Multimedia Department

Introduction to text-to-speech synthesis

Krzysztof Marasek

Motivation

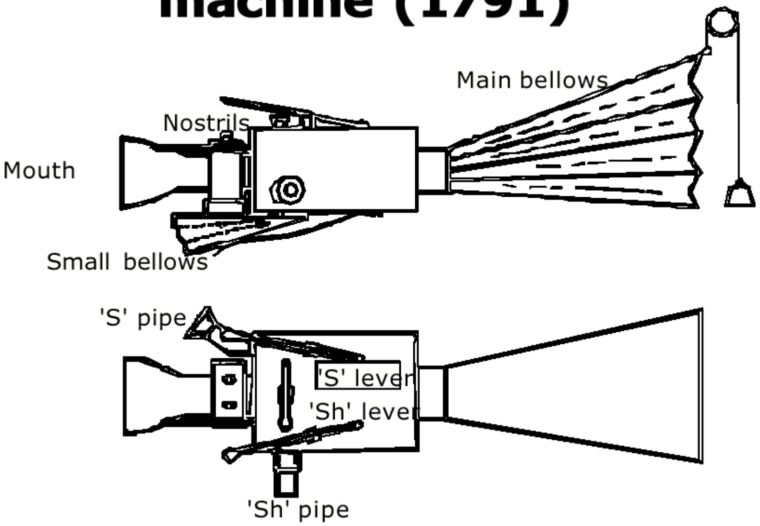
- Increasing and growing popularity of *interactive voice response* (IVR) systems makes the use of *text-to-speech* (TTS) systems more appealing
- *unified messaging systems* (UMS) make use of oral access to any written information such as fax, e-mail, textual databases
- growing demand from *dialog systems*, including robots and agents; dialog systems use natural speech input and user expects the answer naturally sounding response, too
- voice access to databases (price list, events)
- read aloud systems for people at work or visually impaired

Lecture outline

- speech synthesis systems
- criteria of quality evaluation
- methods of speech signal generation
- prosody description
- elements of text-to-speech systems
- Polish speech synthesis
- future of TTS (how far we are to HAL2001?)

At the beginning...

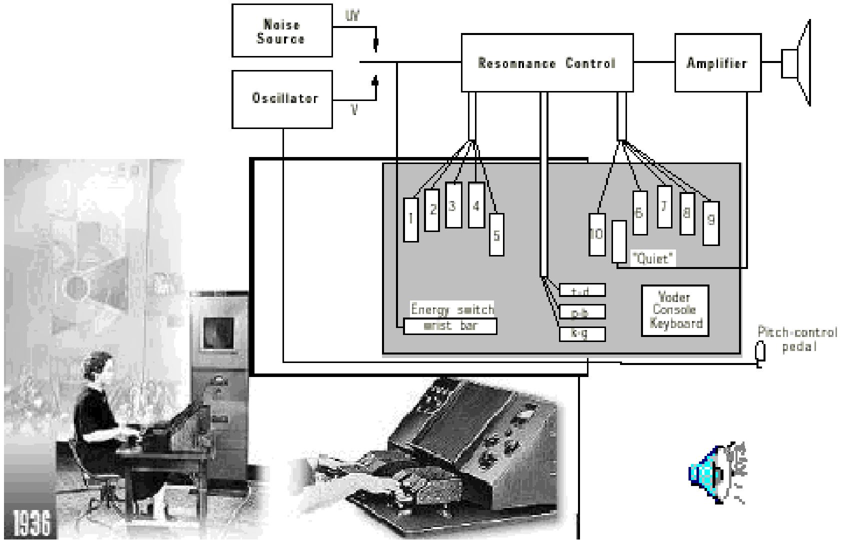
Von Kempelen's talking machine (1791)



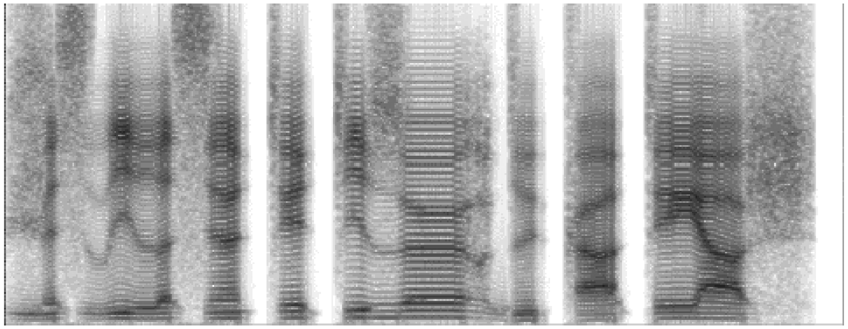
DAVO articulatory synthesis

Comparison of original and synthesized phrase

Omer Dudley's Voder (Bell Labs, 1936)

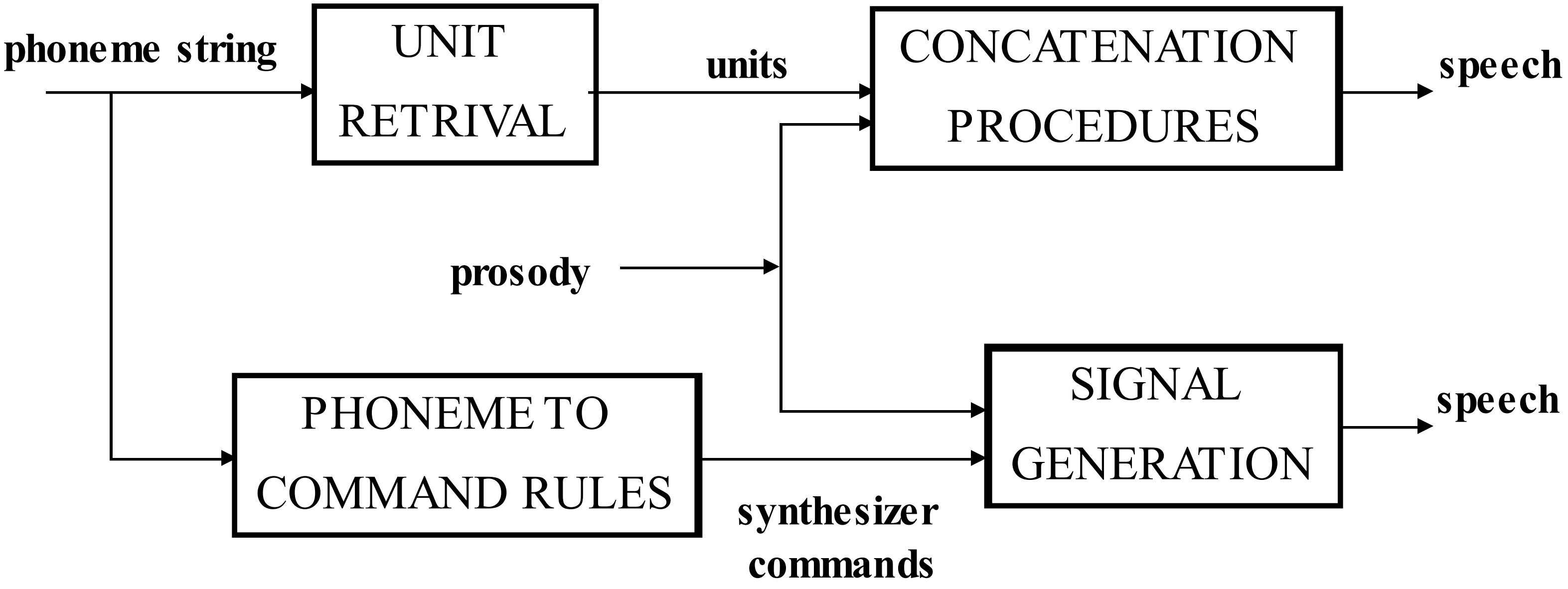


John Holmes' formant synthesizer (1964)



Speech synthesis systems

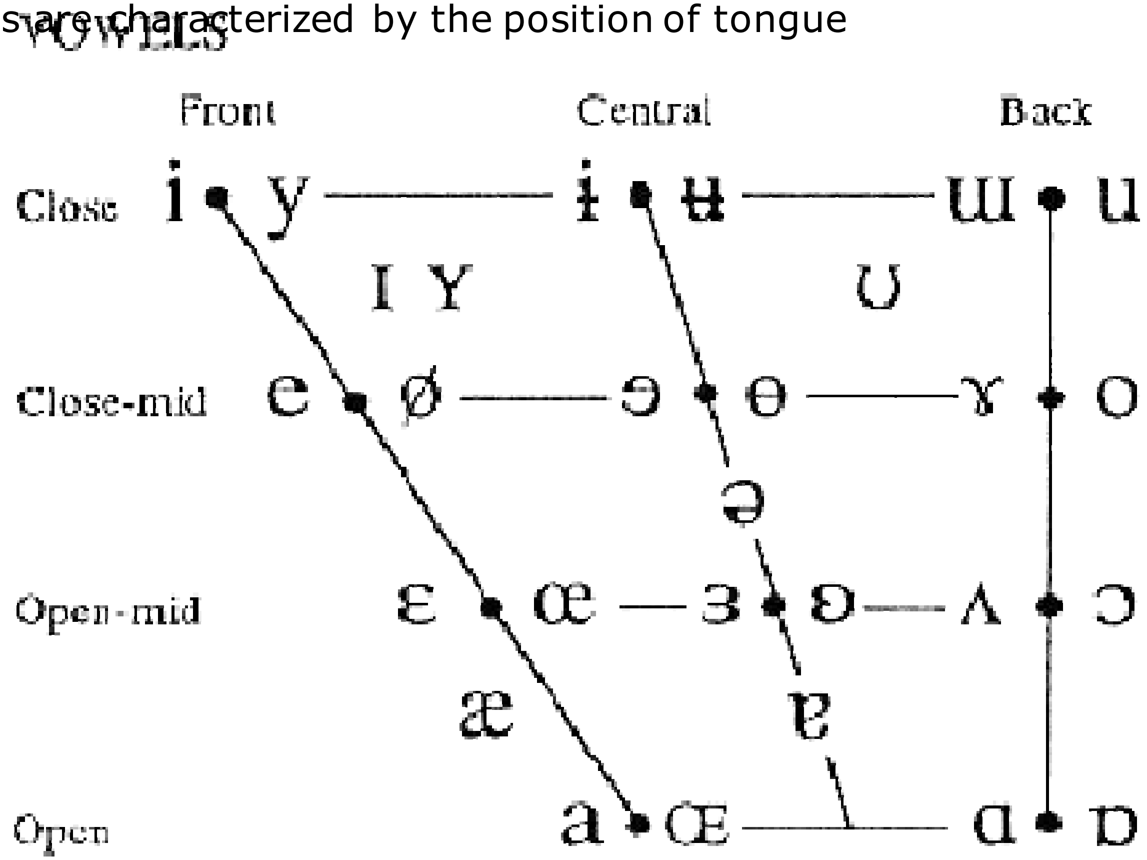
- *are those which can convert a string of phonemes and pauses into a speech signal*
- two schemes



Grapheme-to-phoneme conversion I

- The **phone** is the smallest sound element, which can be segmented. It represents the typical kind of sound and sound nuance for a certain sound. Sounds (phones), which are phonetically similar, belong to the same phoneme.

- Vowels are characterized by the position of tongue

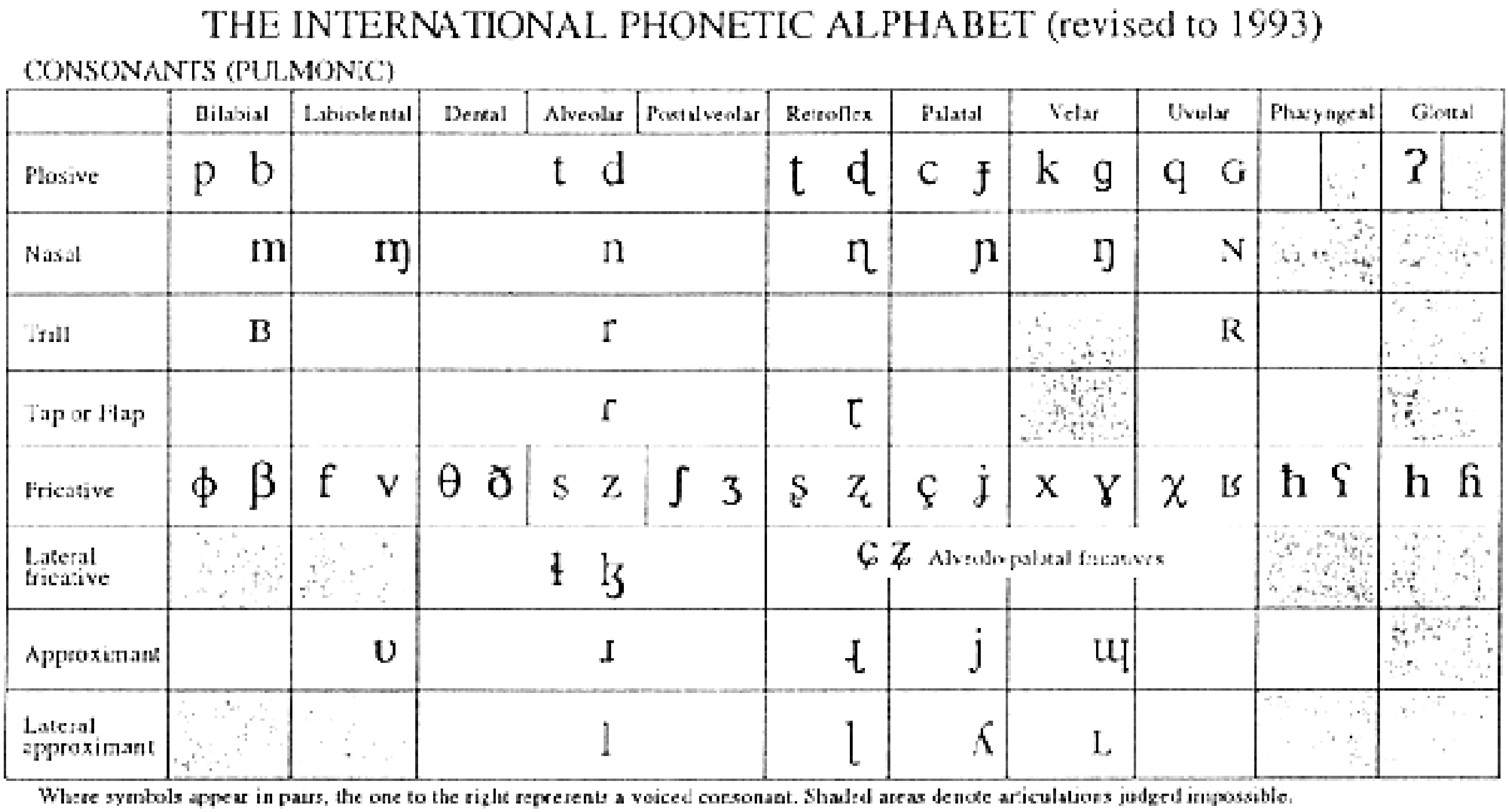


Where symbols appear in pairs, the one to the right represents a rounded vowel.

Grapheme-to-phoneme conversion II

Multimedia Department

- Consonants: placement of constriction



THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993)

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ	▨	ʔ ▨
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ	▨	▨
Trill	ʙ			r				▨	ʀ		▨
Tap or Flap				ɾ		ɽ		▨			▨
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative	▨	▨		ɬ ɮ		ɕ ʑ Alveolo-palatal fricatives				▨	▨
Approximant		ʋ		ɹ		ɻ	j	ɰ			▨
Lateral approximant	▨	▨		l		ɭ	ʎ	ʟ		▨	▨

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

Grapheme-to-phoneme conversion III

- ⌘ Lexicon approach: for all words the phonetic transcription, word stress, syllables and morphological information are given
  - ☒ hand-corrected
  - ☒ many items, expensive
- ⌘ Letter-to-sound conversion:
  - rules
  - decision trees
  - ANNs
  - fast & easy, but lower accuracy

Lexicon

Kalisz	k a l i S
Kamienna	k a m j e n n a
Kaszuby	k a S u b I
Katowice	k a t o v i ts e
Kazimierz	k a zi i m j e Z
Kielce	kj e l ts e
Klakson	k l a k s o n
Kolor	k o l o r
Konopnickiej	k o n o p ni i ts k j e j
Konstytucji	k o n s t I t u ts j i
Koszalin	k o S a l i n
Kościuszki	k o si tsi u S k i
Krakowska	k r a k o f s k a
Krakowsko	k r a k o f s k o
Kraków	k r a k u f
Krzyki	k S I k i
Kujaw	k u j a f
Kutno	k u t n o

g2p

norwescy i rosyjscy płetwonurkowie usiłują się dostać do wnętrza rosyjskiego okrętu podwodnego.
n o r v e s ts I i r o s I j s ts I p w e t f o n u r k o v j e u s' i w u j o~ s' e~ d o s t a ts'
d o v n e n t S a r o s I j s k j e g o o k r e n t u p o d v o d n e g o _sil_

Quality criteria for synthesis evaluation

- Intelligibility
  - words
  - sentences
- Naturalness
  - modeling quality (units)
  - prosody, accentuation, hesitations, etc.
- Fluidity
- Prosody matching
  - phrases
  - continuations, questions
- Auditive tests - complicated, psychological aspects, multi-dimensional scaling

Taxonomy of speech synthesis systems

- synthesis method
  - rule-based:
    - formant synthesis
    - articulatory synthesis
  - concatenation of units
    - monophone
    - diphone
    - poly-phone, semi-syllables
    - micro-segmental
    - unit selection
- concatenation technique
  - TD-PSOLA, FD-PSOLA
- coding of speech units
  - LPC, hybrid harmonic/stochastic, sinusoidal model, etc.
- Mono- or multi-lingual
- Footprint
  - small, for embedded application
  - big, stand-alone application

Rule-based: formant synthesis

- Human-expert formulate the rules of sound generation based on the inspection of the database
- Digital filters used to model the behavior of vocal tract
  - excitation signal
  - formant frequencies and bandwidths
  - durations
  - up to 60 parameters
- wide-spread use in study of characteristic of natural speech
- can be used not only for speech 🔊

🔊 Fant's formant synthesizer 1953

DEC talk 🔊

🔊

Dazy, Haskins 1951 🔊

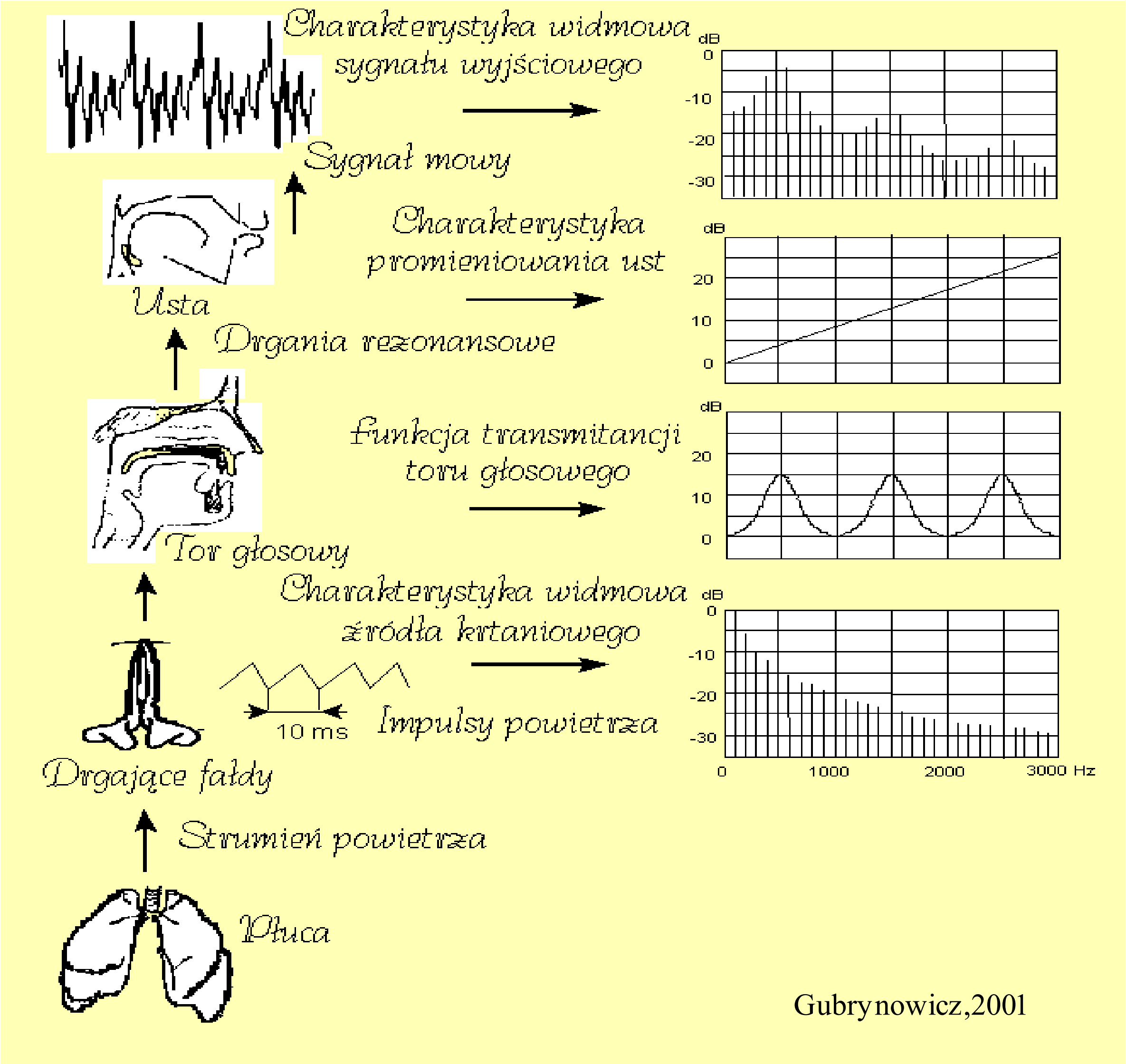
Multimedia Department

Rule-based: ariticulatory synthesis

- ⌘ Full model of human sound generation

Flanagan

Haskins Lab.



Gubrynowicz,2001

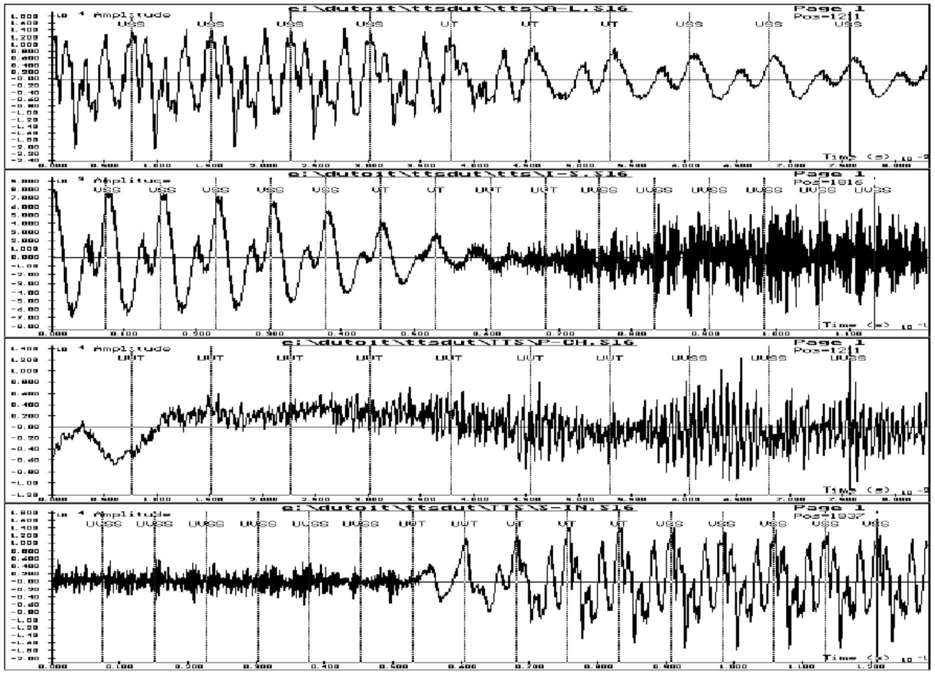
Concatenation synthesis

- Existing speech synthesis systems use different sound elements. The most common are:
    - **phones**
    - **diphones** — Olive 1976.
    - **phone clusters**
    - **half syllables** — Browman 1980.
    - **syllables**
- **Phone-based**: too few segments, low intelligibility
- no account for coarticulation
- **Phone clusters** are sequences of vowels or consonants. According to the position of the sound sequences phone clusters are splitted into initial, medial and final cluster.
- **micro-segmental synthesis**: http://www.webspeech.de/index1.php
    - over 600 context-dependent units, concatenation by rule
    - very small footprint (**less than 1 MB**)

Diphone - what is this?

- A ***diphone*** begins at the second half of a phone (stationary area) and ends at the first half of the next phone (stationary area). Thus, a diphone always contains a sound transition. Diphones are very suitable as sound elements for speech synthesis. Compared with phones, a segmentation is simpler. The time duration of diphones is longer and the segment boundaries are easier to detect.
- Size of diphone database:

  6-20 MB
- quality of synthesis
- depends on quality of dbase



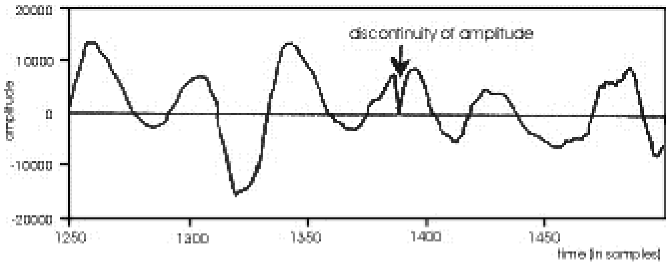
Multimedia Department

K.Marasek
05.07.2005

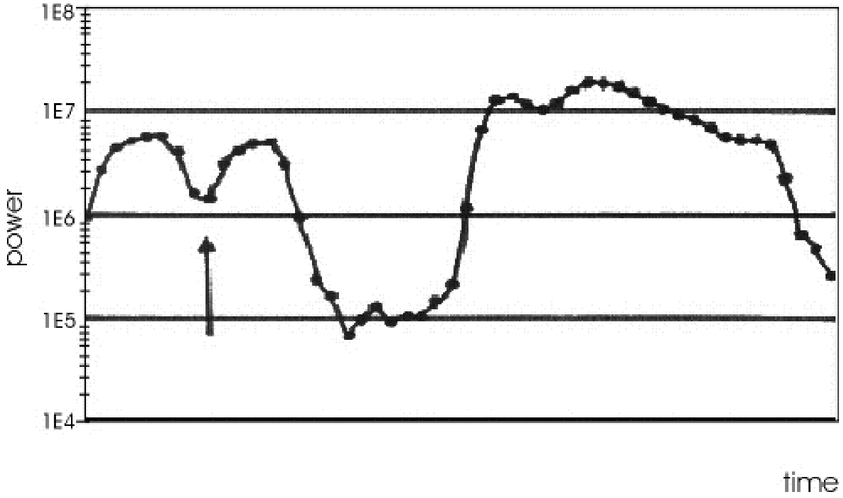
Polish-Japanese Institute of Information Technology

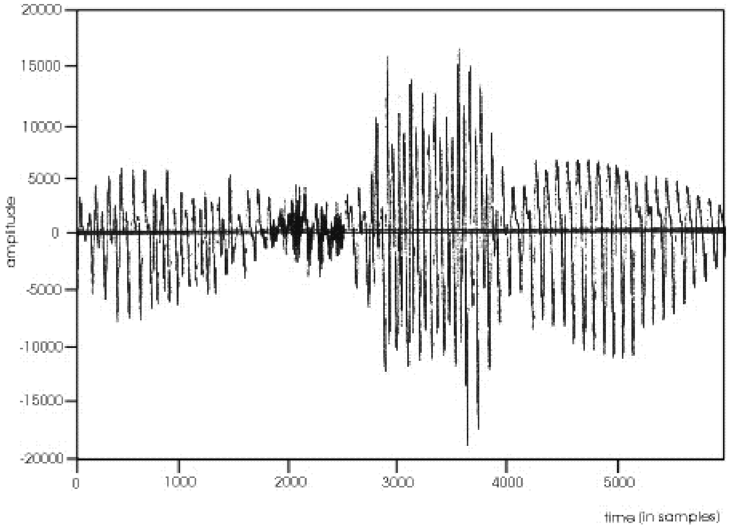
Problems of signal segmentation I

Discontinuity of the amplitude

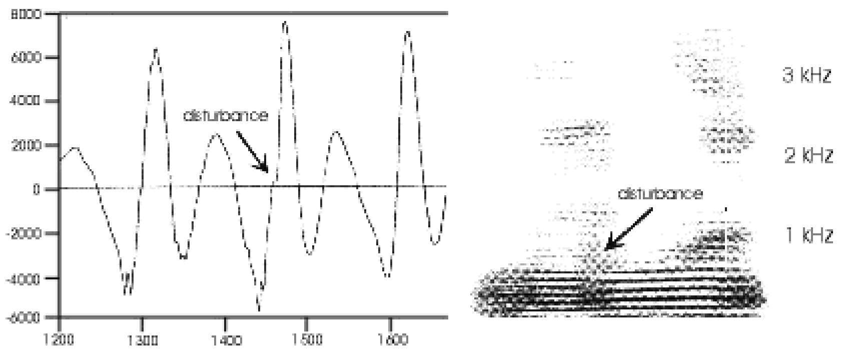


Discontinuity of the energy (time domain)

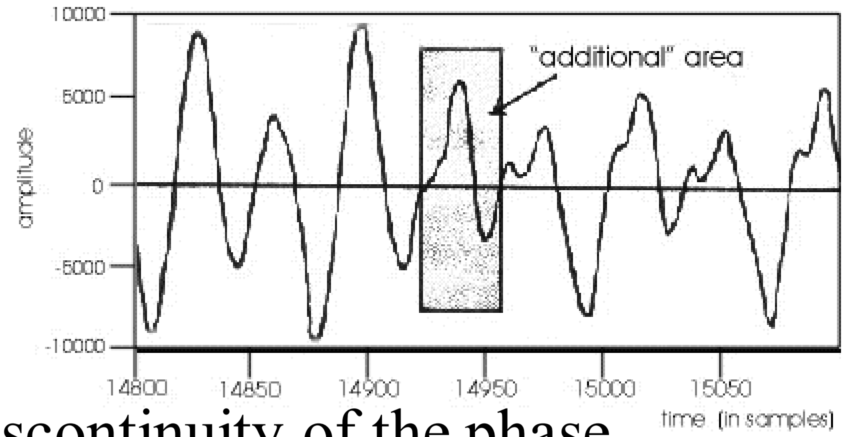




Problems of signal segmentation II



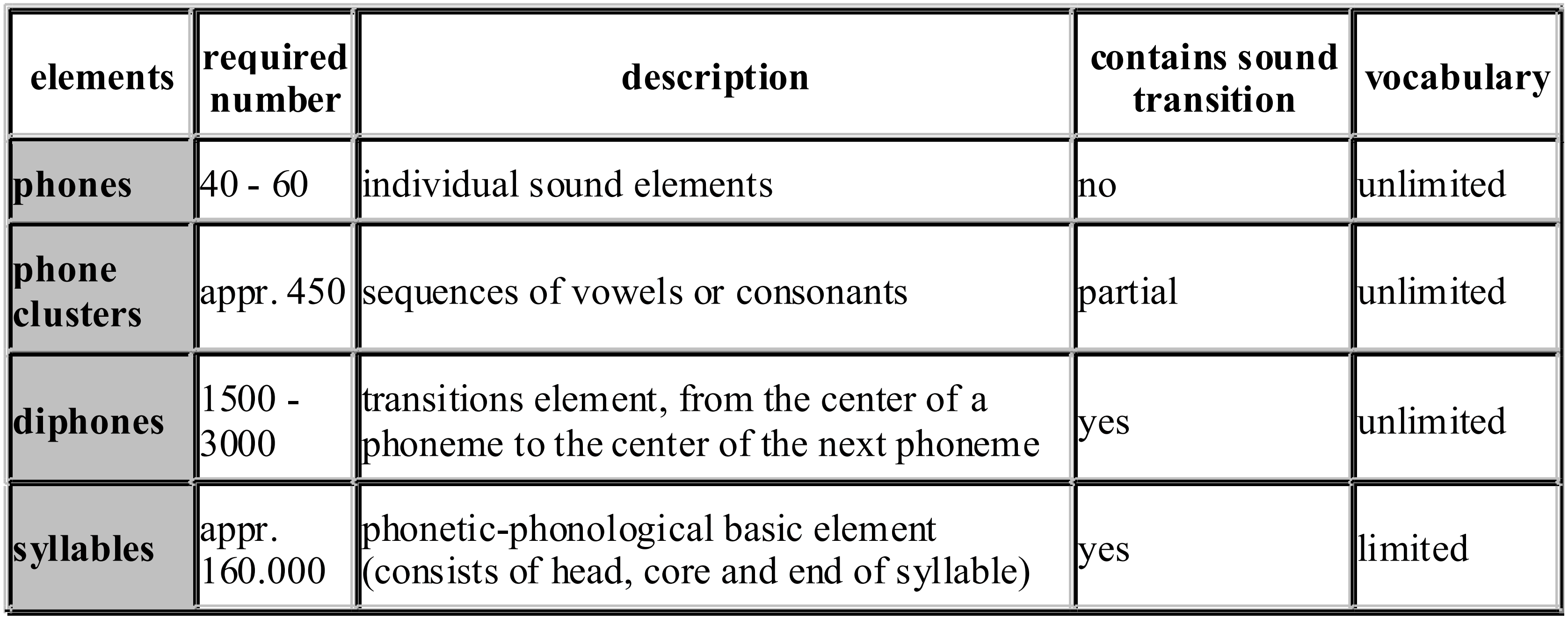
Discontinuity of the frequency (time and frequency domain)



Discontinuity of the phase

These disturbances can be widely reduced or avoided by a careful segmentation

Units of synthesis: summary

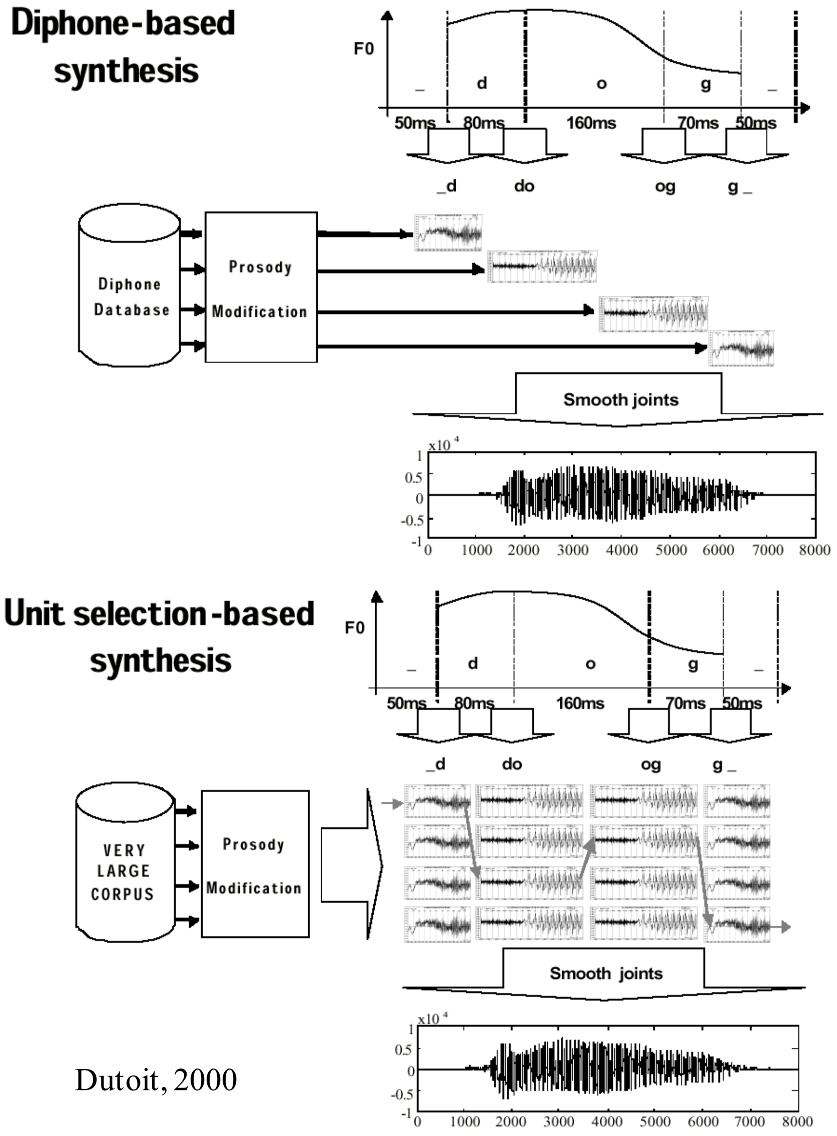


elements	required number	description	contains sound transition	vocabulary
phones	40 - 60	individual sound elements	no	unlimited
phone clusters	appr. 450	sequences of vowels or consonants	partial	unlimited
diphones	1500 - 3000	transitions element, from the center of a phoneme to the center of the next phoneme	yes	unlimited
syllables	appr. 160.000	phonetic-phonological basic element (consists of head, core and end of syllable)	yes	limited

Examples of diphone synthesis

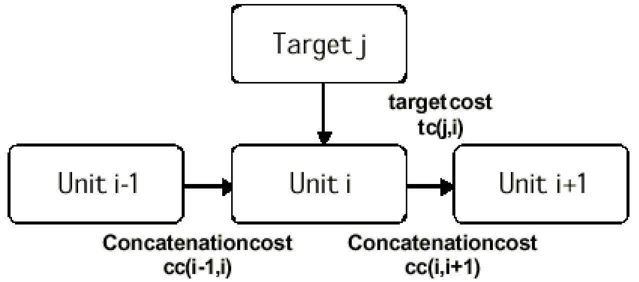
- MBROLA Czech 🔊
- MBROLA Telugu 🔊
- MBROLA German 🔊
- Diphone German 🔊
- Diphone English 🔊 🔊
- Diphone Polish 🔊

Unit-selection



Dutoit, 2000

How to get the best sequence of units for a given utterance? **Viterbi search**



Automatic unit selection

Costs ? Open problem

- Concatenation cost ?
- Target cost ?
- Weights? Trained by resynthesizing the corpus and trying to minimize the difference between original and synthetic

⌘ Target: to pass Turing test

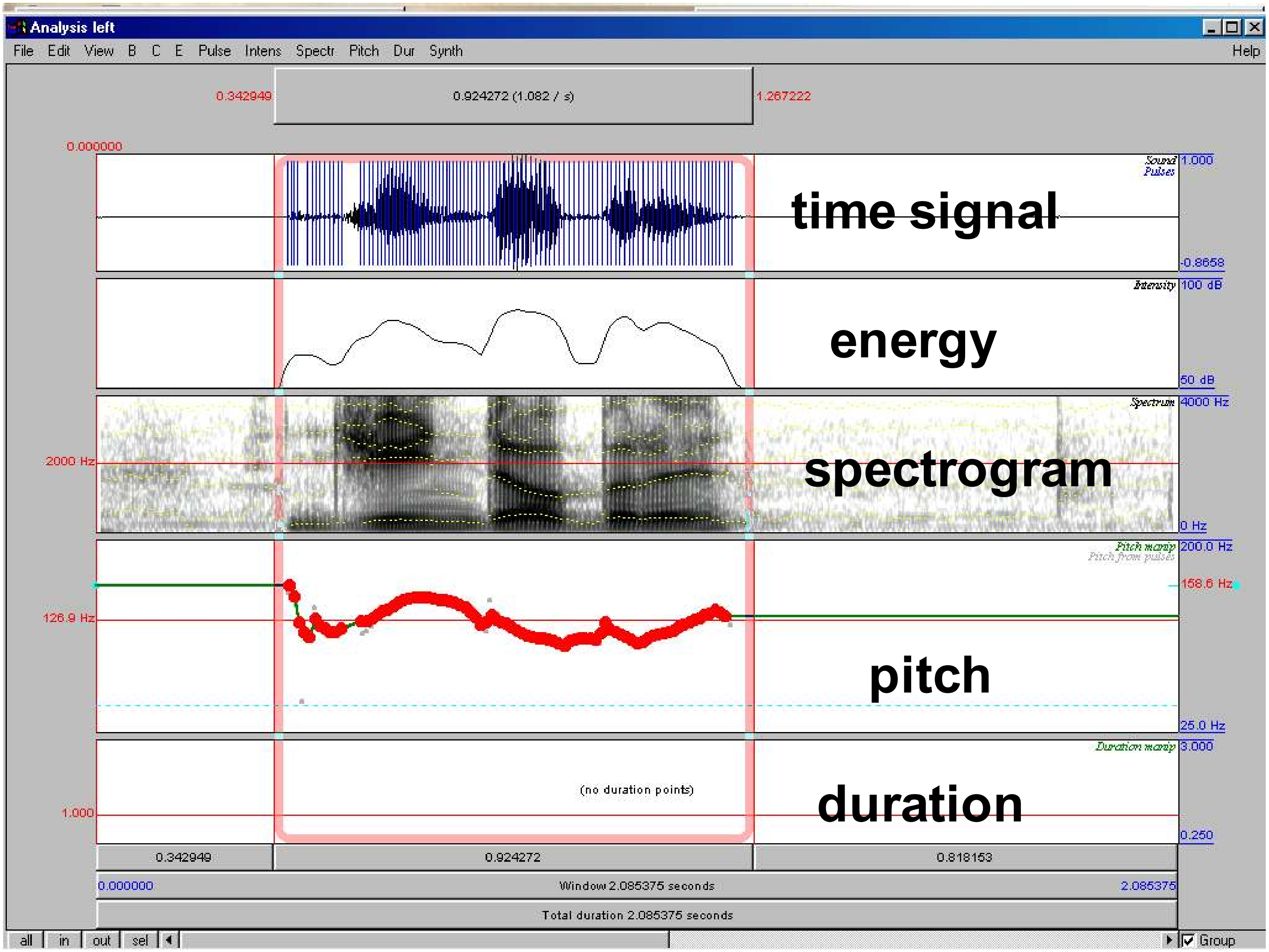
ENG JAP GER

How to make synthesized sound natural?

- Account for coarticulation:
  - derive an optimized set of segments from speech database-> unit selection
  - corpus-based models of speech segments - acoustic data for a given segment in a given context
  - unifying rule-based and concatenation synthesis
- Add prosody
  - make certain fragments more prominent, stressed
- Add variability
  - not random, but changes of articulation, prosody, speaking rate
- Add personality
  - speaking style

  - http://www.ims.uni-stuttgart.de/%7Emoehler/synthspeech/

Description of speech signal



Dimensions of prosody

- Word stress and sentence intonation
  - each word has at least one syllable which is spoken with higher prominence
  - in each phrase the stressed syllable can be accented depending on the semantics and syntax of the phrase
- Stress can be manifested by changed:
  - Pitch
  - loudness
  - rhythm (duration)
- Prosody relies on each and every level on linguistic competence of the reader
  - syntax mainly
  - semantics
  - pragmatics: personal reflection of the reader

Models of intonation: Pitch contour

- **Tonetics** (the British school)
  - tone groups composed of syllables {unstressed, stresed, acented or nuclear}.
  - nuclear syllables have nuclear tones {fall, rise, fall-rise, rise-fall}
- **ToBI** (Tones and Break Indices)
  - Intonational phrases splited into intermediate phrases composed of syllables.
  - Relative tone levels: high (H) or low (L) (plus diacritics) at every intonational or intermediate phrase boundary (%) and on every accented syllable
- **SAMPROSA** (SAM PROsodic Alphabet)
- **stylization method** (prosodic pattern measured from natural speech)

Haskins Lab.
🔊

Klatt – phonological rules
for sentences
🔊

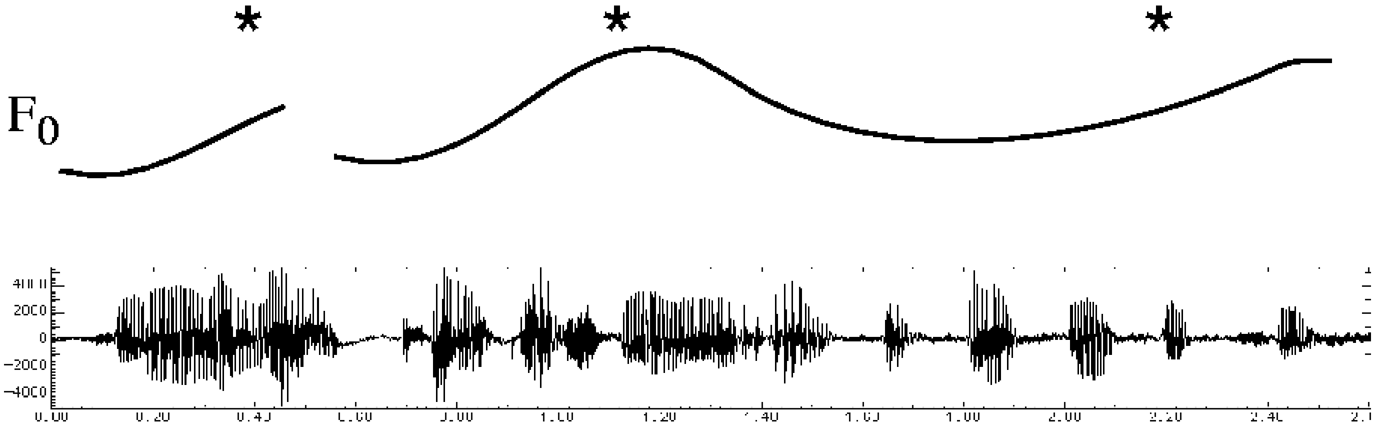
Intonation example

Hr. Müller, kommt er schon um 11:45 h?

Herr Müller$_{\text{Komma}}$ kommt er schon um elf Uhr fünfundvierzig$_{\text{Fragezeichen}}$

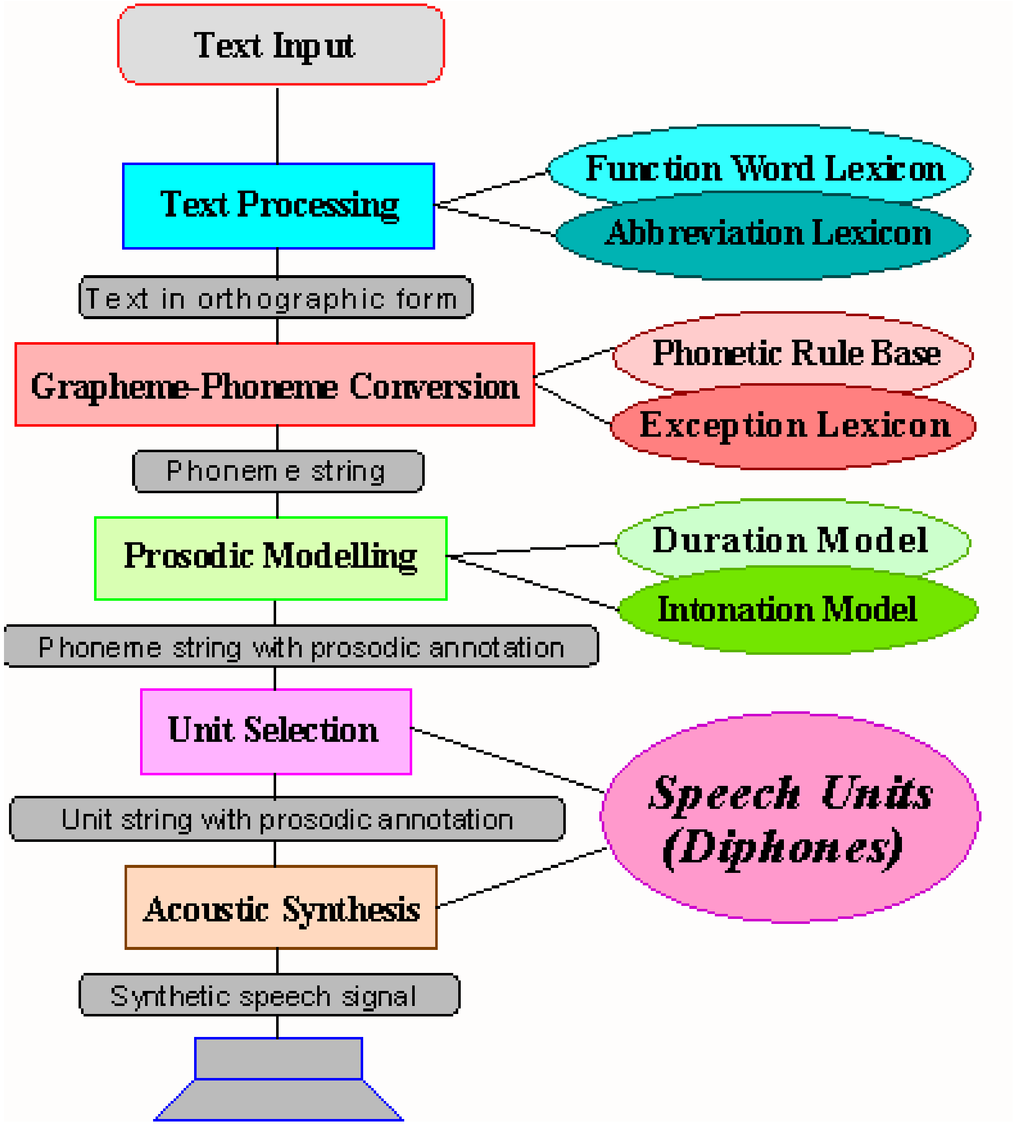
hER mYl6 kOmt e:6 So:n ?Um ?Elftsu: fYnf?UntfIRtsIC

[[_hER mYl6] [kOmt e:6 So:n ?Um ?Elftsu: fYnf?UntfIRtsIC_]]



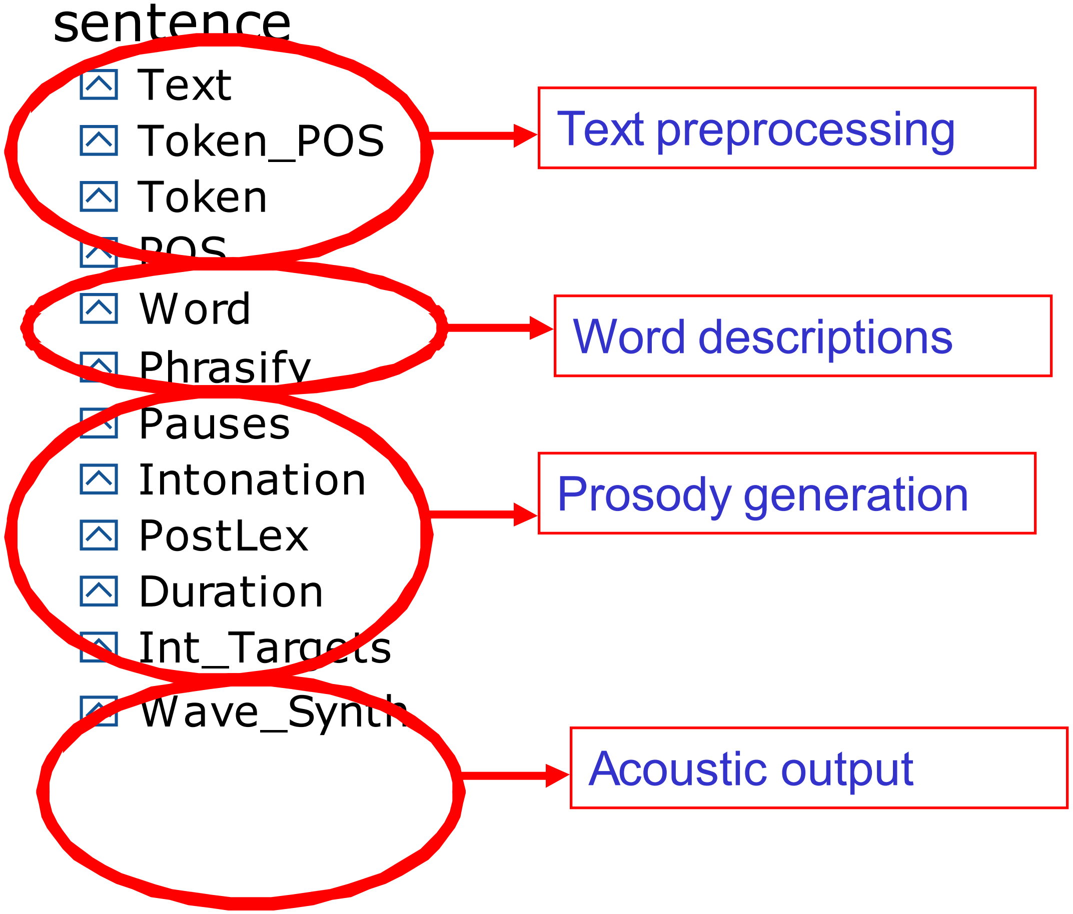
Moeller, 2000

Text-to-speech synthesis



Festival Text-To-Speech

- Festival Speech Synthesis - steps to synthesize a sentence



- Text
- Token_POS
- Token
- POS
- Word
- Phrasify
- Pauses
- Intonation
- PostLex
- Duration
- Int_Targets
- Wave_Synth

Text preprocessing

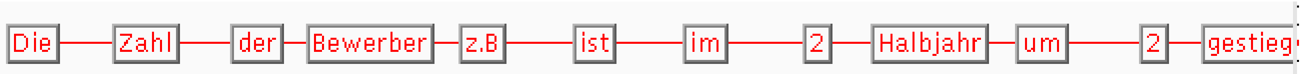
Word descriptions

Prosody generation

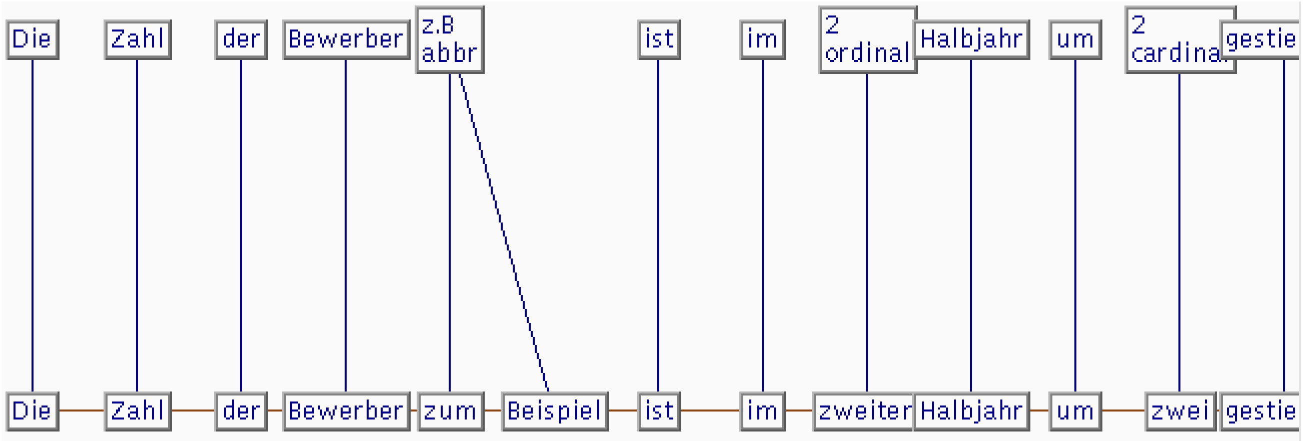
Acoustic output

Multimedia Department

Text preprocessing I

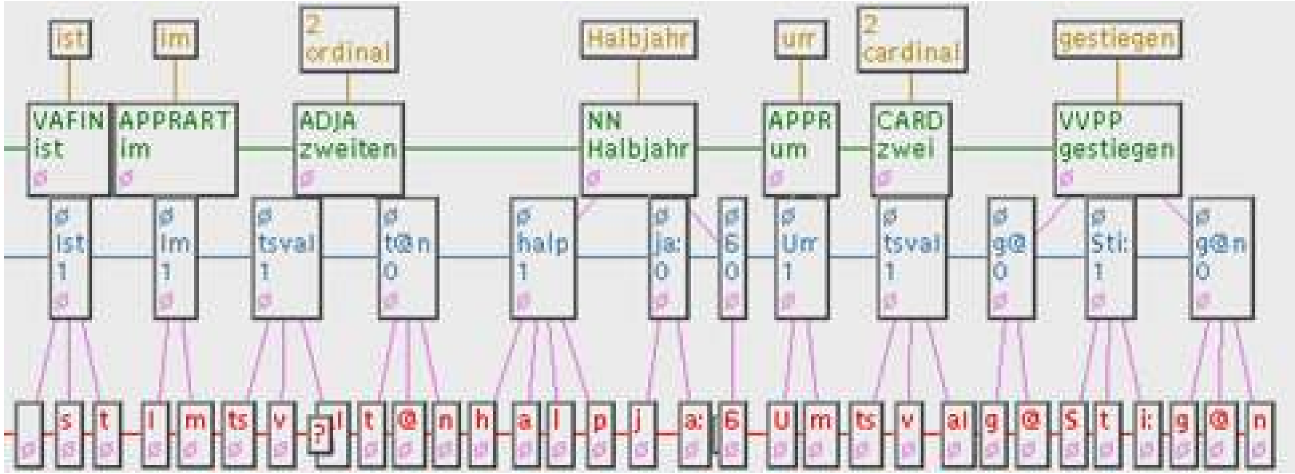


- Text
  - splits the input into a sequence of tokens by separating the input where white space occur, deletes word-final punctuation marks
- Token
  - abbreviation recognition and expansion, determination of a token type (e.g. ordinal vs. cardinal number)



Text preprocessing II

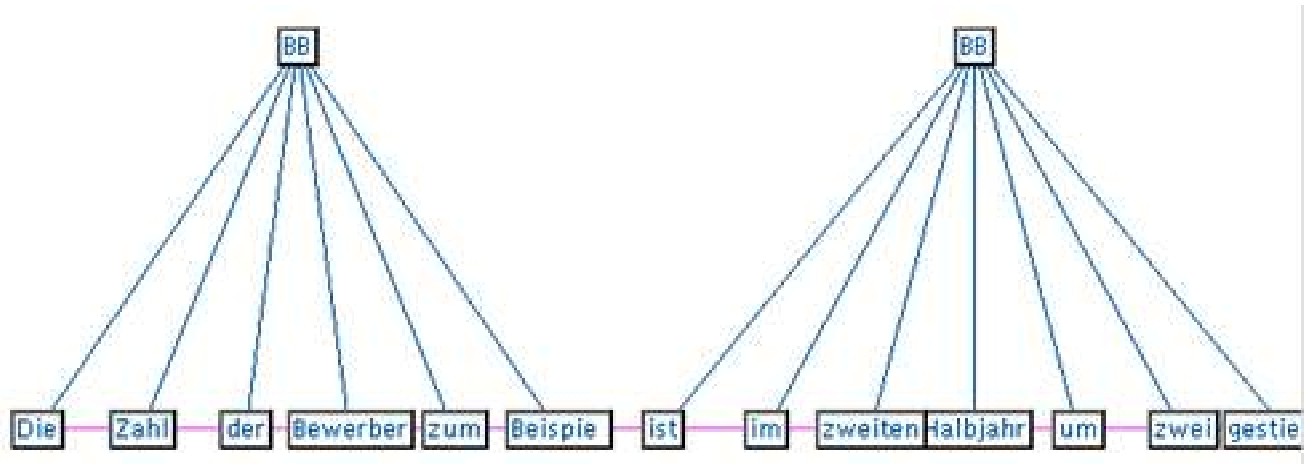
- POS
  - **part of speech tagger** determines the word class of each word
  - some word classes are usually accented, some not
  - the more detailed word classes, the more linguistic analysis is possible and the better intonation will be
- Word
  - morphological analysis
  - uses a lexicon to look up the phonetic transcription, the syllable structure and the word stress for each word



Phrasing

- The module determines where phrase boundaries occur
  - insert pauses on phrase boundaries
  - determined by CART tree trained on big data corpus

Multimedia Department

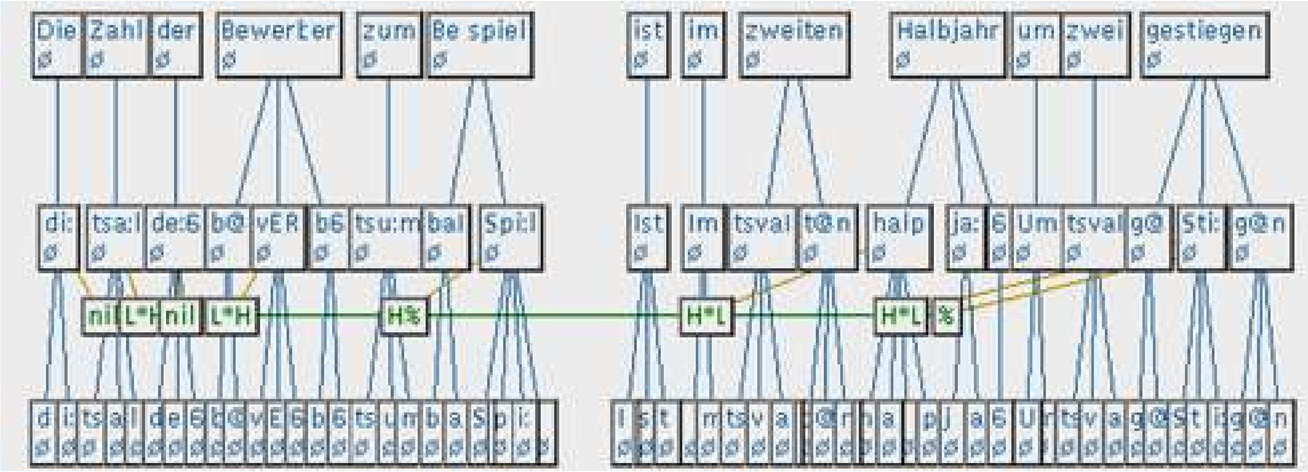


BB BB

Die Zahl der Bewerber zum Beispie ist im zweiten Halbjahr um zwei gestie

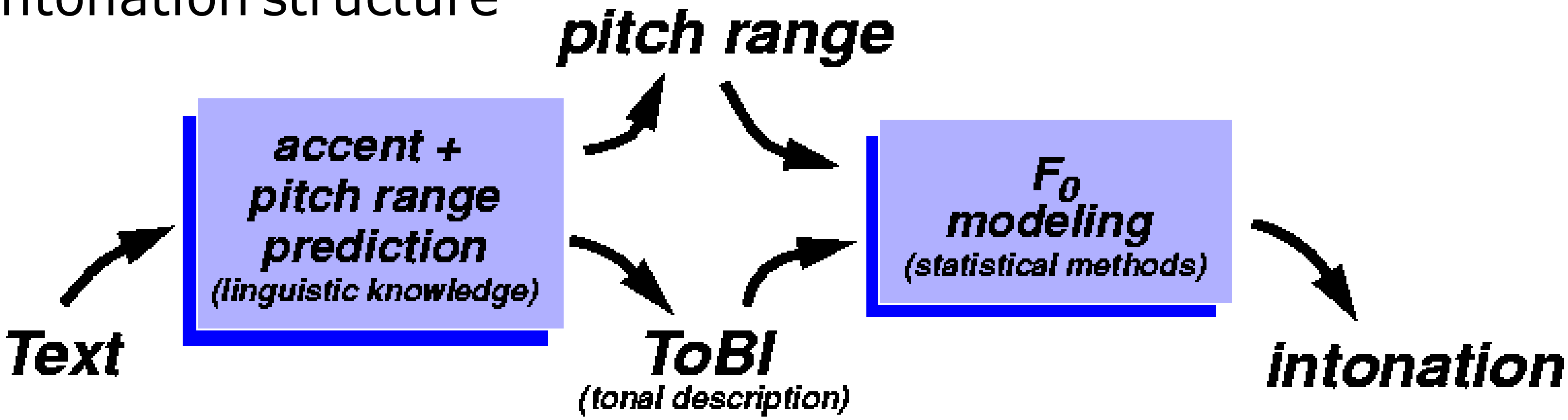
Intonation

- Depending on word class, position of the word in the sentence an in the phrase and depending on word classes of preceding and following words, for each syllable of each word it is decided if it is accented or not, and if so, which type of accent is to be realized



Wave synthesis

- PostLex
  - modifies the phone string by rule
- Duration
  - modifies the duration of the phone, depending on the phone context, etc.
- Wave_Synth
  - concatenate units and modify F0 according to the intonation structure



Text

accent + pitch range prediction (linguistic knowledge)

pitch range

ToBI (tonal description)

F_0 modeling (statistical methods)

intonation

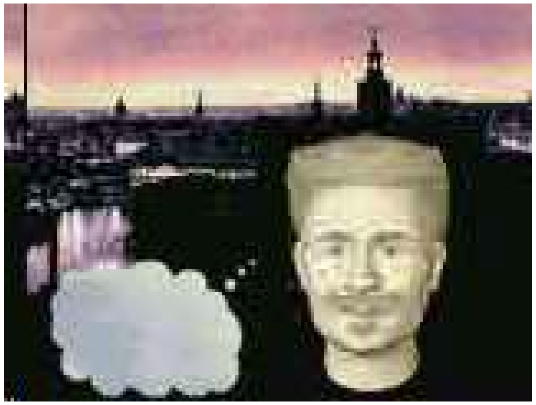
Polish synthesis: current status

- At least two commercial systems available:
  - Harpo: for visually impaired people, previously formant synthesizer used, now the one from Neurosoft
  - Neurosoft:
  - Politechnika Warszawska ?
  - Politechnika Poznanska

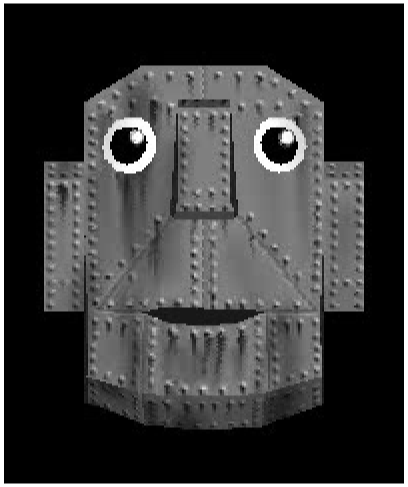
- L&H: system will be available in 6-12 month
- PJWSTK system
- PJWSTK limited domain

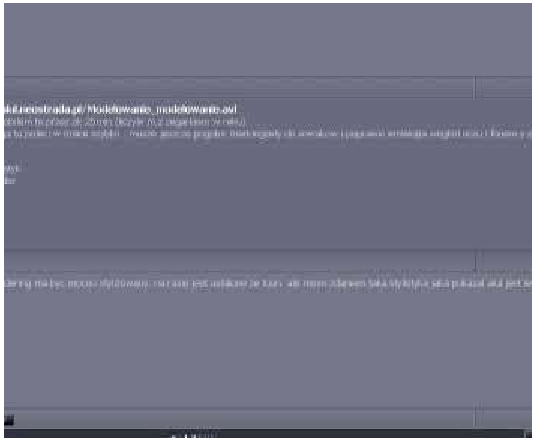
Future challenges

- Emotion synthesis (KTH Stockholm):
  - neutral 🔊
  - angry 🔊
  - happy 🔊
  - sad 🔊
- Dialog systems 🔊
- Avatars and artificial personality



Examples




kl.neostrada.pl/Modelowanie_modelowanie.avi
[illegible]