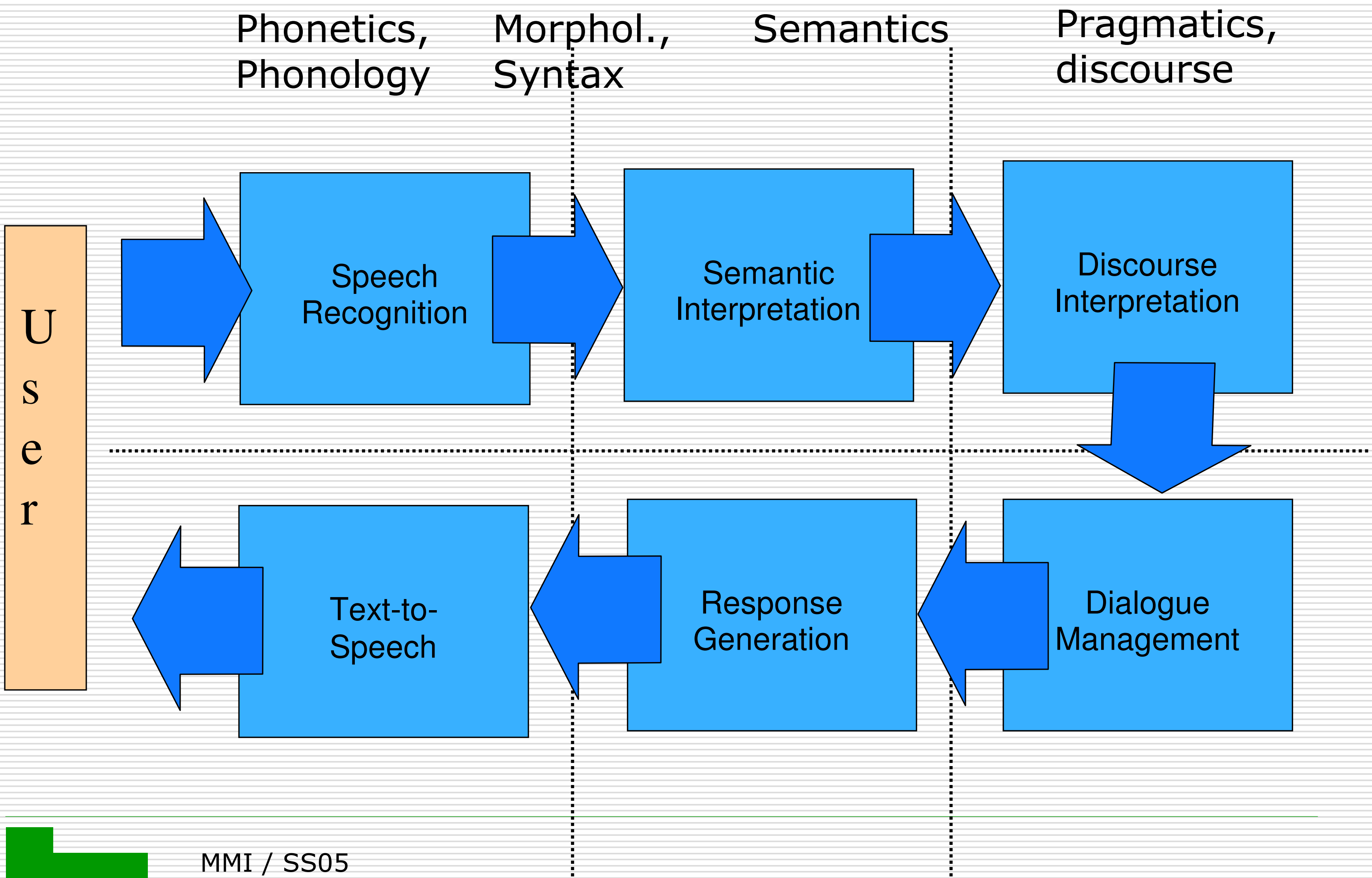
Spoken Language Interaction

Text-to-Speech (TTS)

Spoken Dialogue Systems

- A system that allows a user to *speak* his queries in natural language and receive useful spoken *responses* from it
- Provides an interface between the user and a computer-based application that permits *spoken interaction* with the application in a “relatively natural manner”

Spoken Dialogue System - overview

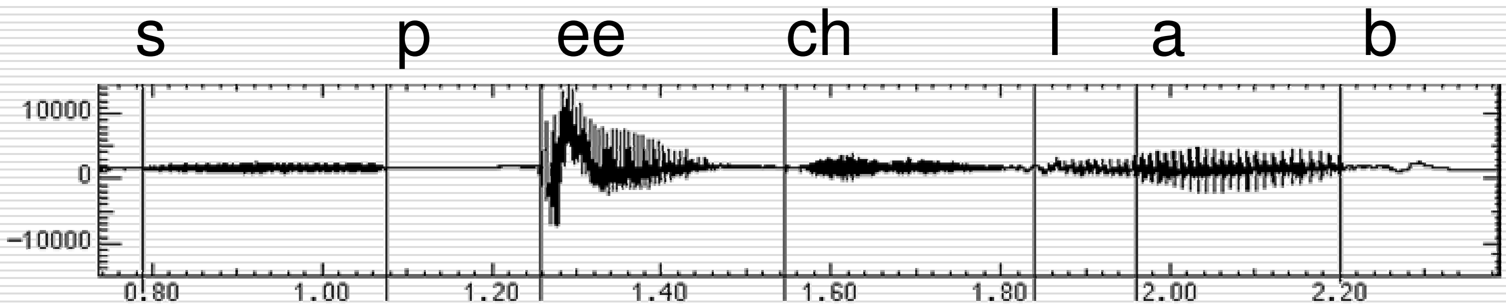


Spoken Dialogue System - overview

- Speech Recognition:
  - Decode the sequence of feature vectors into a sequence of words.
- Syntactic Analysis and Semantic Interpretation:
  - Determine the meaning of the words.
- Discourse Interpretation:
  - Understand what the user intends by interpreting the utterances in context.
- Dialogue Management:
  - Determine system goals in response to user utterances based on user intention.
- Response Generation:
  - Express the system goals in natural utterances
- Text-to-speech:
  - Generate synthetic speech audio for the words

Starting and end point: acoustic waves

- Human speech generates a wave
- A wave for the words “speech lab”:



Text-to-speech

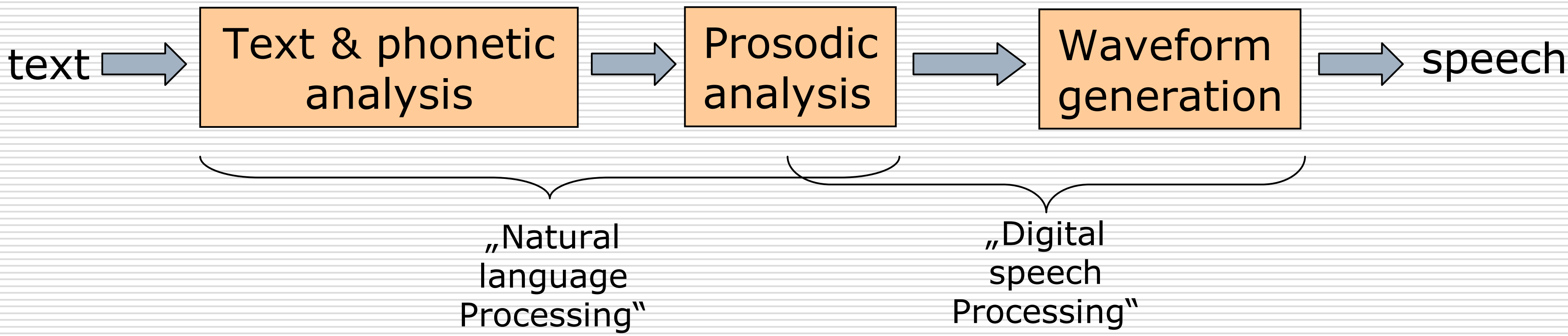
- Task: mapping text to sound waves

- The simplest (and most common) solution is to record prompts spoken by a (trained) human
  - e.g. in-car navigation systems
- Produces human quality voice
- Limited number of prompts
- Problems with re-combination

Text-to-speech

Central steps:

1. Analyse text and select sound *segments*
2. Determine prosody and how to model it with single segments
3. Turn into acoustic waveform (*speech synthesis*)



Segments candidates

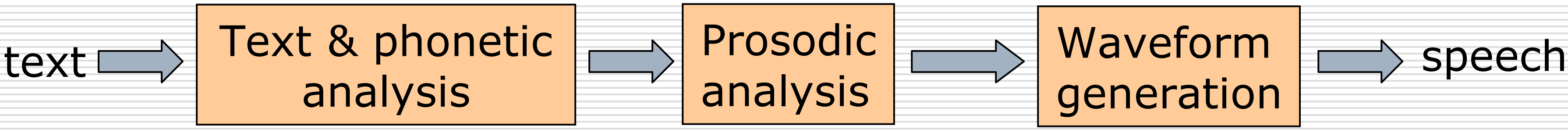
- Phoneme
  - Smallest meaning-distinctive, but *not meaningful* articulatory unit
  - Example: Sounds (phones) [b] (e.g. in bill) and [ph] (e.g. in pill) discriminate two meanings
    → different phonemes /b/ und /p/
  - Subsume different elemental sounds under one phoneme, e.g. [p] in spill and [ph] in pill → /p/
  - Every language has ist own set of phonemes and combination rules
  - Concatenating phonemes for TTS is problematic, co-articulatory effects make result sound unintelligible

Co-articulation = change in segments due to movement of articulators in neighboring segments

Segment candidates

- Allophones
  - Variants of a phoneme in specific contexts
  - Example: Phoneme /p/ → [p] in spill and [ph] in pill
- Diphones („Zweilautverbindungen")
  - Diphones start half-way thru 1st phone and end half-way thru 2nd
  - ⇒ critical phone transition is contained in the segment itself, need not be calculated by synthesizer
  - Example: diphones for German word „Phonetik":
    f-o, o-n, n-e, e-t, t-i, i-k

TTS Architecture – more closely



1. Text analysis
   - Text Normalization, part-of-Speech tagging, homonym Disambiguation
2. Phonetic Analysis
   - Dictionary Lookup, letter-to-Sound mapping
3. Prosodic Analysis
   - Boundary placement, pitch accent assignment, duration computation
4. Waveform synthesis

Text analysis

from text to words

- Analysis of raw text into pronounceable words
- Sample problems:
  - `He stole $100 million from the bank`
  - `It's 13 St. Andrews St.`
  - `The home page is http://www.stanford.edu`
  - `yes, see you the following tues, that's 11/12/01`
- Steps
  - Identify tokens in text
  - Chunk tokens into reasonably sized sections
  - Map tokens to words
  - Identify types for words (“part-of-speech tagging“)

Phonetic analysis
from words to sounds

- Look in pronunciation dictionary!
- Words/wordforms
  - e.g. CMUdict: ~125.000 wordforms
  - primary stress, secondary stress, no

  http://www.speech.cs.cmu.edu/cgi-bin/cmudict

- Problem: a lot of words, some of them unknown
  - Morphological productivity
    - Turkish is an extreme example of multiple affixation: "*uygarlaStIramadIklarImIzdanmISsInIzcasIna*" = "(behaving) as if you are among those whom we could not civilize"
  - Proper names, numbers, foreign words, …

- So, need to map graphemes (letters) to sounds
- Also homograph disambiguation: same spelling, different sounds ("wind", "live", "read", ..)

Letter-to-Sound (LTS) Rules

- Letter-to-sound rules
  - Early systems all LTS
  - MITalk (1987) was radical in having a huge 10.000 rules repository: p – [p]; ph – [f]; phe – [fi]; phes – [fiz]; … … …

- *Festival* LTS rules take account of co-articulation http://www.cstr.ed.ac.uk/projects/festival.html
  - (LEFTCONTEXT [ITEMS] RIGHTCONTEXT = NEWITEMS)
  - Example:
    - (# [c h] C = k)
    - (# [c h] = ch)
  - # denotes beginning of word
  - C means all consonants
  - Rules apply in order
    - "christmas" pronounced with [k]
    - But word with 'ch' followed by non-consonant pronounced [ch], e.g., "choice"

Dictionaries aren't always sufficient

Modern systems have 3-part phonetic analysis

- Big pronunciation dictionary (word forms)
- Special code for handling names and acronyms
- Machine-learned LTS system for other unknown words (not in dictionary)

Learning LTS rules automatically

- Learn LTS rules from a dictionary of the language
- Two steps:
  1. find alignments
  2. learn rule-induction

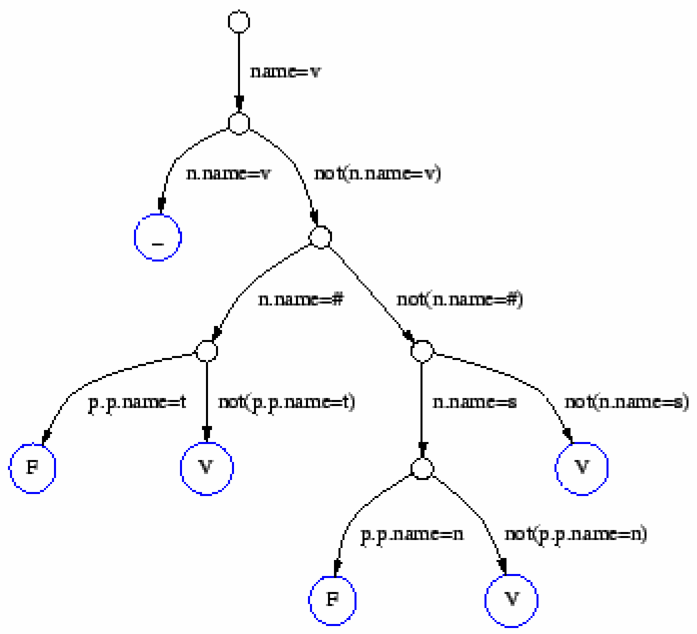
Alignment problem: letters can map to zero, one, two or very exceptionally to three phones

- Letters: c h e c k e d
- Phones: ch _ eh _ k _ t

- Approach by (Black et al. 1998)
  - specify which letters can be rendered as which phones
    - C goes to k/ch/s/sh, W goes to w/v/f, etc.
  - find all valid alignments
  - find probability of a letter to be pronounced by a phone, P(letter|phone)
  - score all alignments, take best

Rule induction

- *CART* = classification and regression tree
- From all alignments, build a tree for each letter in alphabet (26 plus accented) using context of 3 letters
  - # # # c h e c -> ch
  - c h e c k e d -> _
- This produces 92-96% correct *letter* accuracy for English



- Improvements:
  - names and acronyms don't follow the systematicity of other words, take out of training data
  - special-purpose tools for names and acronyms

Prosodic analysis

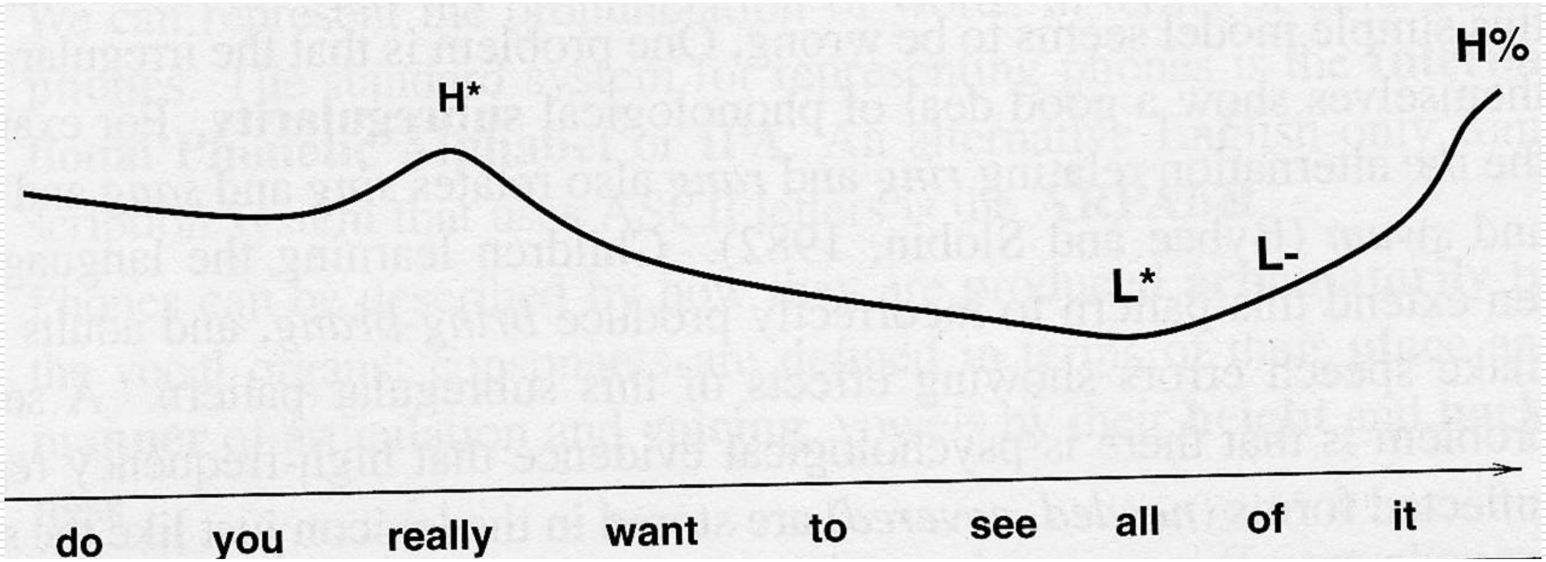
from words+phones to boundaries, accent, F0, duration

TTS systems need to create proper prosody by adapting pitch, duration, and loudness of the segments.

- Prosodic phrasing/boundaries:
  - Need to break utterances into units (intonation phrases and maybe even intermediate phrases)
  - Punctuation and syntactic structure useful, but not sufficient
- Intonation/accents:
  - Predictions of accents: which syllables should be accented?
  - Realization as F0 contour: for given accents/tones, generate F0 contour
- Duration:
  - Predict duration of each phone
  - Helps to create prominence

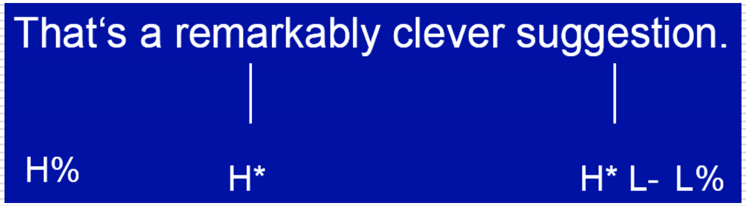
Pitch: tone sequence models

- Idea: generate fundamental frequency (F0) contours from phonologically distinctive tones (High or Low)
- Defines sequence of tonal *pitch targets*
- Complete F0 contour creates pitch accents and phrasing



ToBI - Tone and Break Indices

- *Pitch Accent* (*): H*, L*, H*+L, H+L*, L*+H, L+H*
- *Phrase Accent* (-): H-, L-
- *Boundary Tone* (%): H%, L%
- *Intonation Phrase* = <Pitch A.>$^+$<Phrase A.><Bound.T.>



- Allows to build grammars for pitch accents, phrase accents, and boundary tones. Example:

$$\left(\begin{Bmatrix} \%H \\ \%L \end{Bmatrix} \left(\begin{Bmatrix} H* \\ L* \\ H*+L \\ H+L* \\ L*+H \\ L+H* \end{Bmatrix}^{+} \begin{Bmatrix} H- \\ L- \end{Bmatrix} \right)^{+} \begin{Bmatrix} H\% \\ L\% \end{Bmatrix} \right)^{+}$$

(B. Möbius, U. Stuttgart)

Pitch *accents*

- In the first place, properties of *words*
- Decisive for how words are interpreted!
  - „mark the interpretations of words as contributing to the *distinction* between the speaker's actual utterance and other things that he might be expected to have said in the context" (Steedman, 2004)
- Used to…
  - emphasize new information ("Then I saw a **church**.")
  - contrast parts („I like **blue** tiles better than **green** tiles.")
  - explicitly focus parts („I said I saw a **church**.")

Question: Which tone sequence to choose for which purpose?

Pitch accents

- Limited number of pitch contours typically found in every language, usually six (cf. ToBI for English)
- Accents of German according to GToBI (Reyelt et al.,1996)

H*		Gipfelakzent Emphase
L+H*		Steiler Anstieg auf Akzentsilbe Kontrastakzent
L*		Tonhöhe fällt auf Akzentsilbe ab oder bleibt dort

L*+H		Tieftonige Akzentsilbe gefolgt von Gipfel
H+L*		Abfall aus Höhenlage auf tieftonigeAkzentsilbe
H+!H*		Abfall auf abgesenkten Hochton *(downstep)*

abgesenkt

Which to choose depends on content and discourse
→ „concept-to-speech, content-to-speech“

Intonation and *information structure*

- *Theme:* Part of a proposition that create cohesive connection to previous propositions („discourse cohesion")
- *Rheme:* Part of a proposition that contirbutes new information

Example: Who is he? He is a student.

Theme: "He is" — Rheme: "a student."

(Bolinger; Halliday, 60's)

- *BEAT* (Cassell et al., 2001): assigns accents and boundary tones for English based on theme-rheme analysis:
  - Within THEME:
    - Suggest L+H* accent for NEW objects
    - Suggest LH% boundary tone at end of THEME
  - Within RHEME:
    - Suggest H* accent on NEW objects
    - Suggest LL% boundary tone at end of RHEME
  - H* accents on CONTRAST objects
  - Pauses on CLAUSE boundaries

Intonation and speaker's discourse model

- Information units are rendered as phrasal units
- In the current situation, the *speaker attributes* certain discourse status to each information unit (see Steedman, 2004):
  1. Theme (θ) or rheme (ρ)
  2. Mutually agreed (+) or not (-)
  3. Speaker ([S]) or hearer ([H]) *committed to* it, i.e. responsible for "owning" the information unit
- Meanings of pitch-accent types can be distinguished along the first two dimensions
- Boundaries distinguishable along the third dimension

	+	−
θ	L+H*	L*+H
ρ	H*, (H*+L)	L*, (H+L*)

Meaning of pitch accents

[S]	L, LL%, HL%
[H]	H, HH%, LH%

Meaning of boundaries tones

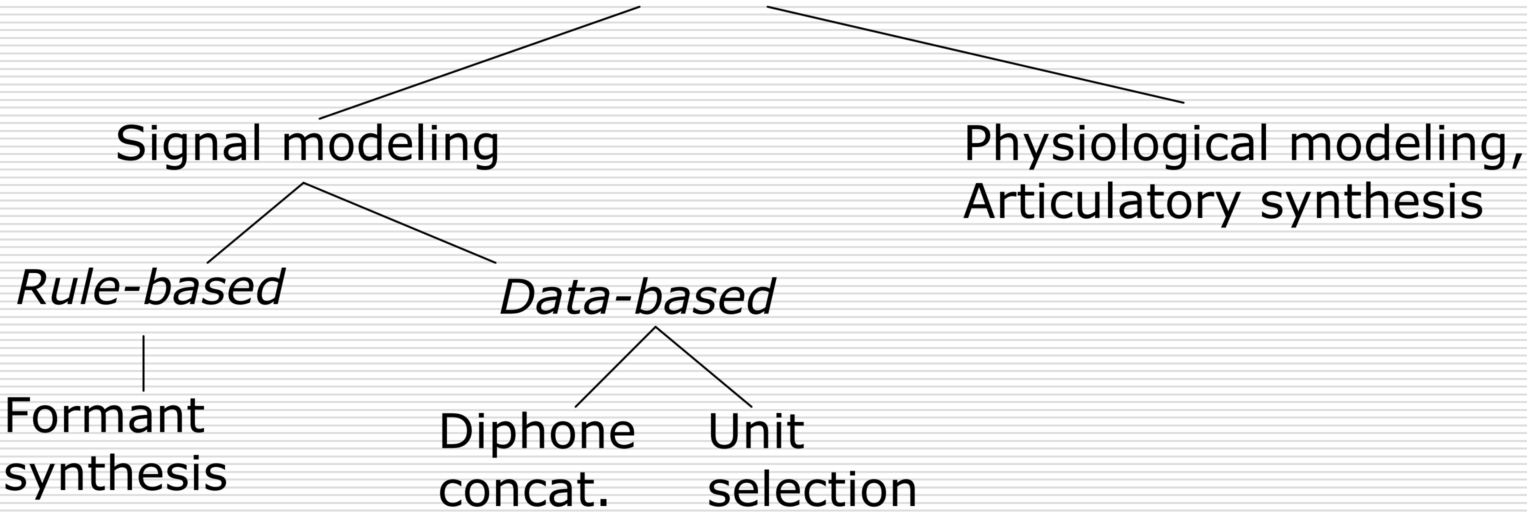
Duration

Generate segments with appropriate duration. Influenced by

- Segmental identity
  - /ai/ in ‘like’ twice as long as /I/ in ‘lick’
- Surrounding segments
  - vowels longer following voiced fricatives than voiceless stops
- Syllable stress
  - stressed syllables longer than unstressed
- Word “importance”
  - word accent with major pitch movement lengthens
- Location of syllable in word
  - word ending longer than starting longer than word internal
- Location of the syllable in the phrase
  - phrase final syllables longer than in other positions

Waveform synthesis

from segments, f0, duration to waveform



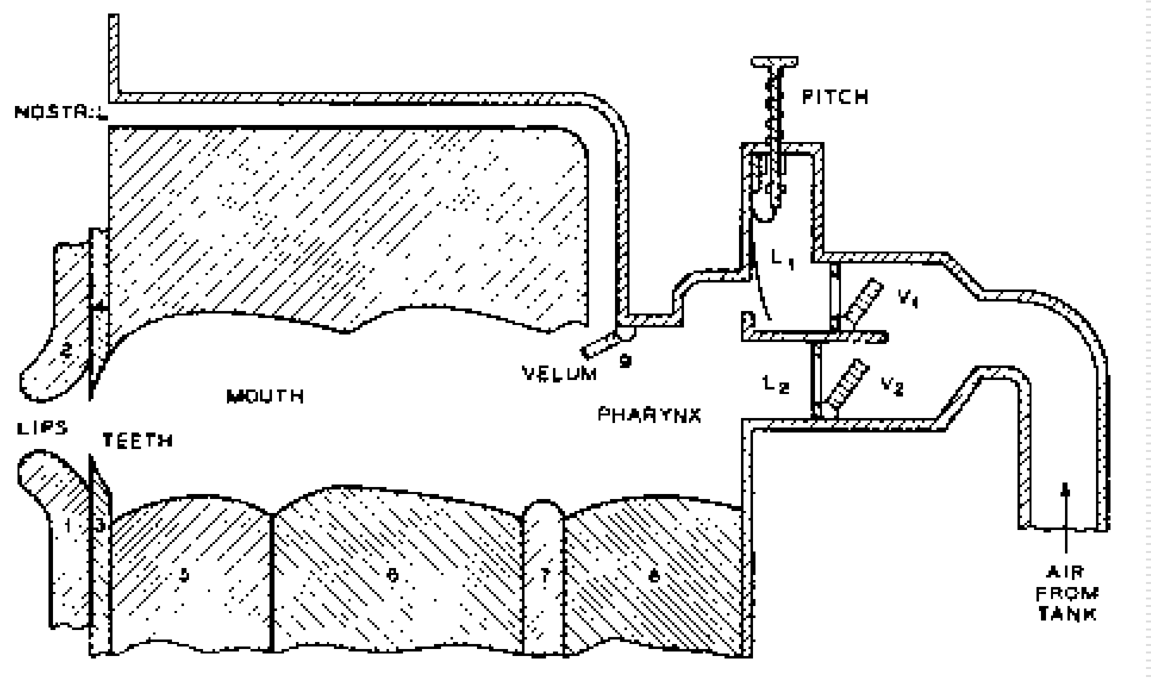
Start with acoustics, rules to create formants

Use databases of stored speech to assemble new utterances

Model movements of articulators and acoustics of the human vocal tract

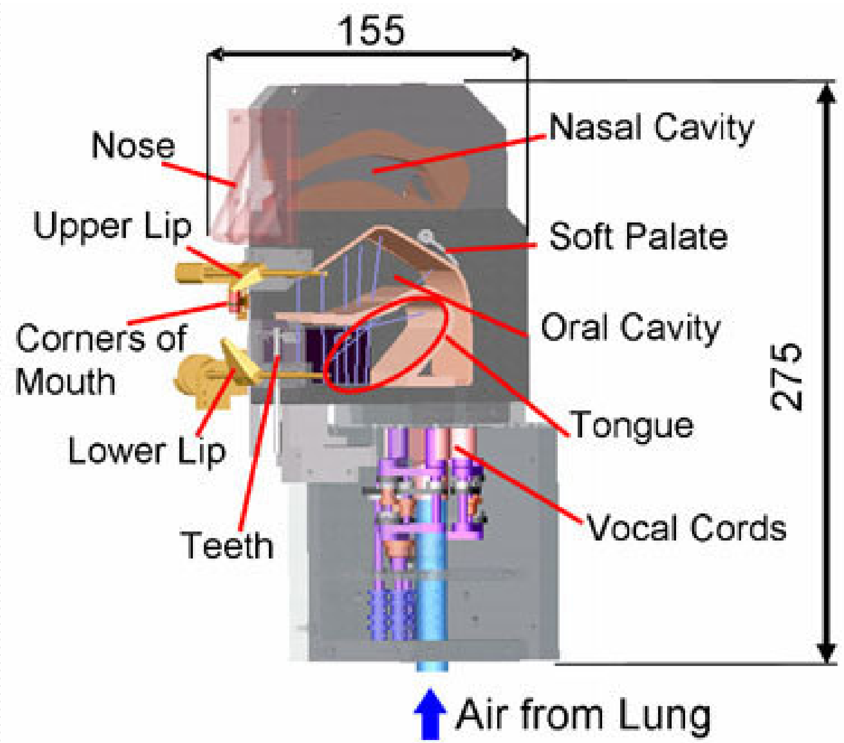
Articulatory synthesis

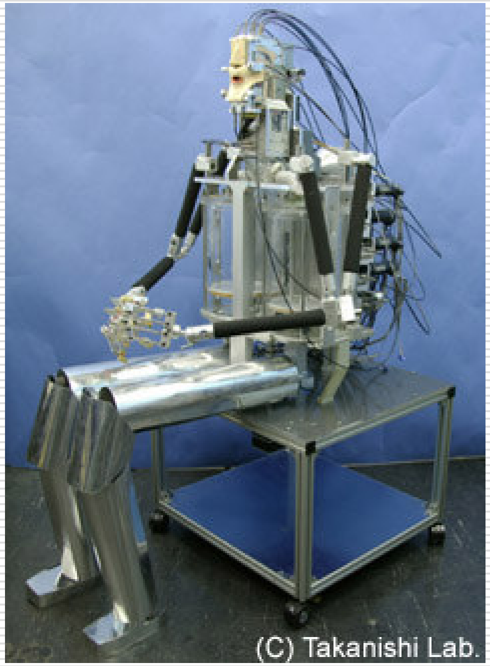
- based on physical or nowadays computational models of the human vocal tract and the articulation processes occurring there
- few of them currently sufficiently advanced or computationally efficient



Articulatory synthesis

Talking robot WT-4
Waseda University, Tokyo

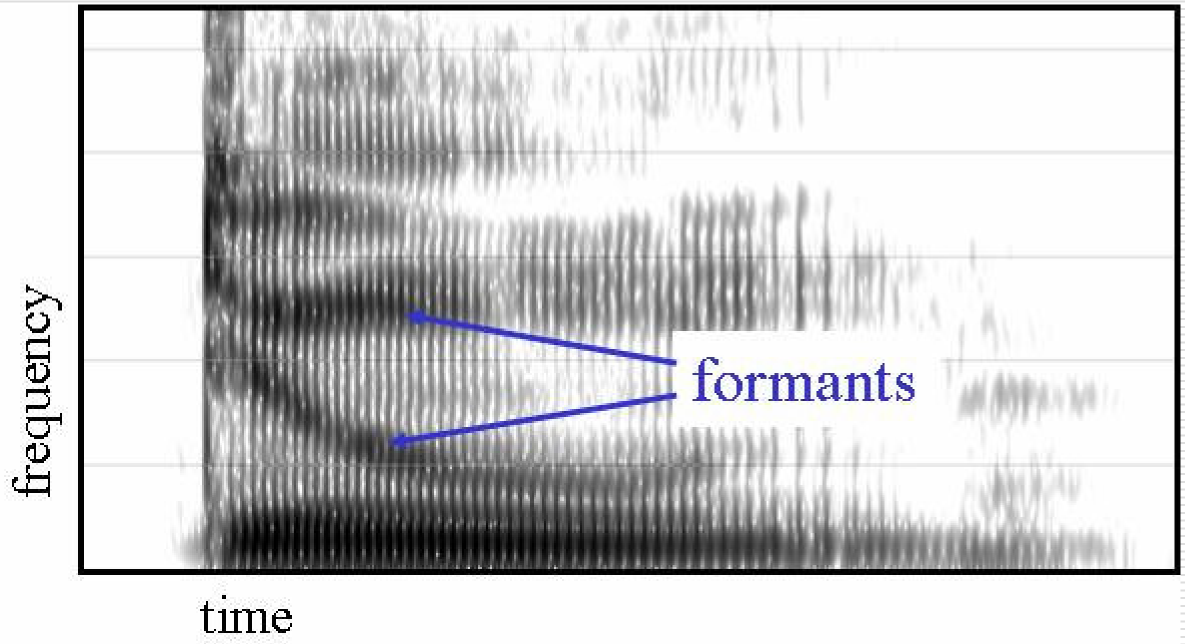




„sasisuseso"

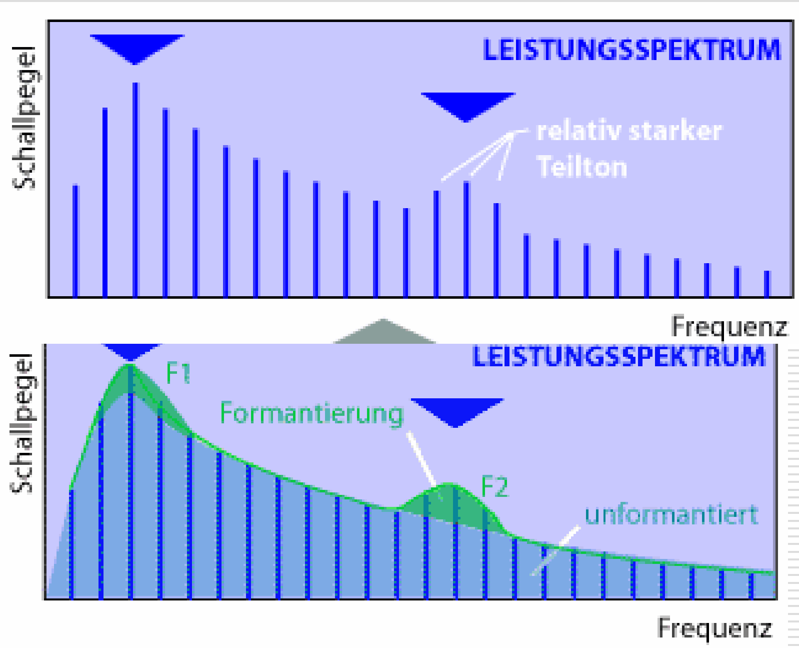
Formant synthesis

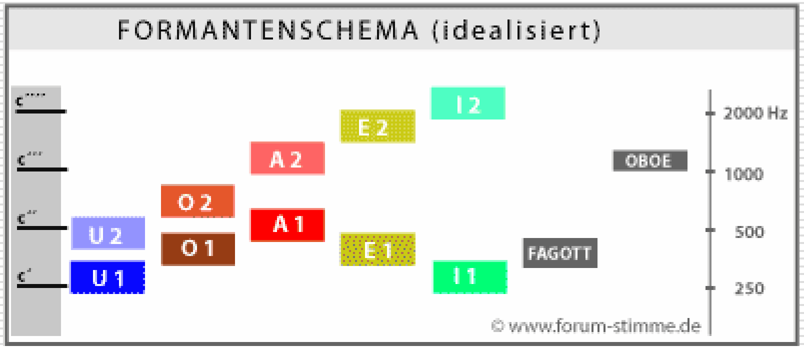
- *Formant*: Frequenzregion, in der die dort hineinfallenden Teiltöne (Obertöne) besonders stark sind
- Wesentlichen Elemente der Klangbildung, je nach Lage und Stärke verschiedene Vokale und Timbre

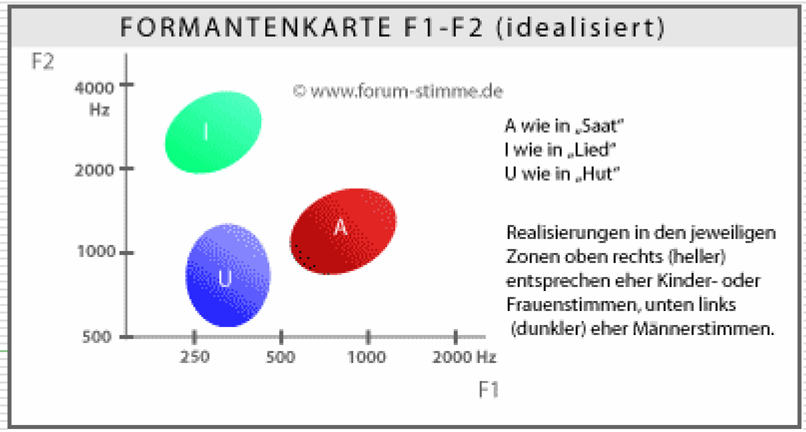


Formant Synthesis

- Annahme: Die für die menschliche Perzeption wesentliche Information ist durch die Töne in den Formanten kodiert
- Dabei prägen vor allem die beiden am tiefsten gelegenen Formanten (F1, F2) die Lautwahrnehmung, mitunter reicht zur Wahrnehmung bestimmter Vokale auch nur ein Hauptformant



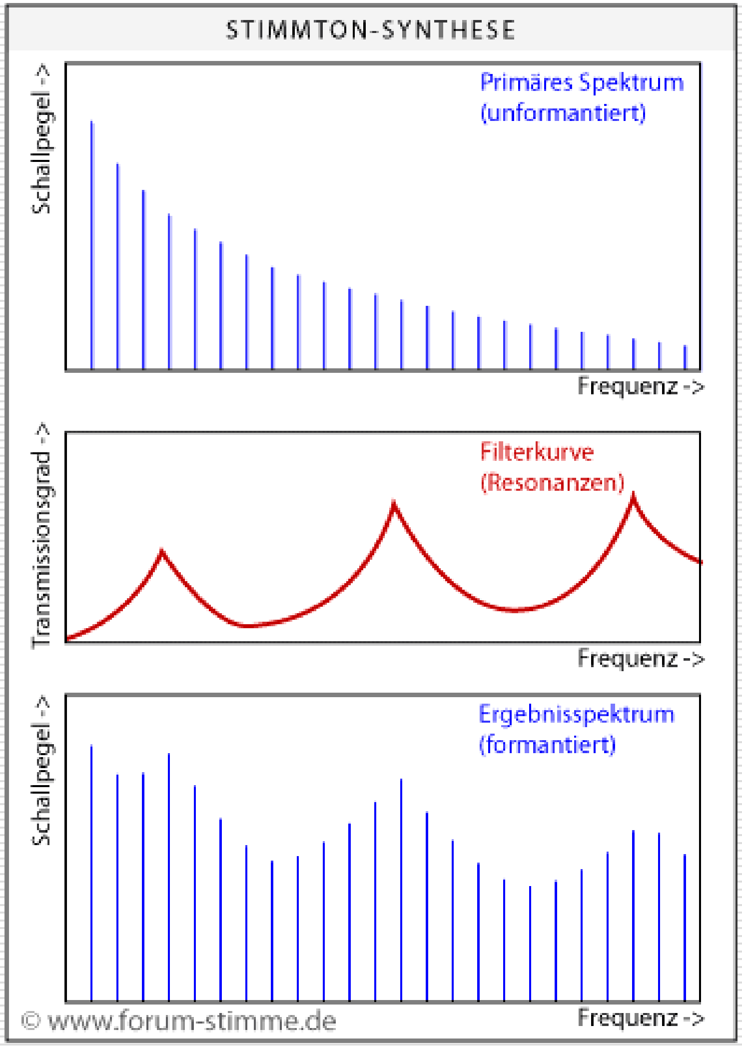




www.forum-stimme.de

Formant Synthesis

- Rules to model relations between tones and acoustic features
- Advantages
  - flexibilty
  - not much storage space needed
- Disadvantages
  - Sounds mechanical
  - Complicated rule sets
- Most common systems while computers were relatively underpowered
  - mitalk — 1979 MIT MITalk (Allen, Hunnicut, Klatt),
  - klatt — 1983 DECtalk system, 'Klatt synthesizer'



Data-based synthesis

- Now, all current commercial systems (1990’s-)
- Steps:
  1. Record basic inventory of sounds (offline)
  2. Retrieve sequence of units at run time (at run-time)
  3. Concatenate and adjust prosody (at run-time)

- What kind of units?
  - Minimize context contamination, capture *co-articulation*
  - Enable efficient search
  - Segmentation and concatenation problems
- How to join the units?
  - dumb (just stick'em together)
  - PSOLA (Pitch-Synchronous Overlap and Add), MBROLA (Multi-band overlap and add)

Einheitenlänge	Einheit	#Einheiten (Englisch)	#Regeln	Qualität
kurz	Allophone	60-80	hoch	gering
	Diphone	$<40^2-65^2$		
	Triphone	$<40^3-65^3$		
	Halbsilben	2K		
	Silben	11K		
	Doppelsilben	<11K2		
	Wort	100K-1.5M		
	Phrasen	∞		
lang	Satz	∞	gering	hoch

Source: E. Andre

Diphone synthesis

- Units = diphones
  - Phones are more stable in middle than at the edges
  - Diphones start half-way thru 1st phone and end half-way thru 2nd
- Typically 1500-2000 diphones, reduce number
  - *phonotactic constraints*: constraints on the way in which phonemes can be arranged to form syllables
  - collapse in cases of no co-articulation
- Record 1 speaker saying each diphone
  - "Normalized": monotonous, no emotions, constant volume

- Example: MBROLA (Dutoit & Leich, 1993) http://tcts.fpms.ac.be/synthesis/mbrola.html

MBROLA

- Input:

  Phonemes (SAMPA)

  Duration

  Pitchmarks = time (%)+F0

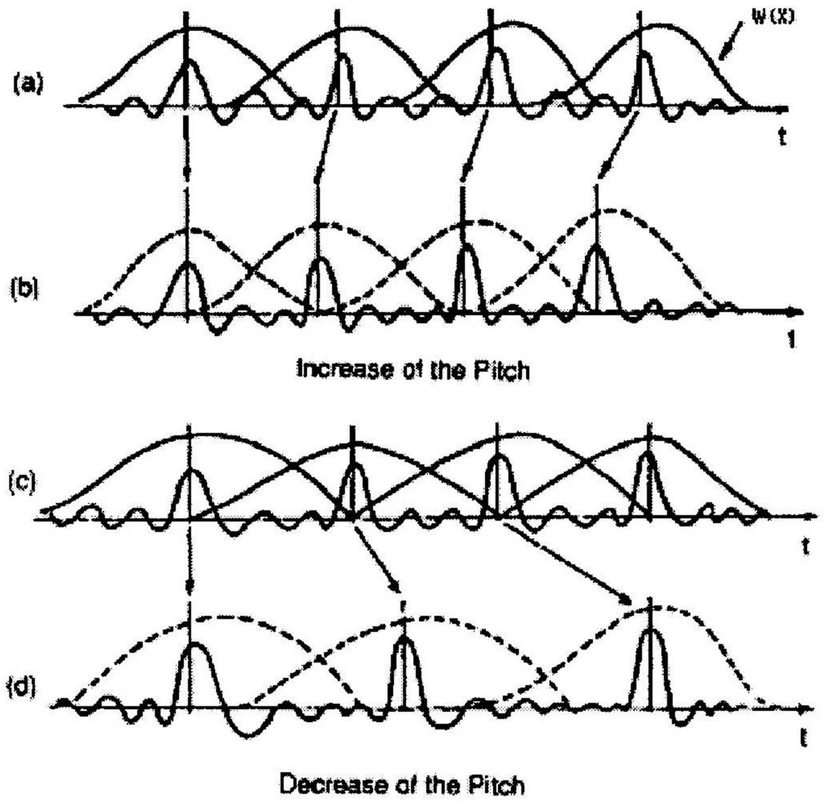
  Phonetic text:

  S 105 18 176 32 ...
  P 90 8 153
  a: 104 4 150 100 ...
  s 71 28 145 85 143

- No real synthesis:
  - Features are extracted from recorded units (diphones)
  - Manipulate features to smooth boundaries where units are concatenated
  - Change prosodic features through “re-synthesis”

MBR-PSOLA (in short, MBROLA)

- *Multiband resynthesis pitch-synchronous overlap & add*
  - Split up tones in frames centered around *pitchmarks*
  - Recombine frames at new set of pitchmarks, with varied distance (changes pitch) & number (duration)
  - *"like an old tape recorder with variable speed"*

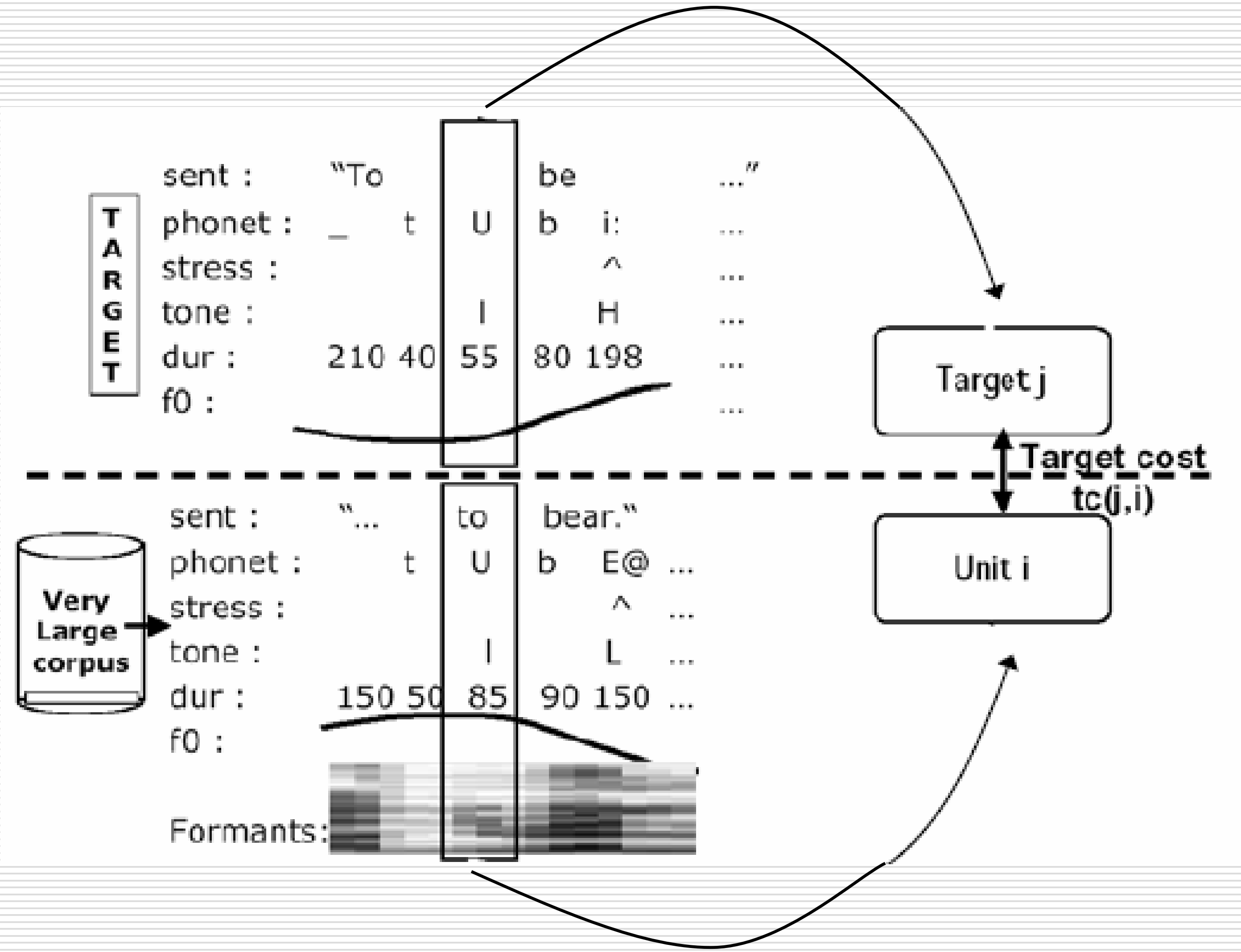


(Lemmetty, 1999)

Unit selection

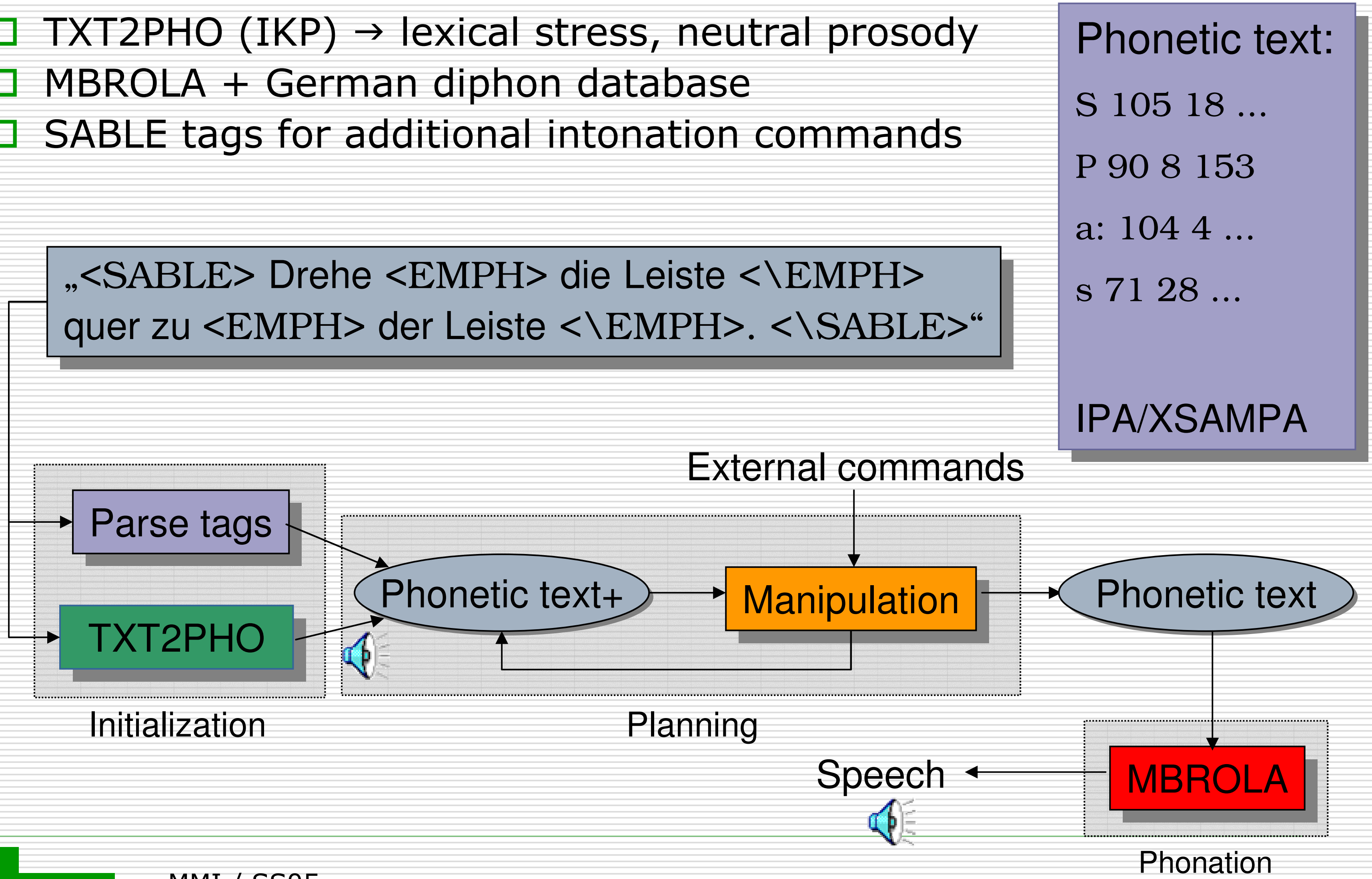
- One example of a diphone is not enough!
- Unit selection:
  - Record multiple copies of each unit with different pitches and durations
  - How to pick the right units? Search!
  - Example (Hunt & Black, 1996):
    - Input: three F0 values per phone
    - Database: phones+duration+3 pitch values
    - Cost-based selection algorithm (like Viterbi)
- Non-uniform unit selection
  - Units of variable length
  - Reduced need of automatic prosody modeling

Unit selection



Example: TTS for *Max*

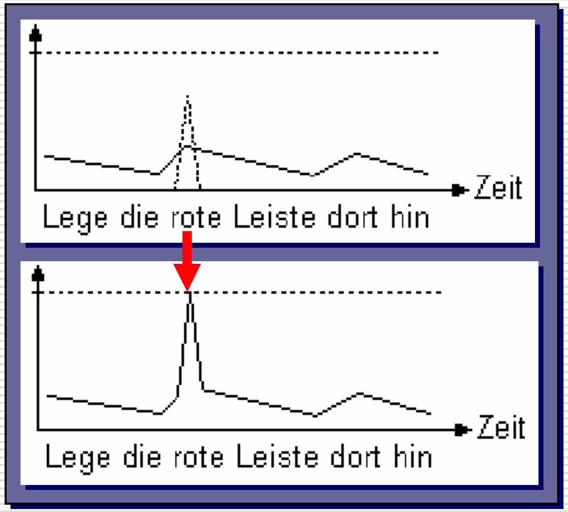
- TXT2PHO (IKP) → lexical stress, neutral prosody
- MBROLA + German diphon database
- SABLE tags for additional intonation commands

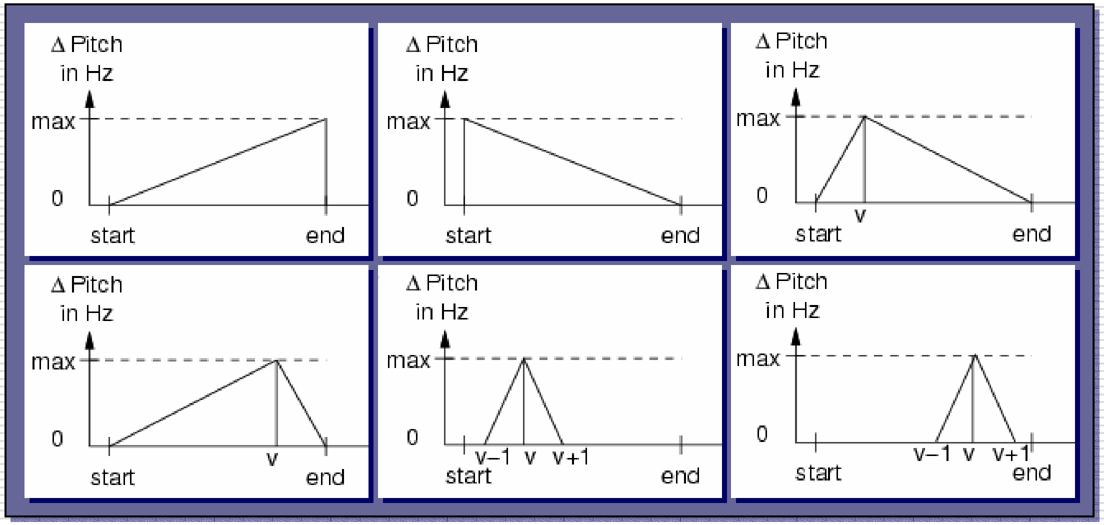


Example: TTS for *Max*

Manipulation of phonetic text

- Overlay stereotyped contours to create accents
- No segmental analysis
- Flexible form, height, duration



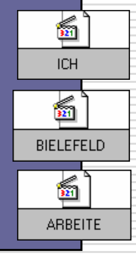


Beispiel: Kontrastierung

Wer arbeitet in Bielefeld?

Wo arbeitest du?

Was tust du in Bielefeld?



Commercial TTS systems - demos

BabelTech Babil	Diphone concat., MBROLA-like	Mp3 (2000)
AT&T	non-uniform unit-selection	Mp3 (1998)
BabelTech BrightSpeech	non-uniform unit-selection	Mp3 (2003)
IBM ctts	non-uniform unit-selection	Mp3 (2002)
Loquendo	non-uniform unit-selection	Mp3 (2003)
Nuance	non-uniform unit-selection	Mp3 (2001)
SVox	Diphone concat.	Mp3 (2000)

Academic TTS systems - demos

BOSS (IKP, Bonn)	non-uniform unit-selection	Mp3 (2001)
IMS Stuttgart	Diphone concat., Festival+MBROLA	Mp3 (2000)
Infovox	Formant synthesis	Mp3 (1994)
Mary (DFKI)	Diphone synthesis, MBROLA	Mp3 (2000)
VieCtoS (ÖFAI, Wien)	Halbsilben, schlechte Tobi-Labelung	Mp3 (1998)
SVox (ETH Zürich)	Diphone concat.,	Mp3 (1998)
HADIFIX (IKP, Bonn)	HAlbsilben, DIphone und sufFIXe	Mp3 (1995)

- Comparison of state-of-the-art TTS systems
  http://ttssamples.syntheticspeech.de/deutsch/index.html
- Janet Cahn‘s Master Thesis, PhD Thesis
  http://xenia.media.mit.edu/~cahn/
- Demos and links for speech synthesizers
  http://felix.syntheticspeech.de/
- Lecture on speech synthesis of Bernd Möbius
  http://www.ims.uni-stuttgart.de/~moebius/teaching.shtml