How to build a Speech Synthesis System?

New Media & Language Technologies

Jozef Stefan International Postgraduate School

November 2005

Jerneja Žganec Gros

jerneja@alpineon.com

Speech Synthesis

- Concatenation of prerecorded speech units :
  - small vocabulary, simple syntax
  - limited application domains: naturally sounding output
- Text-to-speech synthesis :
  - automatic conversion of arbitrary text into speech using GTP
  - unrestricted application domain
- Concept-to-speech synthesis :
  - entry: semantic concepts
  - IVR, speech-to-speech translation

Prerecorded speech

- database structure

DATE	[December 29]
TYPE	[maple]
LOC	[Javorniki Vrh]
WEIGHT	[6.7 kg]
REMARK	[po plohi]

- template

[Donosi na opazovalnicah DATE TYPE LOC WEIGHT REMARK]

Prerecorded speech

- message construction

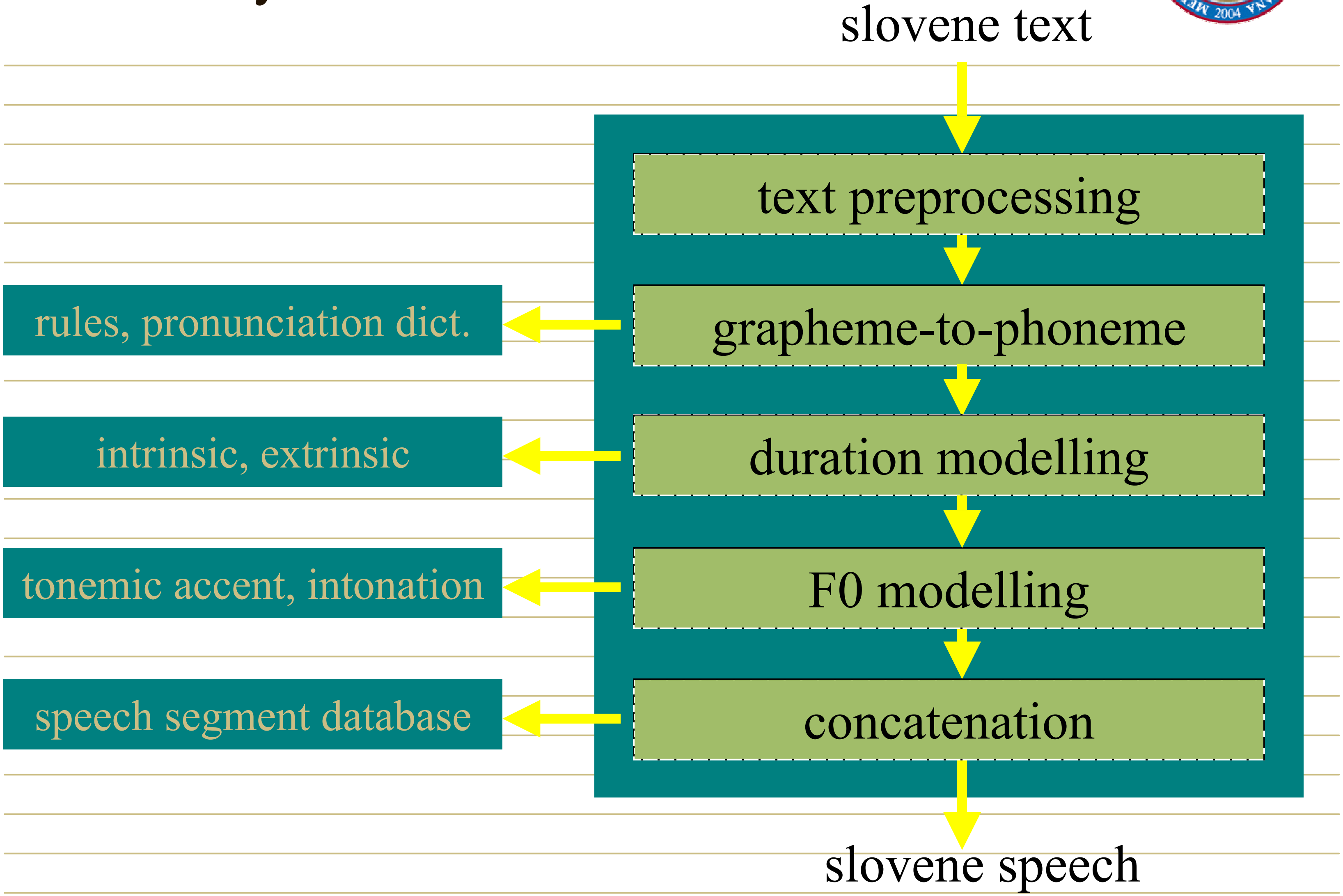
Donosi na opazovalnicah DEVETINDVAJSETEGA DECEMBRA. JAVORJEVA paša. JAVORNIKI VRH. PLUS ŠEST kilogramov SEDEMDESET dekagramov. PO PLOHI.

- speech segment concatenation
  - continuous transitions
  - sentence intonation
  - nearly natural pronunciation

TTS approaches

- Modelling the human vocal tract (hvt):
    - mechanical & electrical models of the hvt…
    - formant frequencies: formant TTS…
- Concatenation methods:
    - PSOLA, MBROLA, unit-selection
    - diphones, poliphones…
- HMM-based methods
- this talk: corpus-driven approaches (AlpSynth)

TTS System Architecture



slovene text

text preprocessing

grapheme-to-phoneme

duration modelling

F0 modelling

concatenation

slovene speech

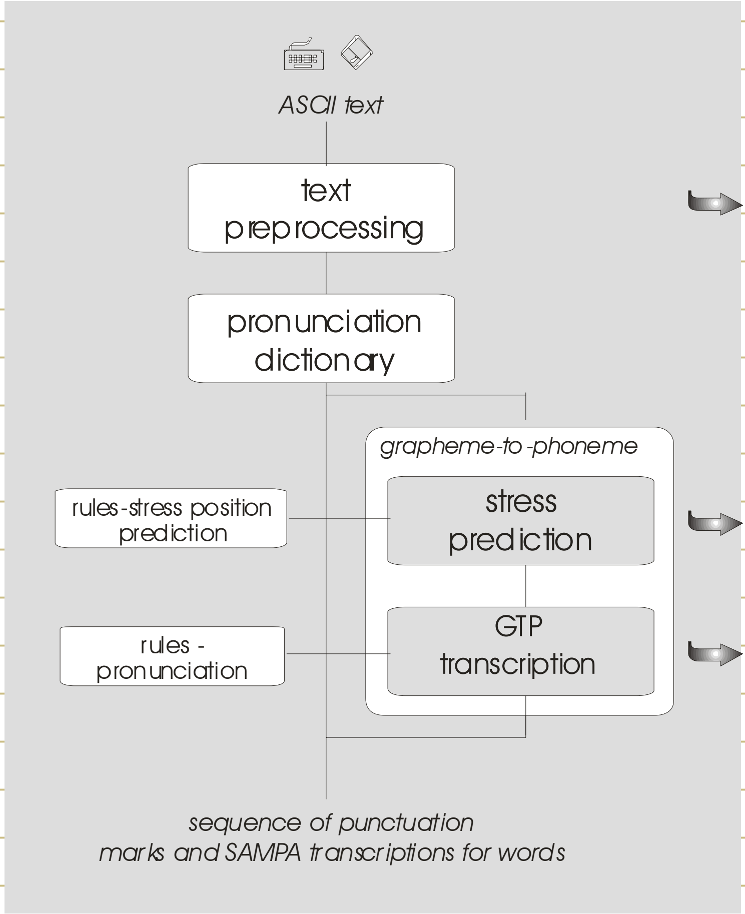
rules, pronunciation dict.

intrinsic, extrinsic

tonemic accent, intonation

speech segment database

Grapheme-to-Phoneme



ASCII text

text preprocessing

pronunciation dictionary

grapheme-to-phoneme

rules-stress position prediction

stress prediction

rules - pronunciation

GTP transcription

sequence of punctuation marks and SAMPA transcriptions for words

special symbols - elimination {,(,-,",...,$

punctuation marks - usage determination

Slovene: free stress position

rules for automatic stress prediction

production rules for automatic grapheme-to-phoneme transcription

Text Normalisation

- **alphanumerical graphemes**
  - tokenization: merging into words
  - sequences of capital letters:

    title / acronym disambiguation

    <AVTOBUSNA POSTAJA>

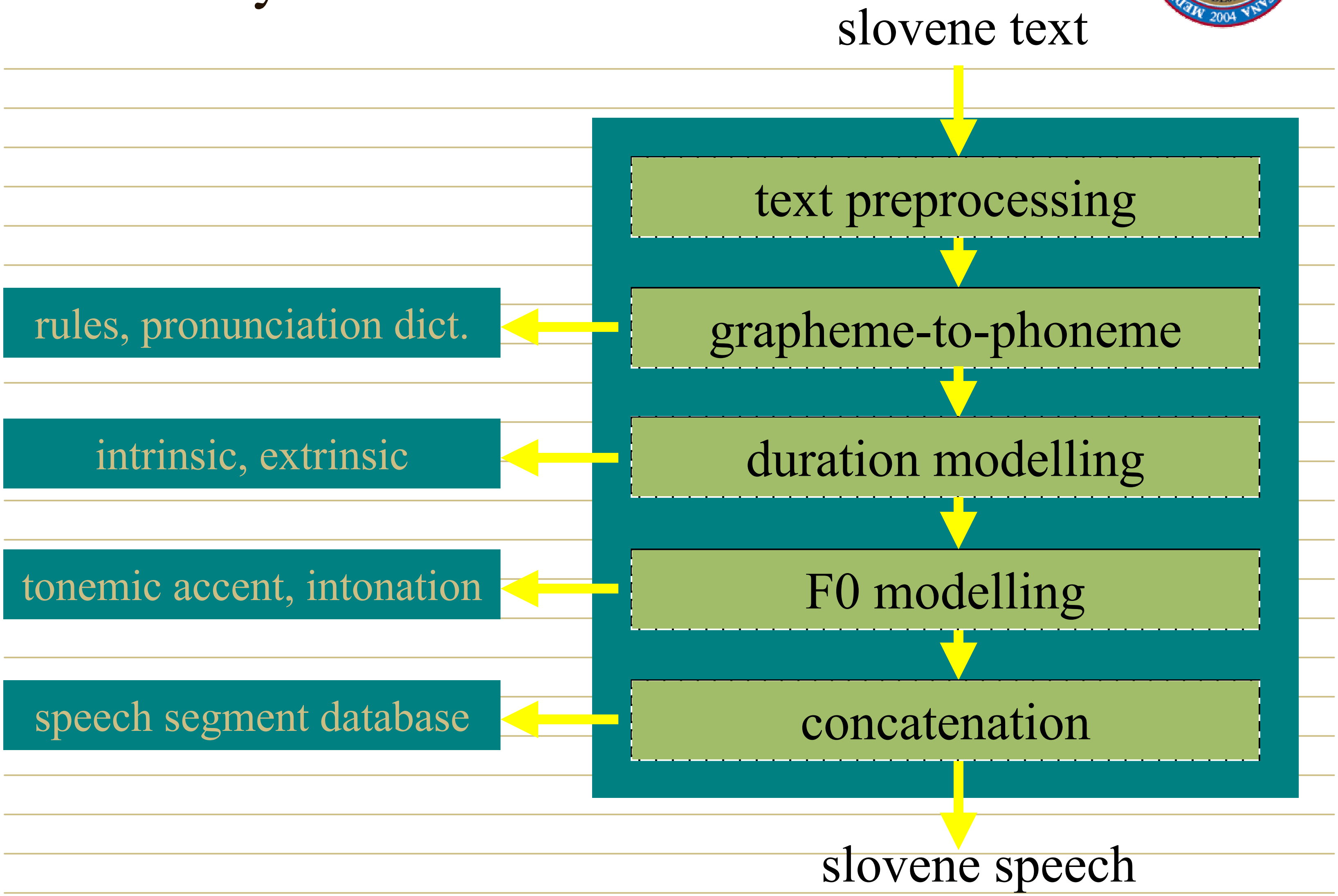
    <ZDA> <NATO>
- **numerals**
  - cardinal / ordinal (**1.** torek ➔ prv**i** torek)
- **ideograms**
  - $, %, &, (,), +, =, /, <, >, ...

Text Normalisation

- **punctuation marks**
- **grammatical usage (e.g. full stop)**
    - followed by a space AND a capitalized word

      <Dopolnil jih je 78. Lepa starost.>
    - followed by 2 line feeds (end of paragraph)
    - not followed by a numeral or space
- **non-grammatical usage**
    - abbreviation stop (as.dr. Simon Dobrišek, dipl.ing.)
    - ordinal numeral (Ob 8. uri zvečer.)
    - decimal (Cena izdelka je 8.12 SIT.)

TTS System Architecture



slovene text

text preprocessing

grapheme-to-phoneme

duration modelling

F0 modelling

concatenation

slovene speech

rules, pronunciation dict.

intrinsic, extrinsic

tonemic accent, intonation

speech segment database

Graphemes-to-Phonemes

- **search in the pronunciation dictionary**
- **coarticulation corrections**

  (word boundaries)
- **stress position prediction**

  (out-of-dictionary words)
- **grapheme-to-phoneme conversion, coarticluation corrections**

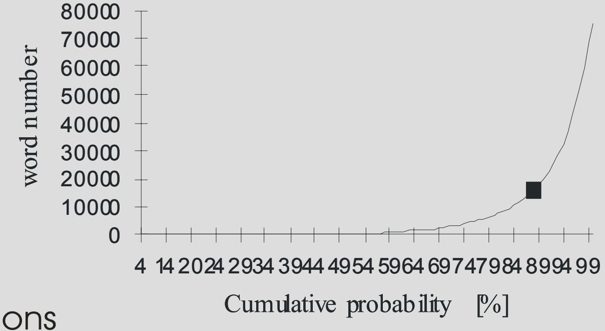
  (out-of-dictionary words)

Pronunciation dictionary

- text database

	Word number
Sveto pismo	152.212
Mikeln, Veliki Voz	162.396
Cankar, Moje `ivljenje	26.916
Slovenec, izbor ~lankov	264.736
Moj Mikro, izbor ~lankov	150.194
Jur~i~, Deseti brat	65.860
total	822.314

- 16.000 most frequent words cover 88.5% input text words



number of most frequent words and their cumulative probabilty

- SAMPA transcription - manual corrections

	word number
Collocations	17
Numerals	234
Words of foreign origin	304
Acronyms	92
Proper names	929
Other frequent words	15.470
Total	16.215

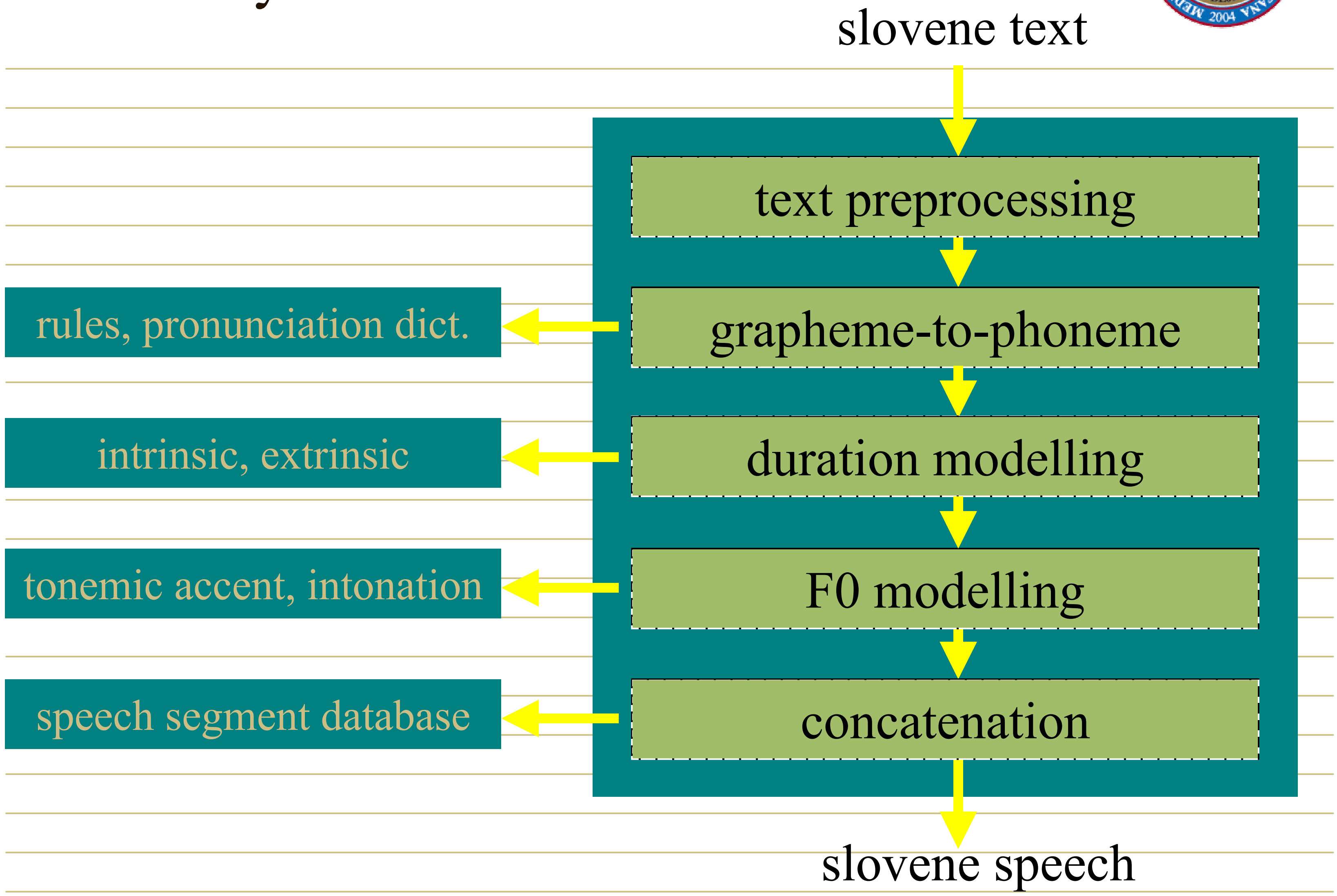
Grapheme-to-Phoneme Rules

- **standard words rule set**
  - **169 context-sensitive rules**

Left context	Grapheme string	Right context	Phonetic transcr.	Example	Rule explanation
$	er	_	[@r]	Gaber	@ occurs after each -r not followed by a vowel (Toporisic91, p.49)
=	m	f	[F]	Simfonija	<m> in front of <f> and <v> is pronounced as a labiodental (Pravopis90, p. 145)

- **names rule set**

TTS System Architecture



slovene text

text preprocessing

grapheme-to-phoneme

duration modelling

F0 modelling

concatenation

slovene speech

rules, pronunciation dict.

intrinsic, extrinsic

tonemic accent, intonation

speech segment database

Duration Modelling

- sequential rule systems (Klatt 73, Van Santen 93)
- neural networks (Campbell 90)
- stochastic modelling (Traber 93), decision trees (Riedi 95), hmms (2000->…)

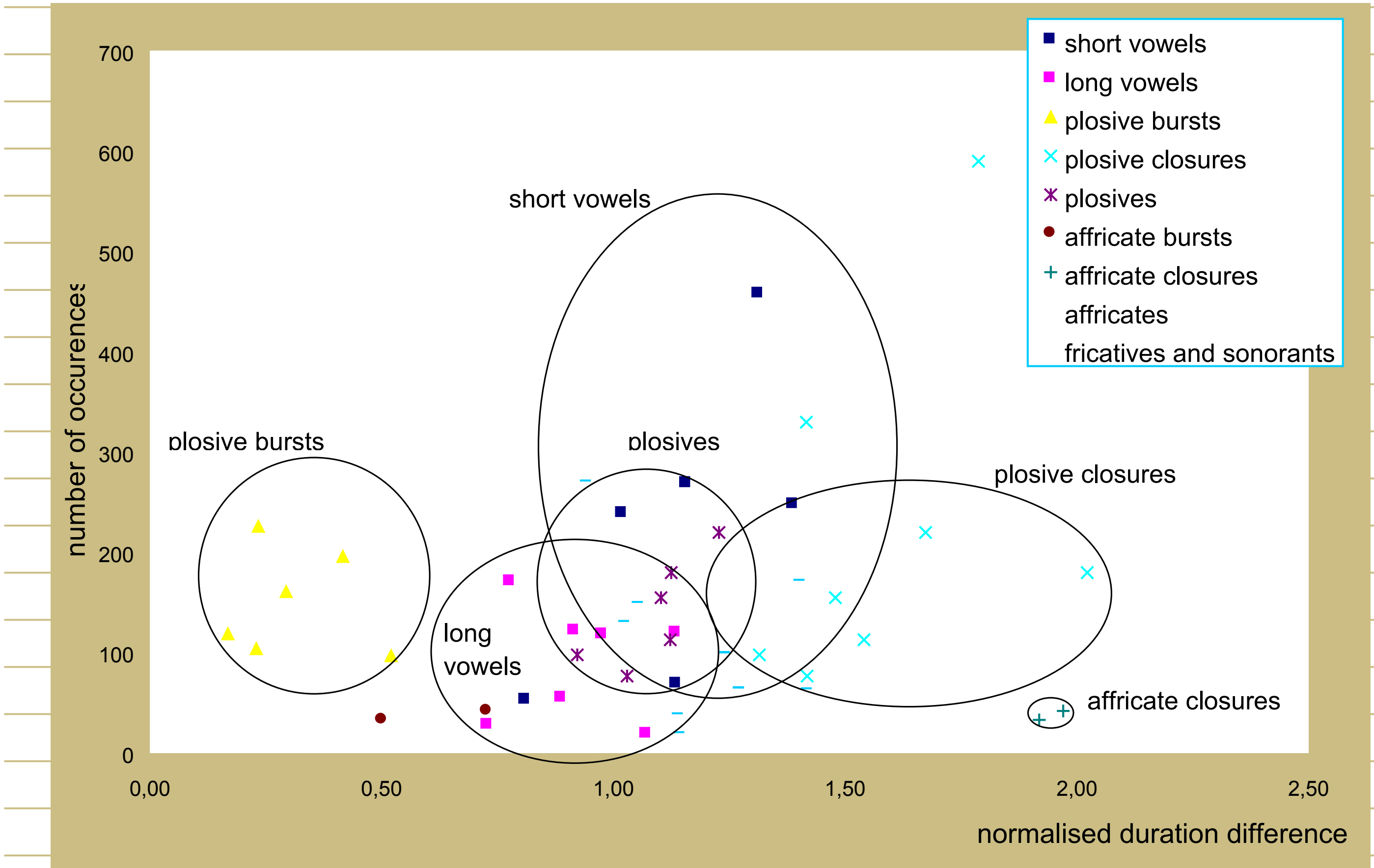
- **two-level approach (Epitropakis 93)**
  - **intrinsic** duration modelling
  - **extrinsic** duration modelling
  - **adaptation** of intrinsic phone duration to extrinsic word duration (Gros 97)

Intrinsic Duration

- phone identity, phone type: C or V
- syllable type: open or closed
- tonic, pretonic, posttonic
- position within the word: initial, medium, final
- phonetic context: CC, VCV

- **Measurements:**
  - **logatoms in neutral intonation position**

Phone Duration



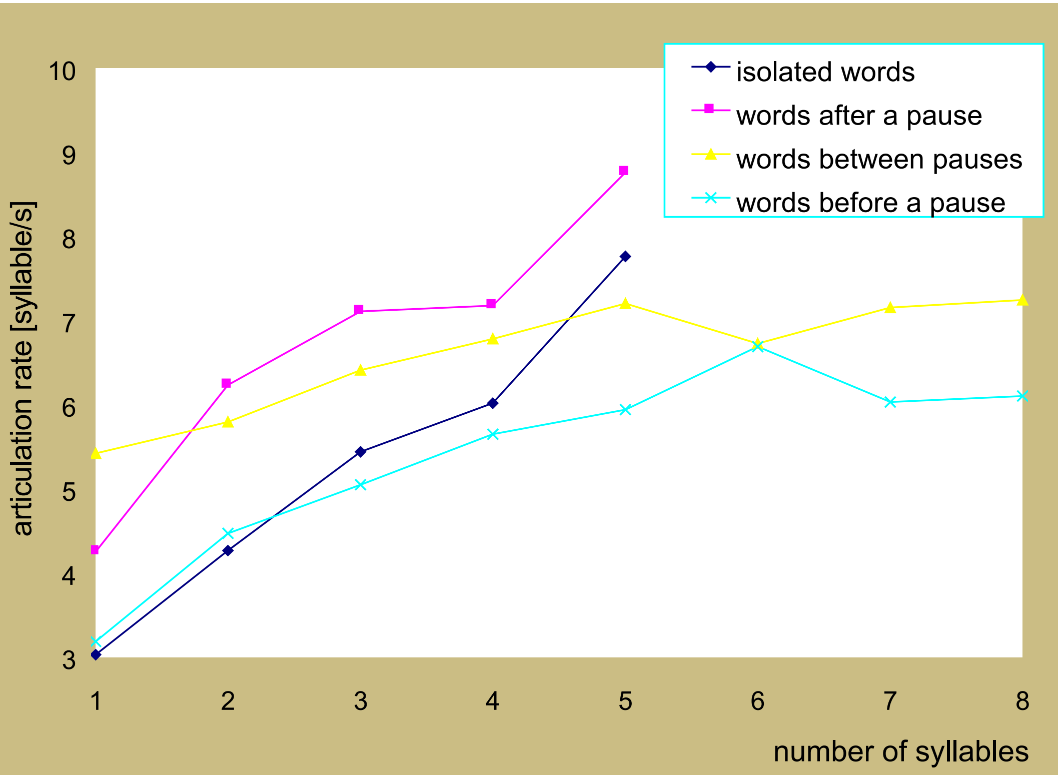
Pair-wise analysis: normal rate - slow rate. Normalised mean duration difference for pairs of phone realisations in the phoneme group context.

Extrinsic Duration

- number of syllables
- word position: phrase initial, medium, final
- requested speaking rate: from slow to normal and fast
- syllable position in a word: initial, medium, final

- **Measurements:**
  - **continuous speech - slow, normal, fast**
  - **duration units!**

Syllable Duration



Articulation rate in number of syllables per second is shown for different word positions within a phrase.

Intrinsic to Extrinisic Dur.

- curves **a**$_i$:

  linear interpolation between average phone duration measurements at different speaking rates

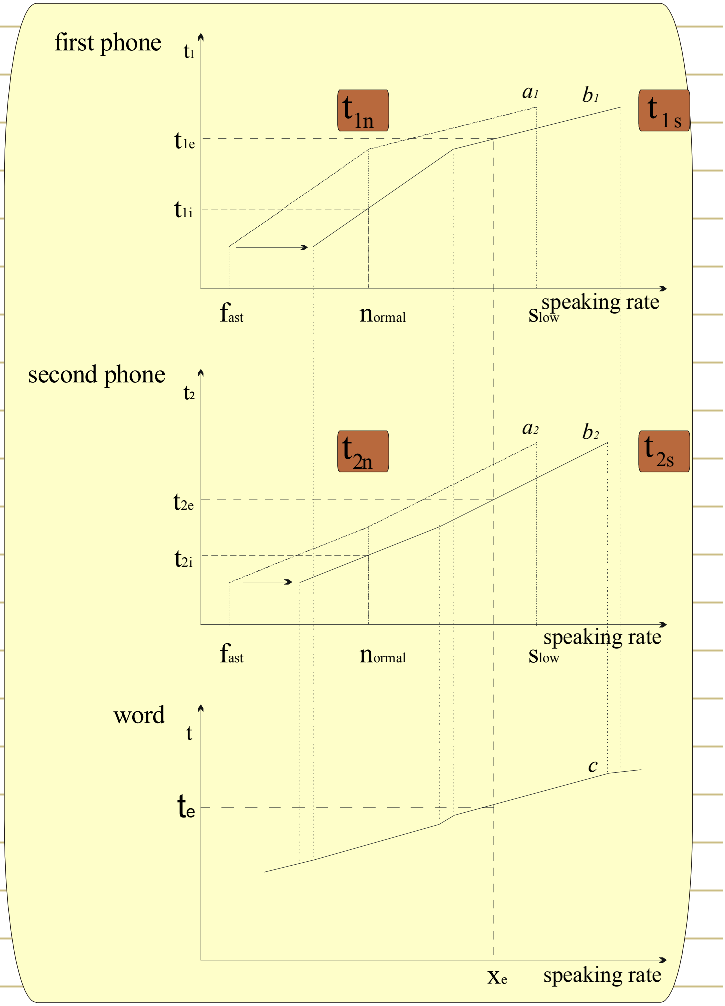
- curves **b**$_i$:

  horizontal translation of a$_i$ in a way that b$_i$ equals the intrinsic phone duration t$_{ij}$ at normal speaking rate

$$t_{je} = t_n + \frac{t_{js} - t_{jn}}{t_p - t_n}(t_e - t_n),\ j = 1,2$$

- curve **c**:

  sum of b$_i$ over all phones;
  extrinsic word duration t$_e$ occurs at the speaking rate x$_e$



Duration Prediction - Eval.

- test base

speech rate	no. of sentences	no. of words	no. of phones
normal speech rate	172	1400	5433
fast rate	49	607	2351
slow rate	60	800	2900

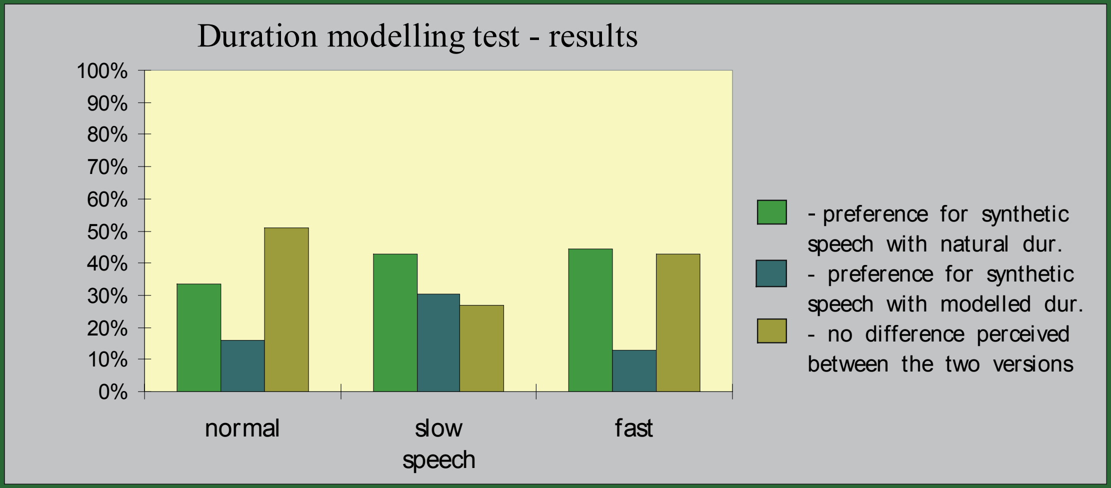
- statistical duration difference evaluation between phone pairs in natural and synthetic speech

	translation	proportio	natural speech
mean absolute difference [ms]	10.97	30.20	5.3
mean absolute diff. [ms] (stressed vowels)	6.89	33.67	
standard deviation [ms]	15.24	26.41	8.2
standard deviation [ms] (stressed vowels)	13.18	28.07	

Duration Prediction - Eval.

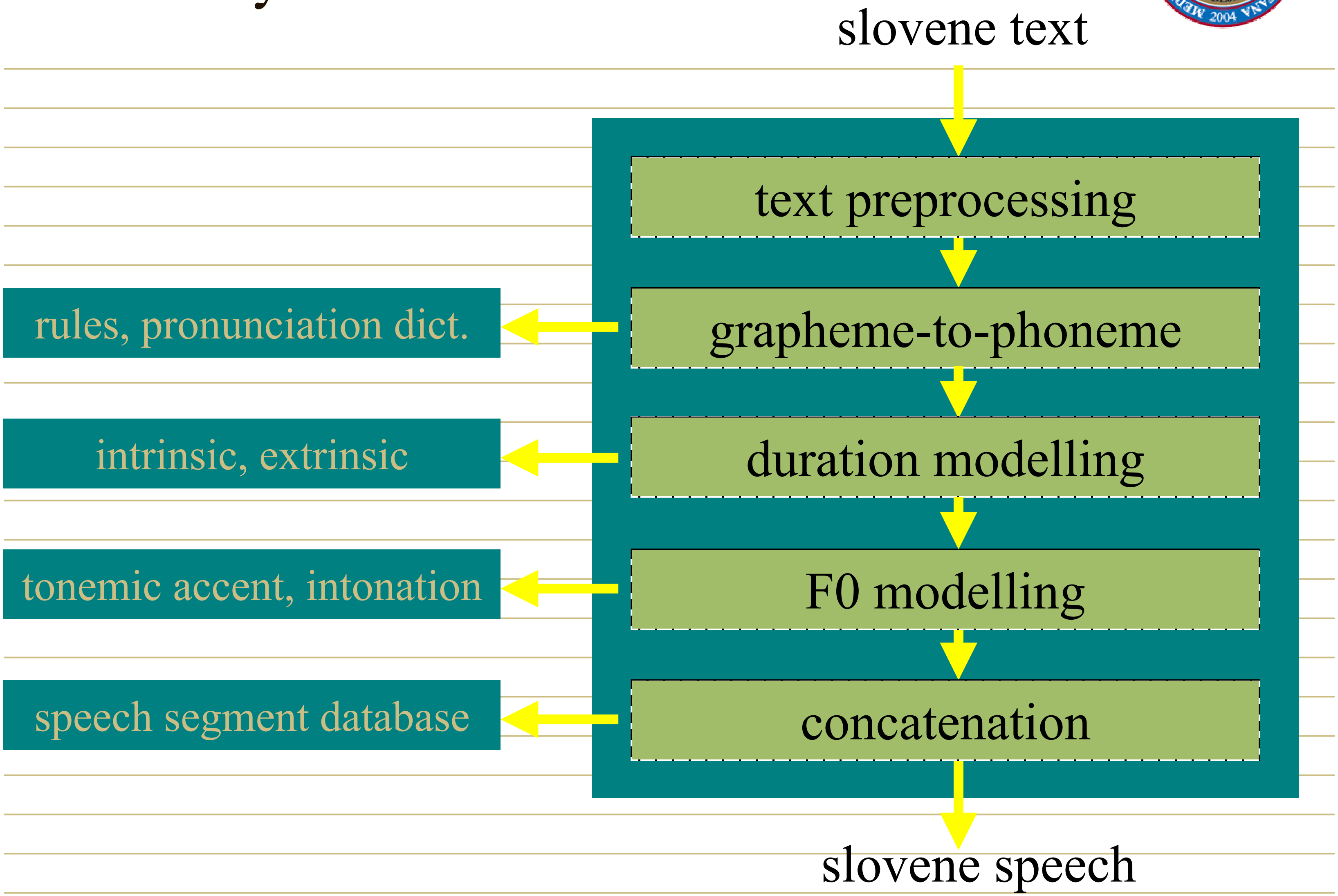
slow, fast and normal speech rate

- phone duration values taken from natural speech
- phone duration values predicted by the 2-level approach



- 20 test subjects, different professional backgrounds
- *ITU/T Recommendation P.85*: A method for subjective performance assessment of the quality of speech voice output devices

TTS System Architecture



slovene text

text preprocessing

grapheme-to-phoneme

duration modelling

F0 modelling

concatenation

slovene speech

rules, pronunciation dict.

intrinsic, extrinsic

tonemic accent, intonation

speech segment database

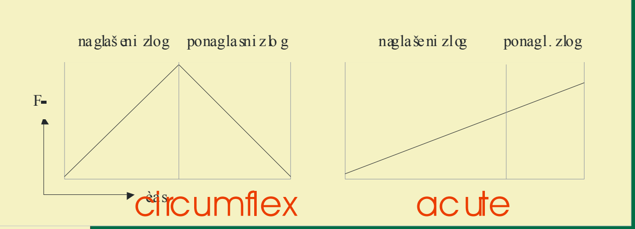
F0 Modelling

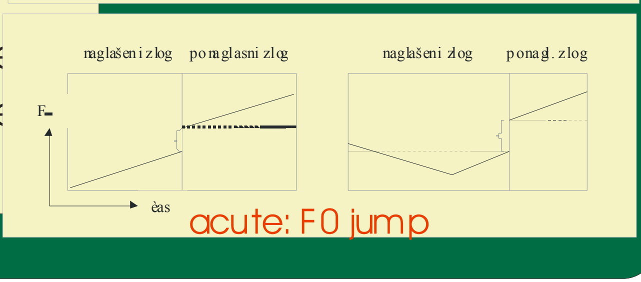
- initial F0 values
- jump
- jump restrictions
- interpolation
- minor random adjustment

- intrinsic pitch frequency
- syllable position: initial/final/mid
- syllable structure: open, closed
- tonic/pretonic/posttonic syllable

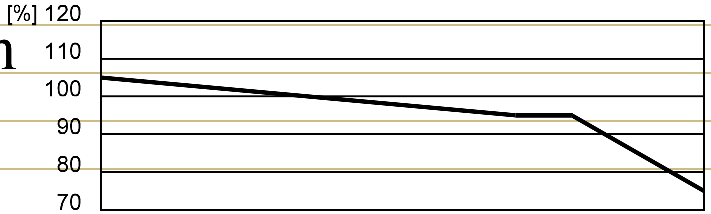
Typical F0 patterns (tonemes)

- barytone acute
- ocsytone acute
- 2-syllabic baritone cirkumfle
- 3-syllabic baritone cirkumfle
- ocsytone cirkumflex

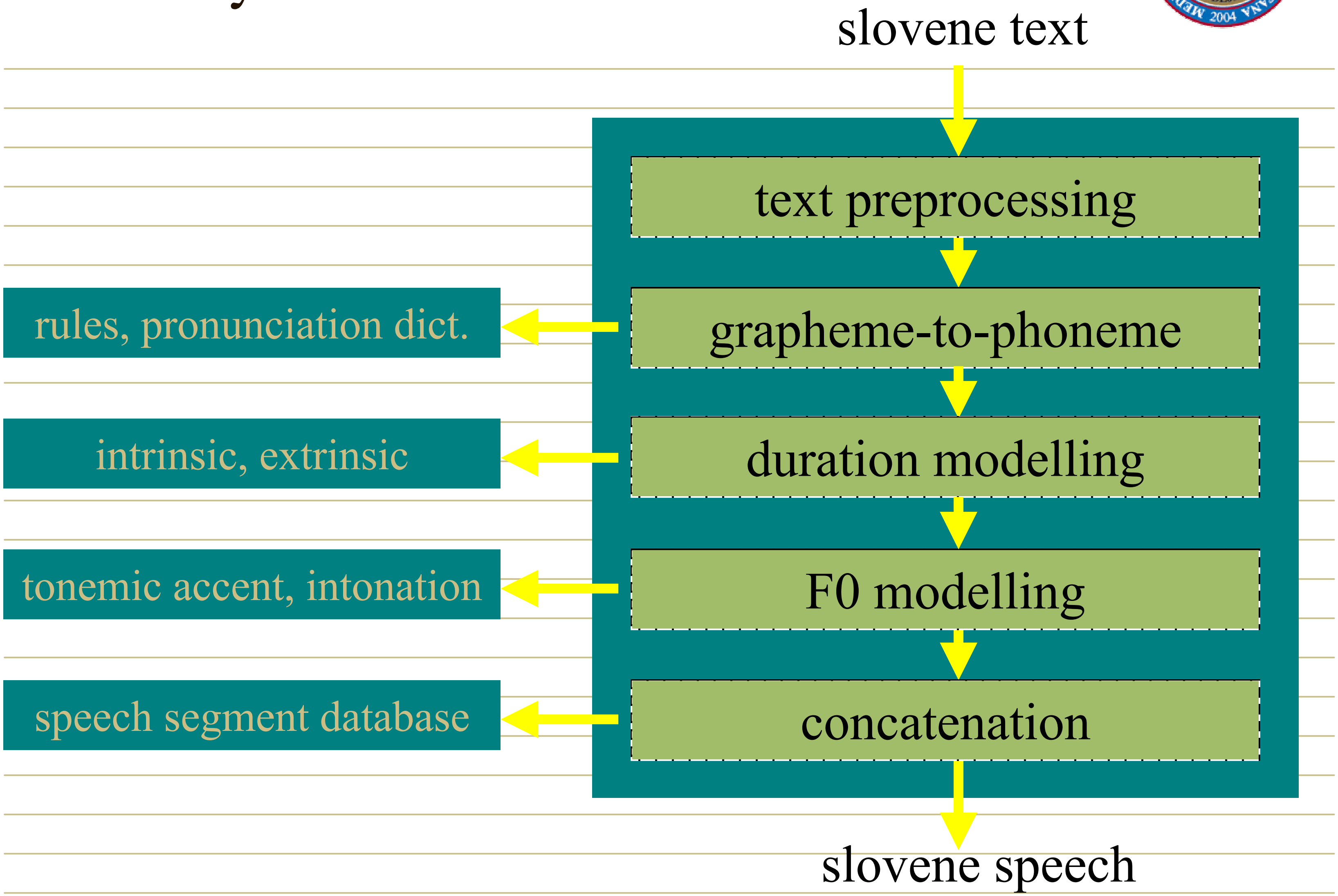




– Sentence intonation



TTS System Architecture



slovene text

text preprocessing

grapheme-to-phoneme

duration modelling

F0 modelling

concatenation

slovene speech

rules, pronunciation dict.

intrinsic, extrinsic

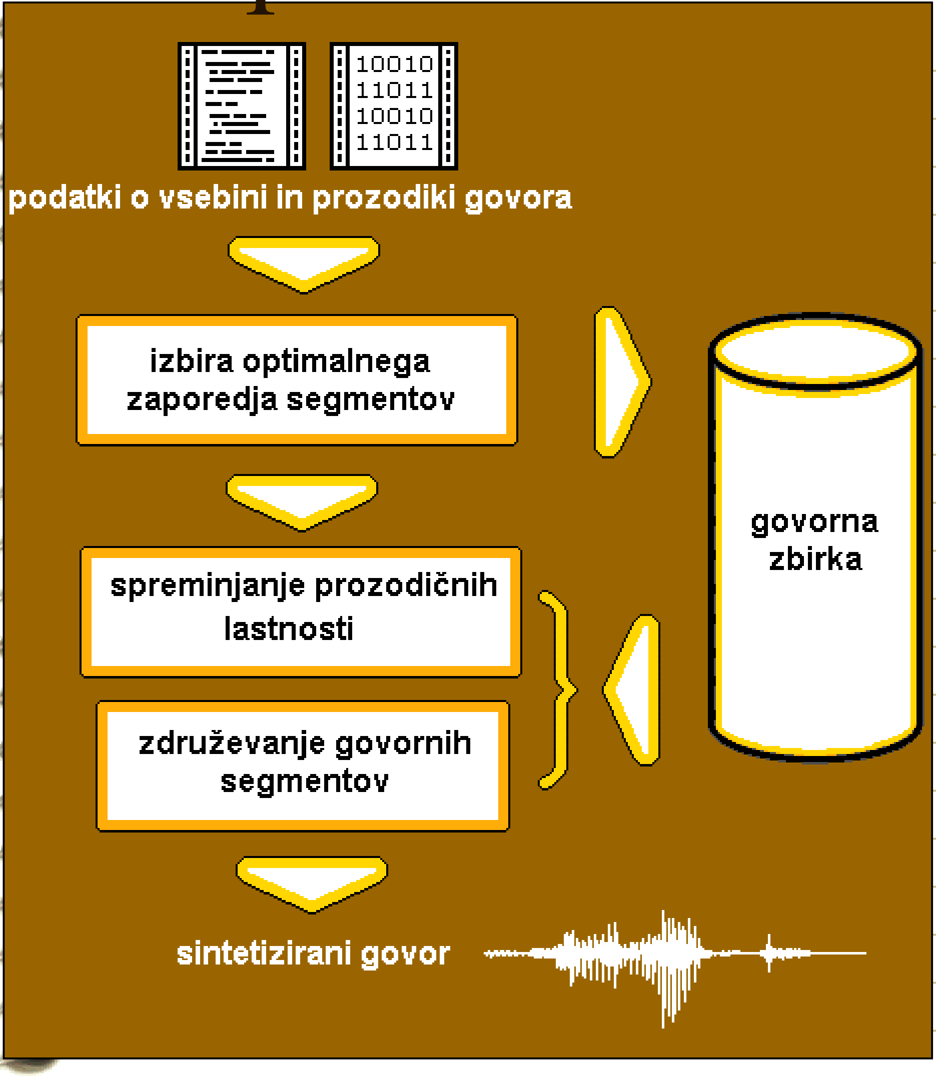
tonemic accent, intonation

speech segment database

Speech segment concatenation

- corpus-driven text-to-speech synthesis
- speech corpus:
  - text selection
    - phonetic transcription of the source text corpus
    - phone frequency analysis
    - algorithm for optimal sentence set selection
  - recording
  - segmentation and labelling

Corpus-driven TTS


10010
11011
10010
11011
podatki o vsebini in prozodiki govora
izbira optimalnega zaporedja segmentov
govorna zbirka
spreminjanje prozodičnih lastnosti
združevanje govornih segmentov
sintetizirani govor

- speech corpus
- optimal speech segment selection

  (dynamic programming)
- speech segment concatenation and prosodic modifications

  (TD-PSOLA,MBROLA)

Corpus – elemental units

- allophones
- diphones
- poliphones
- words
- phrases….

- longer segments:

  - larger corpus
  - more natural speech

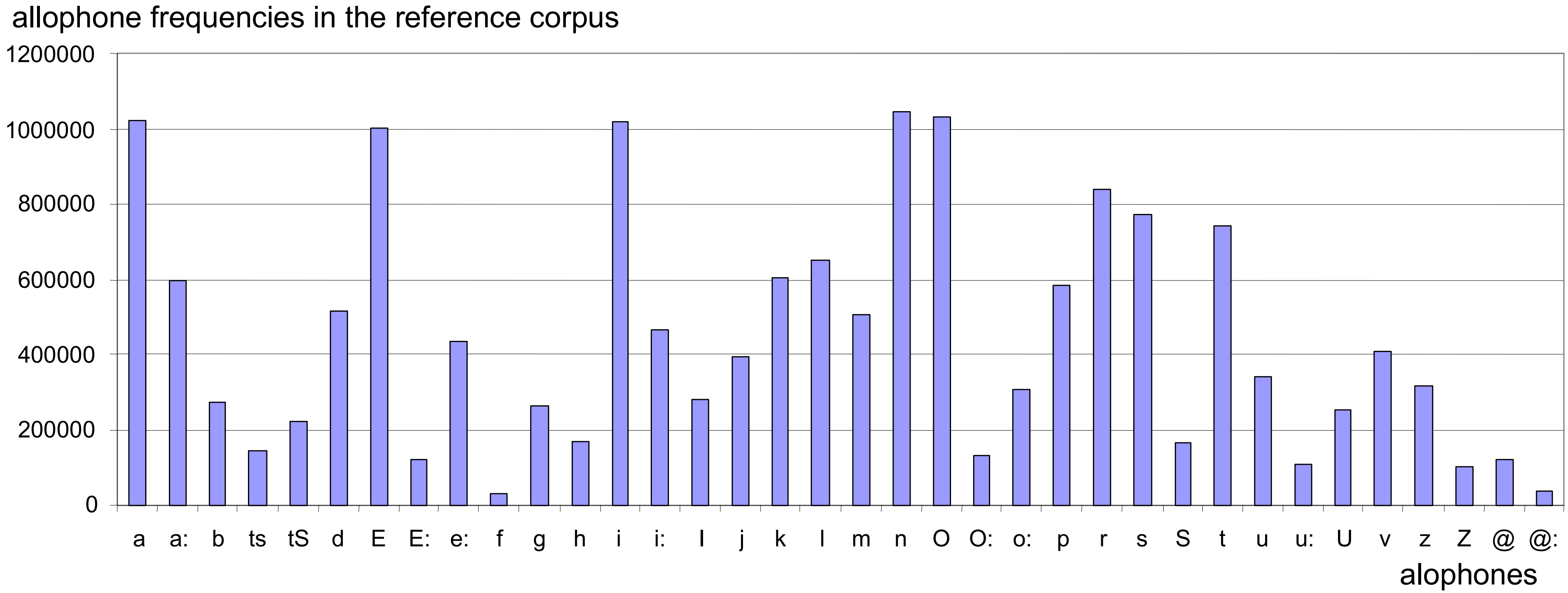
Speech corpus design

- text selection: input reference corpus to resulting text corpus
  - phonetic transcription of the reference text corpus
  - frequency analysis of allophone strings
  - AlpSynth sentence selection method
- recording
- segmentation and labelling
  - initial automatic segmentation
  - manual fine segmentation

Text corpus: phonetic analysis

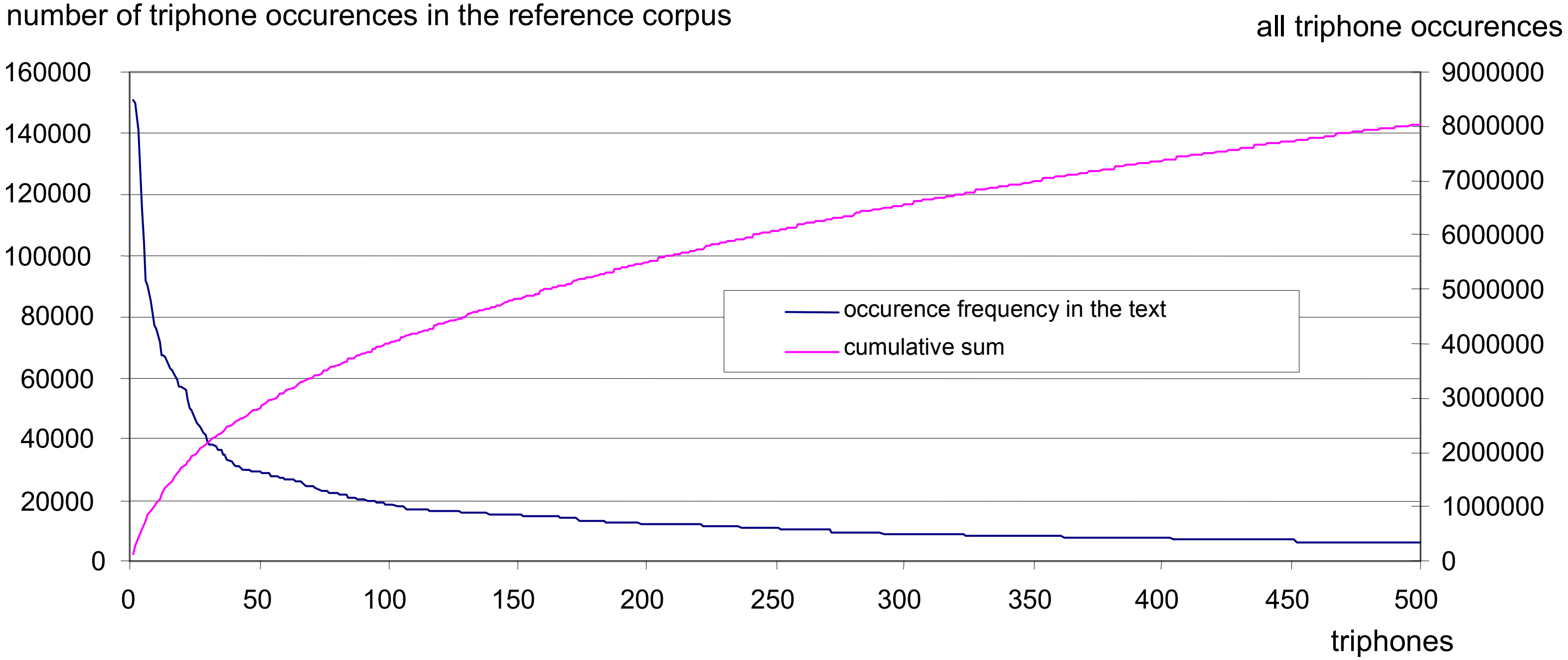
- grapheme-to-phoneme transcription of the initial reference text corpus
- frequency analysis of allophone strings:
  - allophones
  - diphones
  - triphones
  - quadphones

Sentence set selection



allophone frequencies in the phonetic transcription of the reference text corpus

Triphone string frequencies



triphone frequencies in the phonetic transcription of the reference text corpus

Sentence set selection

- goal
  - compact resulting sentence corpus containing all predefined frequent allophone sequences
- method
  - cost evaluation for all sentences
  - cost normalization (to sentence length)
  - ranking and selection of evaluated sentences

Sentence set selection

- features:
  - initial reference text corpus (200.000 sentences)
  - resulting compact text corpus (297 sentences)
  - rich with different allophone sequences
    - 1.132 different diphones
    - 17.784 different triphones
    - 120.425 different quadphones
    - average sentence length: 34.4 allophones oz. 6 words

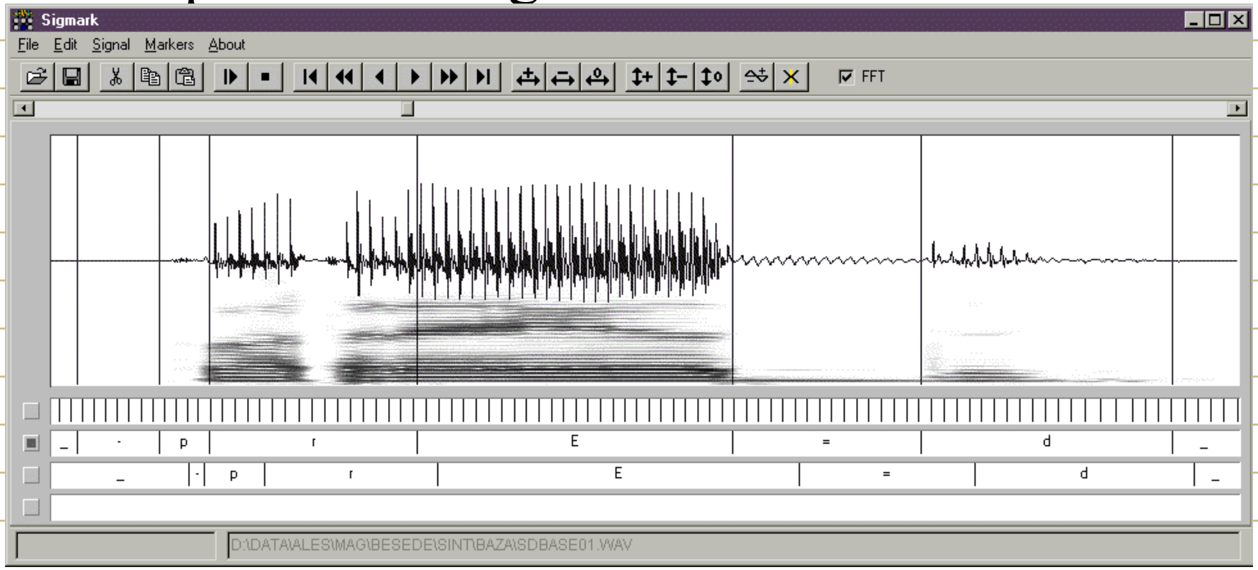
Recording

- male speaker, laboratory conditions
- corpus size:

	duration	number of words		number of phones
		all words	different words	
	natural speech			
A - recorded natural speech	3622 s	1814	1354	10218
	logatoms			
B - complete logatom corpus	1596 s	2837	2837	7342
logotom corpus (no diphtongs)	508 s	1169	1169	2338
logotom corpus (diphtongs only)	1088 s	1668	1668	5004
C - complete TTS speech corpus (A+B)	5218 s	4651	4191	17560

Segmentation and labeling

- Phone segmentation:
  - initial: automatic (HMM)
  - fine: manual - SIGMARK©
- Pitch marking:
  - fine pitch marking: automatic - SIGMARK©

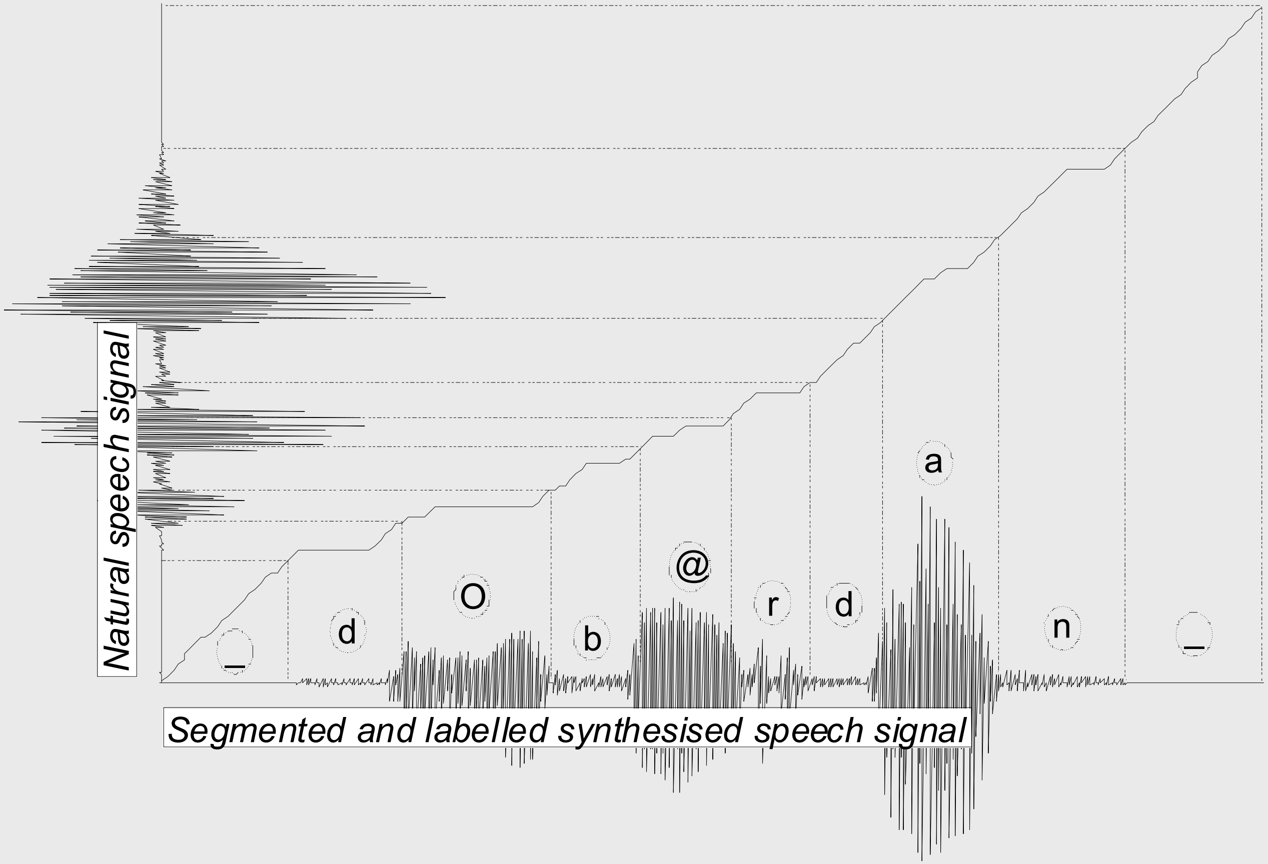


Sigmark
File Edit Signal Markers About
☑ FFT
_ · p r E = d _
_ · p r E = d _
D:\DATA\ALES\MAG\BESEDE\SINT\BAZA\SDBASE01.WAV

Automatic Labelling

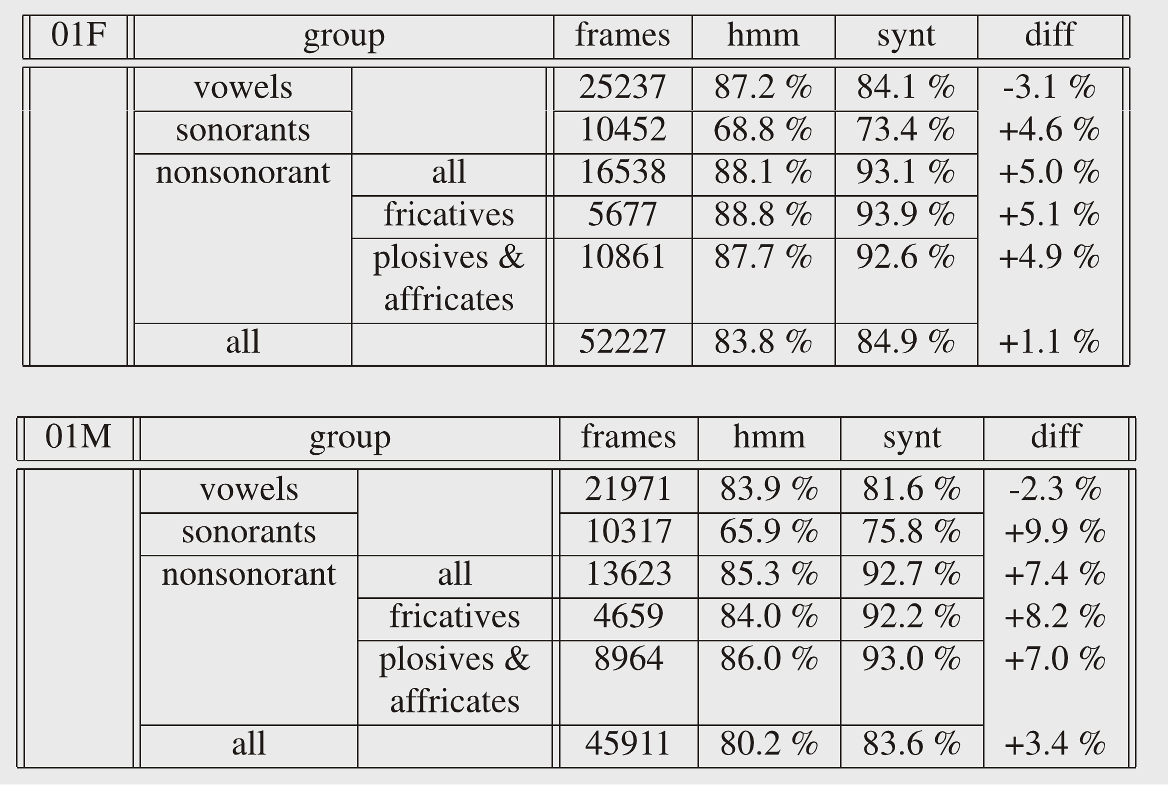
- purpose:
  - basic phonetic research
  - initialisation for the stochastic speech recogniser
- approaches:
  - HMM
  - **DTW alignment of natural and synthetic speech**
- speech synthesis:
  - diphone inventory
- feature vector:
  - loudness, 11 mel-cepstrum coefficients

Automatic Labelling



Automatic Labelling

- average frame match between manual and automatic segmentation



01F	group		frames	hmm	synt	diff
	vowels		25237	87.2 %	84.1 %	-3.1 %
	sonorants		10452	68.8 %	73.4 %	+4.6 %
	nonsonorant	all	16538	88.1 %	93.1 %	+5.0 %
		fricatives	5677	88.8 %	93.9 %	+5.1 %
		plosives & affricates	10861	87.7 %	92.6 %	+4.9 %
	all		52227	83.8 %	84.9 %	+1.1 %

01M	group		frames	hmm	synt	diff
	vowels		21971	83.9 %	81.6 %	-2.3 %
	sonorants		10317	65.9 %	75.8 %	+9.9 %
	nonsonorant	all	13623	85.3 %	92.7 %	+7.4 %
		fricatives	4659	84.0 %	92.2 %	+8.2 %
		plosives & affricates	8964	86.0 %	93.0 %	+7.0 %
	all		45911	80.2 %	83.6 %	+3.4 %

Plans for further work

- reduction of spectral discontinuities
- optimization of the speech segment selection procedure
- selection of optimal intra-segment concatenation locations
- further upgrades of the speech corpus

Evaluating TTS Systems

Jekosch93, Pols94, JEIDA95, Klaus03, ITU-T Recs

- **First experiment**
  - intelligibility
  - naturalness
- **Second experiment**
  - ITU-T Rec. P.81
  - ITU-T Rec. P.85

Text Selection

- Text types:
  - newspaper text (daily newspaper, 264.763 words)
  - The Bible (152.212 words)
  - SUS (semantically unpredictable sentences)
    - basic pattern structures : Subject - Verb – Adverbial, Subject – Transitive Verb - Object, etc.
      - *Hrast gleda morje*
    - word lists from the MULTEXT-EAST lexicon (morpho-syntactic descriptions)
- Text selection methods:
  - 4 text selection methods as proposed by LDC and COCOSDA

Text Selection Methods

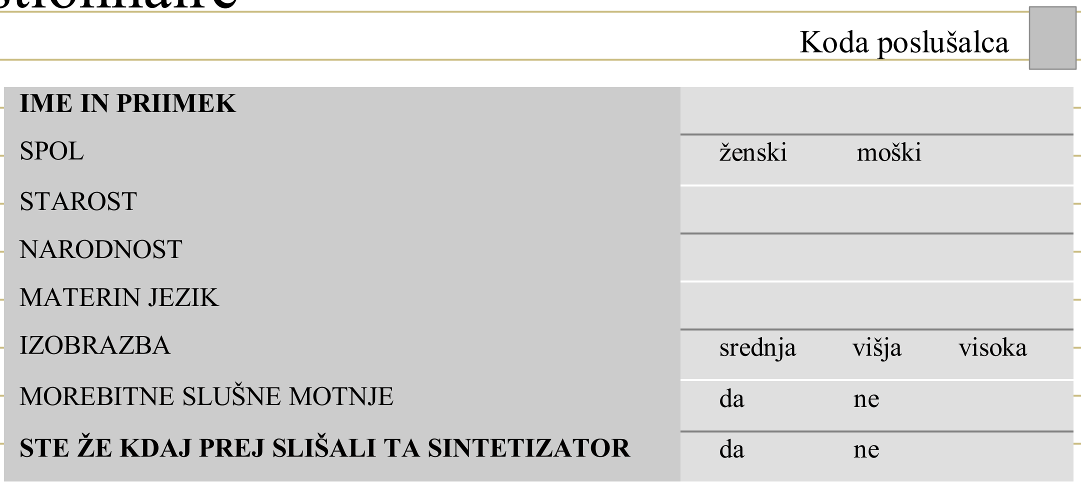
- Random selection
- Minimum word frequency
  - determine number of occurrences (frequency) of each word in the text corpus
  - for each sentence, determine the frequency of the least frequent word
  - sort sentences in descending order by least frequent word frequency
  - randomly select from the top 1, 5, or 10 % of this sorted list

Text Selection Methods

- Overall word frequency
  - determine number of occurrences (frequency) of each word in the corpus
  - for each sentence, add the log frequencies of all its words
  - sort sentences in descending order by log frequency sum
  - randomly select from the top 1, 5, or 10 % of this sorted list
- Overall trigram frequency based selection

Design of the experiments

- laboratory conditions
- 2 sessions, preliminary training session
- various evaluators
- questionnaire



Koda poslušalca ▨

IME IN PRIIMEK			
SPOL	ženski	moški	
STAROST			
NARODNOST			
MATERIN JEZIK			
IZOBRAZBA	srednja	višja	visoka
MOREBITNE SLUŠNE MOTNJE	da	ne	
STE ŽE KDAJ PREJ SLIŠALI TA SINTETIZATOR	da	ne	

Experiment

- TTS system
  - ITU-T Recommendations
  - 21 evaluators

  - acceptability of the synthetic speech for the application
  - naturalness of pronunciation
  - subjective impressions of the synthetic speech

Acceptability

- ITU-T Recommendation P.85

  (a method for subjective performance assessment of speech voice output devices)

- application domain - automatic information retrieval

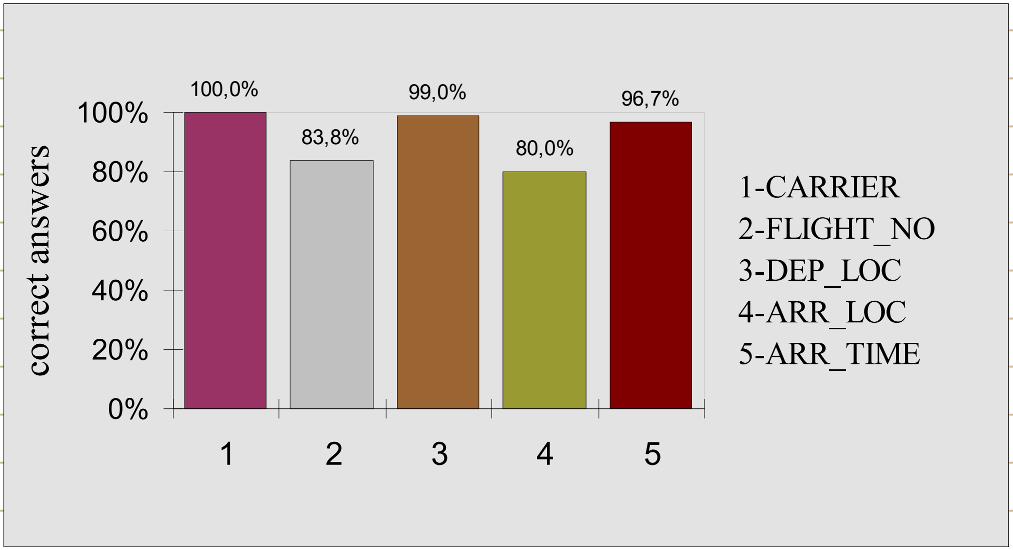
  (for comparison with the test of the S5 TTS system – Gros97)

- message templates

```
CARRIER, flight number FLIGHT_NO, arriving from
DEP_LOC, is about to land at ARR_LOC at ARR_TIME.

Adria Airways, flight number JP743, arriving from
Frankfurt, is about to land in Ljubljana at 13:30.
```

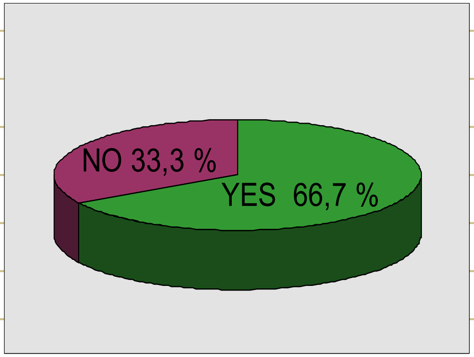
Acceptability



Do you think this TTS system could be used in a automatic information dialog system for airline timetable retrieval?

□ YES □ NO

Comments:



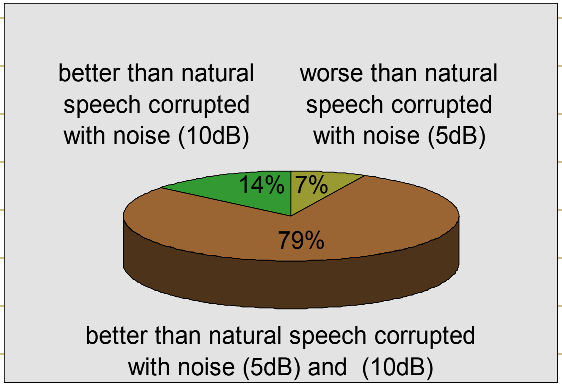
Naturalness

- ITU-T Recommendation P.81

  (Telephone quality subjective transmission tests - Modulated noise reference unit)

- voice sources
  - corrupted natural speech (SNR 5dB, 10dB, 15dB, 30dB)
  - speech synthesiser

- MOS opinion scales
  - overall impression
  - listening effort
  - comprehension problems
  - articulation
  - voice pleasantness



Subjective impressions

- ITU-T Recommendations P.80 and P.85

  ”Methods for subjective determination of transmission quality”

  “A method for subjective performance assessment of the quality of speech voice output devices”

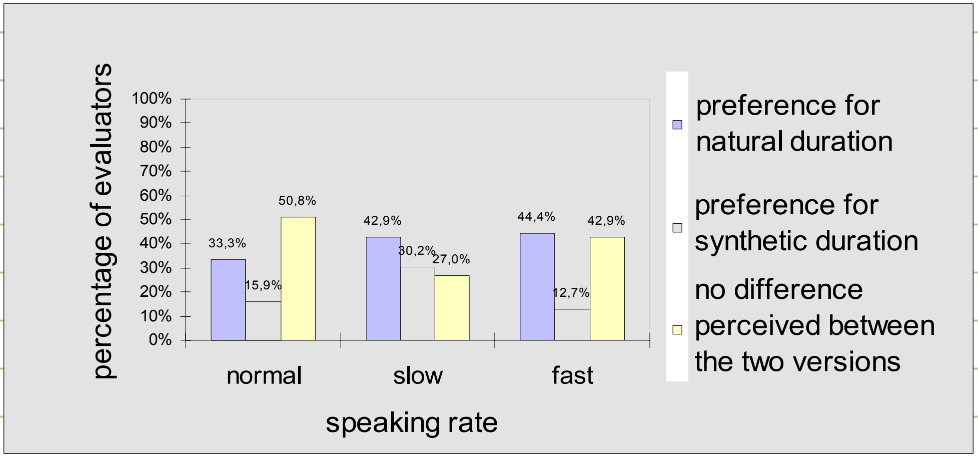
MOS scale	Overall impression	Comprehension problems	Articulation	Speech rate	Voice pleasantness
	How do you rate the quality of the sound?	Did you find certain words hard to understand?	Were the sounds distinguishable?	The average speed of delivery was:	How would you describe the voice?
5	excellent	never	yes, very clear	much faster than preferred	very pleasant
4	good	rarely	yes, clear enough	faster than preferred	pleasant
3	fair	occasionally	fairly clear	preferred	fair
2	poor	often	no, not very clear	slower than preferred	unpleasant
1	bad	all the time	no, not at all	much slower than preferred	very unpleasant

Subjective impressions

- MOS rating scales:
  - overall impression, listening effort, comprehension problems, articulation, pronunciation, speech rate and voice pleasantness
- overall quality of the synthetic speech
- evaluation of individual components of the TTS system:
  - grapheme-to-phoneme: pronunciation dictionary
  - prosody modeling:
    - tonemic accent patterns
    - segment duration prediction methods

Subjective impressions

- Segment duration prediction evaluation:
  - segment duration of the synthetic speech
    - taken from natural speech
    - automatically predicted by the two-level approach (Gros et al, 1997)



Conclusion

- Slovenian TTS system performance evaluation
  - pleasant, quite natural speech, sufficiently rapid, not overarticulated
  - further work: prosody, concatenation, lexical stress assignment
- Slovenian TTS: demo applications