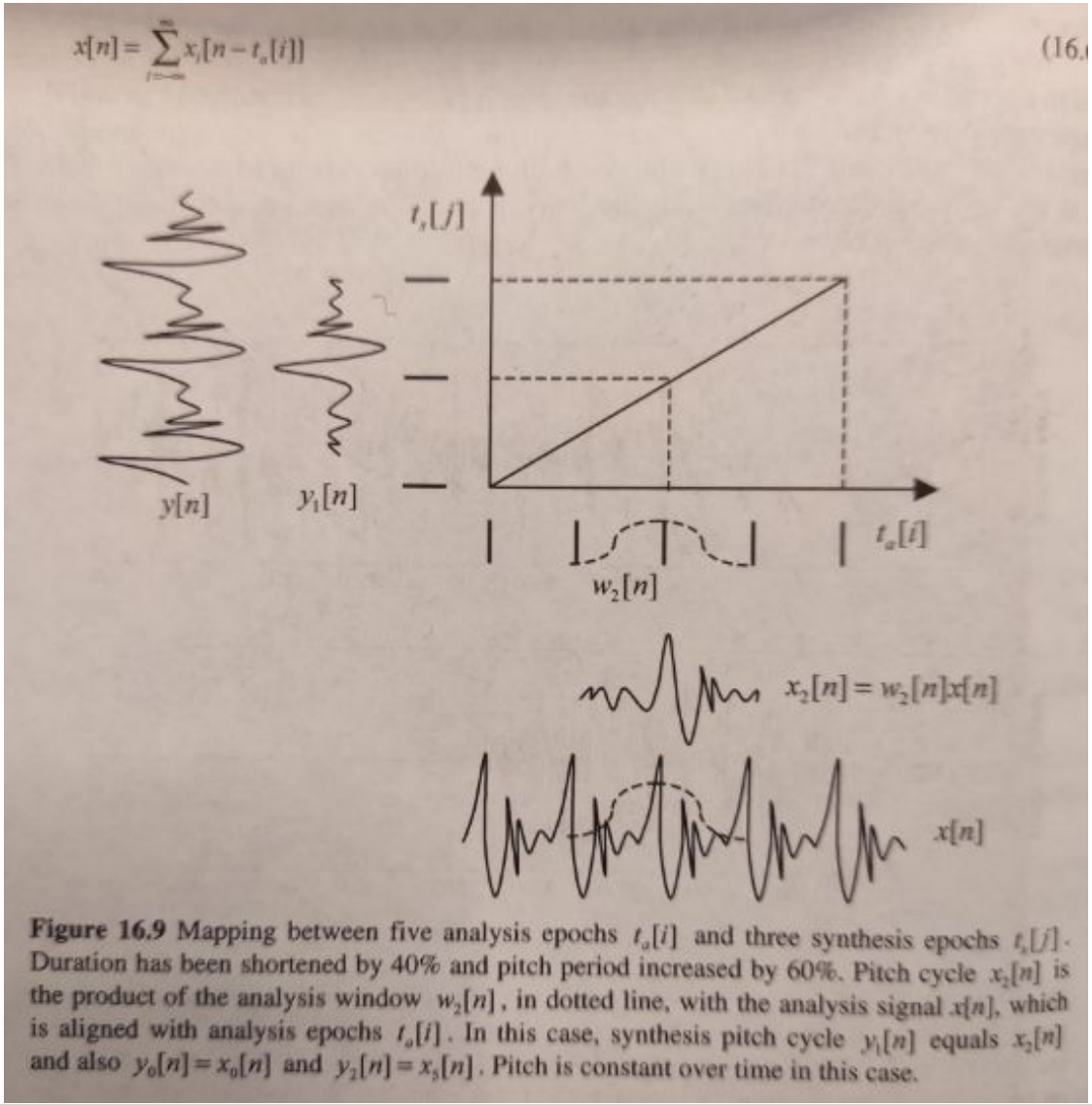


Figure 16.9 Mapping between five analysis epochs $t_a[i]$ and three synthesis epochs $t_s[j]$. Duration has been shortened by 40% and pitch period increased by 60%. Pitch cycle $x_2[n]$ is the product of the analysis window $w_2[n]$, in dotted line, with the analysis signal $x[n]$, which is aligned with analysis epochs $t_a[i]$. In this case, synthesis pitch cycle $y_1[n]$ equals $x_2[n]$ and also $y_0[n] = x_0[n]$ and $y_2[n] = x_3[n]$. Pitch is constant over time in this case.

Prosodic processing

- Pitch Synchronous Overlap and Add (PSOLA) (cont.2)
- Modifying the time scale:

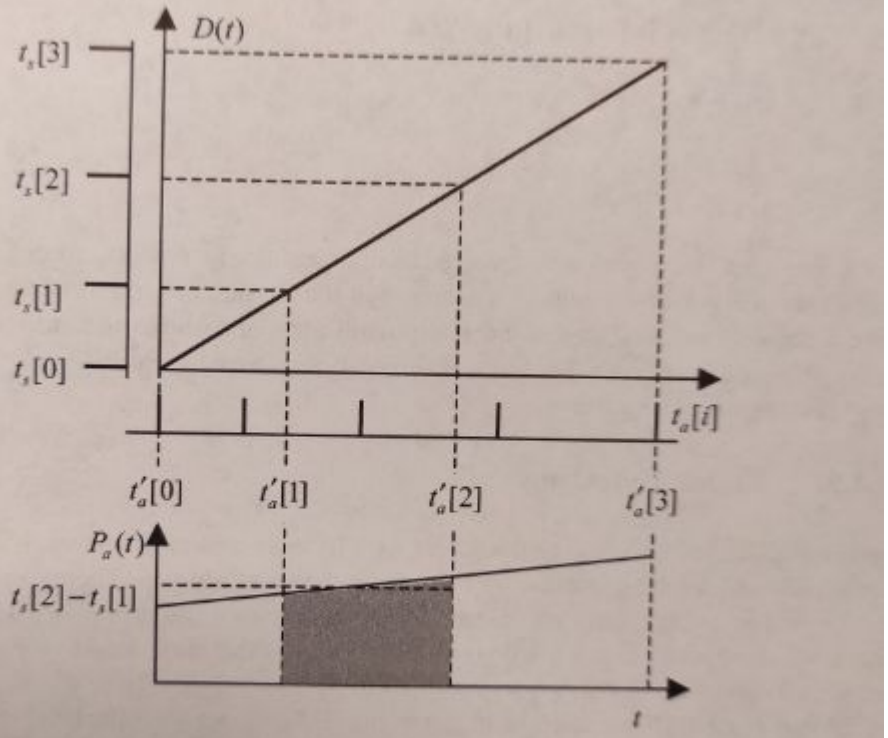


Figure 16.13 Time-scale modification of speech. The five analysis epochs $t_a[i]$ are shown in the x-axis and the four synthesis epochs $t_s[i]$ in the ordinate. Duration is shortened by 25% while maintaining the same pitch period. The corresponding virtual analysis epochs $t'_a[i]$ are obtained through the mapping $D(t)$, a linear transformation with $\alpha = 0.75$.

Prosodic processing

- Pitch Synchronous Overlap and Add (PSOLA) (cont.3)
- Problems with TD-PSOLA:

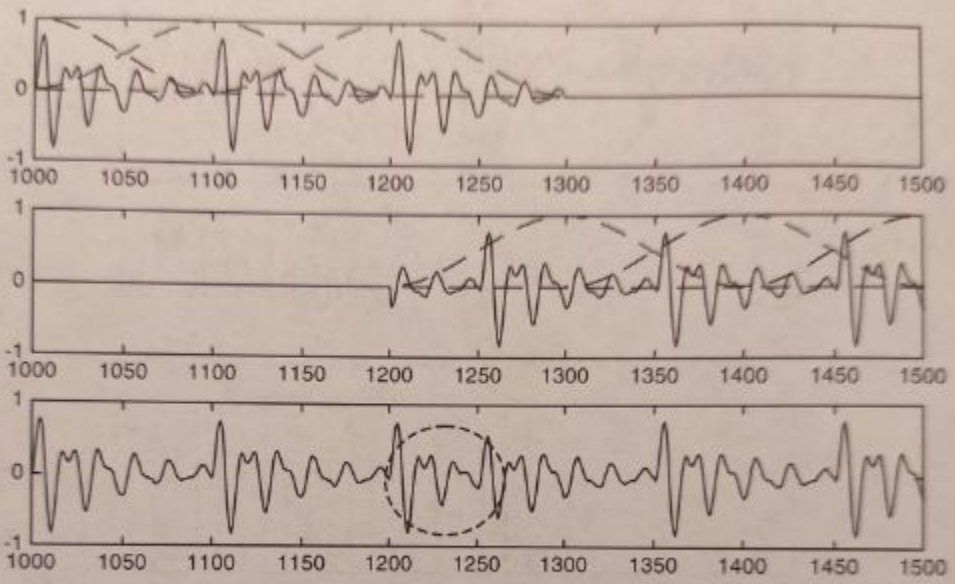


Figure 16.15 Phase mismatches in unit concatenation. Waveforms are identical, but windows are not centered on the same relative positions within periods.

Main references

Reconhecimento da Fala, Diamantino Freitas, Vitor Pêra, apontamentos Processamento da Fala 2009/2010, DEEC-FEUP.
Speech Processing, Chris Rowden, McGraw-Hill
Spoken language processing, Xuedong Huang et. Al., Prentice-Hall.