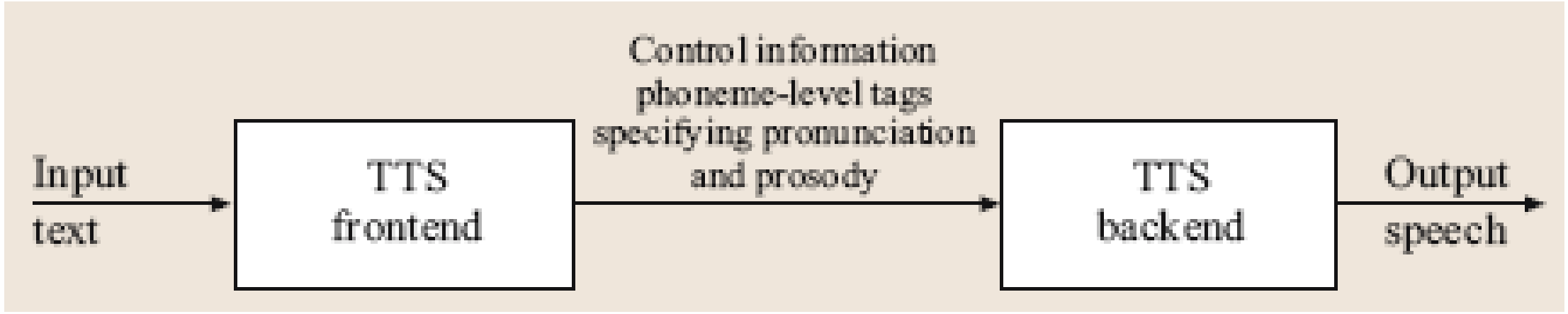
Speech Synthesis
(Text-to-Speech (TTS) Synthesis)

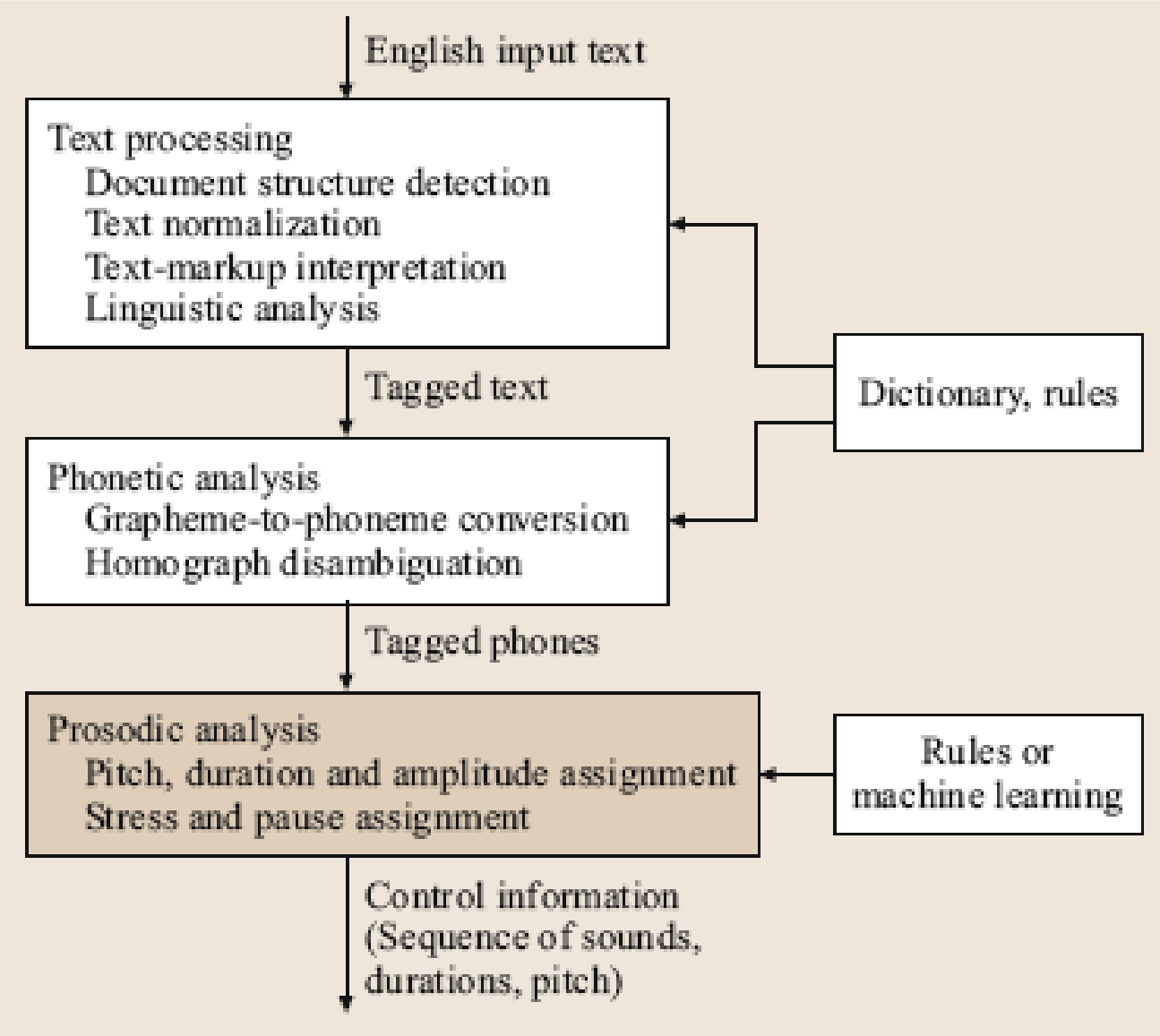
Basic Components of TTS



TTS Front-End

- Document structure detection
  - Interpreting punctuation marks, filtering out email headers, SGML, …
- Text normalization
  - Handles abbreviations and acronyms
- Interprets text markup (SSML)
- Linguistic analysis (text to phone sequence)
  - Morphological analysis (proper word pronunciations)
  - syntactic analysis (accenting, phrasing and to resolve the ambiguities in the written text)
  - Phonetic analysis (what sound to produce and how to produce)
  - Grapheme-to-phoneme conversion
  - Homograph disambiguation
  - Dictionaries and letter-to-sound rules (for deriving pronunciations)
- Prosodic analysis
  - Progression of intonation, speaking rate, rhythm and loudness

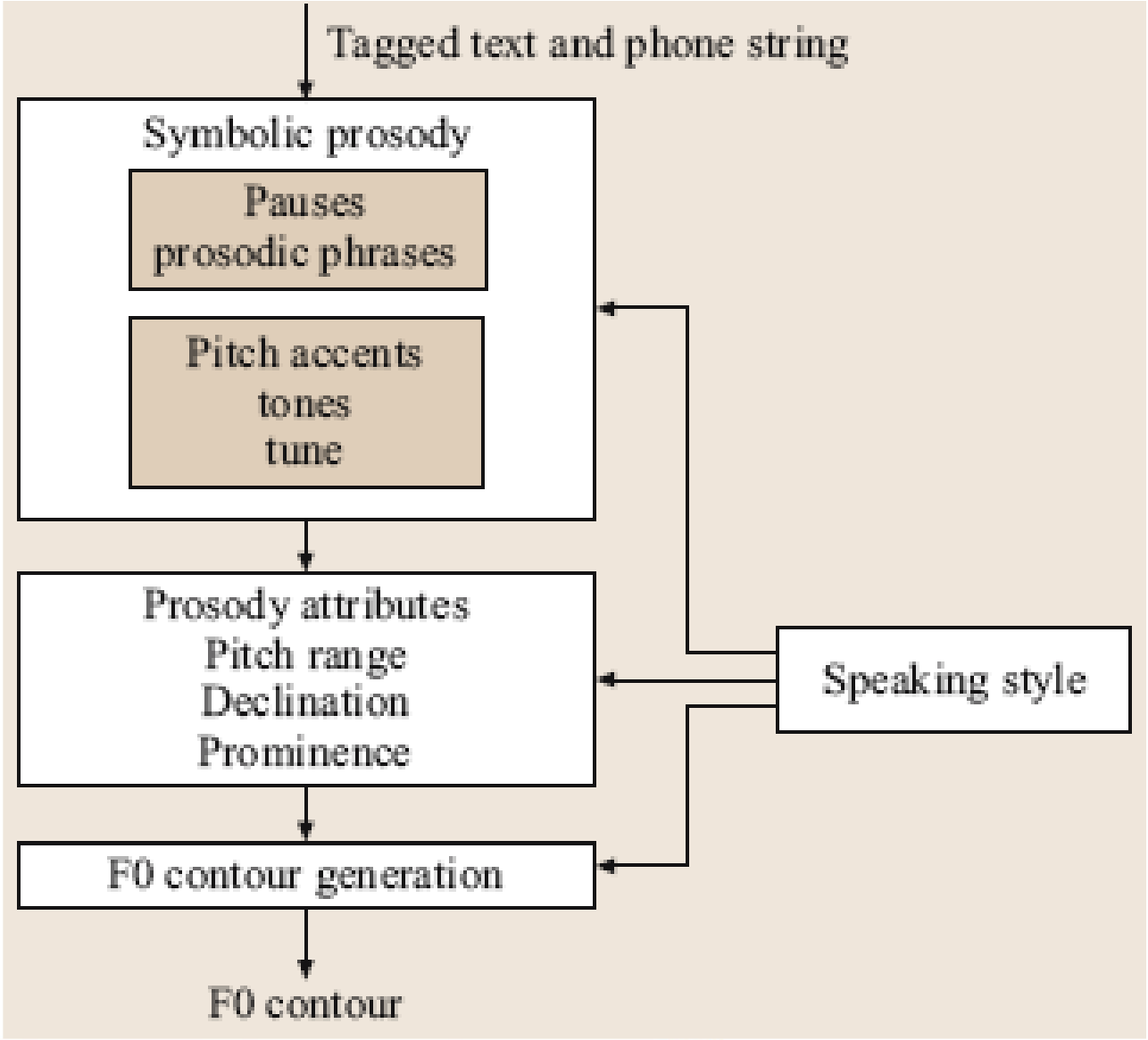
TTS Front-End (Cont..)



Prosodic Analysis

- Duration modeling (timing of the sequence of sounds)
  - Rule based & Data driven approaches
- Intonation modeling
  - Tone sequence model (ToBI)
  - Fujisaki model
  - Tilt model

Generation of a Pitch Contour



TTS Back-End

- Rule-based speech synthesis
  - Articulatory synthesis
  - Formant synthesis
- Corpus-based speech synthesis
  - Concatenative synthesis
- Statistical parametric speech synthesis
- End-to-End speech synthesis

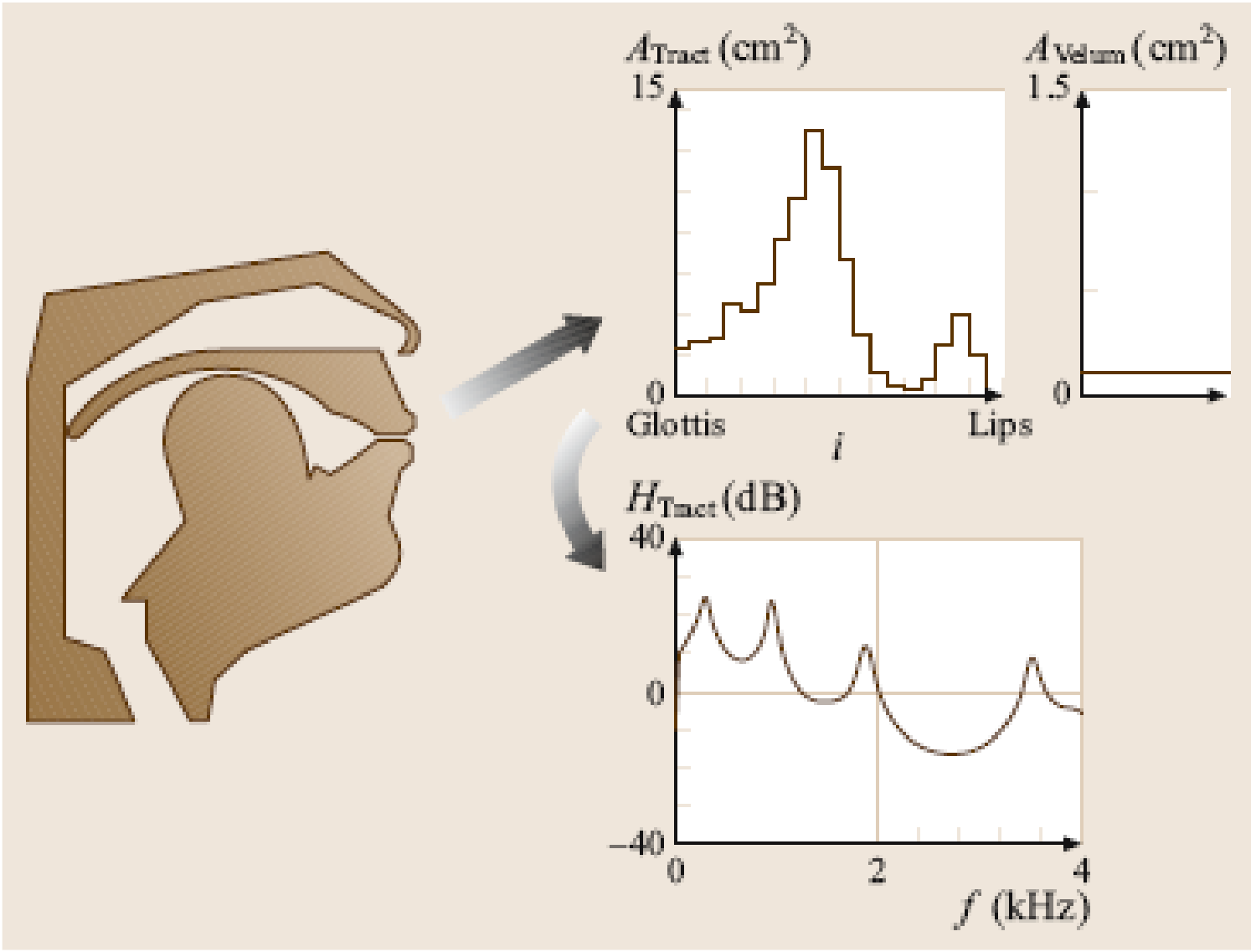
Articulatory Speech Synthesis

- Exploits the mechanical and acoustic models of speech production
- Dynamics of the articulators are represented using partial differential equations
  - Ultrasound pictures of tongue shape
  - Nuclear magnetic resonance

$$\begin{pmatrix} P_L \\ U_L \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} P_G \\ U_G \end{pmatrix}$$

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \prod_{k=1}^{p} \begin{pmatrix} A_k & B_k \\ C_k & D_k \end{pmatrix}$$

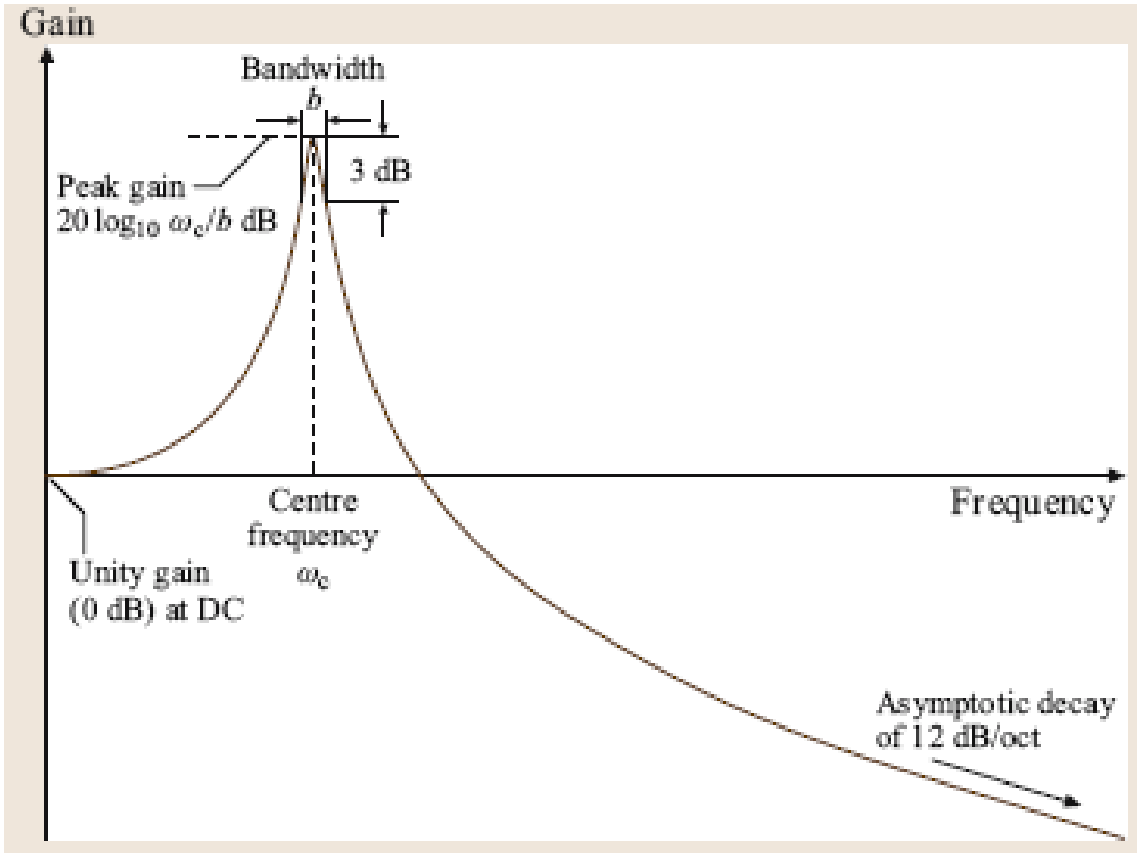
Articulatory Speech Synthesis (Cont..)

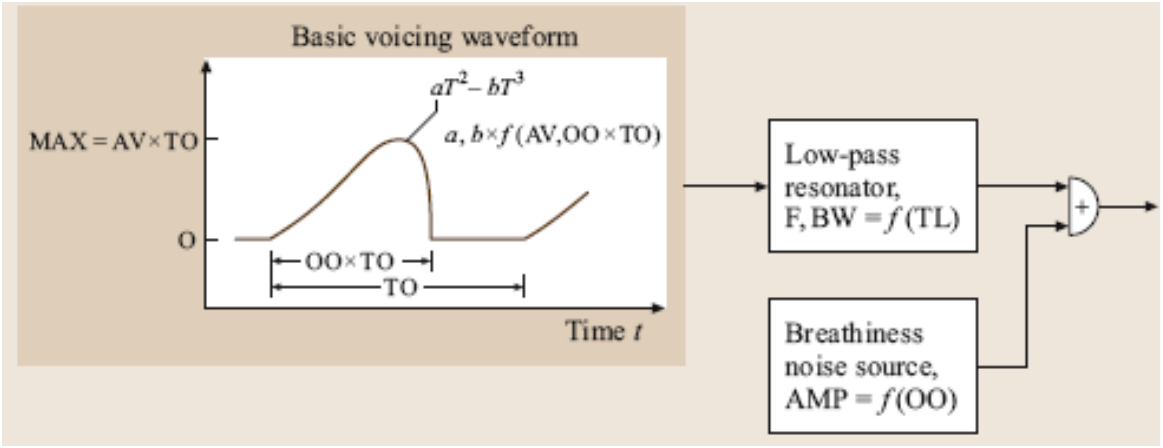


Formant Speech Synthesis

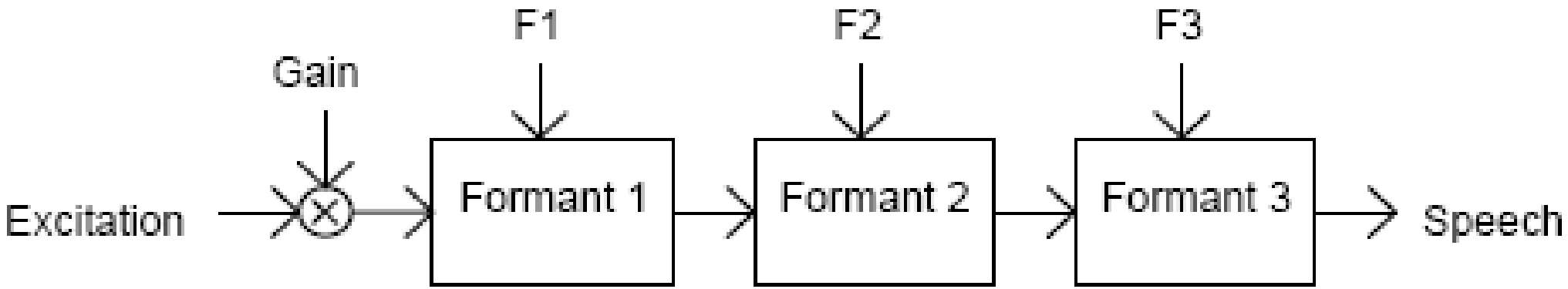
- Cascade (serial) synthesizer
  - Open and non-nasal vocal tract (Vowels and liquids)
- Parallel synthesizer
  - Fricatives and nasals
  - Individual gains can be controlled
- Hybrid synthesizer (combined)
- High intelligibility and low naturalness

Formant Speech Synthesis (Cont..)

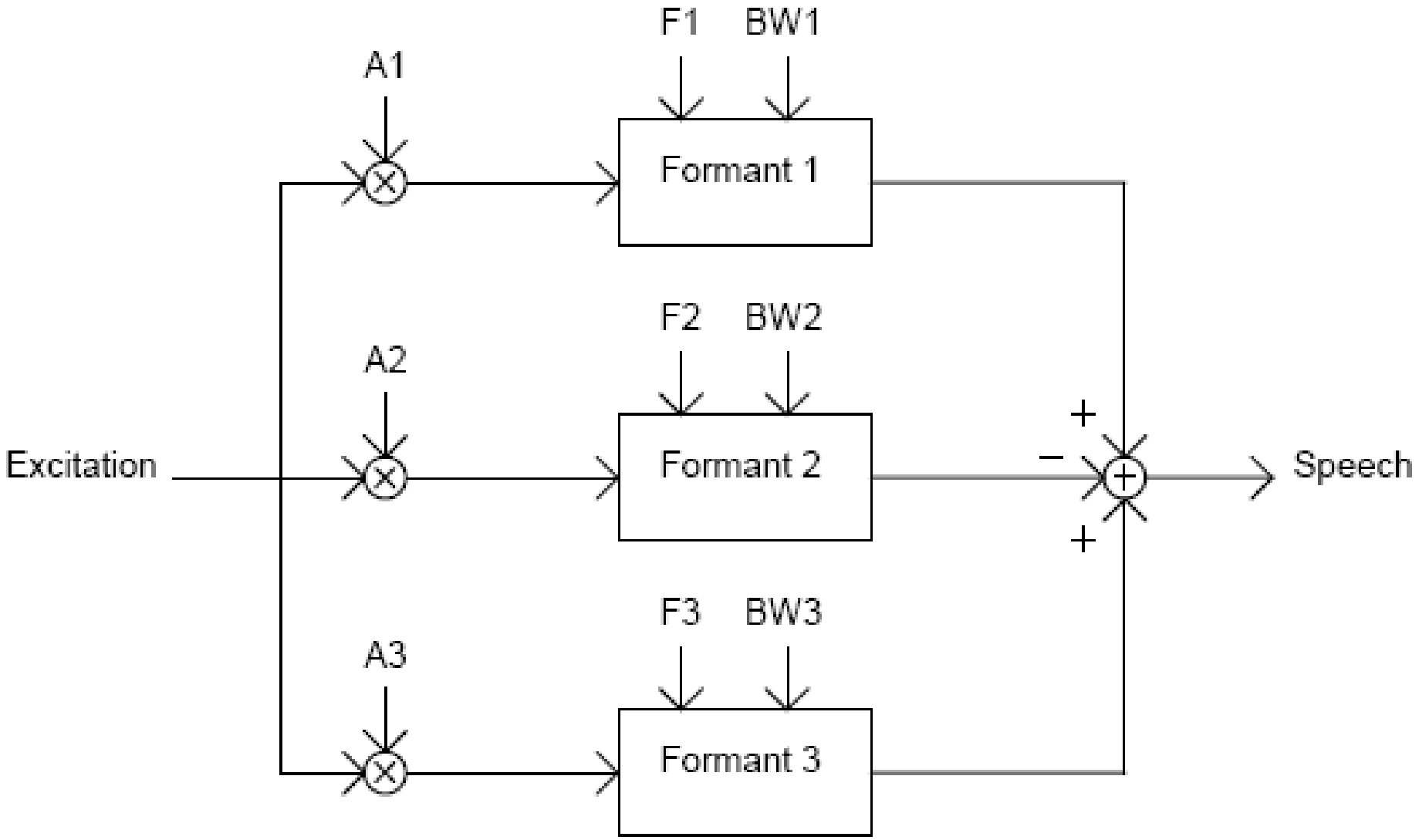




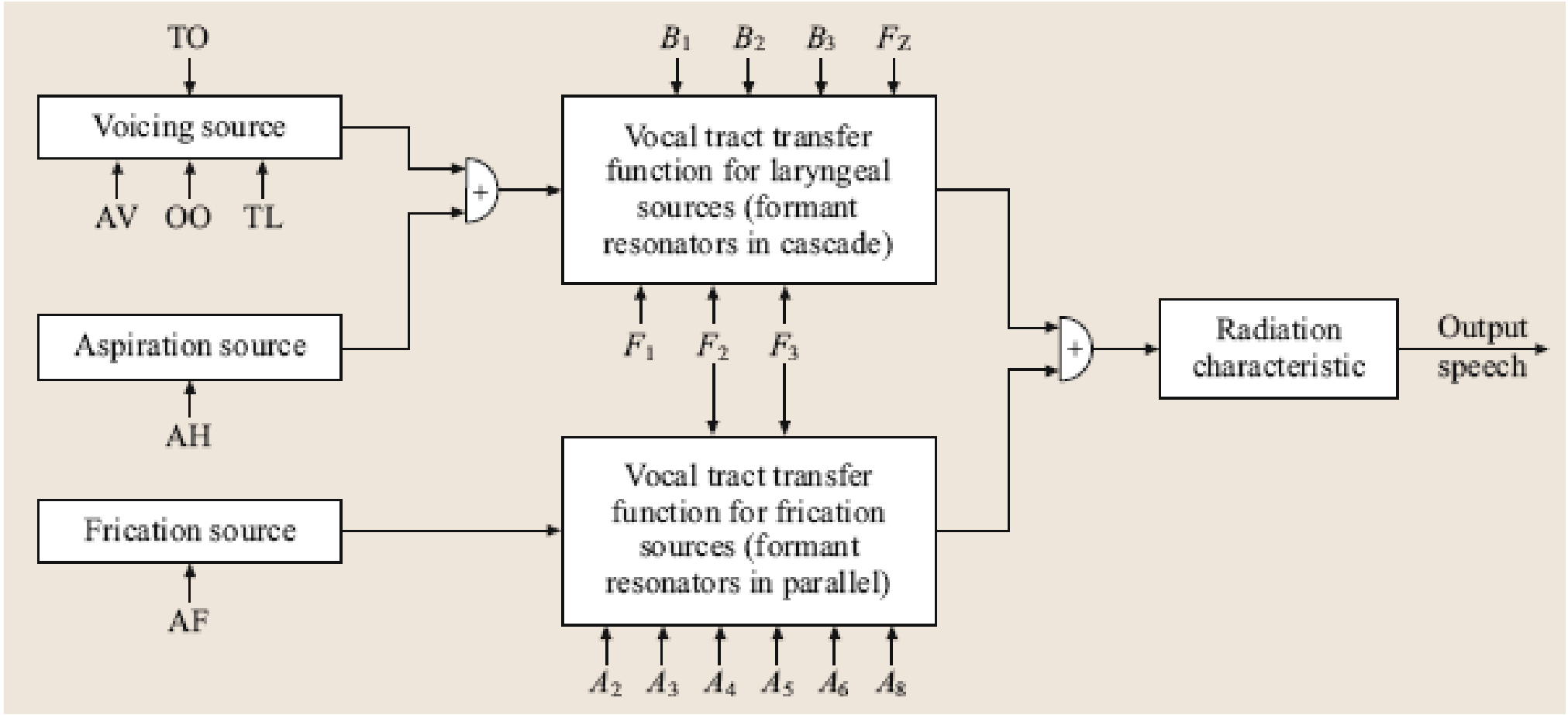
Cascade (serial) Formant Speech Synthesizer



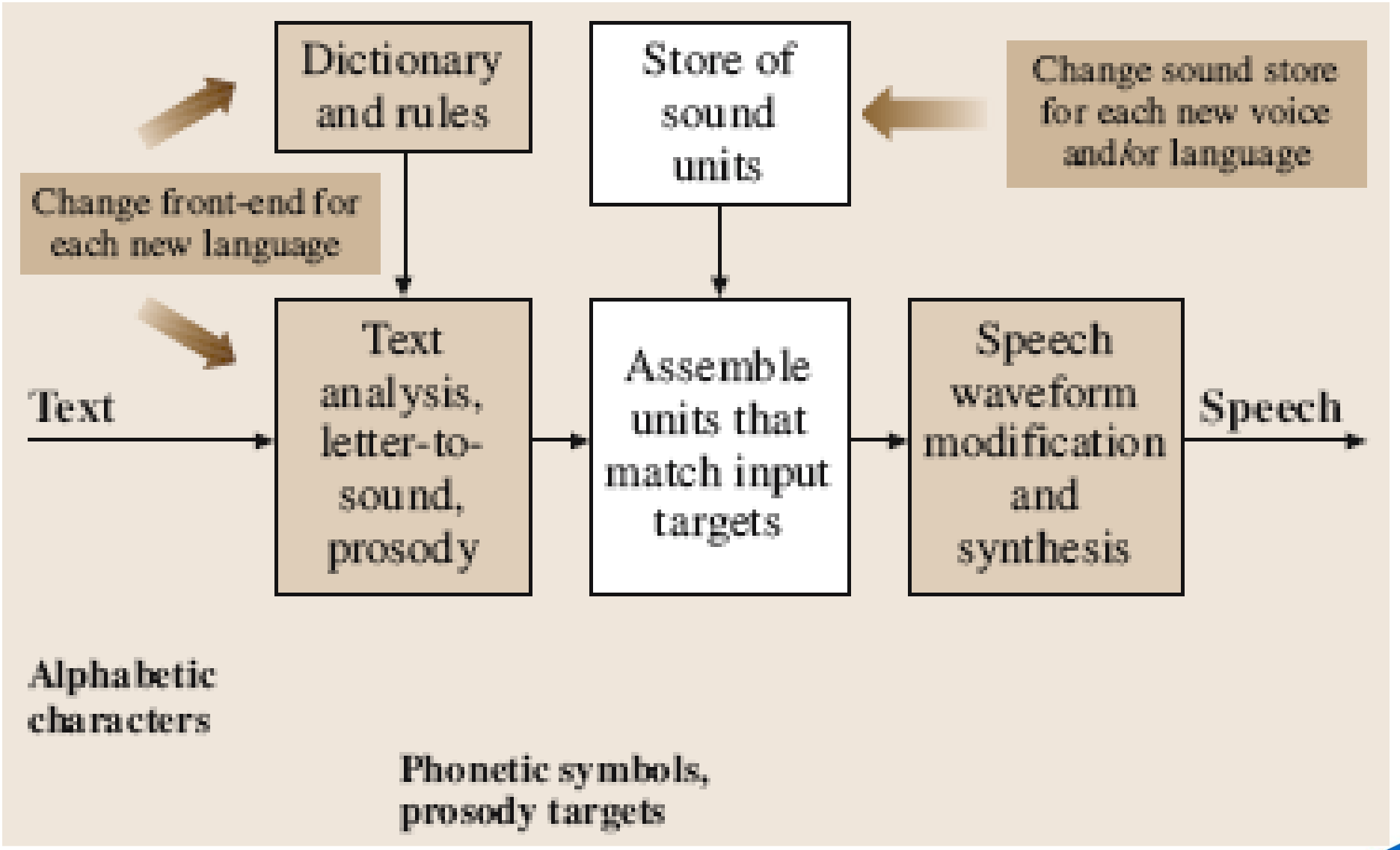
Parallel Formant Synthesizer



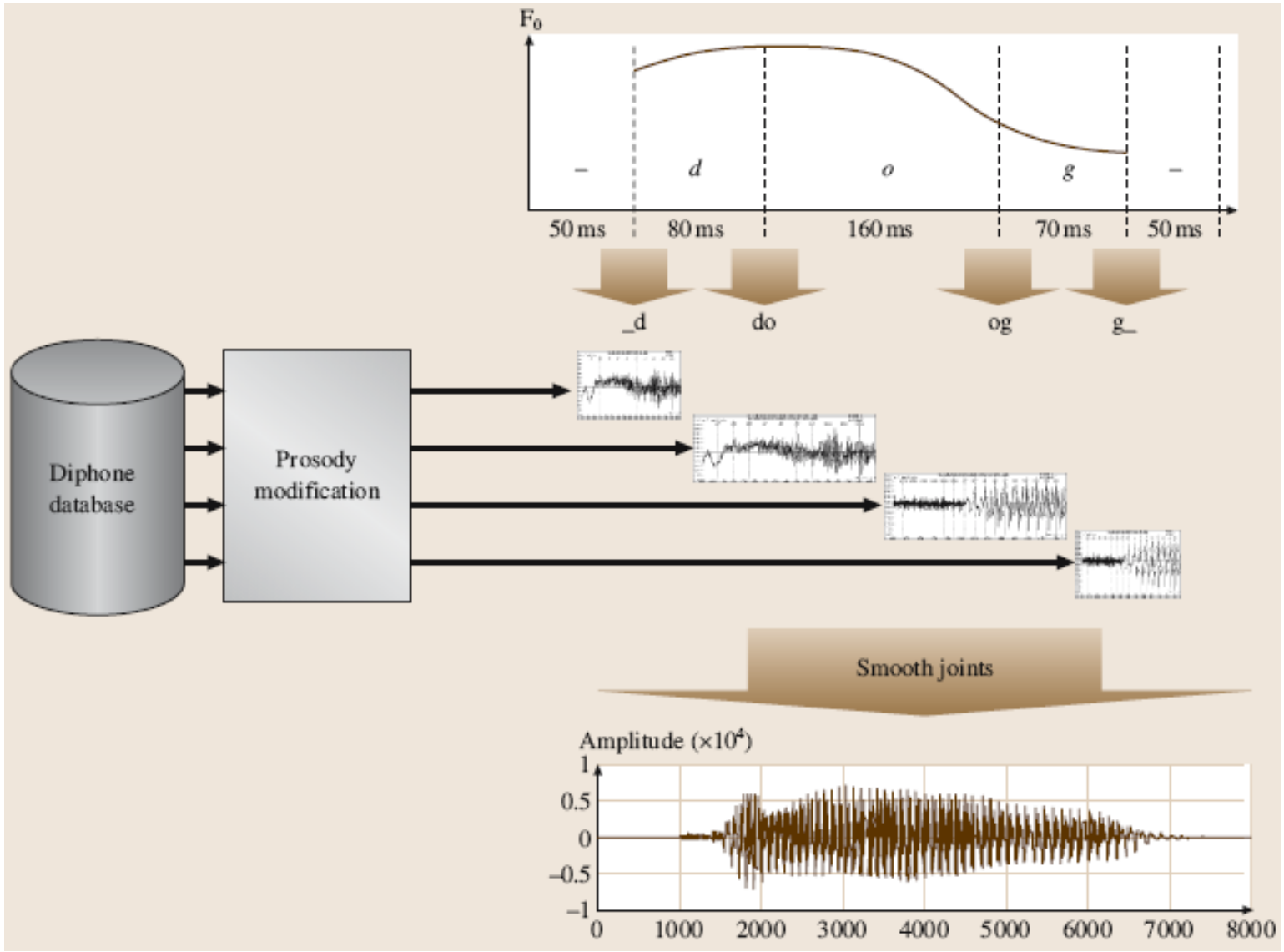
Hybrid Synthesizer



Concatenative Speech Synthesis



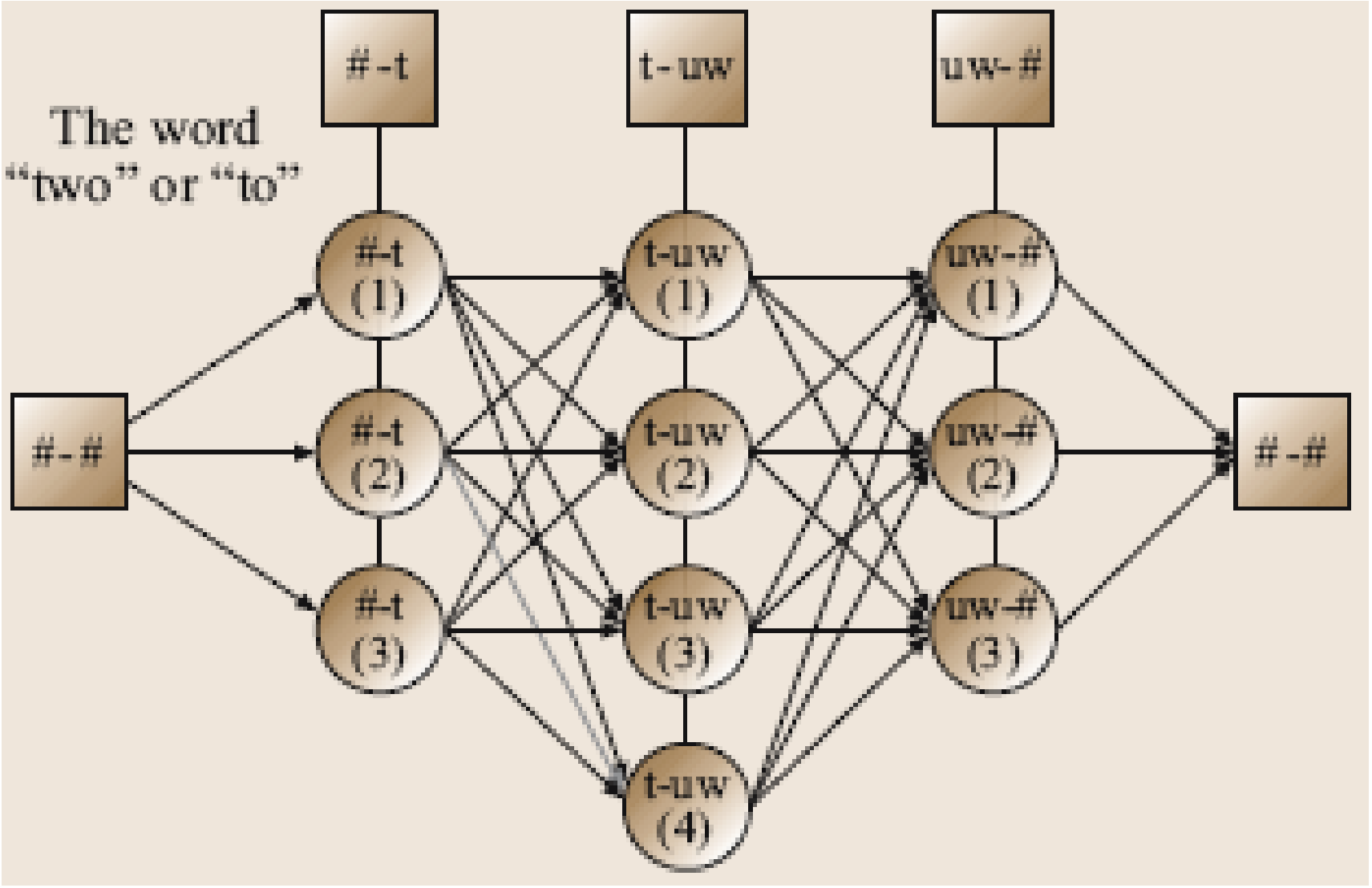
Di-phone based Concatenative Speech Synthesis



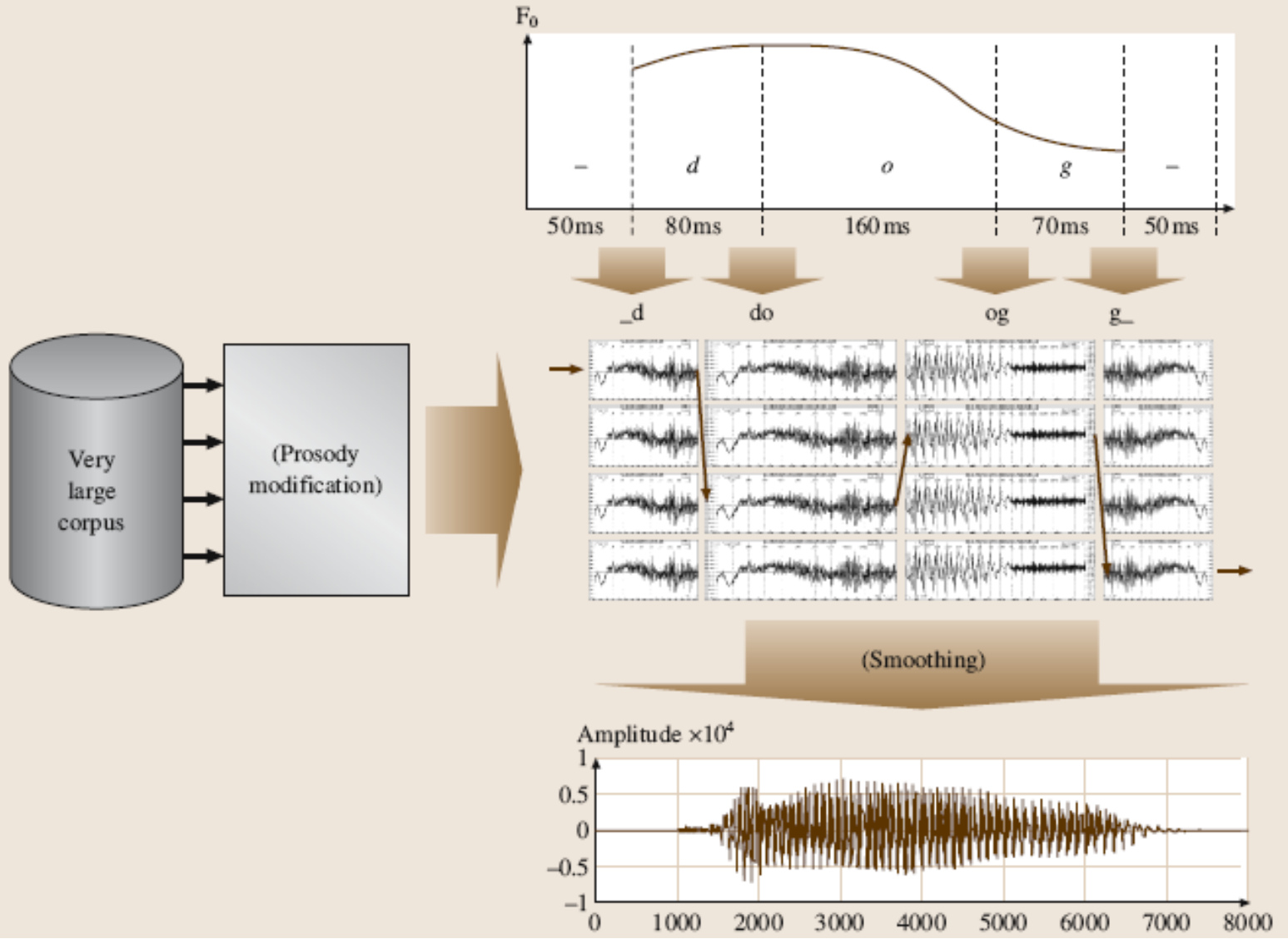
Concatenative Speech Synthesis

- Units for concatenation
  - Diphones & Demisyllables
- Unit selection synthesis
  - Spectral similarities at the boundaries
  - Matching prosodic tags
  - Target cost (unit segmental distortion USD)
    - Example: ‘ah’ in want for synthesizing ‘ah’ in cart
  - Join cost (unit concatenative distortion UCD)
    - Spectral discontinuties across the boundaries

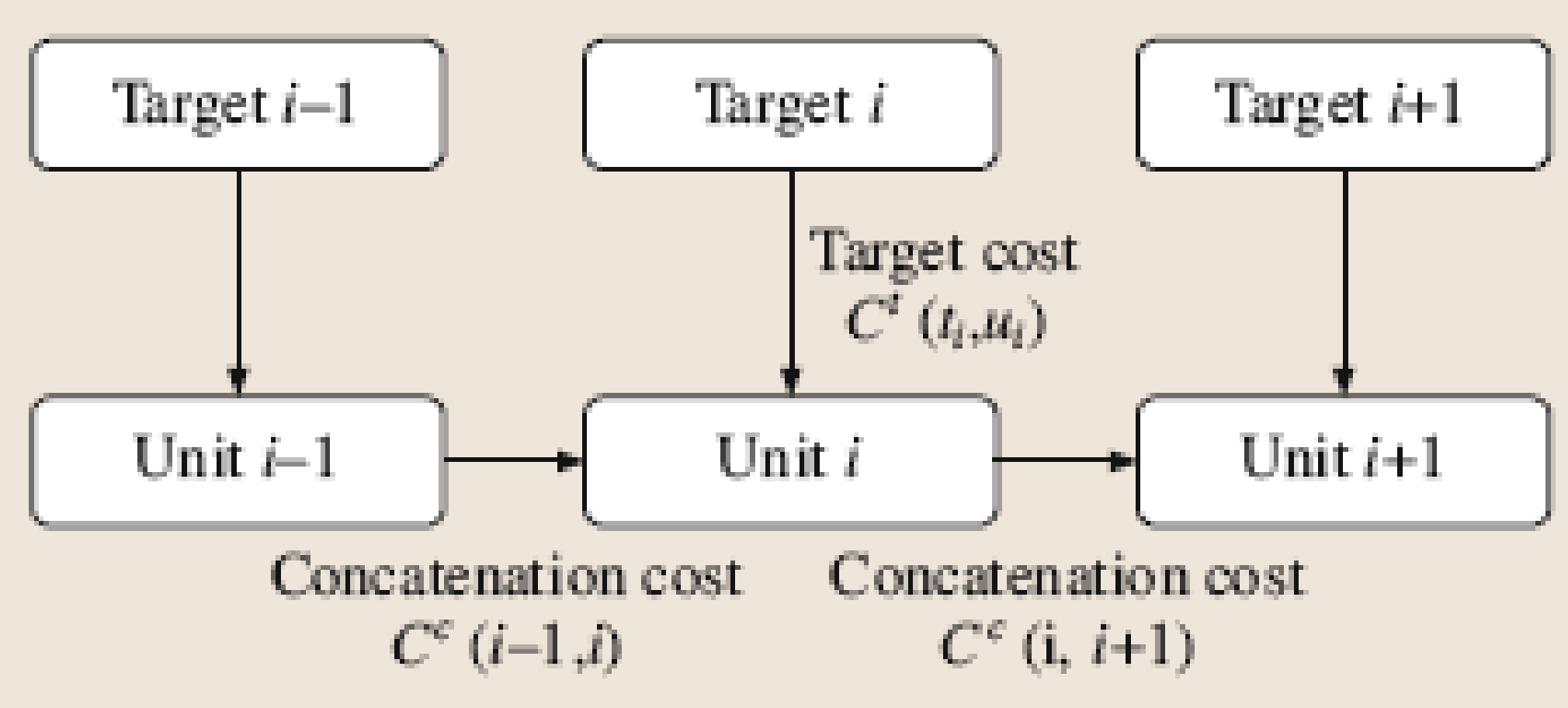
Unit Selection for the Word ‘two’



Unit Selection for the Word ‘dog’



Target and Concatenation costs in USS



Target and Concatenation costs in USS

Optimization cost

$$C(t^n, u^n) = \sum_{i=1}^{n} C^t(t_i, u_i) + \sum_{i=2}^{n} C^c(u_{i-1}, u_i) + C^c(s, u_1) + C^c(u_n, s)$$

Target cost (symbolic & numerical features)
intelligibility

$$C^t(t_i, u_i) = \sum_{j=1}^{p} w_j^t C_j^t(t_i, u_i)$$

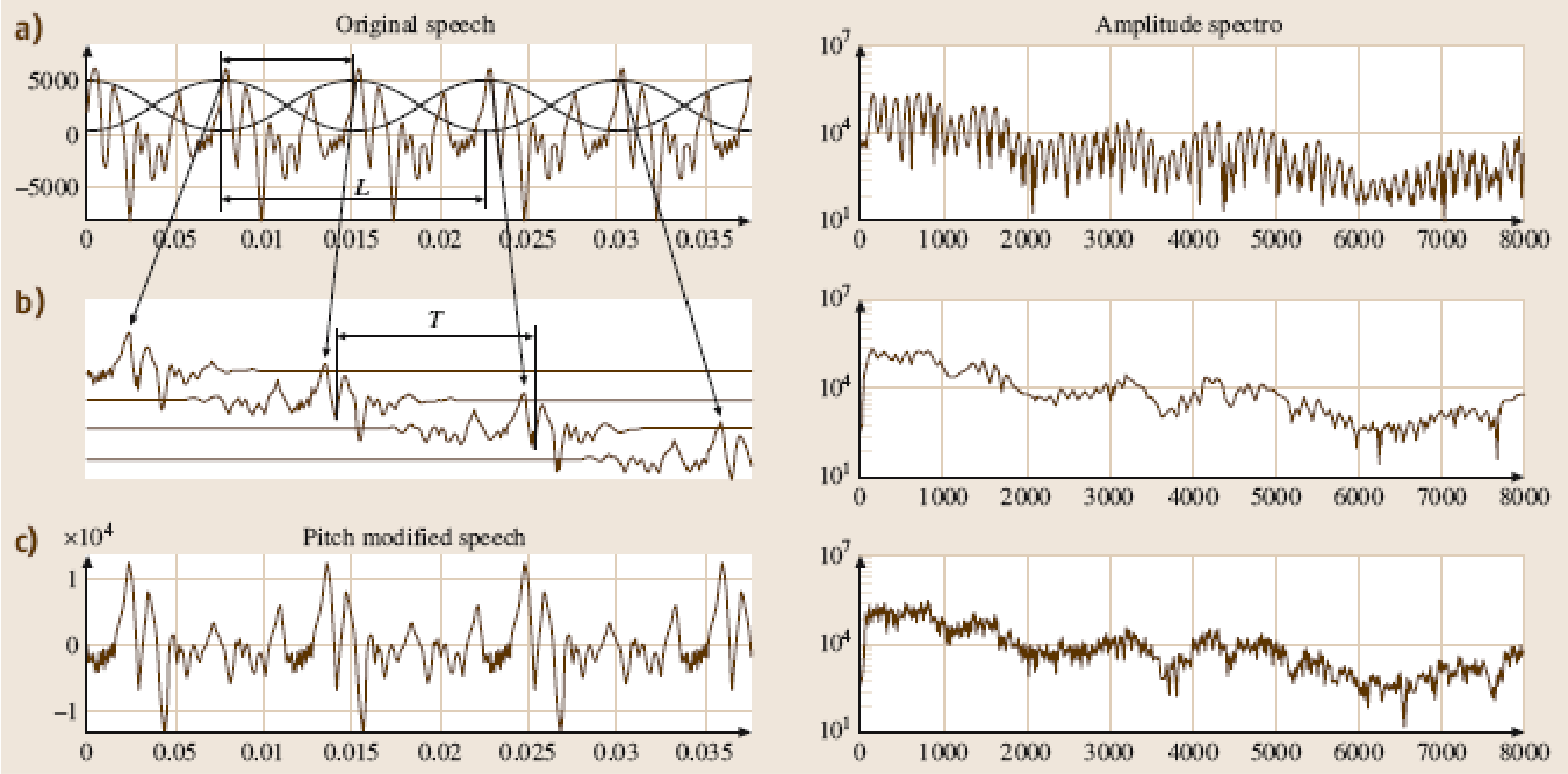
Concatenation cost (formants, LPCs, MFCCs, ..)
naturalness

$$C^c(u_{i-1}, u_i) = \sum_{j=1}^{q} w_j^c C_j^c(u_{i-1}, u_i)$$

Signal Representations and Signal Processing for Concatenative Synthesis

- Can be compressed at high quality
- Low computational cost
- Minimum perceptual distortion
- Selected segments must allow natural modifications of prosody
- TD-PSOLA
- LPC Synthesis
- Sinusoidal synthesis
- HNM model

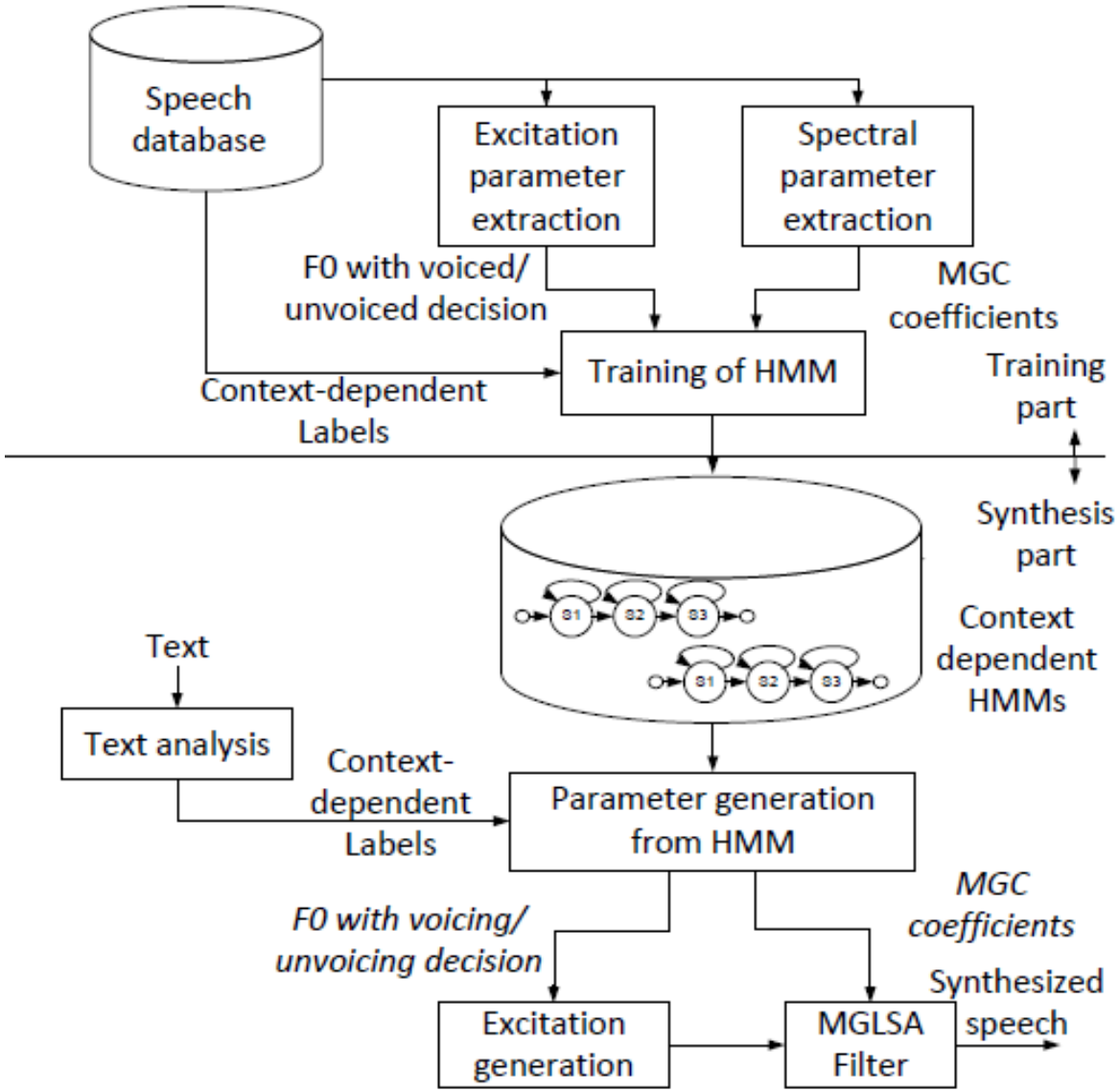
TD-PSOLA


a)
Original speech
5000
0
−5000
L
0
0.05
0.01
0.015
0.02
0.025
0.03
0.035
b)
T
c)
×10^4
Pitch modified speech
1
0
−1
0
0.05
0.01
0.015
0.02
0.025
0.03
0.035
Amplitude spectro
10^7
10^4
10^1
0
1000
2000
3000
4000
5000
6000
7000
8000

Speech Signal Transformation Principles

- Speaker transformation
- Expressive speech generation
- Dialect or accent transformation
  - Pitch, duration, amplitude and spectrum
- Prosody transformation methods
- Changing speaker characteristics
  - Code books, GMM, ANN

Statistical Parametric Speech Synthesis



Speech Synthesis Evaluation

- Accuracy (evaluation of front end)
  - Ability to read the input text like the knowledgeable human
- Intelligibility and Naturalness (all the components are responsible)
  - Formant synthesis: high intelligibility
  - Unit selection synthesis: both intelligibility and naturalness
  - SPSS : Intelligible, flexibility, slightly less natural

Speech Synthesis Evaluation (Cont..)

- Tests to evaluate the accuracy
  - Running the text corpus of task relevant acronyms and abbreviations.
- Listening tests
  - Intelligibility tests (word, sentence, passage levels)
  - Intelligibility in the absence of prosody
  - MOS (overall impression, listening effort, comprehension)

Thank You