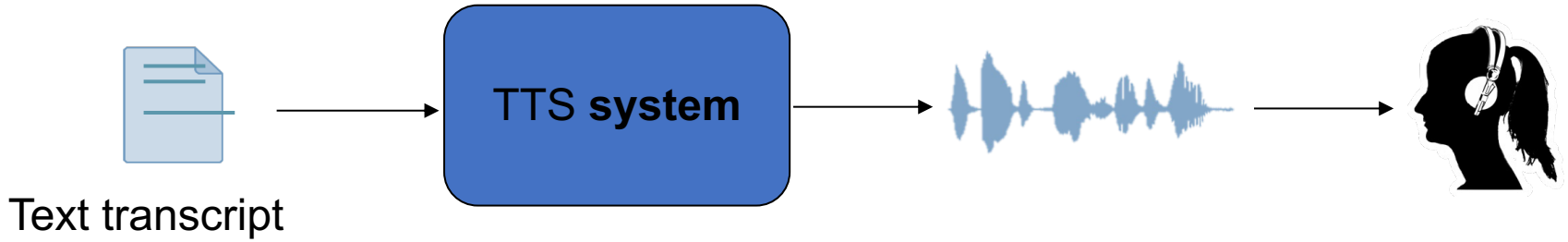
Text-to-Speech Synthesis – Part I

Outline

- Overview
- Natural language processing (NLP) for speech synthesis
- Articulatory speech synthesis
- Formant speech synthesis
- Concatenative speech synthesis
- Statistical parametric speech synthesis
- End-to-end speech synthesis
- Evaluation

What is text-to-speech synthesis (TTS)?



End-use:

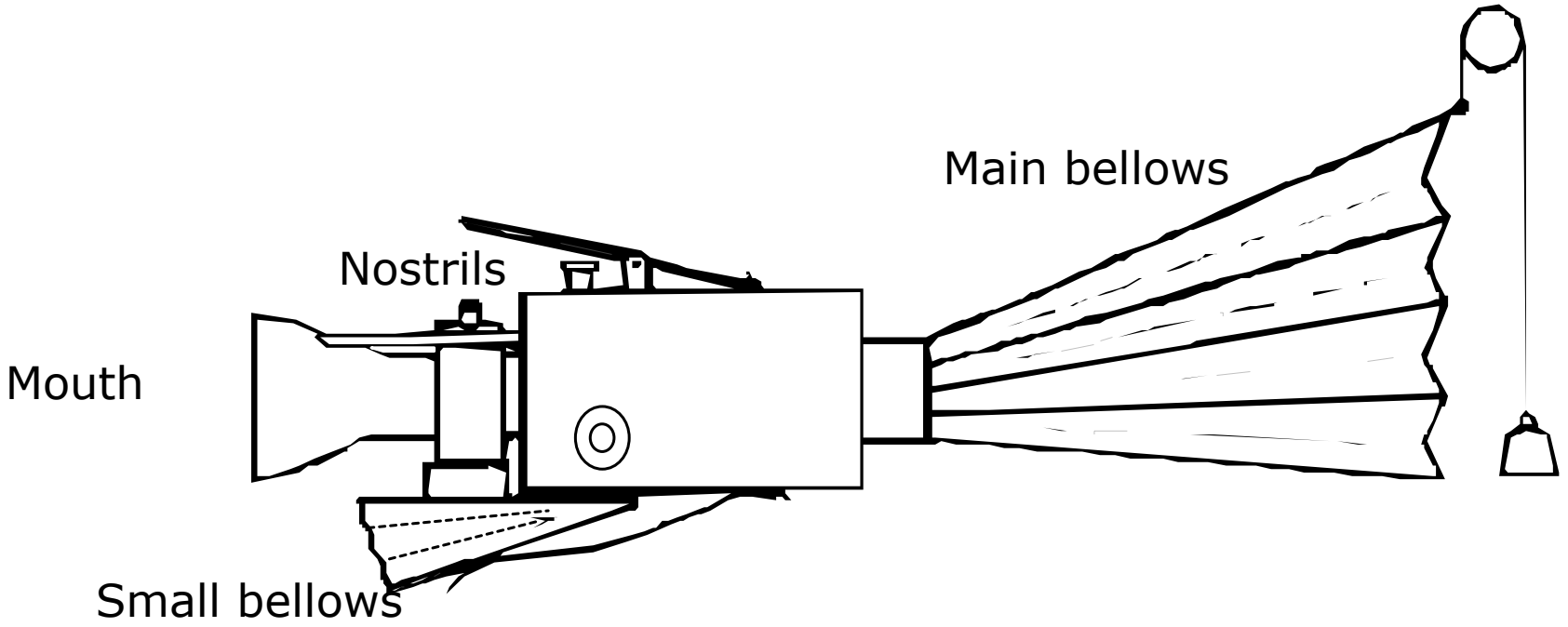
- Announcement systems
- Dialog systems
- Assistive systems for visually impaired and speech impaired persons

Challenges:

- Fast adaptation to new speaker
- Multilingual speech synthesis
- Affective speech synthesis
- Objective evaluation

History (1)

Von Kempelen's speaking machine (18th century)



History (2)

Dudley's Voder (1939)

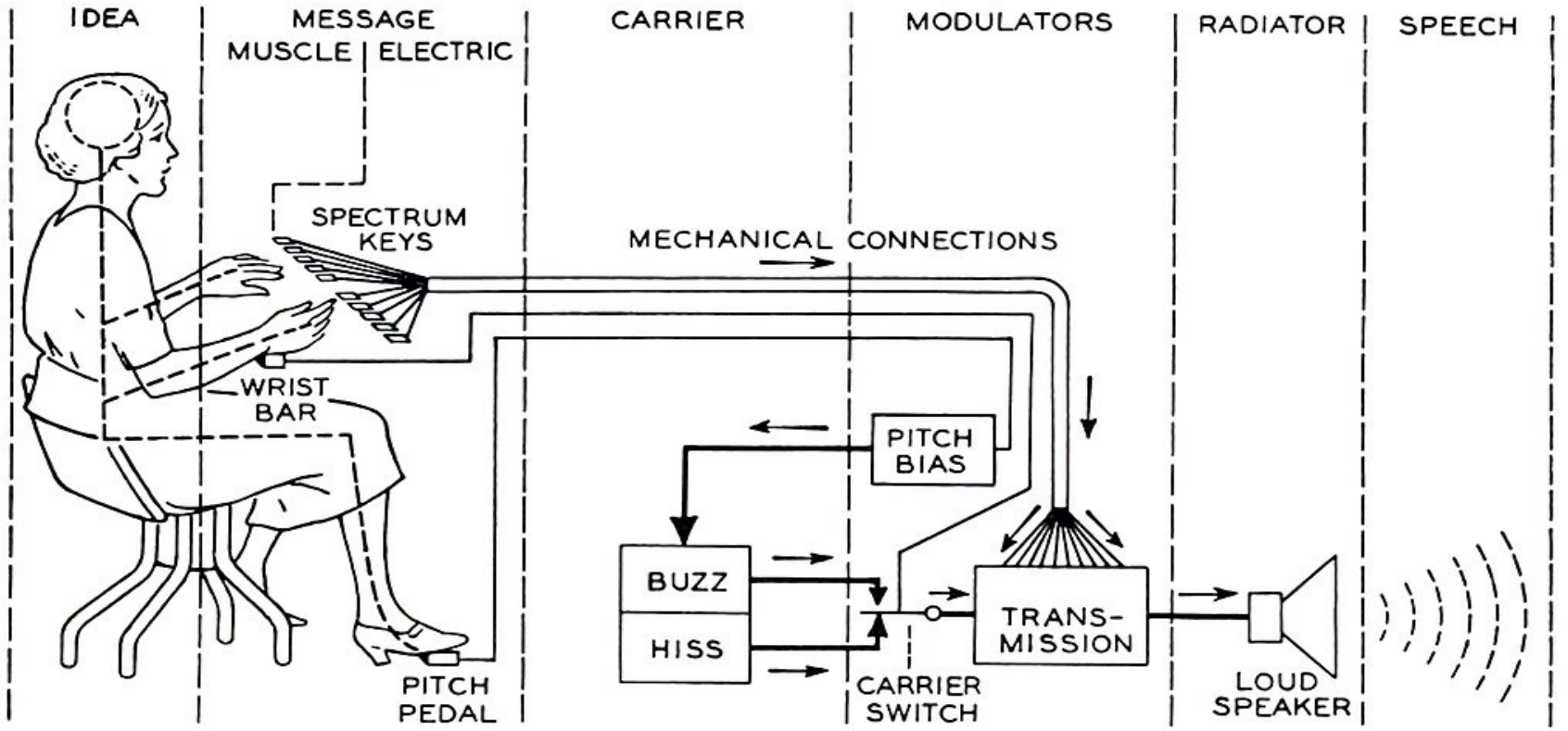
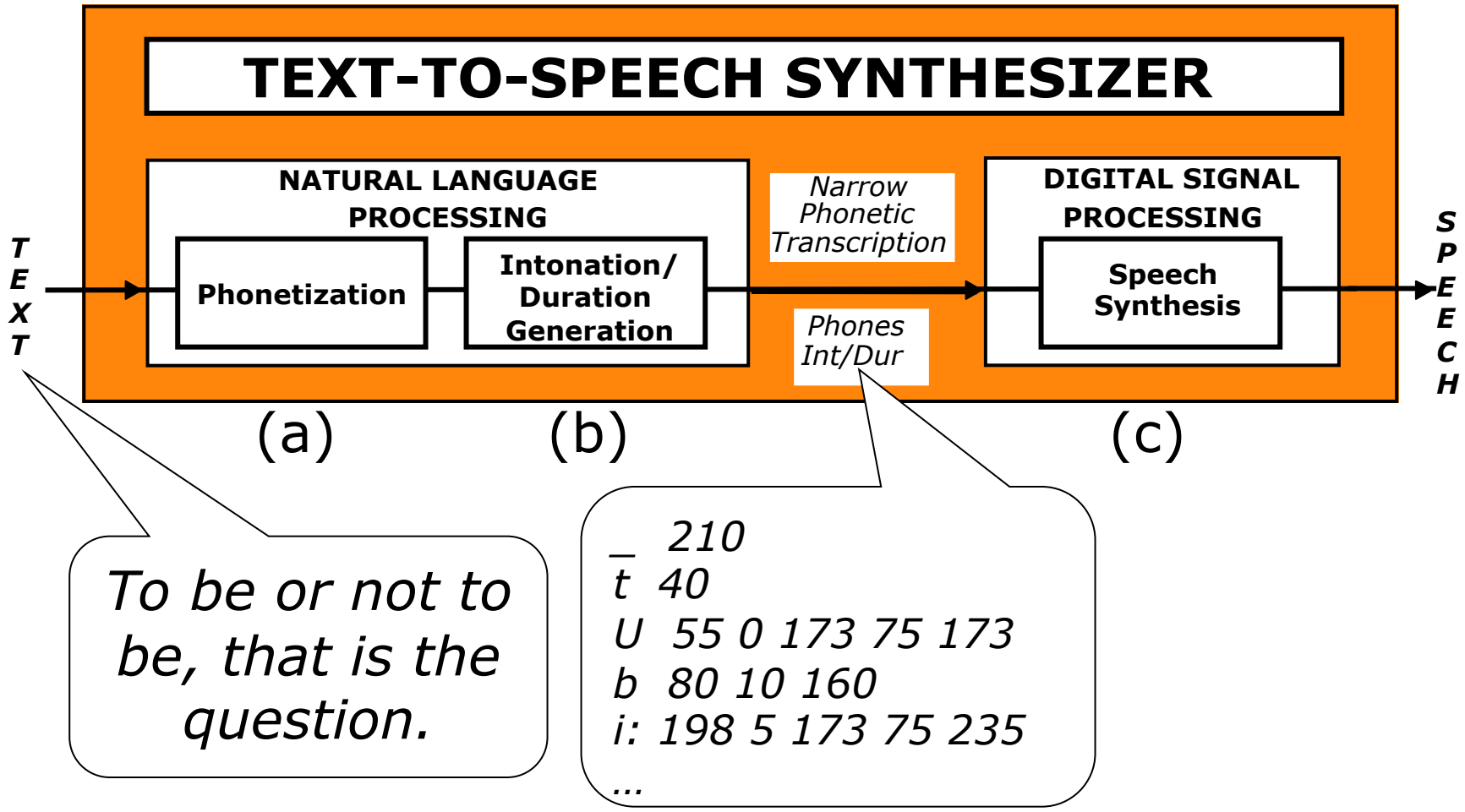


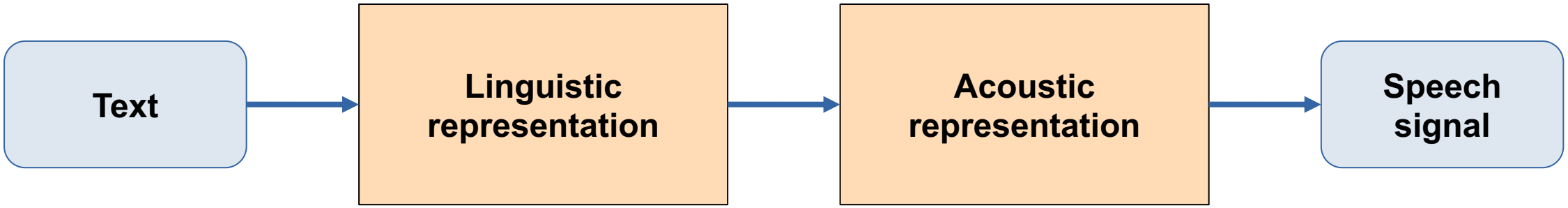
Fig. 8—Schematic circuit of the voder.

Wikimedia

Current TTS systems



TTS – Basic Steps



NLP for Speech Synthesis

Converting input text into a linguistic representation:

1. Text normalization
    - Identify tokens and convert them to words (e.g. **747 years ago** vs. **Boeing 747**)
2. Phonetic analysis
    - Retrieve pronunciations from dictionary or with letter-to-sound rules.
3. Prosodic analysis
    - Determine duration, intonation, location of pauses, stress.

NLP module

end of sentence, abbreviations, numbers, ...

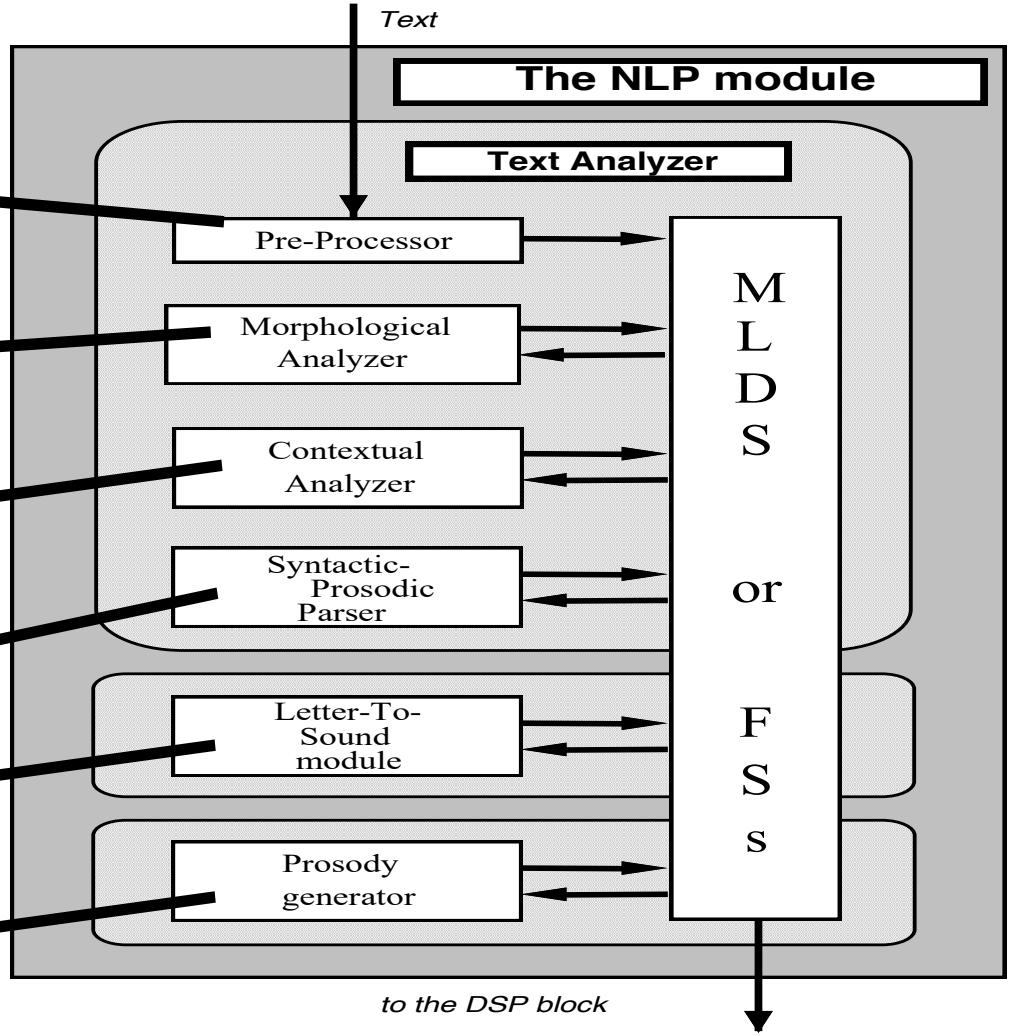
word=stems+affixes
=>possible part-of-speech(word)

pos(context)

hierarchical description

phonetization

intonation , rhythm



Pre-processing

- Text segmentation into broad segmentation units

  (I)( )(know)( )(1)(.)(000)( )(words)(,)( )(Dr)(.)( )(Jones)(.)

- From broad to final segmentation units
  - Sentence end detection

    The man (and he certainly was one !) just said "Maybe. I 'll see. I can't promise."
  - Dealing with abbreviations

    German 'tgl.' = 'täglich', 'tägliche', 'täglichem', 'täglichen', 'täglicher', 'tägliches'

    'Dr. Jones lives at the corner of Jones Dr. and St. James St.
  - Recognizing acronyms

    IBM, BBC, EPFL, ...
  - Processing numbers

    '3.14', '2.16 pm', '13:26', '08.11.94', 'the 16th'

Morphological analysis

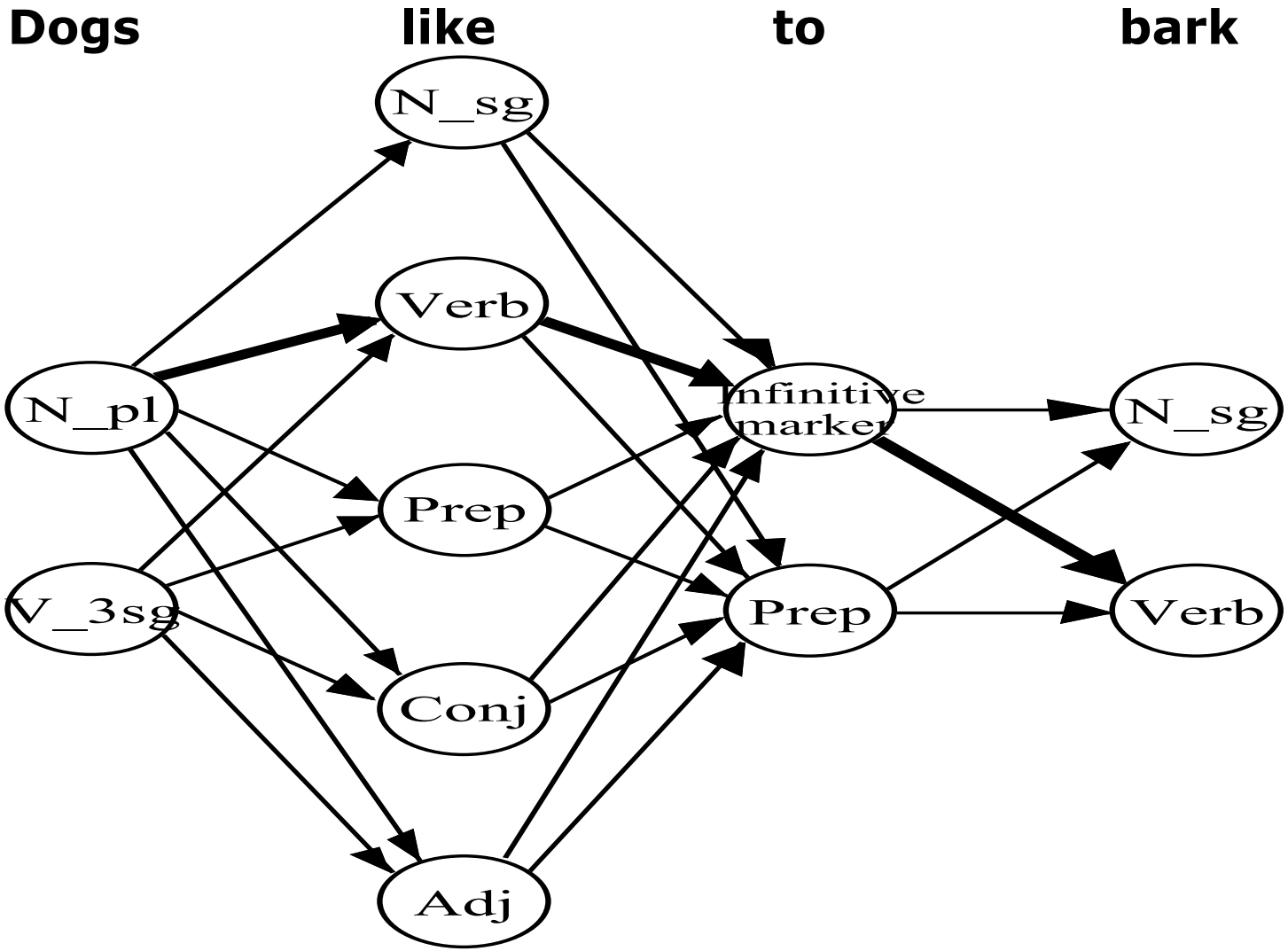
- Why ?
  - Constrain the size of lexicons
  - Morphological features for syntactic processing
  - Morphologically related pronunciation

    'Nebenstrasse', /St{ /↔ 'demonstration' /st{ /, 'hothouse', ...
  - Word-level stress in free stress languages

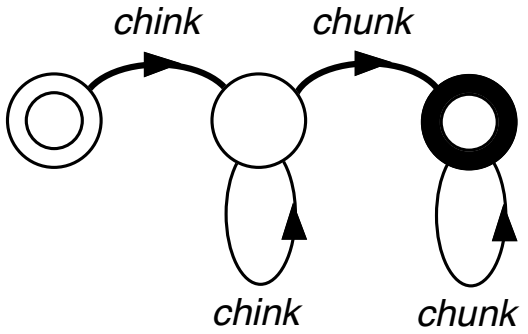
- How ? morphology = highly language dependent

  (English verbs : four to eight forms ; French verbs : 37 to 41 forms !; compounding much more complex in germanic languages : hottentottentottententoonstelling !)
  - Typically : regular rules, finite state automata, organized in a language-dependent way

Contextual analysis



Syntactic prosodic parsing



- Chinks'n chunks

a prosodic phrase =

a sequence of chinks (≈function words)

followed by a sequence of chunks (≈content words)

- Example :

I asked them

if they were going home

to Idaho

and they said yes

and anticipated one more stop

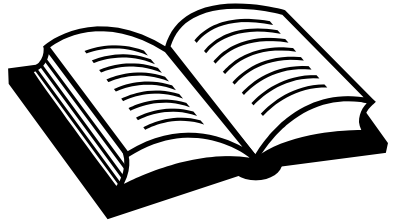
before getting home

From text to phones (1)

To be or not to be, that is the question.

_ t U b i: Q r n Q t t U b i: _ D { t s D @ k w e s tS @ n _

Be	*b i:*
Not	*n Q t*
Or	*O r*
Question	*k w e s tS @ n*
That	*D { t*
The	*D @*
To	*t U*
's	*s*



From text to phones (2)

Not that simple

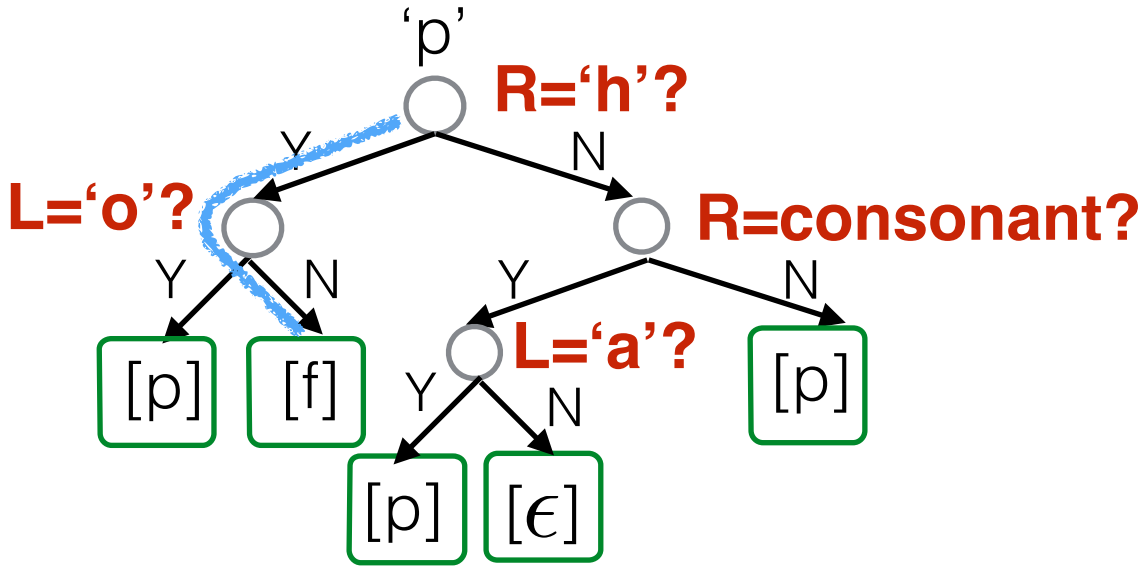
Problem	Example	Level	Information
Assimilation	nasality or sonority assimimation, vocalic harmonization	word/sentence	reading style, pronunciation of neighbors
Heterophonic homographs	**the**, record, contrast , read, est, couvent, portions, etc.	word	part-of-speech, meaning (rare)
Schwa deletion	table rouge, je ne te le redirai pas	sentence	syntactic articulation, pronunciation of neighbors, speaking style
Phonetic liaisons	très utile, deux à deux, plat exquis	sentence	syntactic articulation,
New words	proopiomelancortin	word	spelling analogy
Proper names	*your name here ...*	word	morphology, analogy

Letter-to-sound conversion (Grapheme-to-phoneme conversion)

- Decision tree-based approach
- Hidden Markov model based approach (e.g., joint sequence modeling)
- Neural network based approach (e.g. NETtalk)

$$\mathbf{G}: P \rightarrow H \rightarrow O \rightarrow N \rightarrow E$$

$$\mathbf{F}: /f/ \rightarrow /ow/ \rightarrow /n/$$



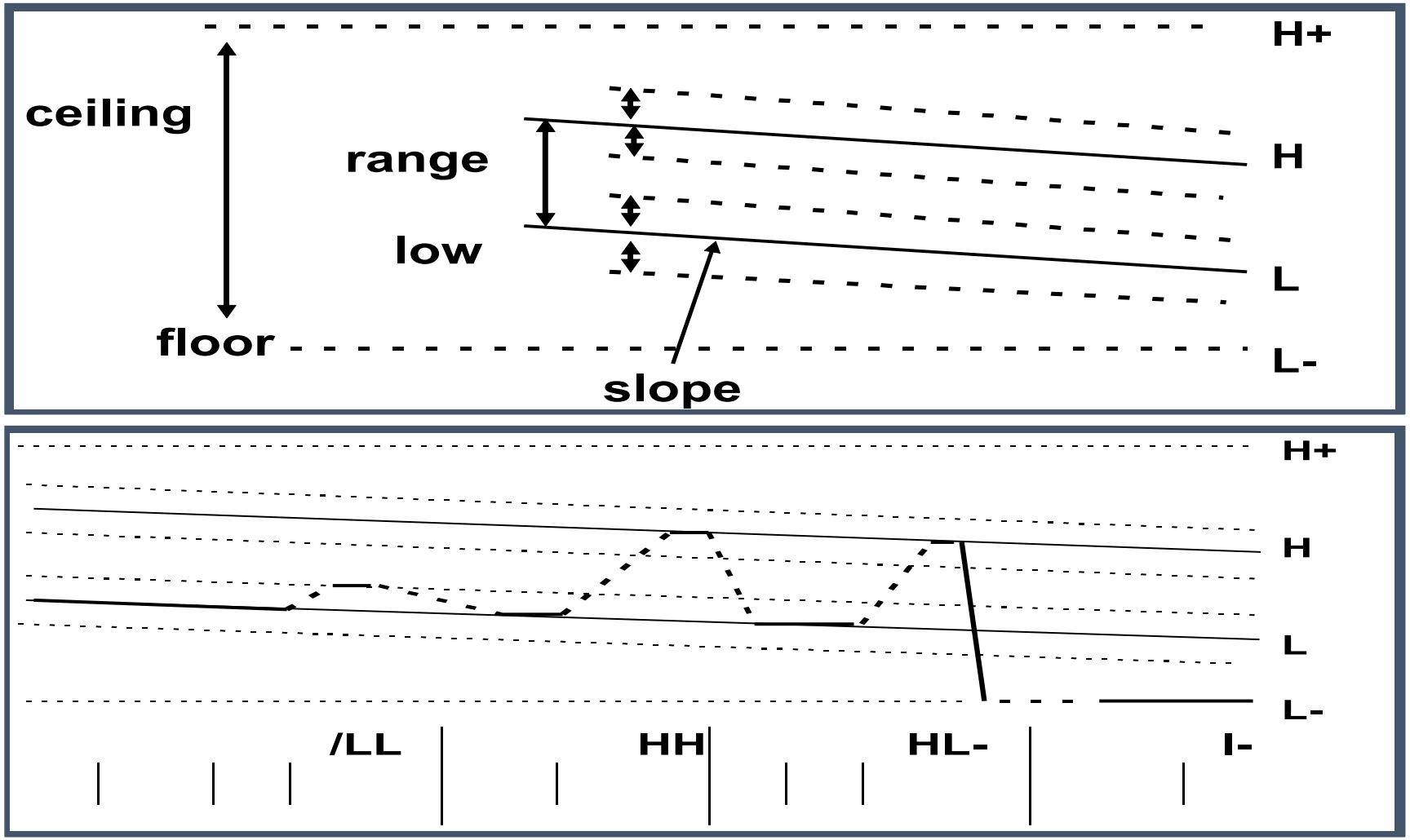
Prosody generation: text-to-tones

	ce personnage grossier, te dérange-t-il
WS	. . . o . o . . o .
SG	(. . . -)(. -) (. . . -)
IG 1	(. . . /LL)(. HH) (. . . H/H)
IG 2	(. . . - . HH) (. . . H/H)

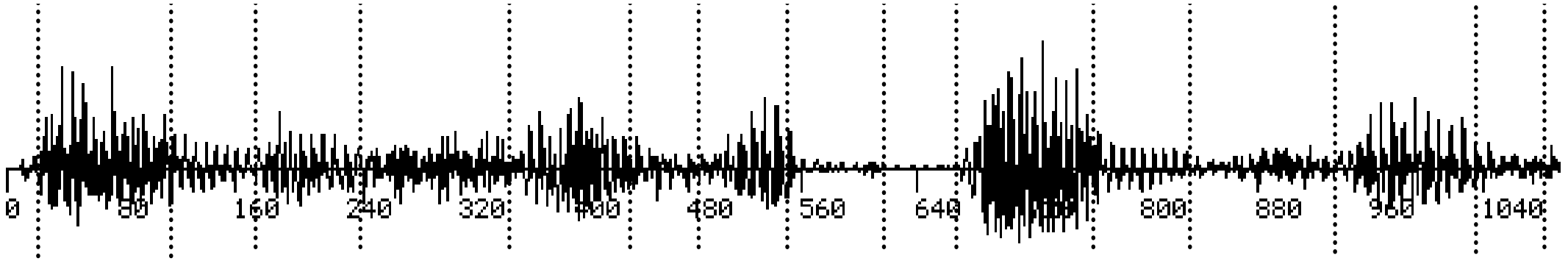
WS = word stress = lexical stress ← *Phonetization*
SG = stress group
IG = intonation group ← *Synt.-Pros. Phrasing*
(only one stressed syllable)

See: Tone and Break Indices (ToBI) , Guidelines for ToBI labeling

Prosody generation: tones-to-F0

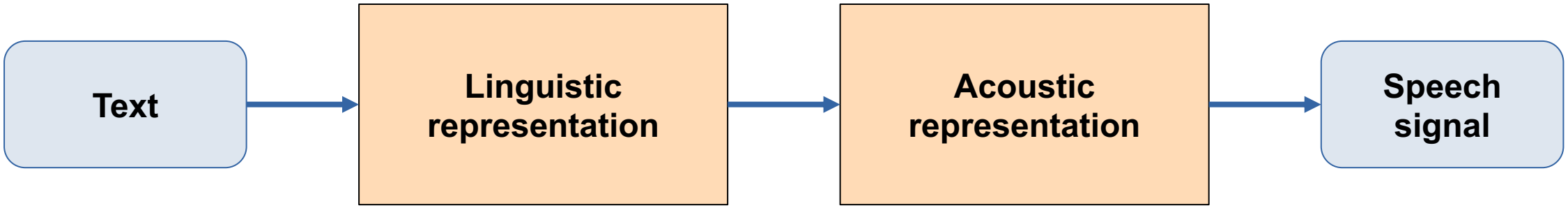


Phoneme duration

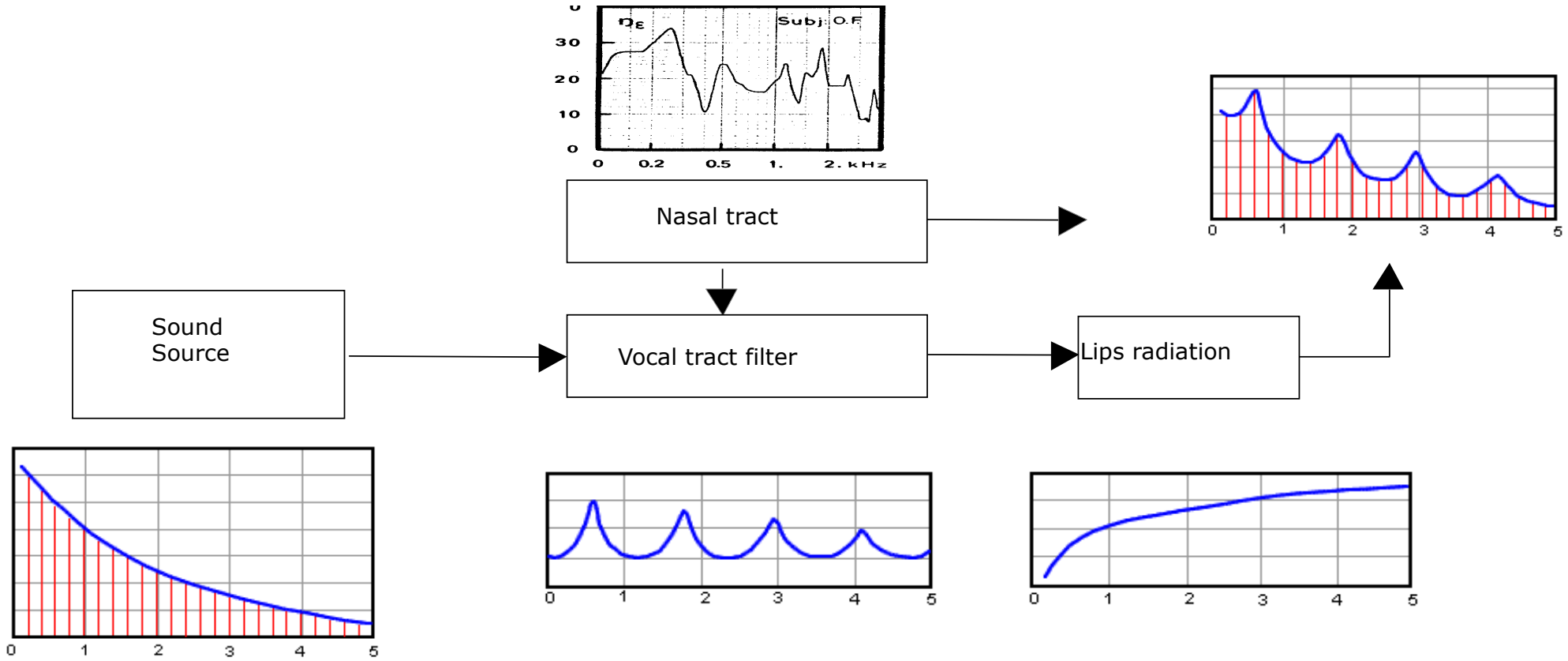


- Not constant
- Not fixed for a given phoneme
- Linked to intonation
  (longer on accented syllables)
- Predicted using rules or decision trees

TTS – Basic Steps



Speech production model



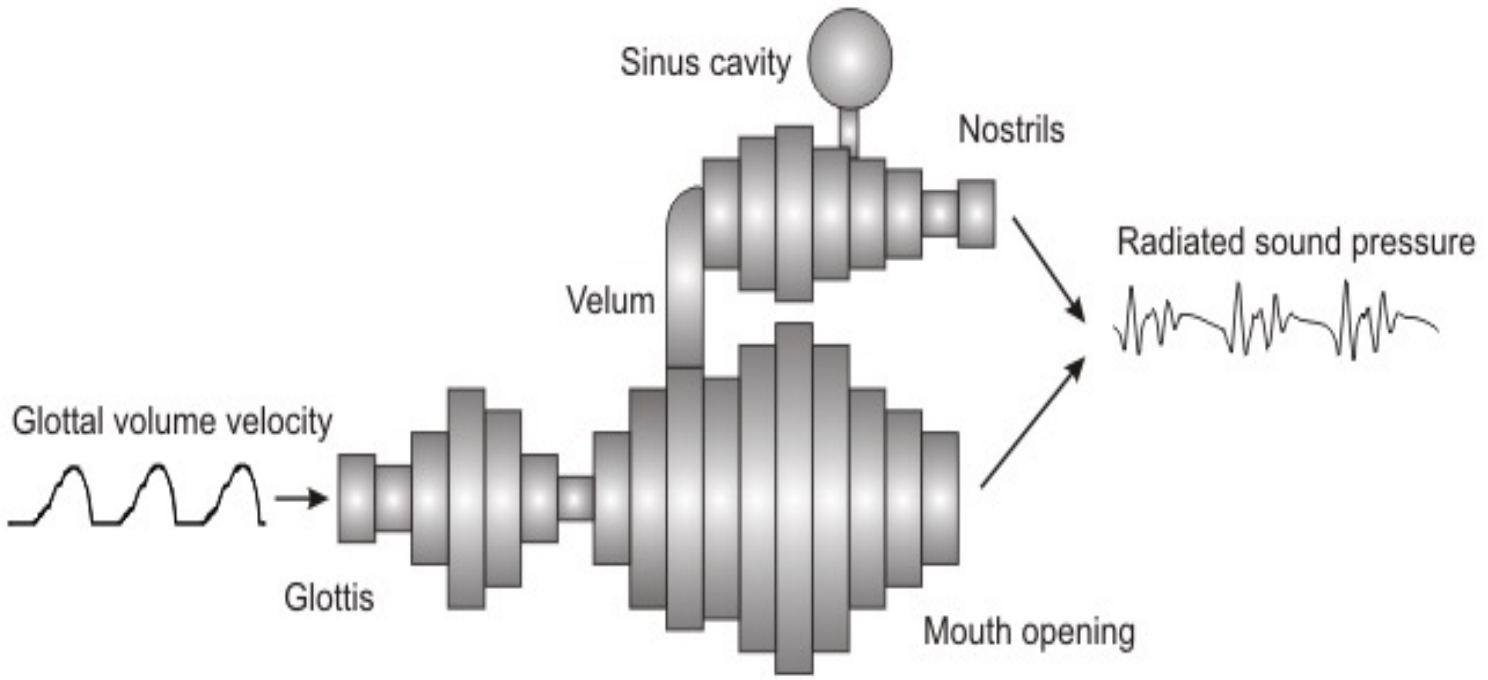
Credits: Lindqvist-Gauffin, Sundberg, Stevens, Mannel

Voice Source

- Two main types:
  - Glottal source (quasi periodic)
  - Constrictive noise source (stochastic)
- Class of phonemes <=> type of source
  - Only glottal source: vowels and semi-vowels
  - Noise source: consonants; type of constriction
    - Fricatives
    - Plosives
    - Approximants

Articulatory Speech Synthesis (1)

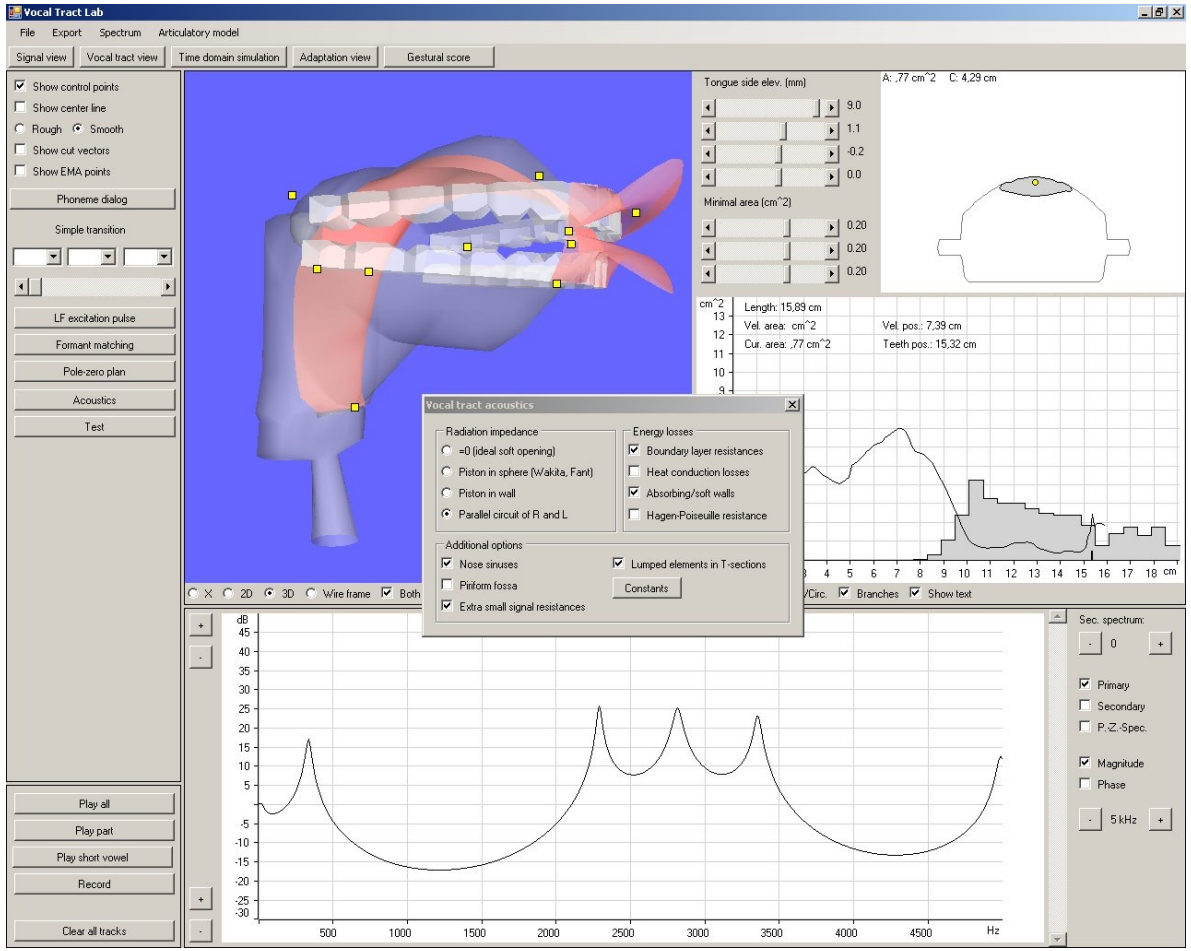
1. Geometric description of vocal tract based on a set of articulatory parameters.
2. A mechanism to control the articulatory parameters in an utterance
3. Acoustic simulation based on an acoustic model



Source: VocalTractLab (Birkholz et al.)

Articulatory Speech Synthesis (2)

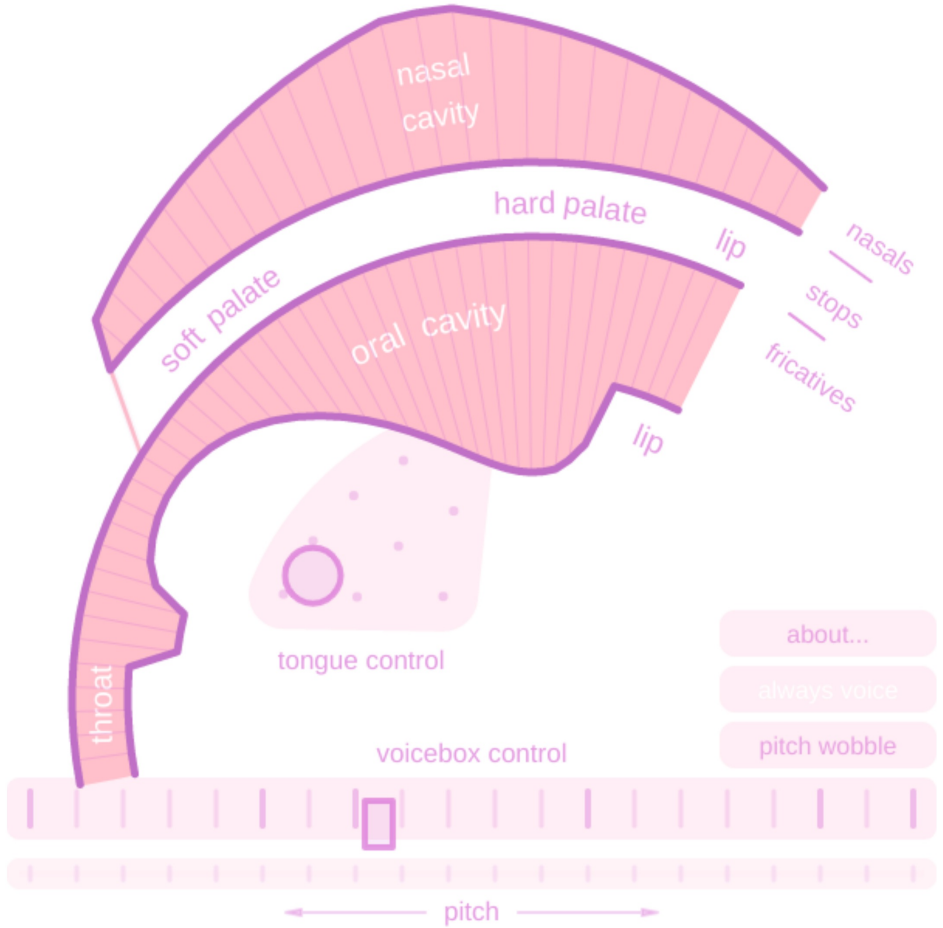
- Emulate the human speech production process
  - Allows fine-grained control
  - Challenging to model all details of the vocal tract
  - Too complex for practical applications
  - Very slow



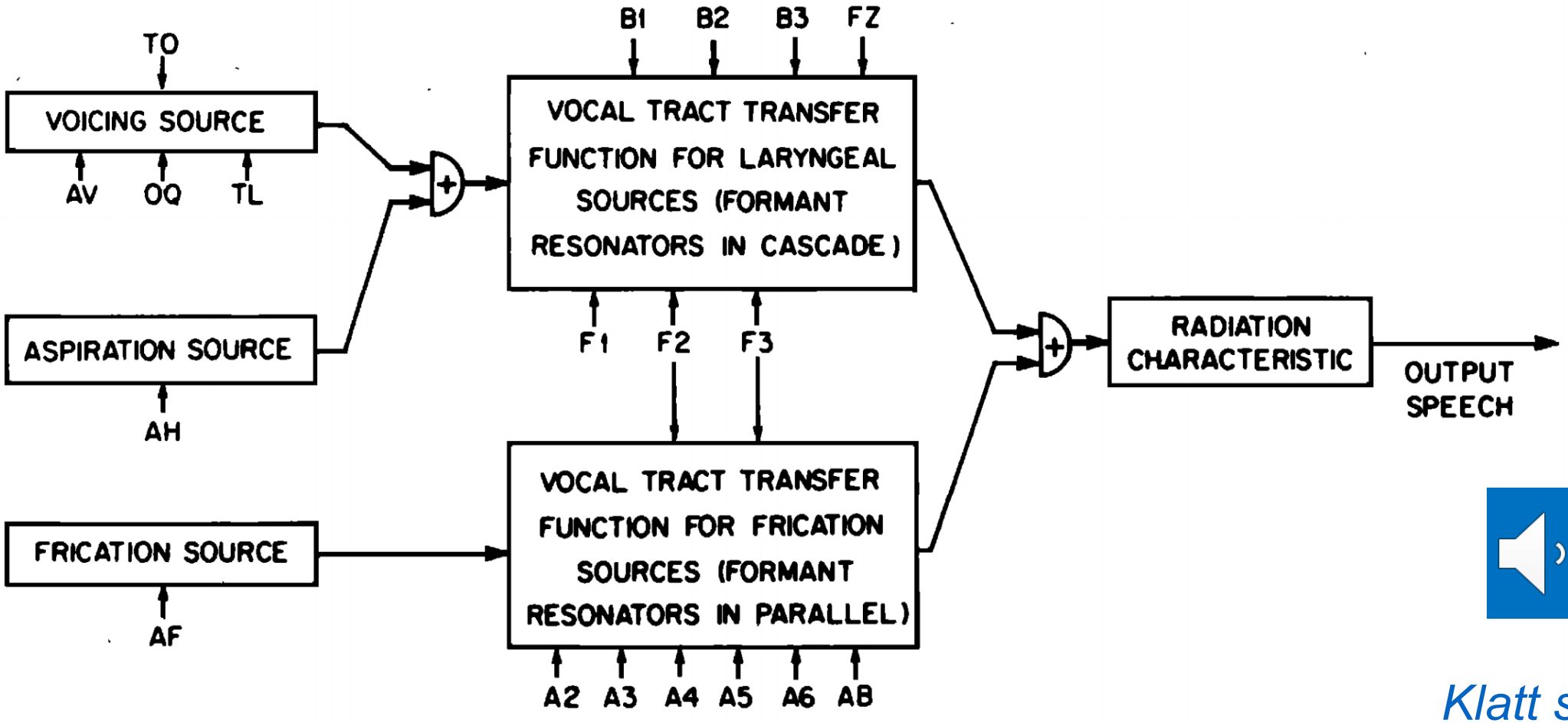
[VocalTractLab (Birkholz et al.)]

Articulatory Speech Synthesis (3)

Interactive online demo (best on multi-touch devices): **Pink Trombone**



Formant Speech Synthesis

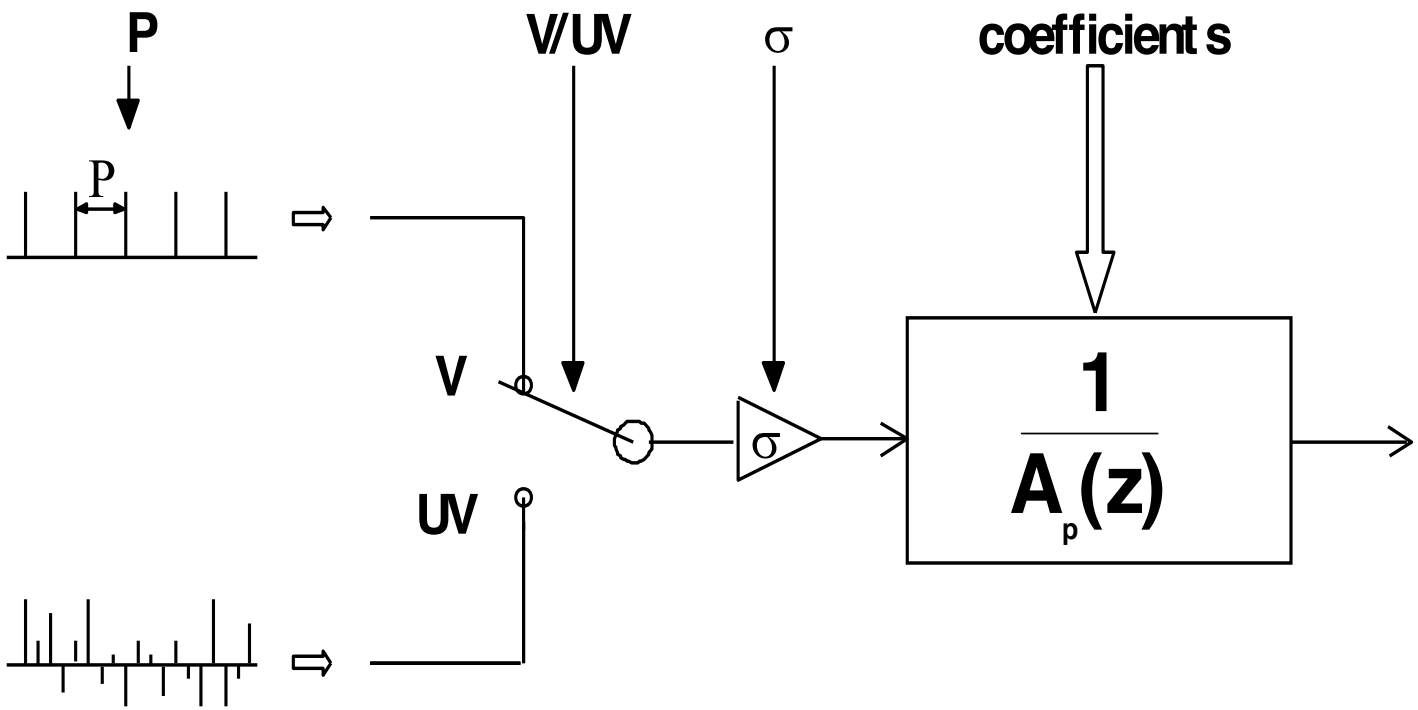


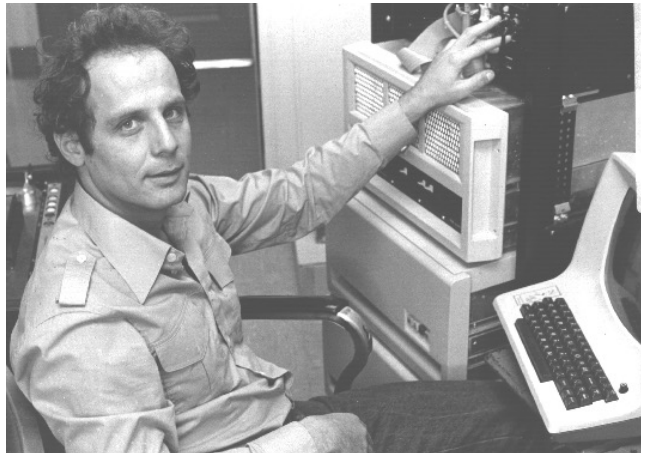
Klatt synthesizer

Formant Speech Synthesis

- Based on the source-filter model by combining:
    1. An excitation signal
    2. Formant resonators that model the vocal tract
- Cascade or parallel structure of resonators or a combination
- Interpretable parameters

LPC synthesizer





Olive(1980)

FPMs (1989)

text (Dr. Bourlard, Bourlard Dr)

↓

semantic pre-processing

phrase parsing

↓

text DOCTOR BOURLARD

pronunciation dictionary and rules → text-to-phone

↓ /d//aa//k//t//er//b//ou//rr//l//ax//r/

timing and intonation

↓ /d//**aa**//k//t//er//sp//b**//OU**//rr//l//ax//r

inventory of acoustic segments → segmental concatenation

↓ {pd}
{paa}{pk}{pt}{per}…
…

synthesizer

↓

speech signal

Raw concatenation

Concatenative Speech Synthesis

Generate speech by concatenating pre-recorded segments.

- Diphone synthesis
- Unit selection synthesis
- Domain-specific synthesis (e.g. in train stations)

Sounds very natural, but can lead to artefacts at segment boundaries and may need a very large recording database. Limited to the recorded speaker.

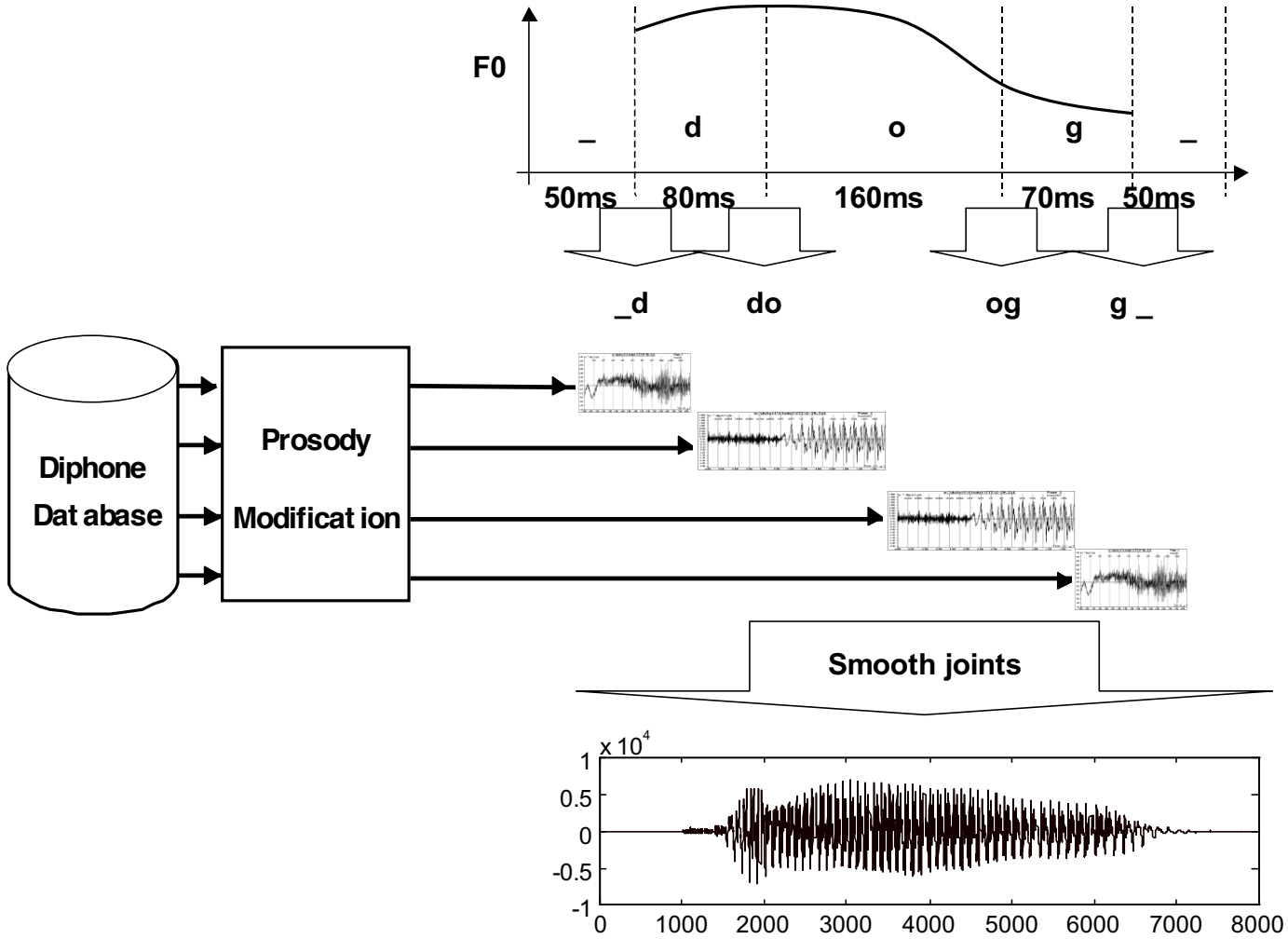
Recording a TTS database

- Recorded speaker = *voice talent*
- Quiet, ideally studio environment (different from ASR)
- Recording prompts should provide coverage of phones and phonetic context
- Generally aiming for neutral speech
  - Emotional speech synthesis is an open problem

Example: CMU_Arctic database

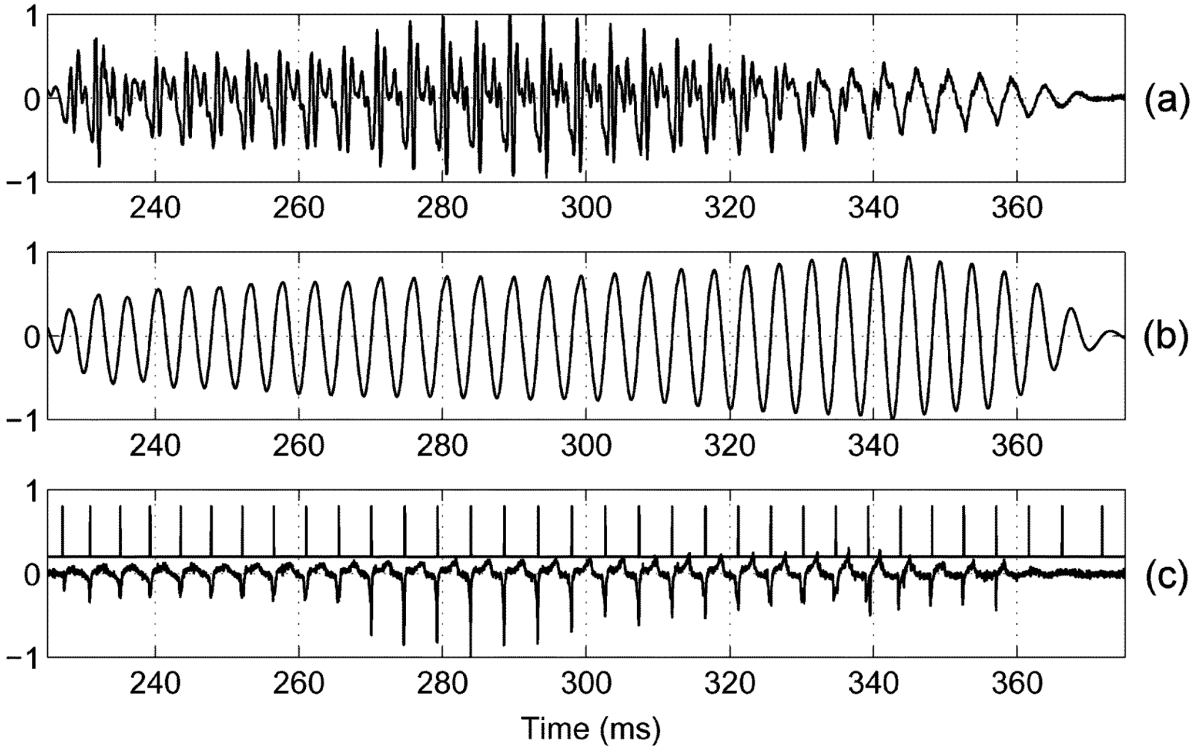
Diphone Synthesis

- **Diphone:** Transition between two phones.
- Record one instance of each diphone and concatenate to form utterances.
- Adjust prosody for naturalness.
- Forced alignment identifies points to cut recordings.



Signal Processing for Concatenative TTS

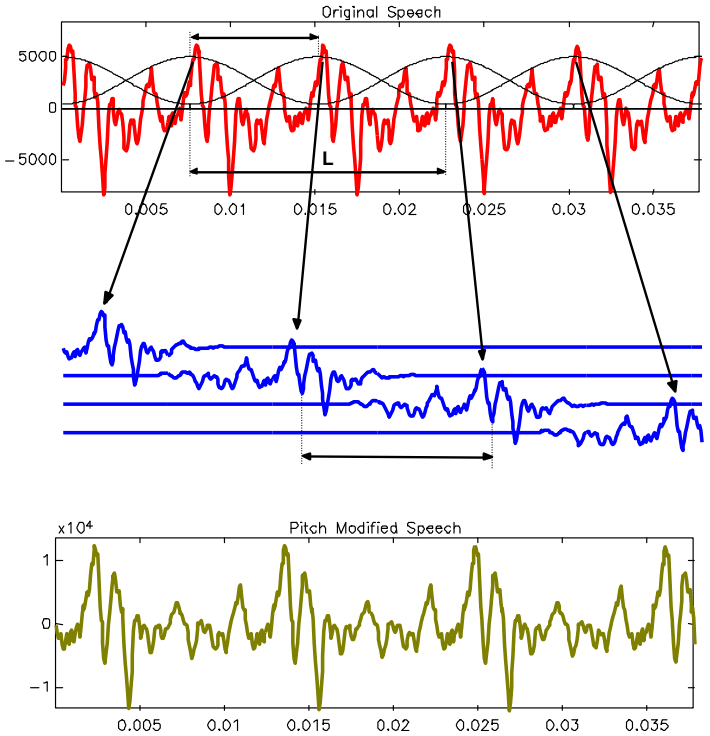
- Need to avoid artefacts when joining segments
  - Pitch-synchronous concatenation (epoch detection)

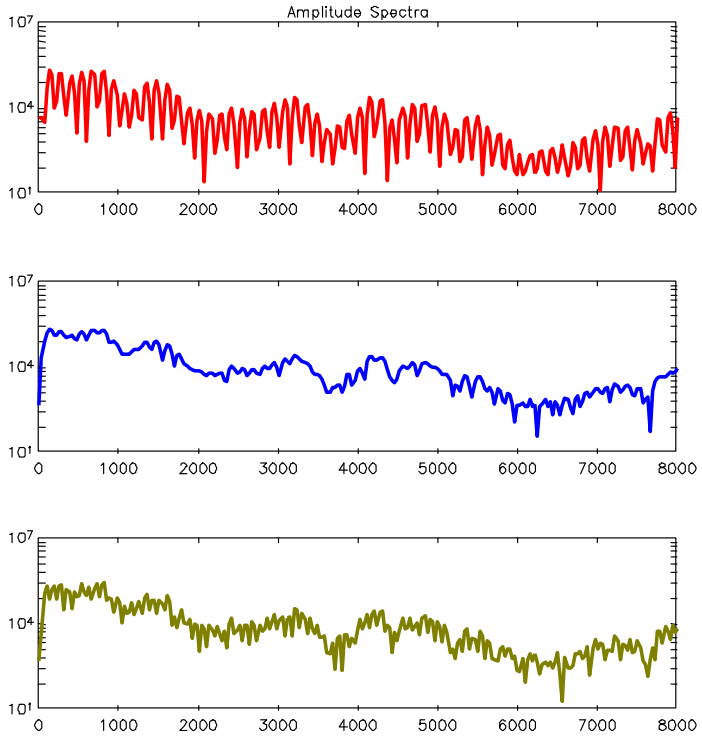


Epoch extraction from speech signals (2008)

Pitch-synchronous Overlap and Add (PSOLA)

- Prosody (duration, intonation) modification for concatenative synthesis
- Obtain pitch-synchronous windows that can then be modified



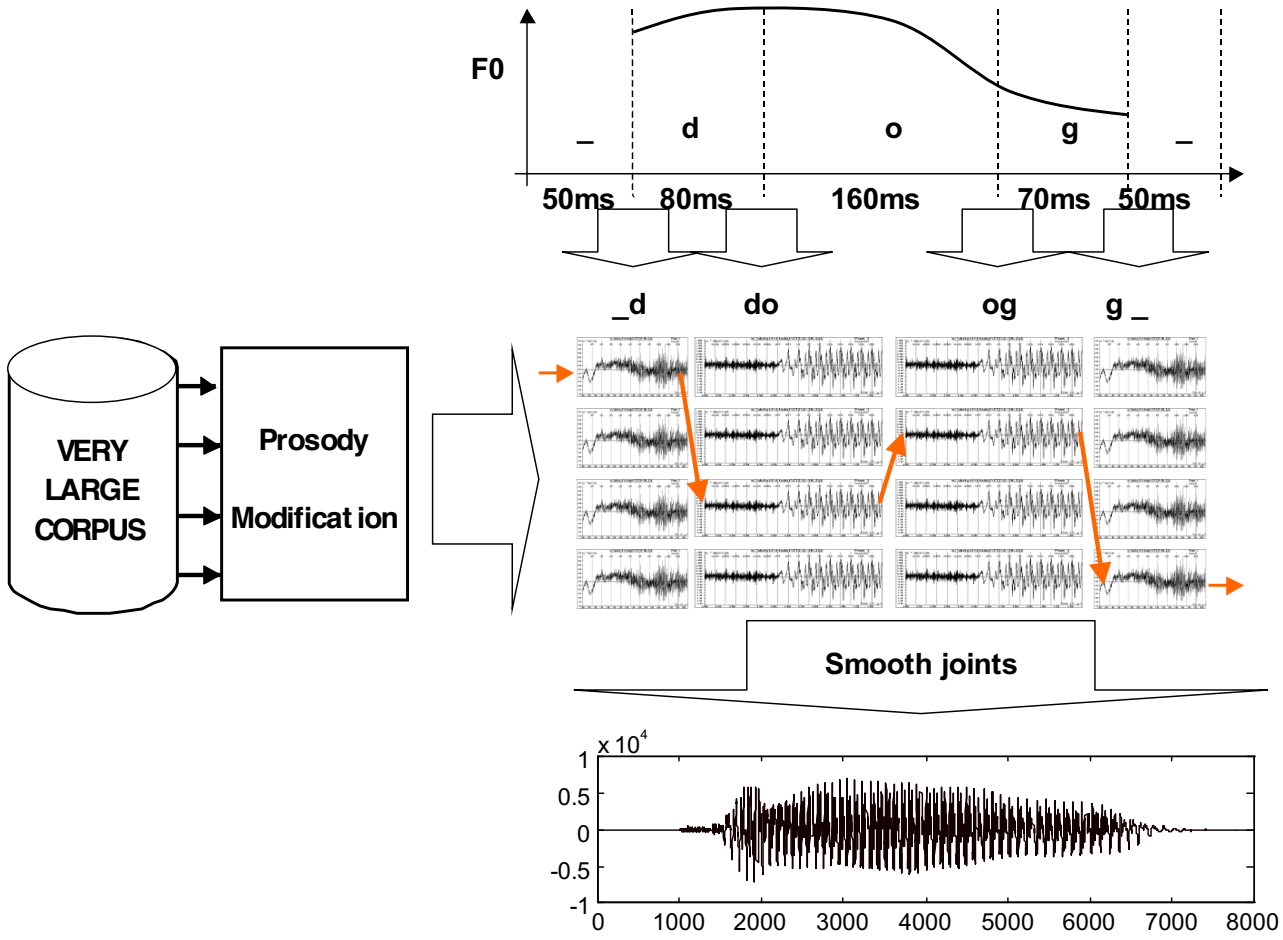


Cnet (1990)

Limsi (1992)

Automatic Unit Selection

- Record a large corpus with many instances of each unit.
- For each utterance, select the best sequence of units through **Viterbi beam search**.
  - **Target cost:** Measures how well a unit fits the context.
  - **Join cost:** Measures how well two units can be concatenated.



Automatic Unit Selection: Target and Join Cost

Target cost: Find best match to the target unit, in terms of

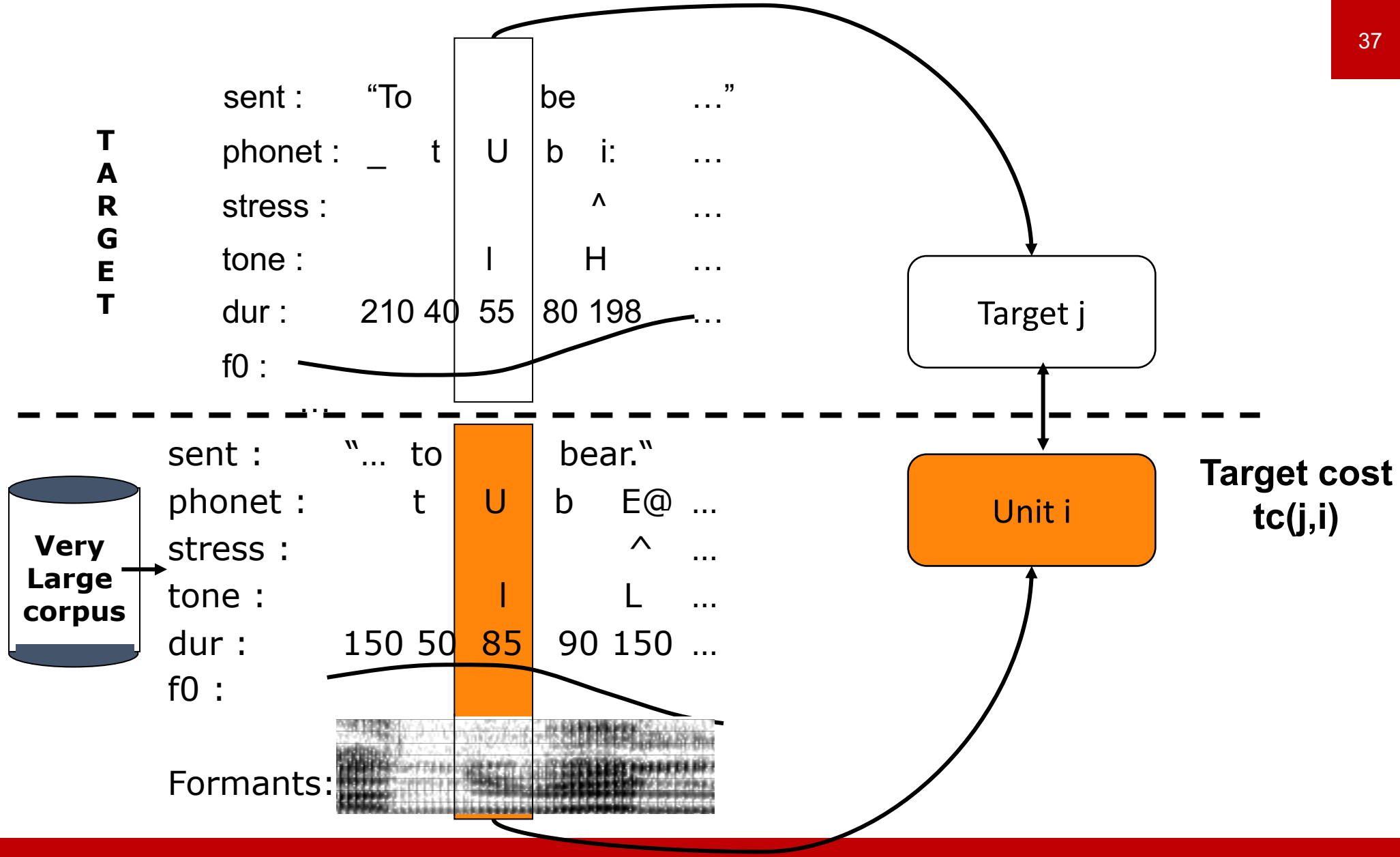
- Phonetic context
- F0, stress, phrase position, duration
- Acoustic distance

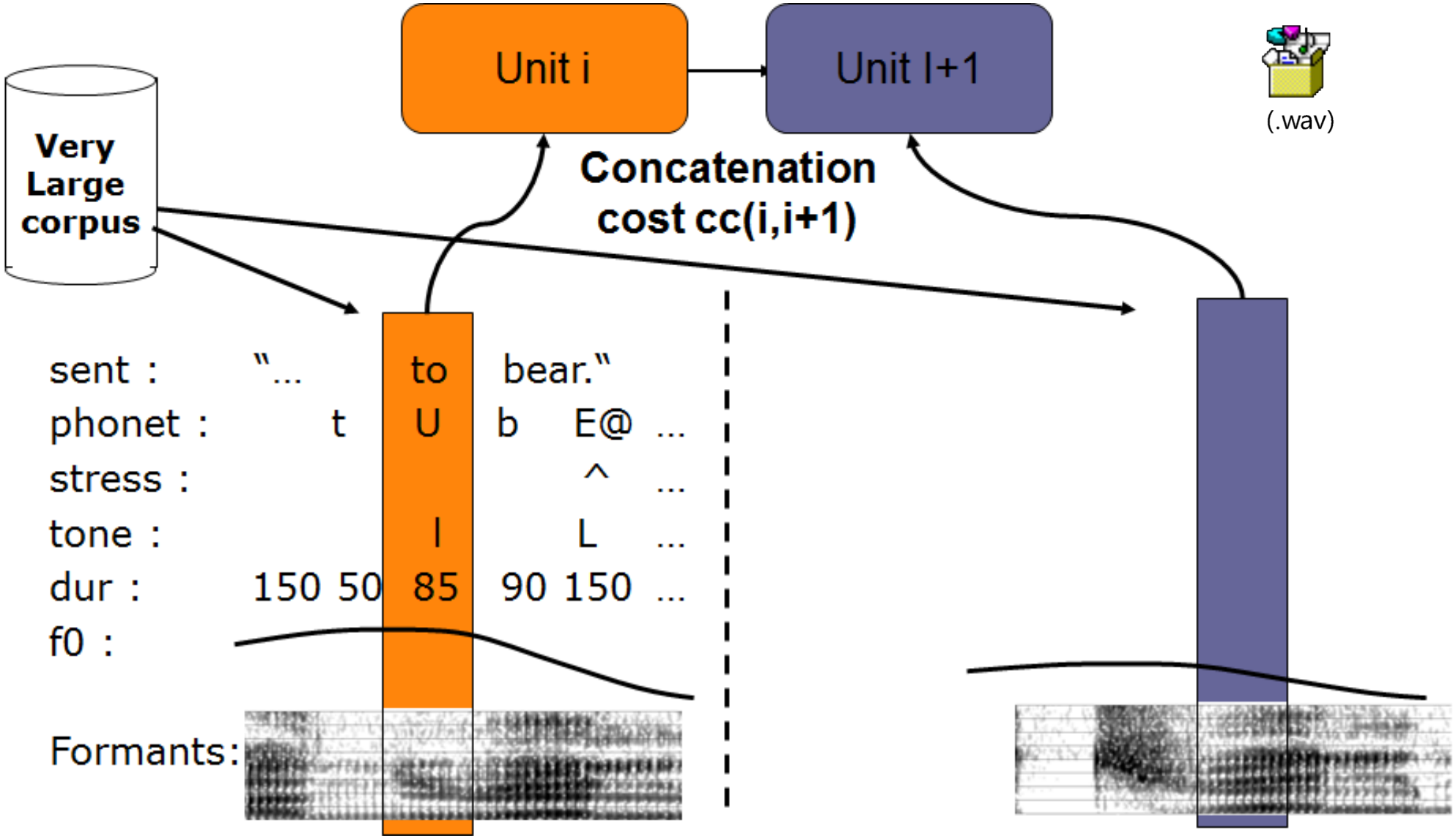
Join cost: Find a unit that can combine well with neighboring units and has

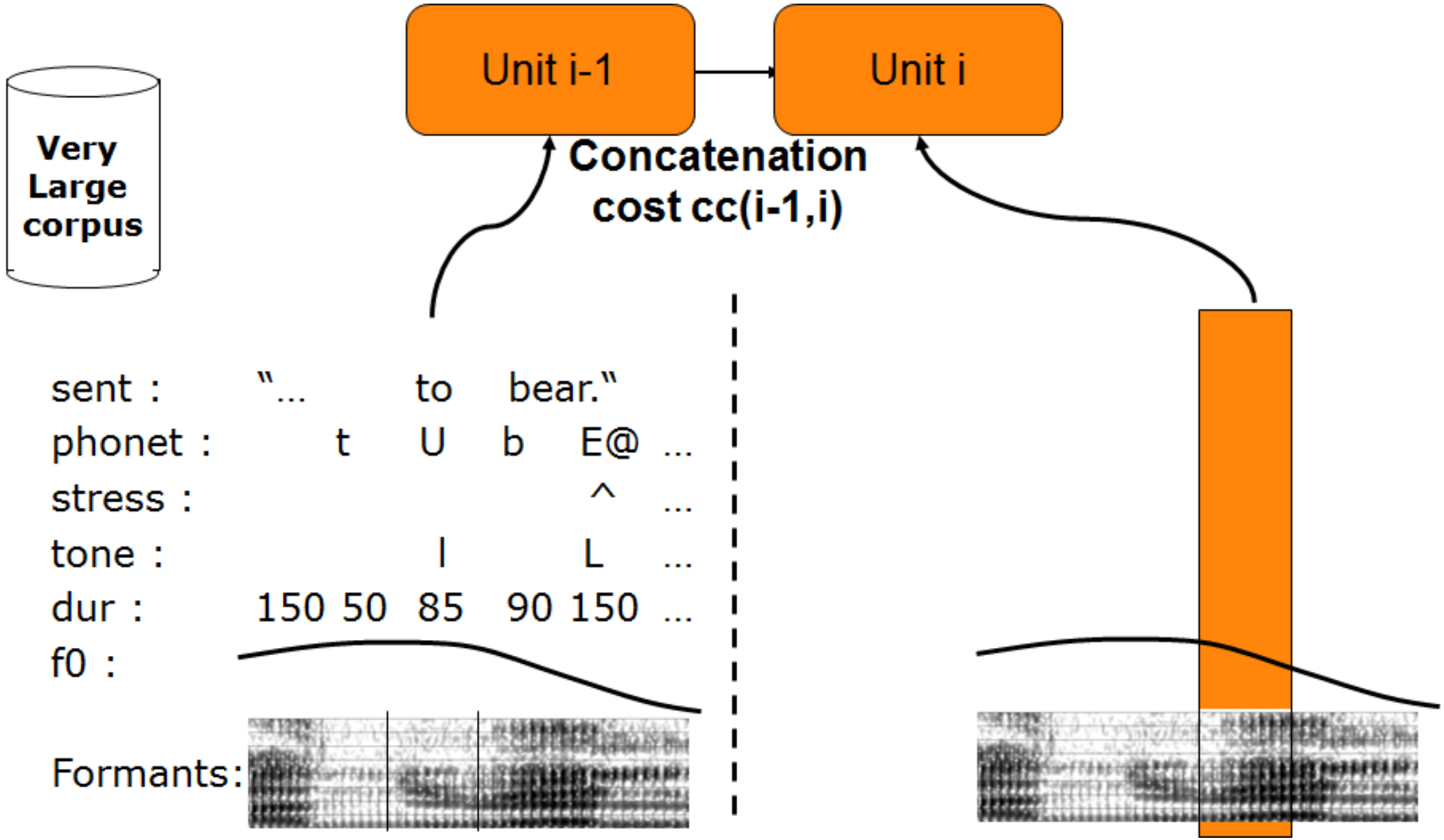
- Matching formants, energy, F0

Build your own: Festival unit selection voice

Voice demo


TARGET
sent : "To be …"
phonet : _ t U b i: …
stress : ^ …
tone : l H …
dur : 210 40 55 80 198 …
f0 :
Target j
Unit i
Target cost
tc(j,i)
Very Large corpus
sent : "… to bear."
phonet : t U b E@ …
stress : ^ …
tone : l L …
dur : 150 50 85 90 150 …
f0 :
Formants:


Unit i
Unit I+1
(.wav)
Very Large corpus
Concatenation cost cc(i,i+1)
sent : "... to bear."
phonet : t U b E@ ...
stress : ^ ...
tone : I L ...
dur : 150 50 85 90 150 ...
f0 :
Formants:


Unit i-1
Unit i
Very Large corpus
Concatenation cost cc(i-1,i)
sent : "... to bear."
phonet : t U b E@ ...
stress : ^ ...
tone : l L ...
dur : 150 50 85 90 150 ...
f0 :
Formants:

Outline

- Overview
- Natural language processing (NLP) for speech synthesis
- Articulatory speech synthesis
- Formant speech synthesis
- Concatenative speech synthesis
- **Statistical parametric speech synthesis (Part II)**
- **End-to-end speech synthesis**
- **Evaluation**

Thank you for your attention!