10. Speech response systems

SmartLab
Intelligent Interactions
smartlab.tmit.bme.hu

BME TMIT

NVIDIA
GPU
EDUCATION
CENTER

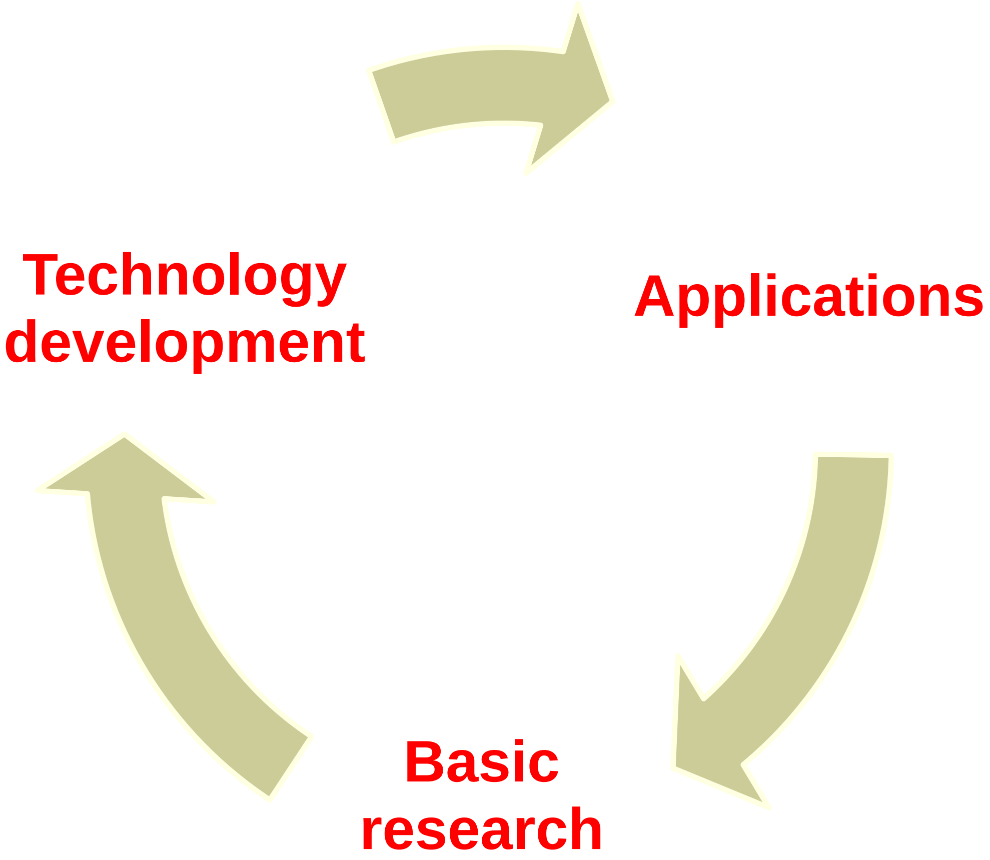
Topic

- The creation of machine generated speech

modern / contemporary technologies

- Modeling approaches
  - Generative models (articulation, source-filter, …)
  - Production models (concatenation, …)
- Technological solutions
- Applications
- Outlook

The research cycle of speech technology

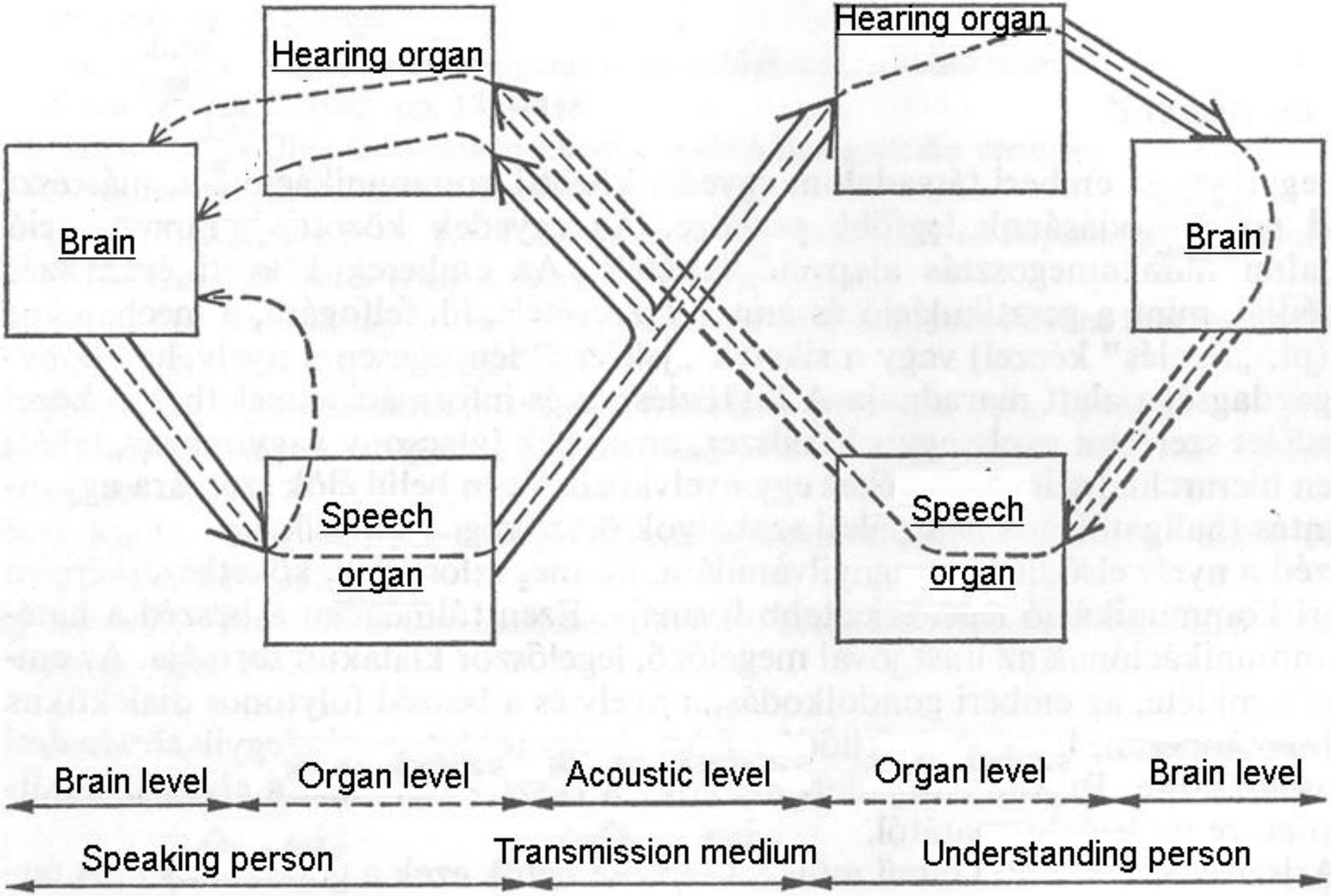


Technology development

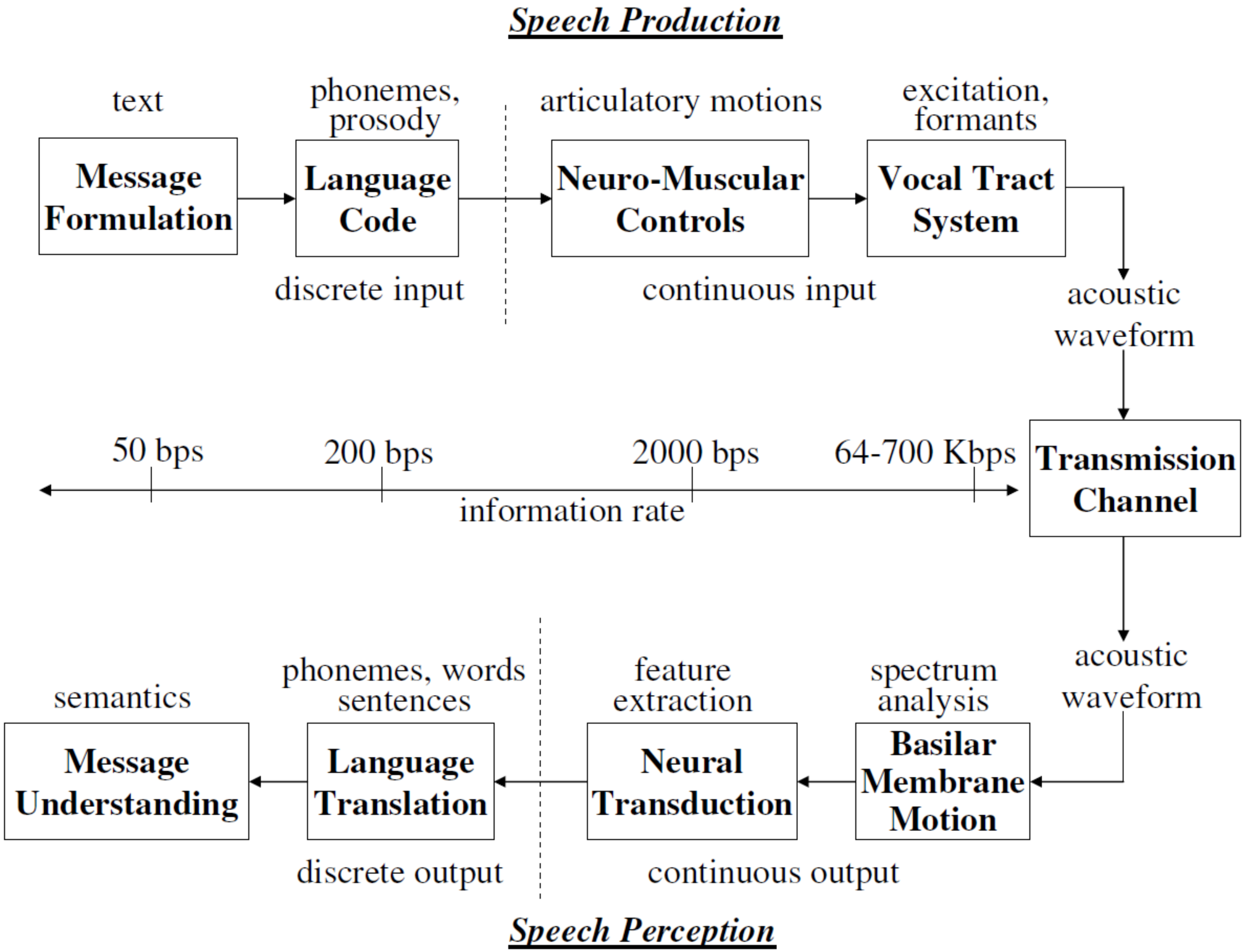
Applications

Basic research

The natural speech chain



The natural speech chain

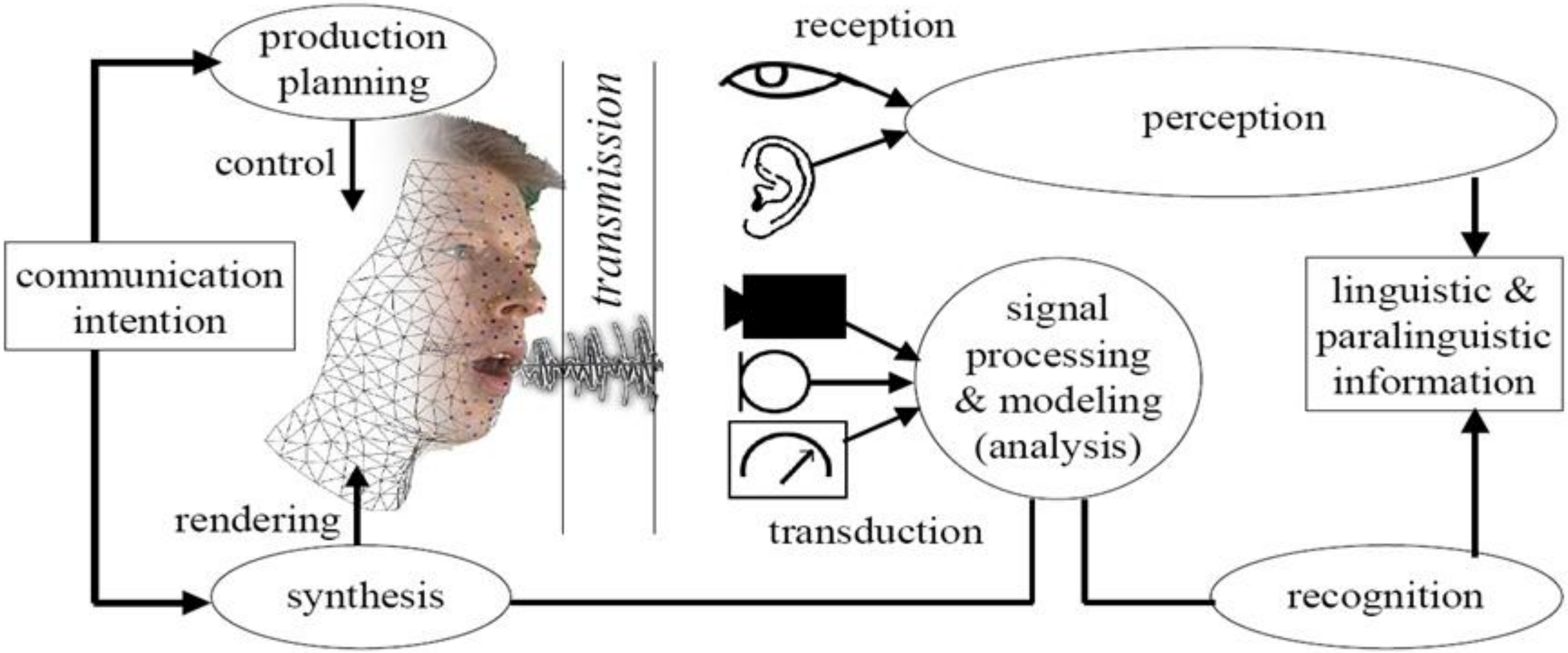


SmartLab
Intelligent Interactions
smartlab.tmit.bme.hu

BME TMIT

NVIDIA GPU EDUCATION CENTER

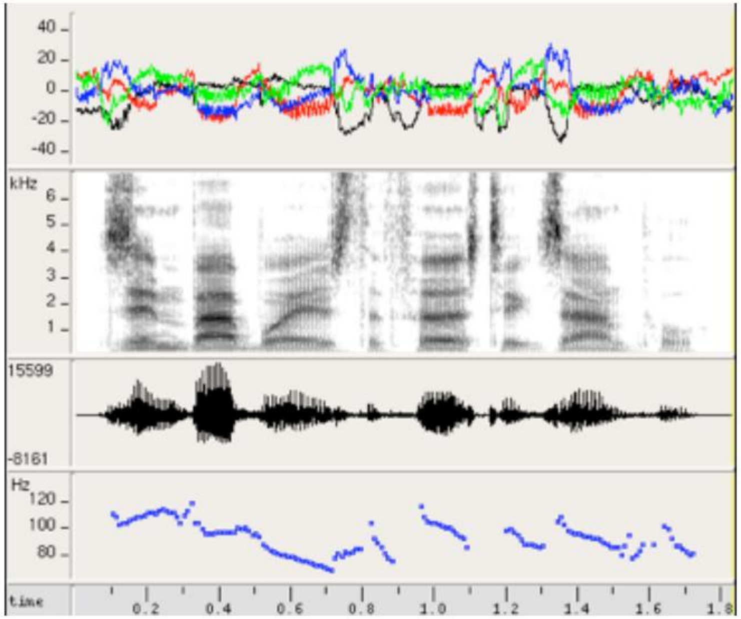
The natural speech chain



Machine speech generation

- Machine speech generation (speech synthesis ???)
  - Playback or concatenation based on simple symbols
    - Gramophone
    - Tape recorder
    - CDs …
  - From text (speech synthesis ???)
    - Text-to-speech (TTS)
    - Text-to-speech with additional information (concept-to-speech - CTS)
  - Audiovisual speech synthesis

Modeling approach



Speech in the acoustic domain

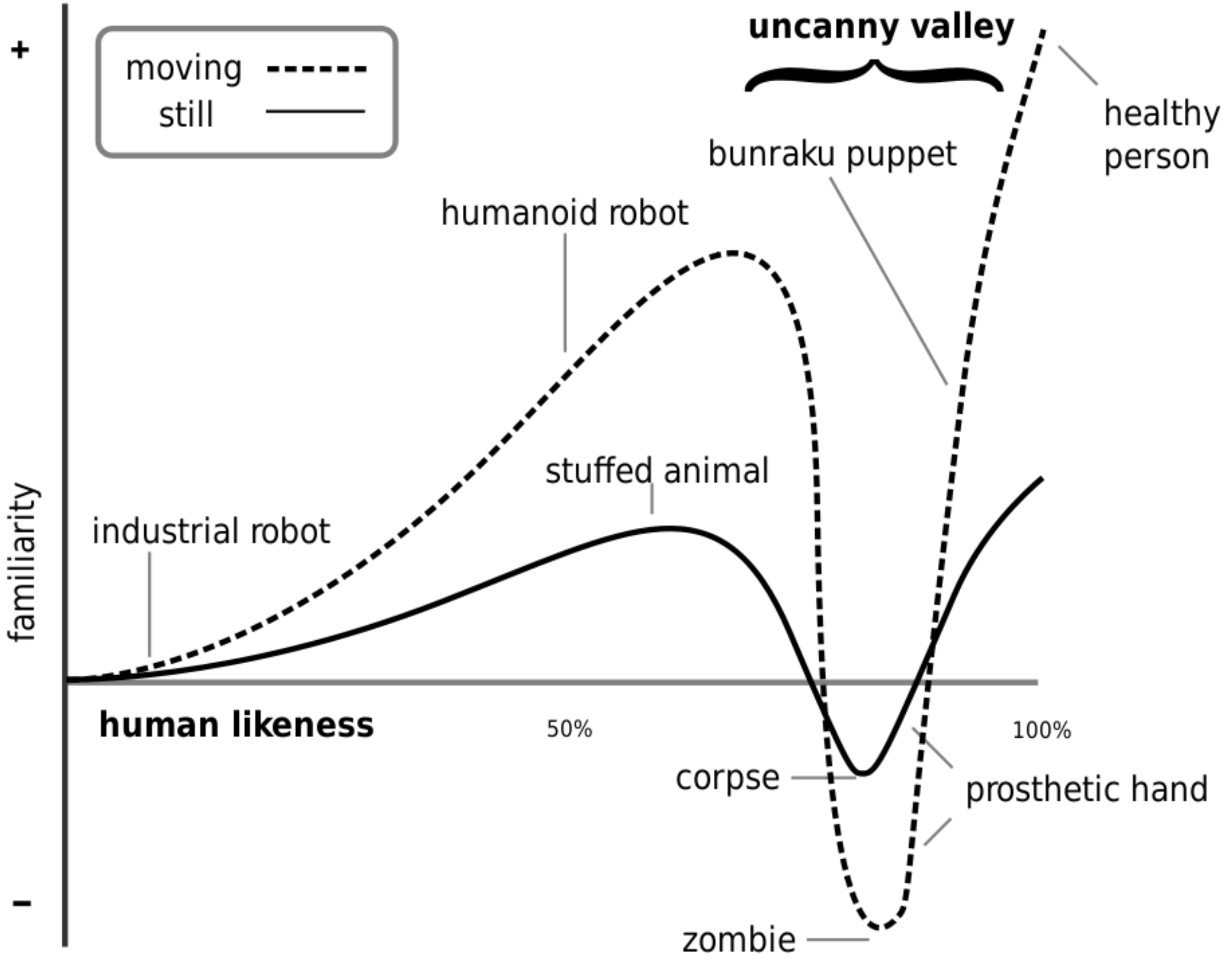
Production model

Speech in the articulatory range

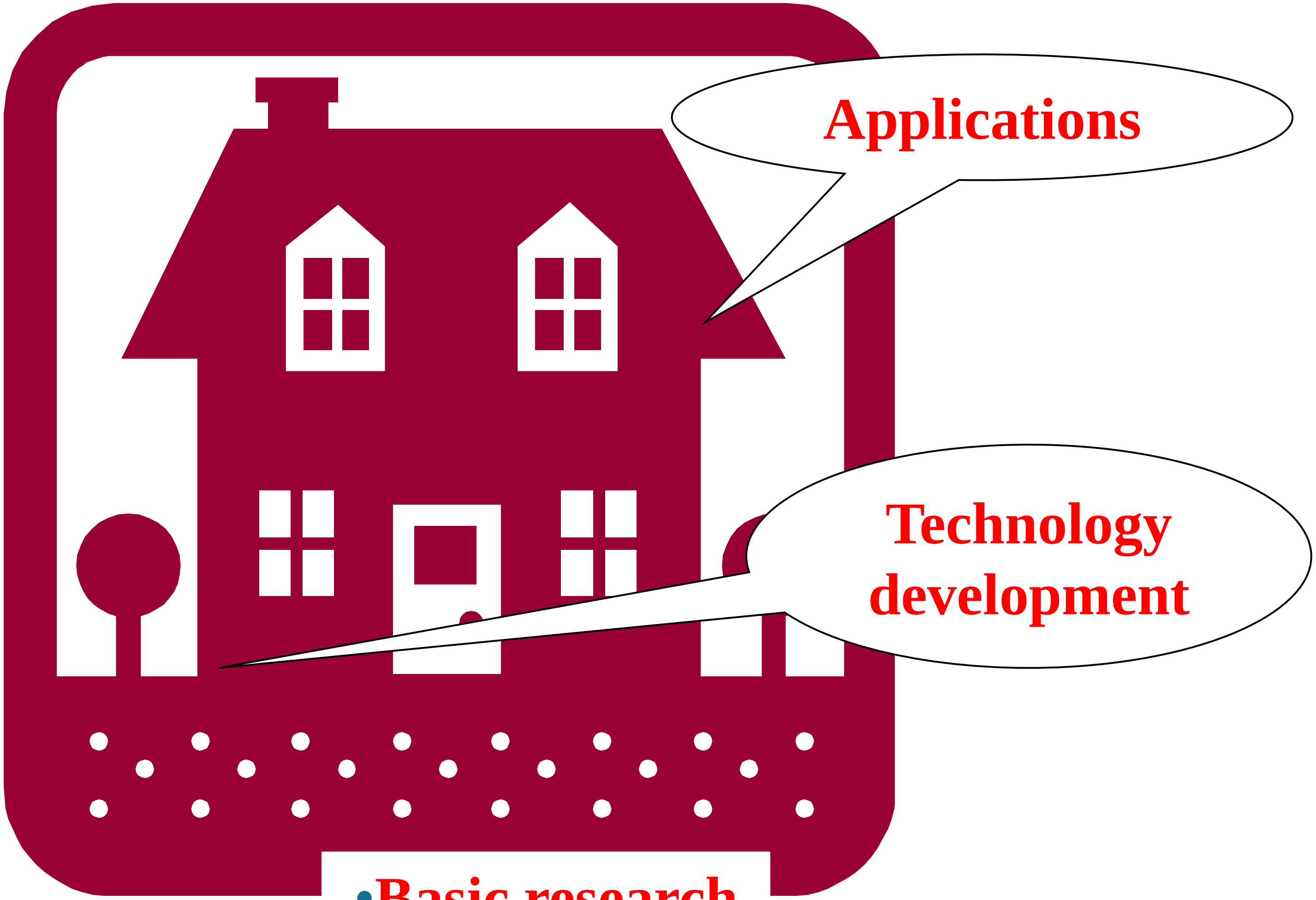
Generative model

Research challenge

Avoiding the uncanny valley



Speech technology research cycle



- Basic research

SmartLab
Intelligent Interactions
smartlab.tmit.bme.hu

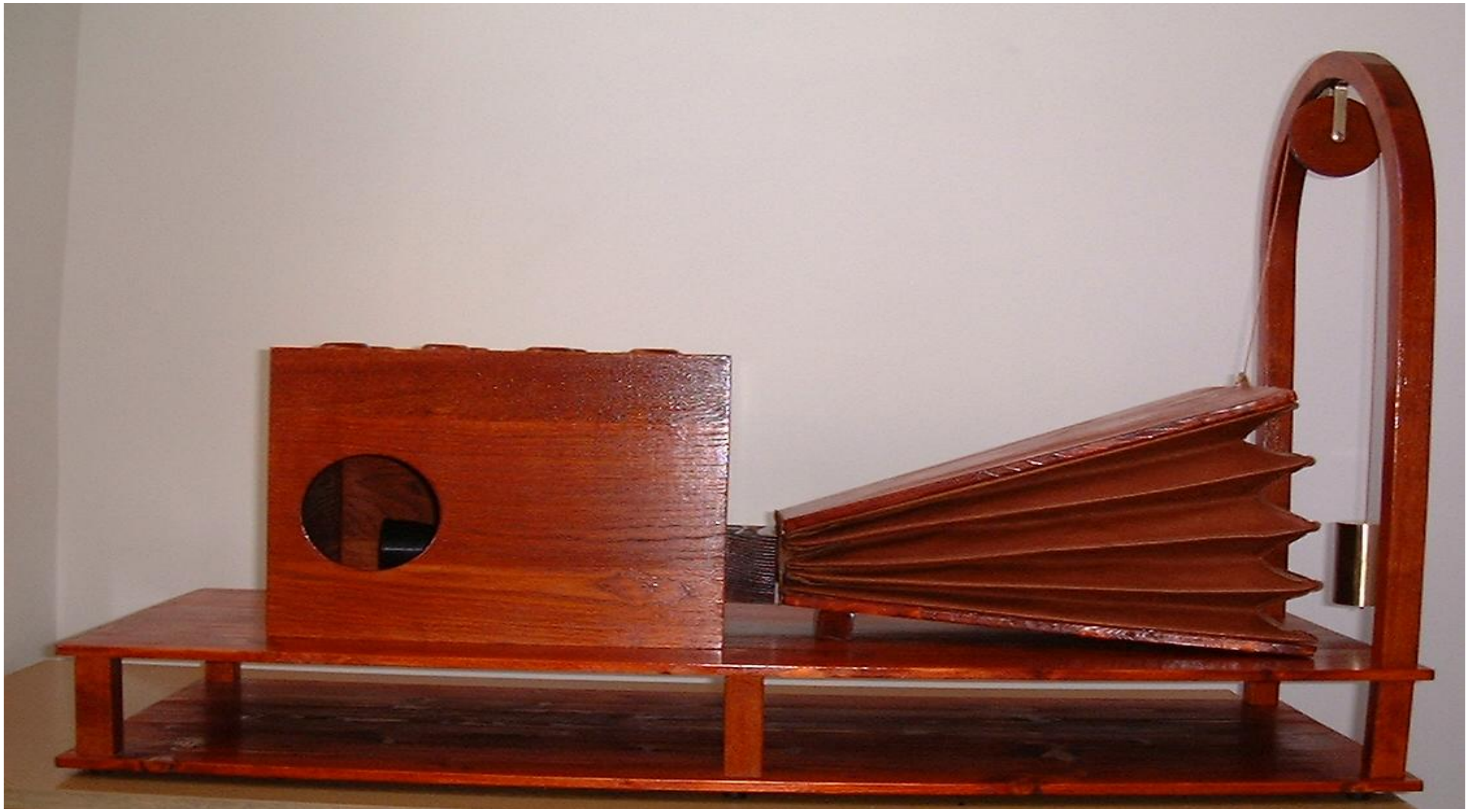
BME TMIT

NVIDIA
GPU
EDUCATION CENTER

Modeling approach

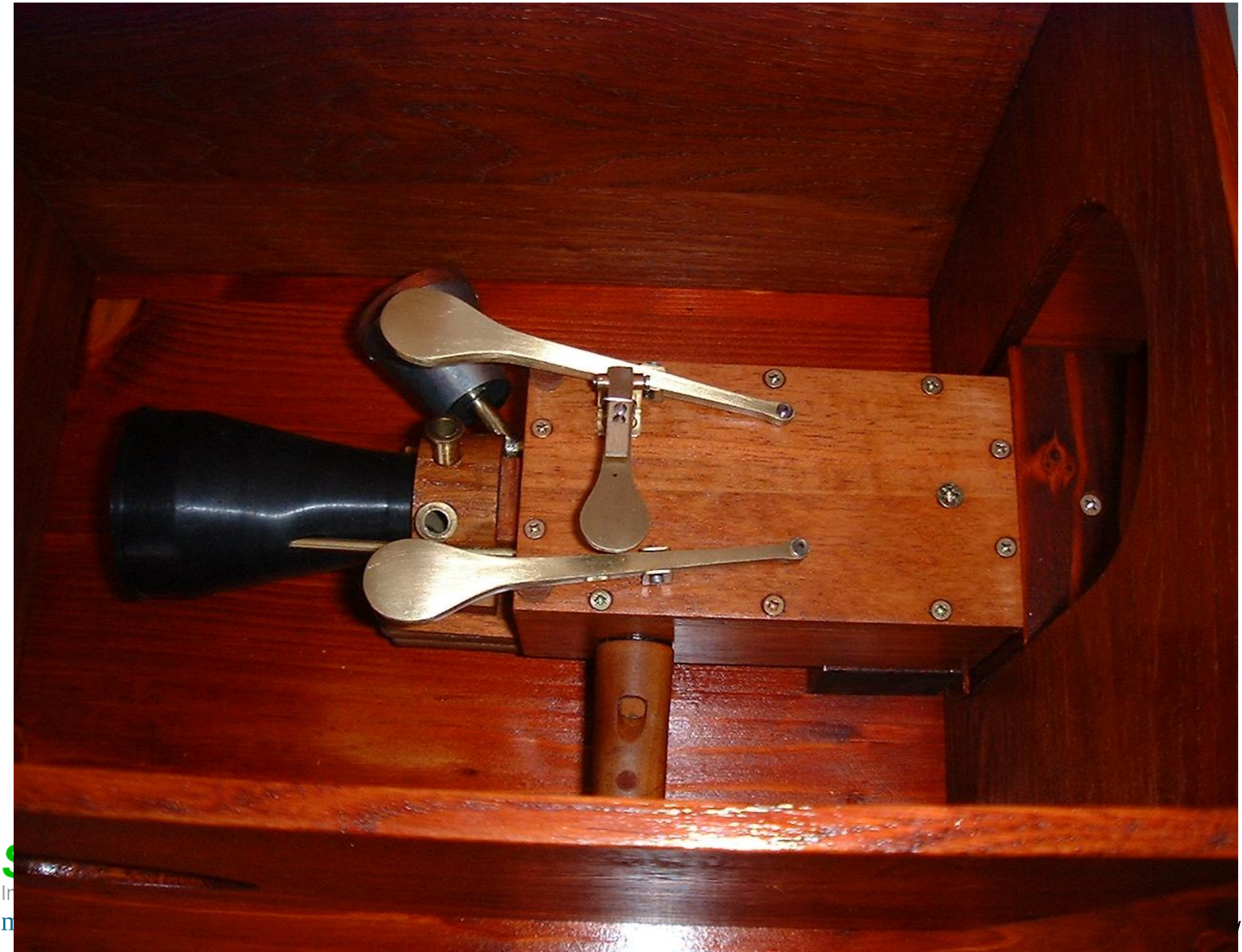
- Generative, articulation model
  - Segmental level
    - intelligibility of sounds
  - Suprasegmental level
    - correct intonation (prosody)
  - Non-linguistic features
    - individual tone, emotion, …

Kempelen's talking machine (1791)



RECONSTRUCTED, WORKING VERSION (2001 MTA Nyl , 2020 BME)

Kempelen's talking machine (1791)



Kempelen's talking machine (1791)

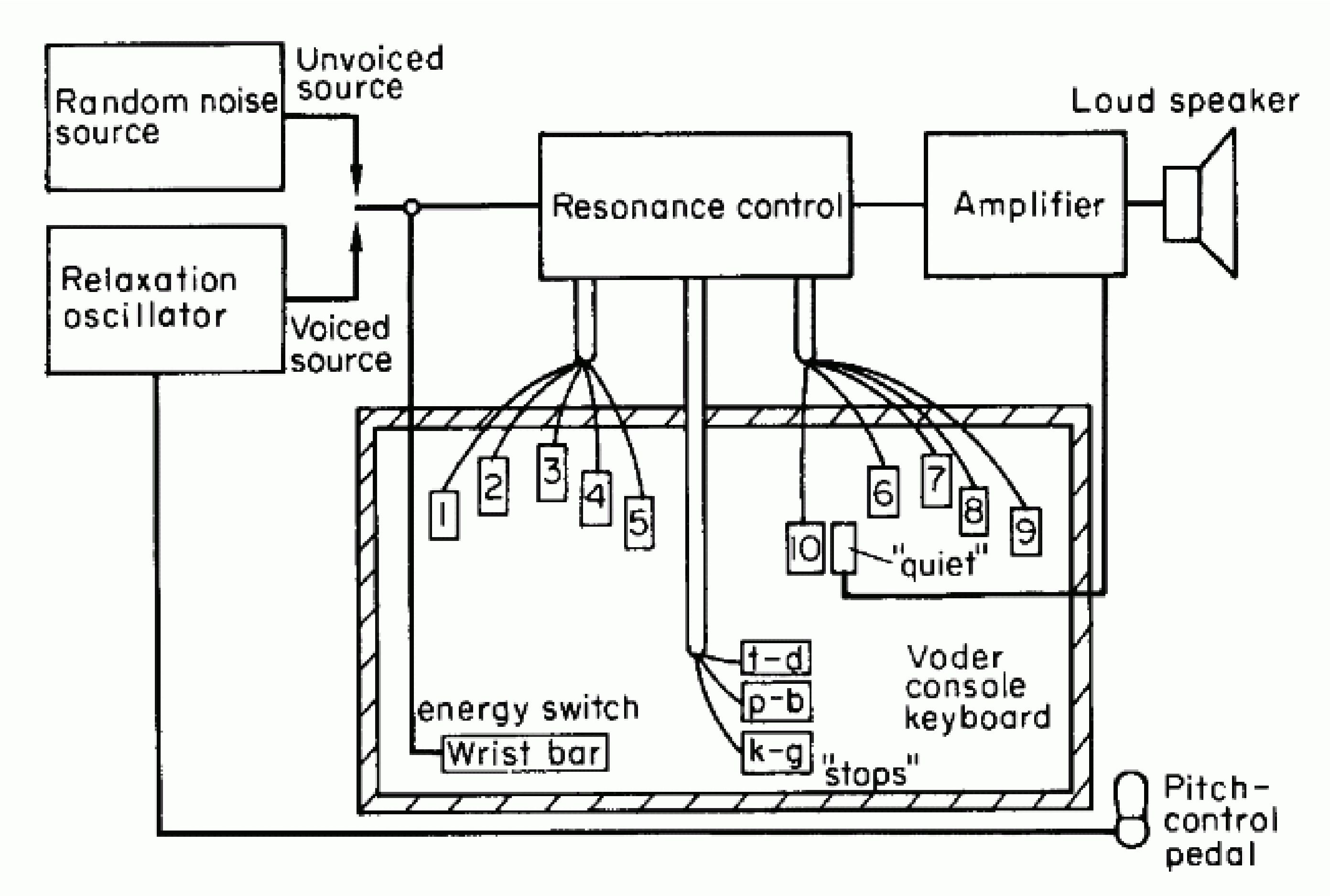
- Resources
  - Mechanical articulation model
  - Direct control of model elements by human hands
- Solution
  - Continuous / sustained sounds 🔈
  - "simple" sound concatanations 🔈

Kempelen's "choir" of machines

Source: Meeting of Kempelen machines, September 2019, Vienna

Dudley-Voder (1939)

Dudley-Voder (1939)



SmartLab
Intelligent Interactions
smartlab.tmit.bme.hu

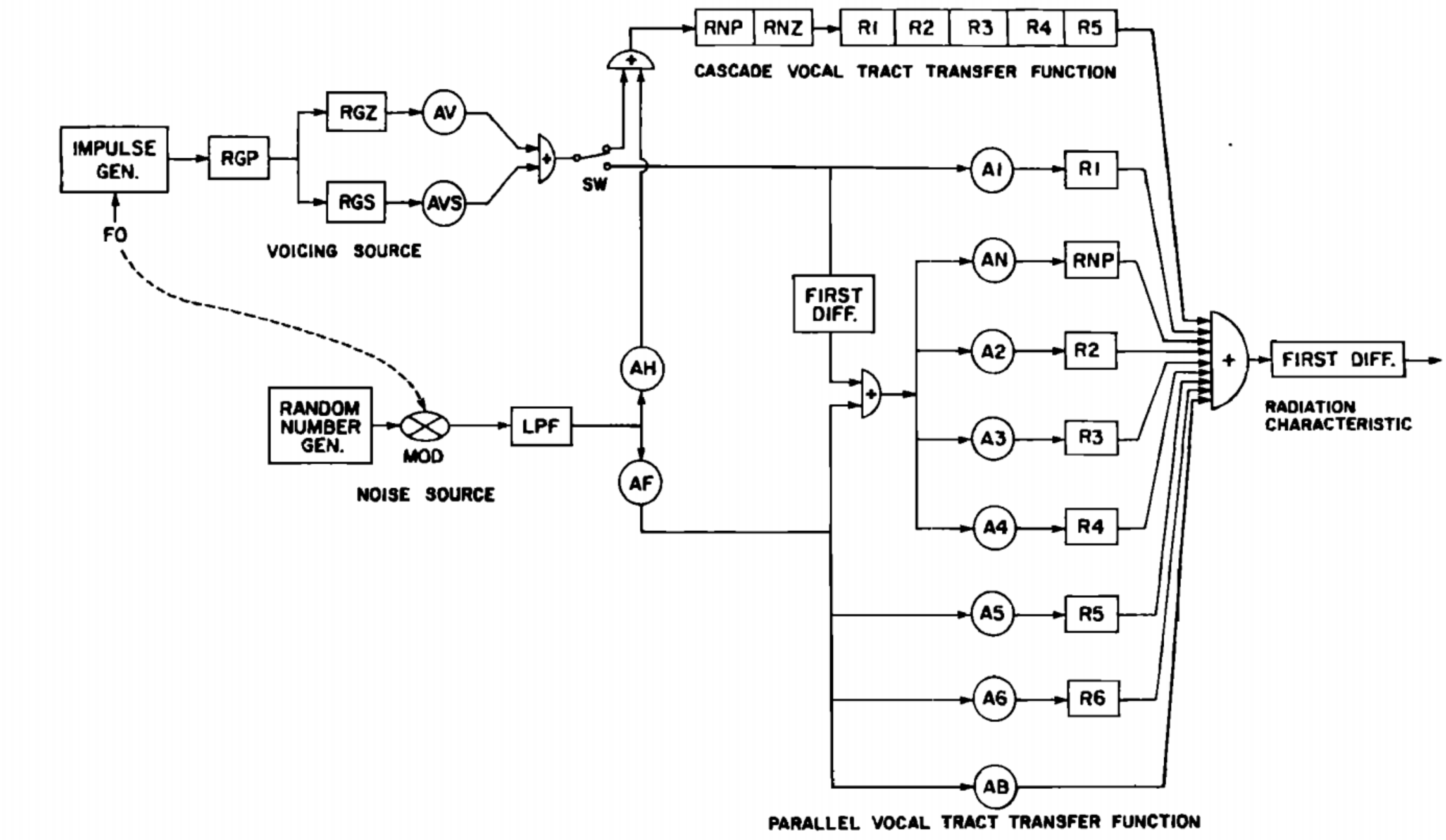
BME TMIT

NVIDIA GPU EDUCATION CENTER

Dudley-Voder (1939)

- Resources
  - Electro -mechanical articulation model
  - Keyboard control with human hand
- Solution
  - In principle, a loose vocabulary
  - Demonstration trick

Serial and parallel formant synthesizer (1983-84)

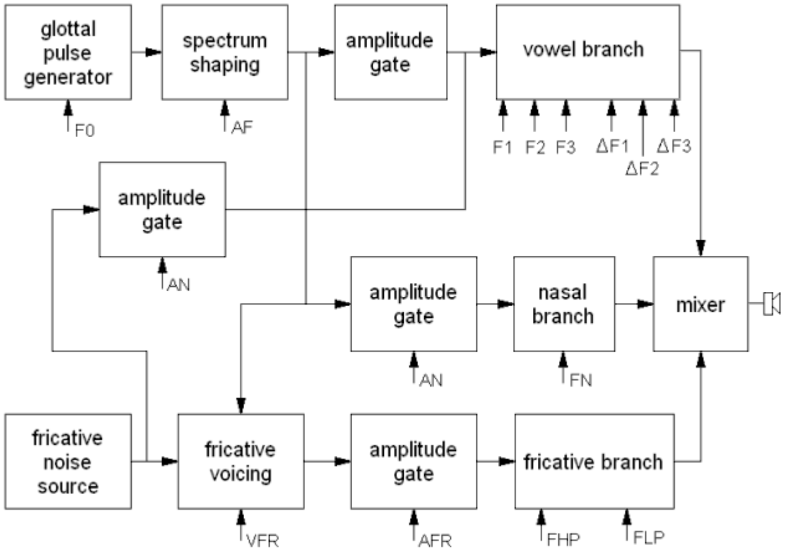


Holmes Olaszy Klatt
(1961 -1982 -84)

- Resources
  - Electronic formant/articulation model (equivalent LPC representation solved)
  - Rule-based processing
  - Control with computer
- Solution
  - In principle, a loose vocabulary
  - Singing
  - Small acoustic database (up to 1kByte)

Electronic formant synthesis solutions

KlatTalk/DecTalk 1982



HungaroVox 1982

TTS
The multilingual text-to-speech synthesizer for Europe by NIKOL Electronics
GERMAN
SPANISH
DUTCH
FINNISH
ITALIAN
HUNGARIAN
ESPERANTO
ARABIC
MULTIVOX
TM
hardware & software
The VOICE SYSTEM for your PC

MultiVox 1986-2002

Modeling approach

- Production model
  - Audio elements of varied details
    - Concatenation (modified by signal processing)
  - Segmental level
    - intelligibility of sounds
  - Suprasegmental level
    - correct intonation (prosody)
  - Non-linguistic features
    - individual tone, emotion, …

Miklós Bánó's talking machine (1916)

- June 21 , 1919. Hungarian Patent Office Office , Patent description , number 74361 , class IX/d

A talking machine suitable for reading any text

- Dr. Miklós Bánó certified engineer and economic engineer In Budapest
- The submission day:

November 30, 1916

Content of the patent
The Machine Reader (1916)

- 5 claims
  - 1. "A speaking machine capable of reproducing any text, characterized by recording individual sounds or groups of sounds separately and by playing these separately recorded sounds or groups of sounds in the order corresponding to them."
  - 2. "An embodiment of the machine protected against the claim of 1., characterized in that each sound of the language is recorded on a separate recording cylinder or disc"

Content of the patent
The Repentant (1916)

- 3. sound storage in a cylinder .
- 4. 1. + 3.
- 5. " Under Claim 1 protected machine implementation shape , characterized by the same recording cylinder or discs activation by typewriter-style buttons which can generate speech."

- (Naive) Example : per cylinders of *r+o+z* activated after each other melt into the word *rose*

Application example
Bánó (1916)

- "Connected to a typewriter...suitable for the visually impaired typist to listen to the written text"
- "The speech impaired typist ... should be able to communicate what he has to say aloud with the help of the device"
- ..."also suitable for the reproduction of prescribed or duplicated texts"

Miklós Bánó's talking machine (1916)

- Resources
  - Electro-mechanical speech concatenation model
  - Keyboard control with human hand
- Solution
  - No working sample
    - Sustained speech sounds
    - In principle, unlimited vocabulary
    - Simulation reconstruction (2010)

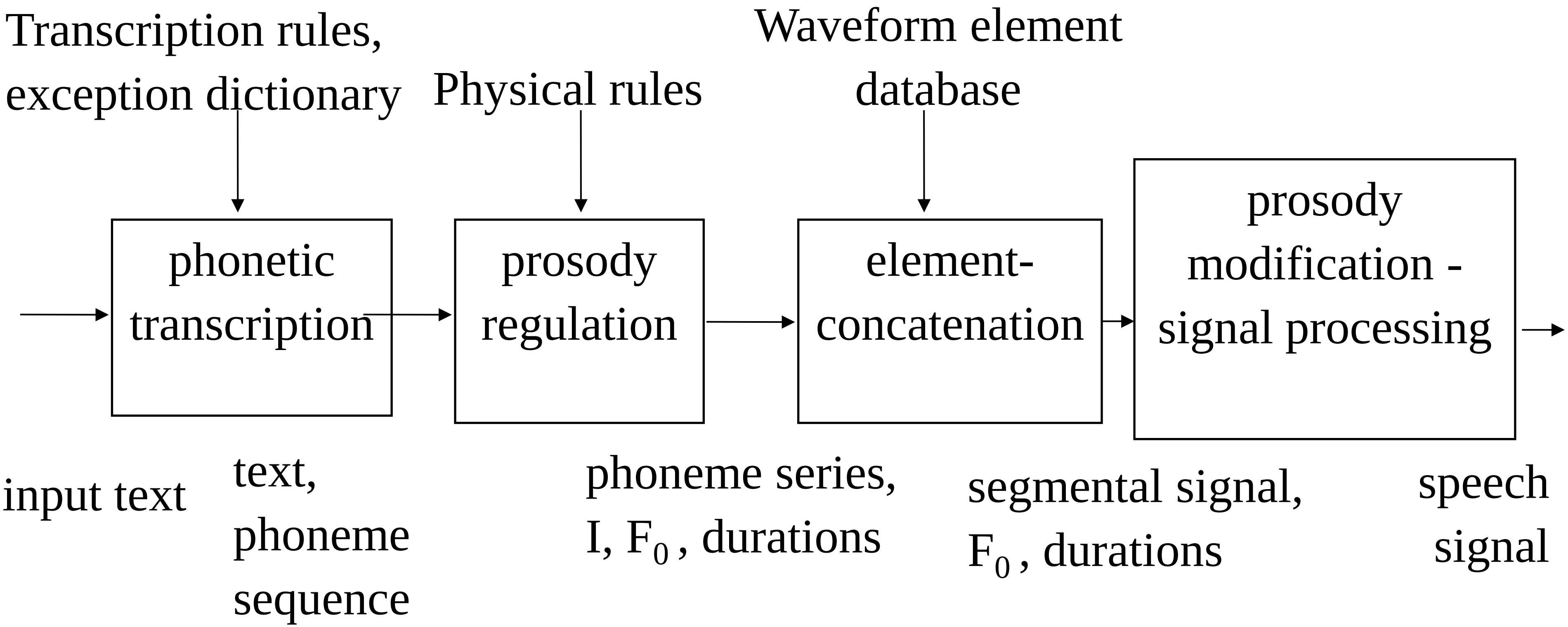
Overview of machine speech generation methods

Speech synthesis method	Prosody production	Speech database type
“classical” formant synthesis (the starting method)	rule-based, with encoder control parameters	parametric (formant filter model)
element-concatenation (diad/diphone)	rule-based, with waveform modification	sound, diad/diphone waveform elements (logatoms)
element-concatenation (triad/triphone)	rule-based, with waveform modification	sound, triad/triphone and diad/diphone waveform elements (logatoms)
element selection (corpus)	indirect, pattern search based on the time scale of the current sentence, typically without waveform modification	large waveform database (from readings) of variable-sized elements (words, word strings, sentences, etc.)
statistical parametric	with a statistical (HMM or DNN) model, which operates based on parameter n-grams at the sentence level	parametric (LPC, harmonic+noise, sinusoidal, etc.)
waveform-based statistical parametric (WaveNet /DNN)	with statistical (DNN) model	neural network parameters (learning from waveform and direct generation)

Database features

Sign	Audio recording length/database length (minutes)	Gender	Language	Goals
DIAD1- DIAD4	28/2.5	man	Hungarian	Element-concatenation (diad-triad / diphone-triphone) research
DIAD5–DIAD8	approx. 28 minutes/2.5 minutes	woman		
TRIAD1	approx. 120min/32 min	man		
TRIAD2	approx. 120min/32 min	woman		
TIME1	630 minutes	woman	Hungarian	Item selection (corpus-based) research
FON1	100 minutes			
PAGE1	110 minutes			Experimental station passenger information system
ARU1	330 minutes			
UGYF1	505 minutes			
SZAM1	10 minutes/XXXX			Reading numbers up to 1 billion
SZAM2	10 minutes/XXXXX	man		
RADIO	516 minutes	3 men		
FON2-5	approx. 130 minutes/person	4 women		Statistician
FON6-10	approx. 130 minutes/person	5 men		parametric (HMM and
BEA1	30 minutes	woman		DNN) research
BEA2	31 minutes	man		
GABOR	3 minutes	man		Spemoticon research

Element concatenation model



Transcription rules, exception dictionary

Physical rules

Waveform element database

phonetic transcription

prosody regulation

element-concatenation

prosody modification - signal processing

input text

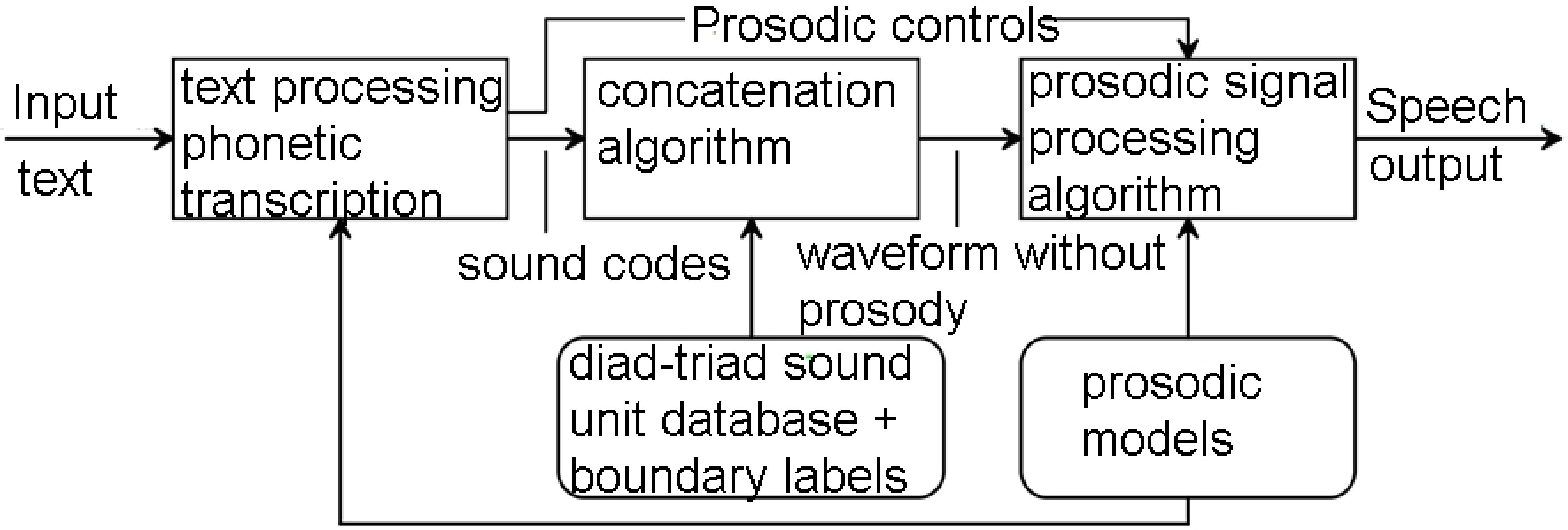
text, phoneme sequence

phoneme series, I, F_0, durations

segmental signal, F_0, durations

speech signal

Diad- triad concatenation ProfiVox (1999-)


Prosodic controls
Input text
text processing phonetic transcription
concatenation algorithm
prosodic signal processing algorithm
Speech output
sound codes
waveform without prosody
diad-triad sound unit database + boundary labels
prosodic models

papa = _p pa ap pa a_

or

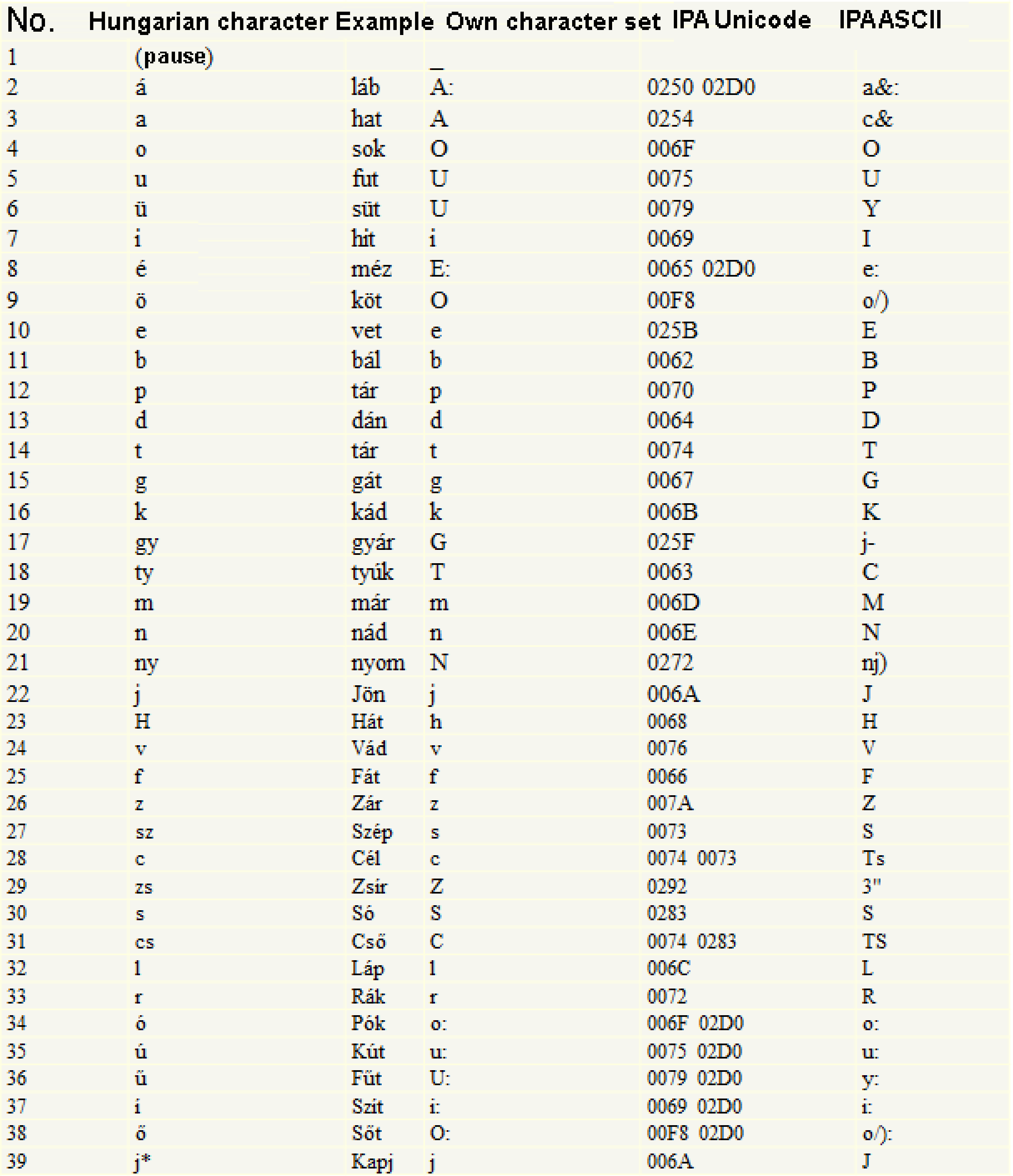
p pap pa

diad + triad

Hungarian sound definition for synthesis

- Hungarian speech with 39 speech sounds
  - 24 consonants (C)
  - 14 vowels (V)
  - plus the pause (_ sign) can be covered
    - long consonants are generated from their short versions using signal processing.
    - We are very sensitive to the quality of vowels, so it is advisable to treat short and long versions separately.

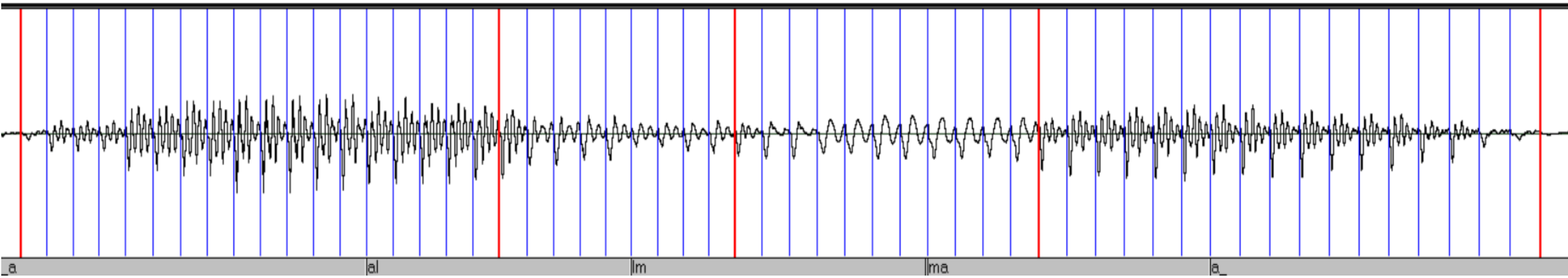
Speech sound set



No.	Hungarian character	Example	Own character set	IPA Unicode	IPA ASCII
1	(pause)		_		
2	á	láb	A:	0250 02D0	a&:
3	a	hat	A	0254	c&
4	o	sok	O	006F	O
5	u	fut	U	0075	U
6	ü	süt	U	0079	Y
7	i	hit	i	0069	I
8	é	méz	E:	0065 02D0	e:
9	ö	köt	O	00F8	o/)
10	e	vet	e	025B	E
11	b	bál	b	0062	B
12	p	tár	p	0070	P
13	d	dán	d	0064	D
14	t	tár	t	0074	T
15	g	gát	g	0067	G
16	k	kád	k	006B	K
17	gy	gyár	G	025F	j-
18	ty	tyúk	T	0063	C
19	m	már	m	006D	M
20	n	nád	n	006E	N
21	ny	nyom	N	0272	nj)
22	j	Jön	j	006A	J
23	H	Hát	h	0068	H
24	v	Vád	v	0076	V
25	f	Fát	f	0066	F
26	z	Zár	z	007A	Z
27	sz	Szép	s	0073	S
28	c	Cél	c	0074 0073	Ts
29	zs	Zsir	Z	0292	3"
30	s	Só	S	0283	S
31	cs	Cső	C	0074 0283	TS
32	l	Láp	l	006C	L
33	r	Rák	r	0072	R
34	ó	Pók	o:	006F 02D0	o:
35	ú	Kút	u:	0075 02D0	u:
36	ű	Fűt	U:	0079 02D0	y:
37	í	Szit	i:	0069 02D0	i:
38	ő	Sőt	O:	00F8 02D0	o/):
39	j*	Kapj	j	006A	J

SmartLab
Intelligent Interactions
smartlab.tmit.bme.hu

The 5 diads of the word alma

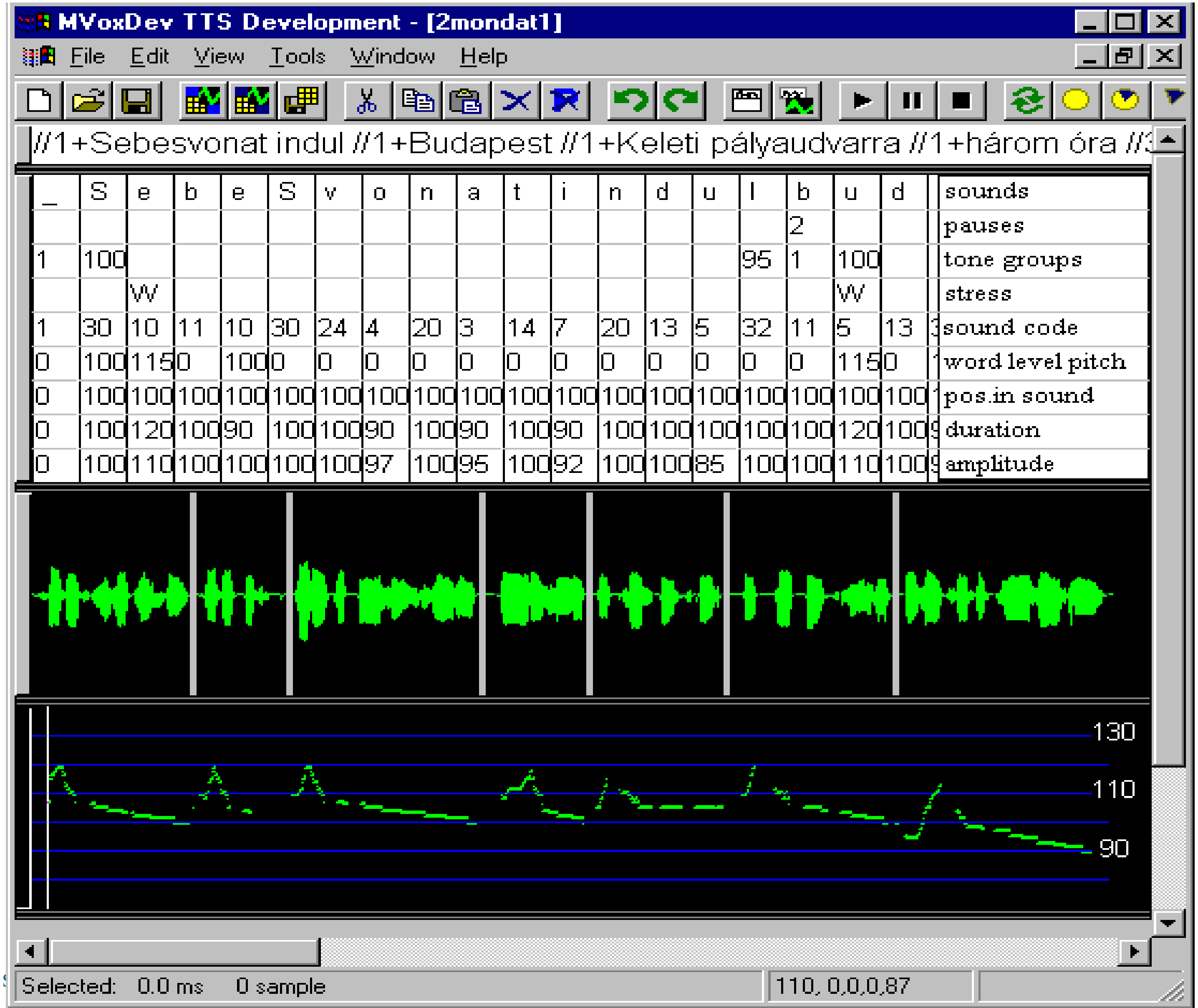


- The boundaries of the sound periods are shown in blue,
- the sound boundary in red
- the boundaries of the diads are the lines next to the markings in the lower gray bar

The number of diad variants required for the synthesis of the Hungarian language

Audio	CV	VC	CC	VV	_V and V_	_C and C_	Total
Quantity	336	336	576	196	28	48	1520

Prosodic units


MVoxDev TTS Development - [2mondat1]
File Edit View Tools Window Help

//1+Sebesvonat indul //1+Budapest //1+Keleti pályaudvarra //1+három óra //3

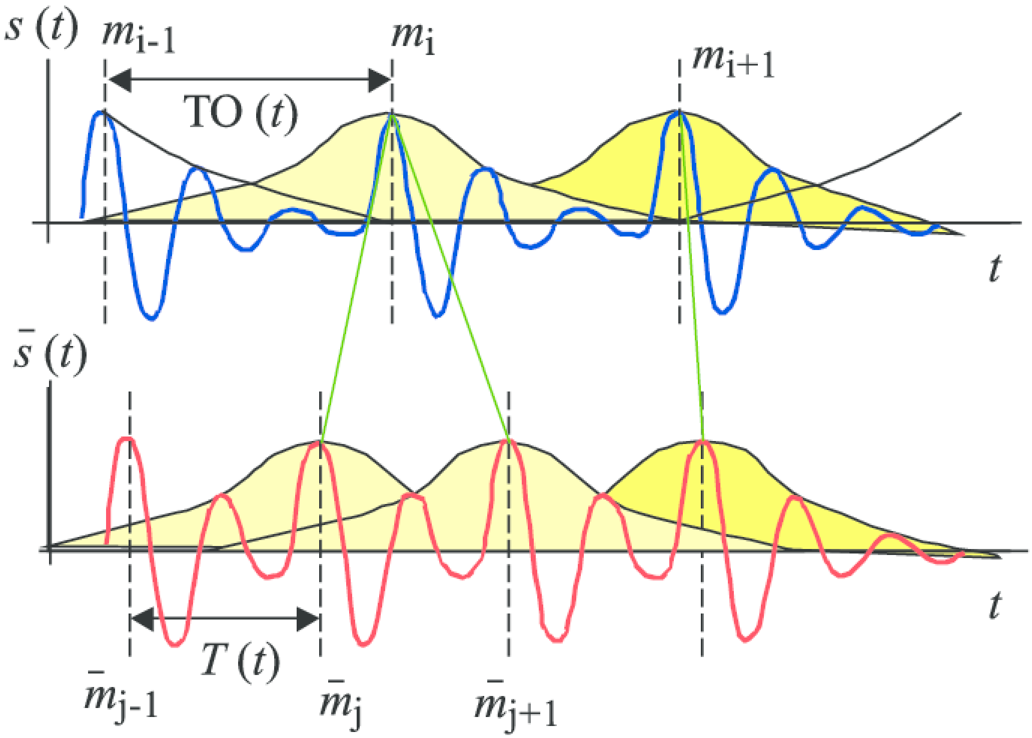
_	S	e	b	e	S	v	o	n	a	t	i	n	d	u	l	b	u	d		sounds
																2				pauses
1	100														95	1	100			tone groups
		W															W			stress
1	30	10	11	10	30	24	4	20	3	14	7	20	13	5	32	11	5	13	3	sound code
0	100	115	0	100	0	0	0	0	0	0	0	0	0	0	0	0	115	0	1	word level pitch
0	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	1	pos.in sound
0	100	120	100	90	100	100	90	100	90	100	90	100	100	100	100	100	120	100	9	duration
0	100	110	100	100	100	100	97	100	95	100	92	100	100	85	100	100	110	100	9	amplitude

Selected: 0.0 ms 0 sample

110, 0,0,0,87

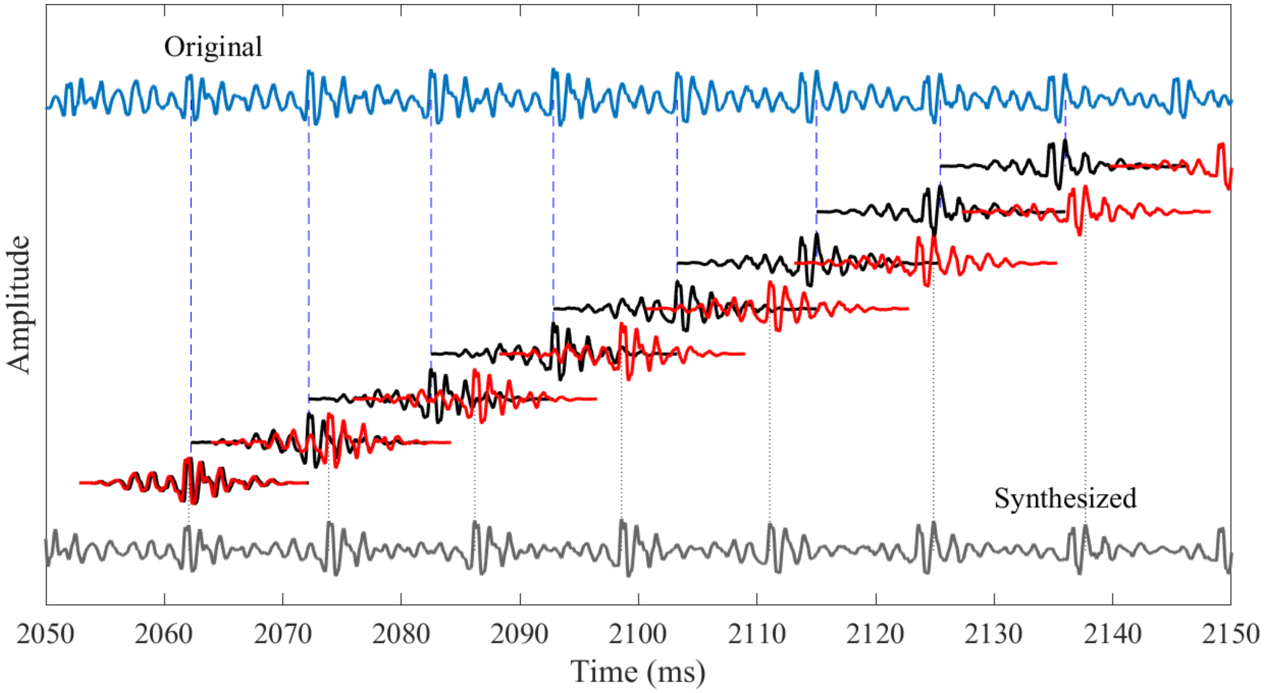
F0/sound duration modification PSOLA

Pitch Synchronous OverLap-Add



https://www.researchgate.net/publication/252824047_Voice_Conversion_using_Pitch_Shifting_Algorithm_by_Time_Stretchingwith_PSOLA_and_Re-Sampling/figures?lo=1

F0/sound duration modification PSOLA



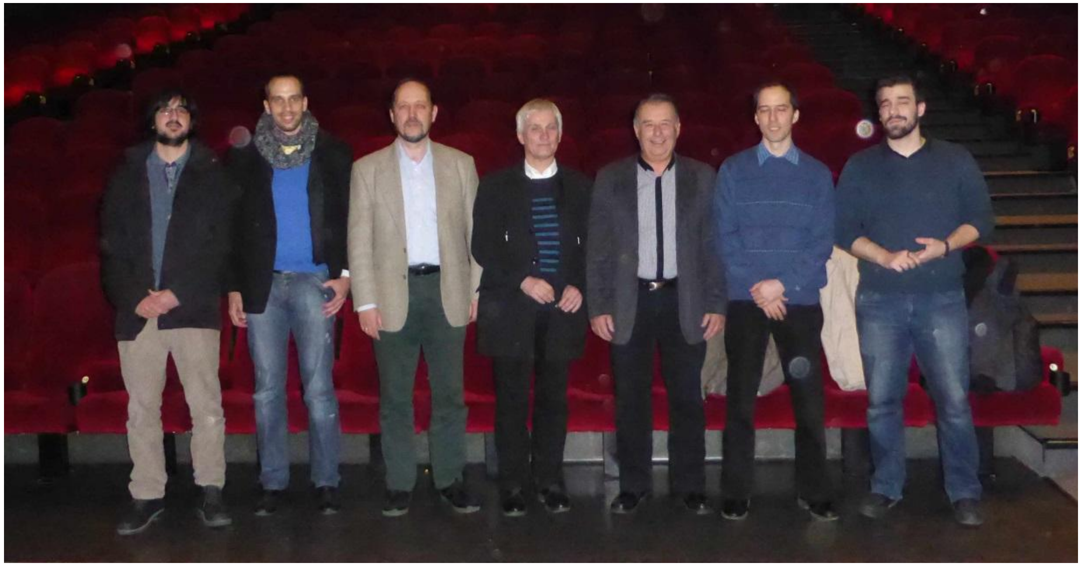
https://wiki.aalto.fi/pages/viewpage.action?pageId=155477136

Diad- triad concatenation ProfiVox (1999-)

- Resources
  - Digital diad - triad voice concatenation model
  - Rule-based processing
  - Control with computer
- Solution
  - Practically unlimited vocabulary
  - Flexible prosodic control
  - Acoustic database 1.5-60Mbyte

Stephen Hawking's machine voice (English - Hungarian)

Dectalk 1982



ProfiVox 2000 – 2014

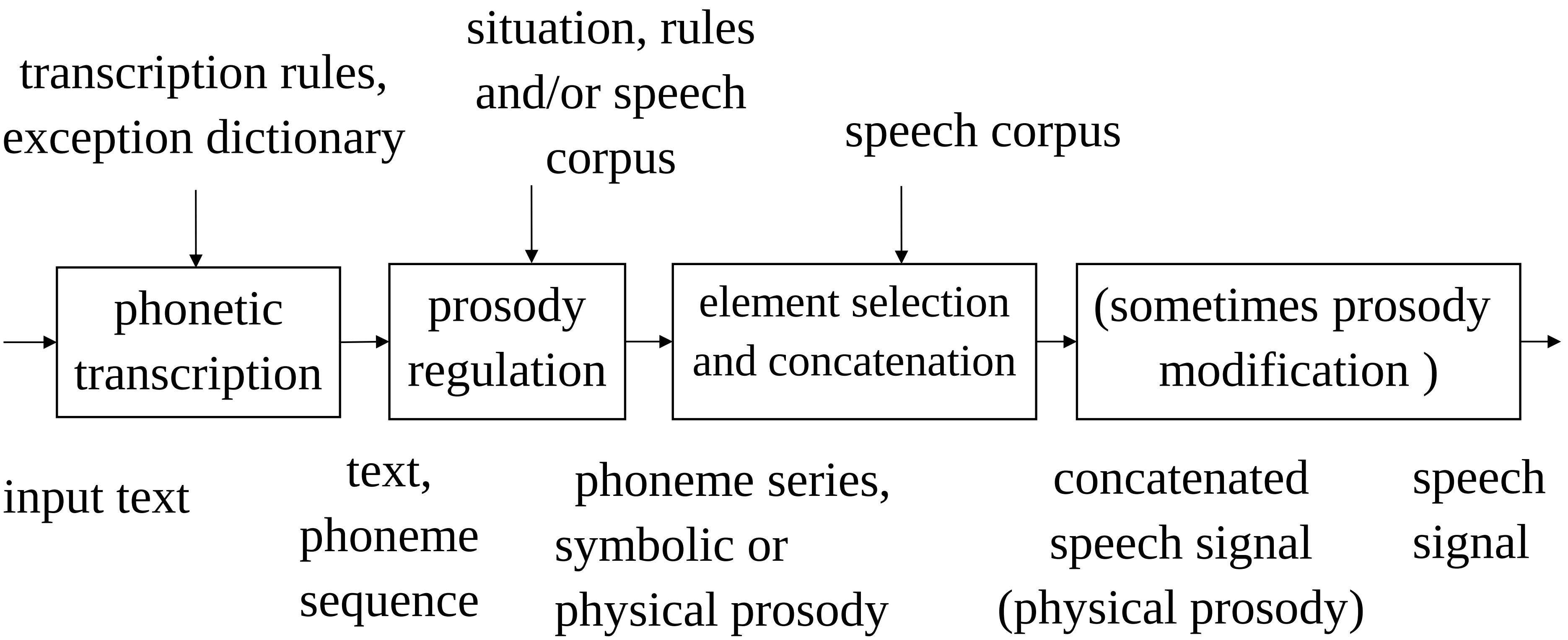
Hungarian dubbing: Theory of Everything movie (Theory of Everything)

Corpus approach
(2003-)

- N * Gbytes of storage space available
  - Store as many items as possible
  - Limit: limits of human reading with the same tone
- Corpus-based concatenation synthesis

  (corpus-based , unit-selection synthesis)
  - Prosody can be controlled better indirectly
  - Exceptions are difficult to implement

Corpus-based, waveform element selection speech synthesizer model



transcription rules, exception dictionary

situation, rules and/or speech corpus

speech corpus

phonetic transcription

prosody regulation

element selection and concatenation

(sometimes prosody modification)

input text

text, phoneme sequence

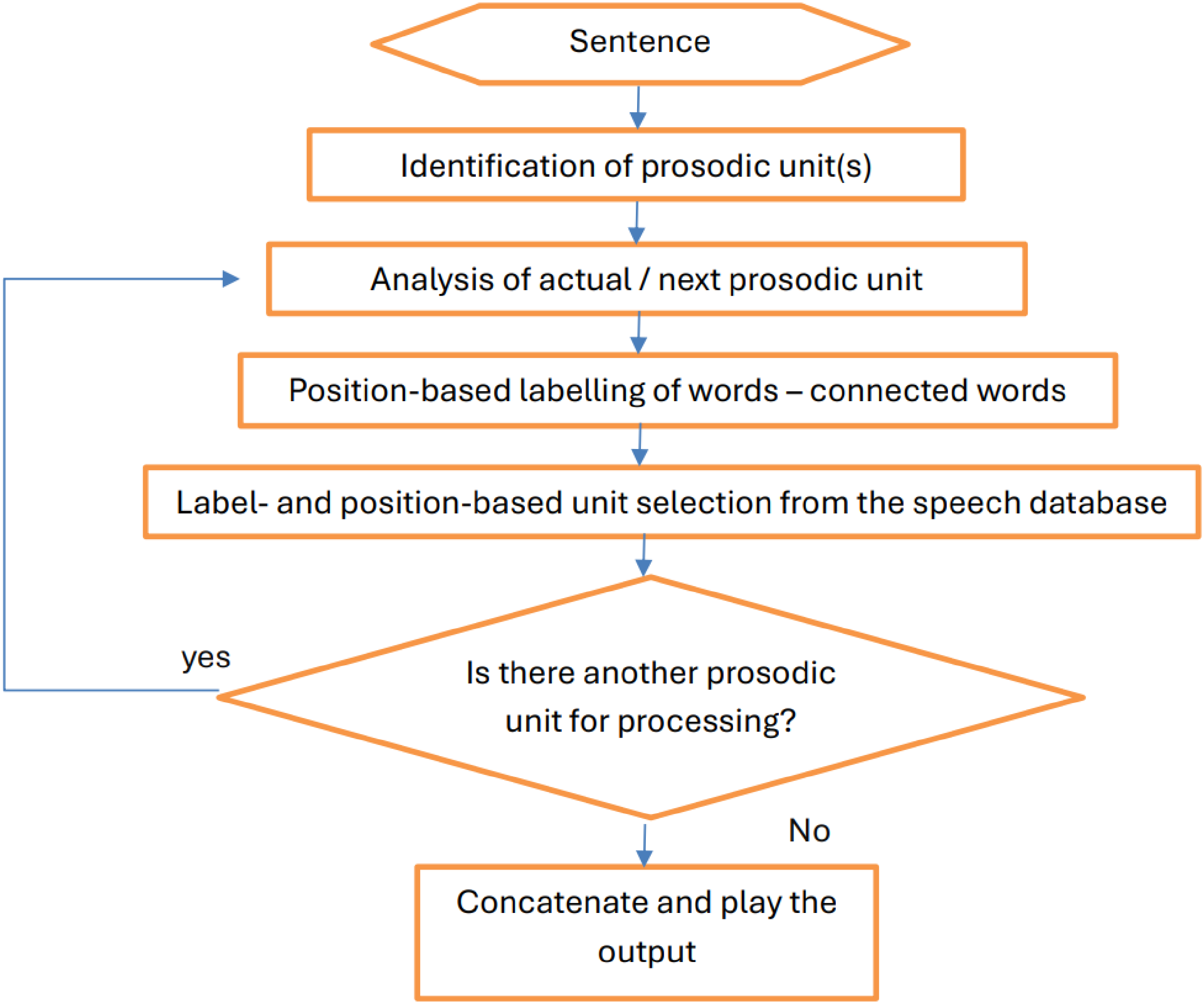
phoneme series, symbolic or physical prosody

concatenated speech signal (physical prosody)

speech signal

smartlab.tmit.bme.hu

Indirect prosodic model



Sentence

Identification of prosodic unit(s)

Analysis of actual / next prosodic unit

Position-based labelling of words – connected words

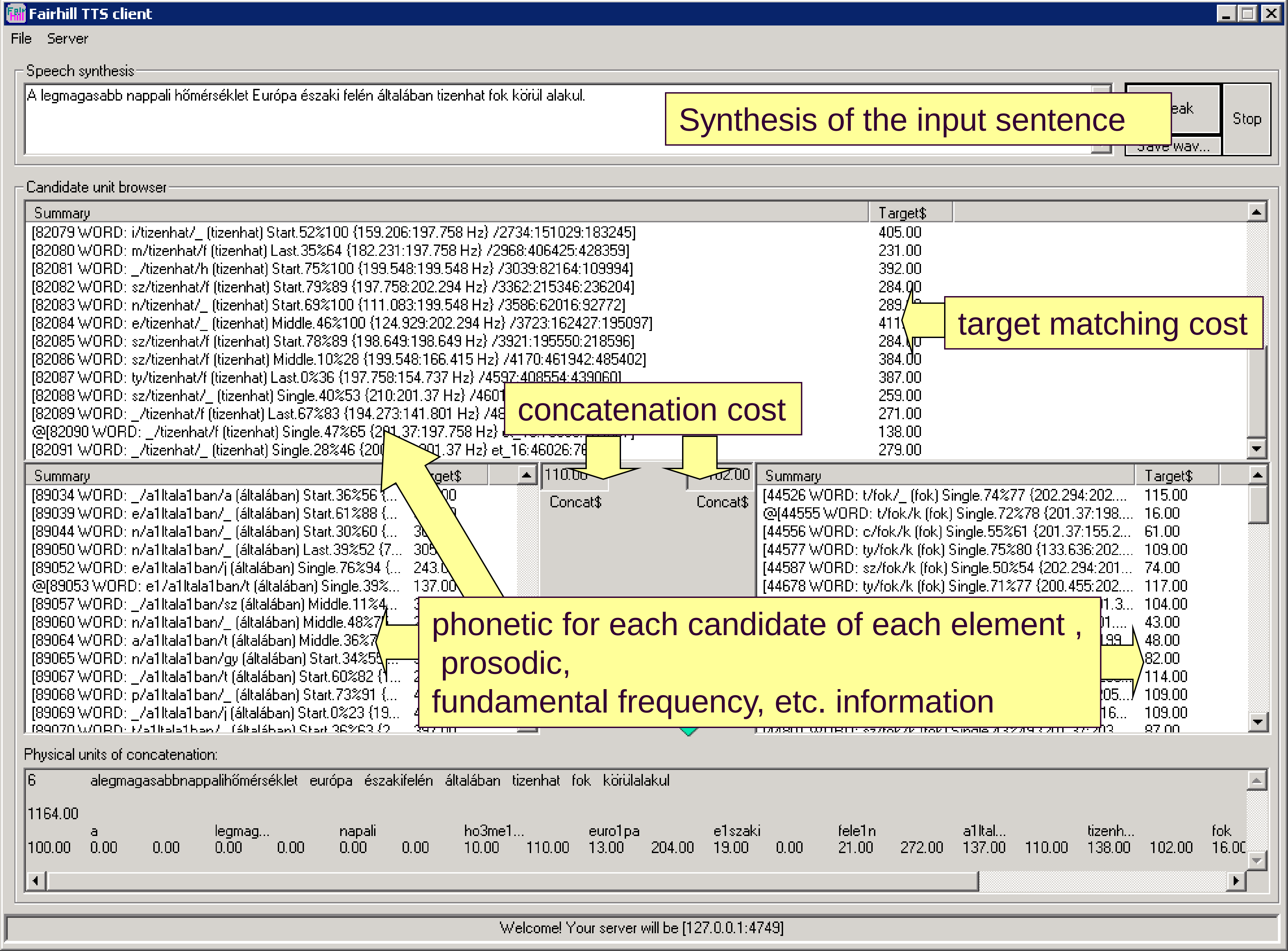
Label- and position-based unit selection from the speech database

Is there another prosodic unit for processing?

yes

No

Concatenate and play the output



A legmagasabb nappali hőmérséklet Európa északi felén általában tizenhat fok körül alakul.

Synthesis of the input sentence

```
Summary                                                                                      Target$
[82079 WORD: i/tizenhat/_ (tizenhat) Start.52%100 {159.206:197.758 Hz} /2734:151029:183245]  405.00
[82080 WORD: m/tizenhat/f (tizenhat) Last.35%64 {182.231:197.758 Hz} /2968:406425:428359]    231.00
[82081 WORD: _/tizenhat/h (tizenhat) Start.75%100 {199.548:199.548 Hz} /3039:82164:109994]   392.00
[82082 WORD: sz/tizenhat/f (tizenhat) Start.79%89 {197.758:202.294 Hz} /3362:215346:236204]  284.00
[82083 WORD: n/tizenhat/_ (tizenhat) Start.69%100 {111.083:199.548 Hz} /3586:62016:92772]    289.
[82084 WORD: e/tizenhat/_ (tizenhat) Middle.46%100 {124.929:202.294 Hz} /3723:162427:195097] 411
[82085 WORD: sz/tizenhat/f (tizenhat) Start.78%89 {198.649:198.649 Hz} /3921:195550:218596]  284.
[82086 WORD: sz/tizenhat/f (tizenhat) Middle.10%28 {199.548:166.415 Hz} /4170:461942:485402] 384.00
[82087 WORD: ty/tizenhat/f (tizenhat) Last.0%36 {197.758:154.737 Hz} /4597:408554:439060]    387.00
[82088 WORD: sz/tizenhat/_ (tizenhat) Single.40%53 {210:201.37 Hz} /4601                     259.00
[82089 WORD: _/tizenhat/f (tizenhat) Last.67%83 {194.273:141.801 Hz} /48                     271.00
@[82090 WORD: _/tizenhat/f (tizenhat) Single.47%65 {201.37:197.758 Hz}                       138.00
[82091 WORD: _/tizenhat/_ (tizenhat) Single.28%46 {200 201.37 Hz} et_16:46026:76             279.00
```

target matching cost

concatenation cost

```
Summary                                                      Target$
[89034 WORD: _/a1ltala1ban/a (általában) Start.36%56 {...
[89039 WORD: e/a1ltala1ban/_ (általában) Start.61%88 {...
[89044 WORD: n/a1ltala1ban/_ (általában) Start.30%60 {...    30
[89050 WORD: n/a1ltala1ban/_ (általában) Last.39%52 {7...    305
[89052 WORD: e/a1ltala1ban/j (általában) Single.76%94 {...   243.0
@[89053 WORD: e1/a1ltala1ban/t (általában) Single.39%...     137.00
[89057 WORD: _/a1ltala1ban/sz (általában) Middle.11%4...     3
[89060 WORD: n/a1ltala1ban/_ (általában) Middle.48%7         2
[89064 WORD: a/a1ltala1ban/t (általában) Middle.36%7
[89065 WORD: n/a1ltala1ban/gy (általában) Start.34%55...     3
[89067 WORD: _/a1ltala1ban/t (általában) Start.60%82 {1...   2
[89068 WORD: p/a1ltala1ban/_ (általában) Start.73%91 {...    4
[89069 WORD: _/a1ltala1ban/j (általában) Start.0%23 {19...   4
```

```
110.00          102.00
Concat$         Concat$
```

```
Summary                                                        Target$
[44526 WORD: t/fok/_ (fok) Single.74%77 {202.294:202....     115.00
@[44555 WORD: t/fok/k (fok) Single.72%78 {201.37:198....     16.00
[44556 WORD: c/fok/k (fok) Single.55%61 {201.37:155.2...     61.00
[44577 WORD: ty/fok/k (fok) Single.75%80 {133.636:202....    109.00
[44587 WORD: sz/fok/k (fok) Single.50%54 {202.294:201...     74.00
[44678 WORD: ty/fok/k (fok) Single.71%77 {200.455:202....    117.00
                                                01.3...      104.00
                                                01....       43.00
                                                199          48.00
                                                             82.00
                                                             114.00
                                                205...       109.00
                                                16...        109.00
```

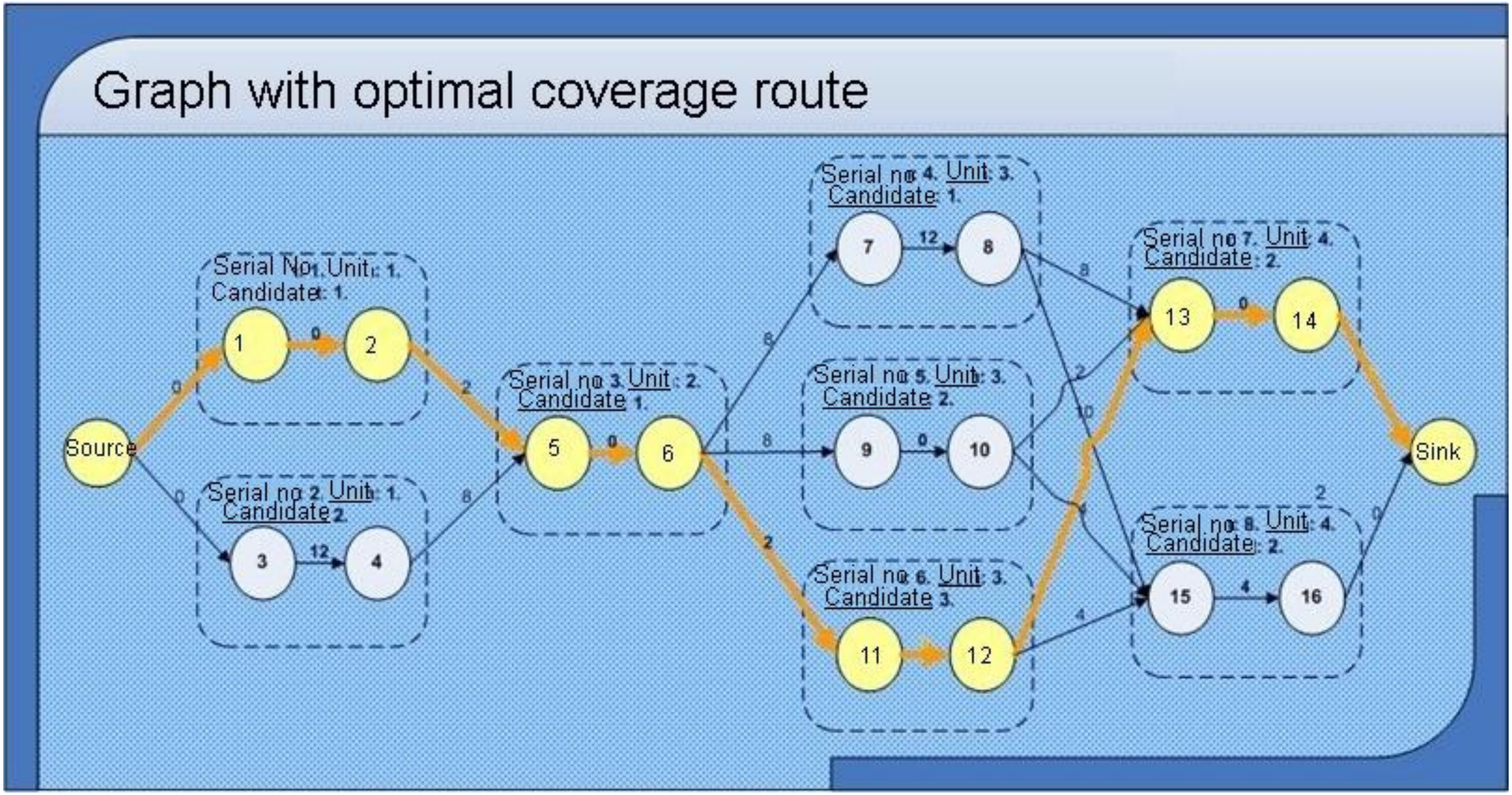
phonetic for each candidate of each element ,
prosodic,
fundamental frequency, etc. information

```
6         alegmagasabbnappalihőmérséklet   európa   északifelén   általában   tizenhat   fok   körülalakul

1164.00
          a         legmag...          napali          ho3me1...          euro1pa          e1szaki          fele1n          a1ltal...          tizenh...          fok
100.00    0.00      0.00     0.00      0.00     0.00   10.00     110.00   13.00    204.00  19.00    0.00    21.00   272.00  137.00    110.00   138.00    102.00   16.00
```

Welcome! Your server will be [127.0.0.1:4749]

Corpus approach (2003-)

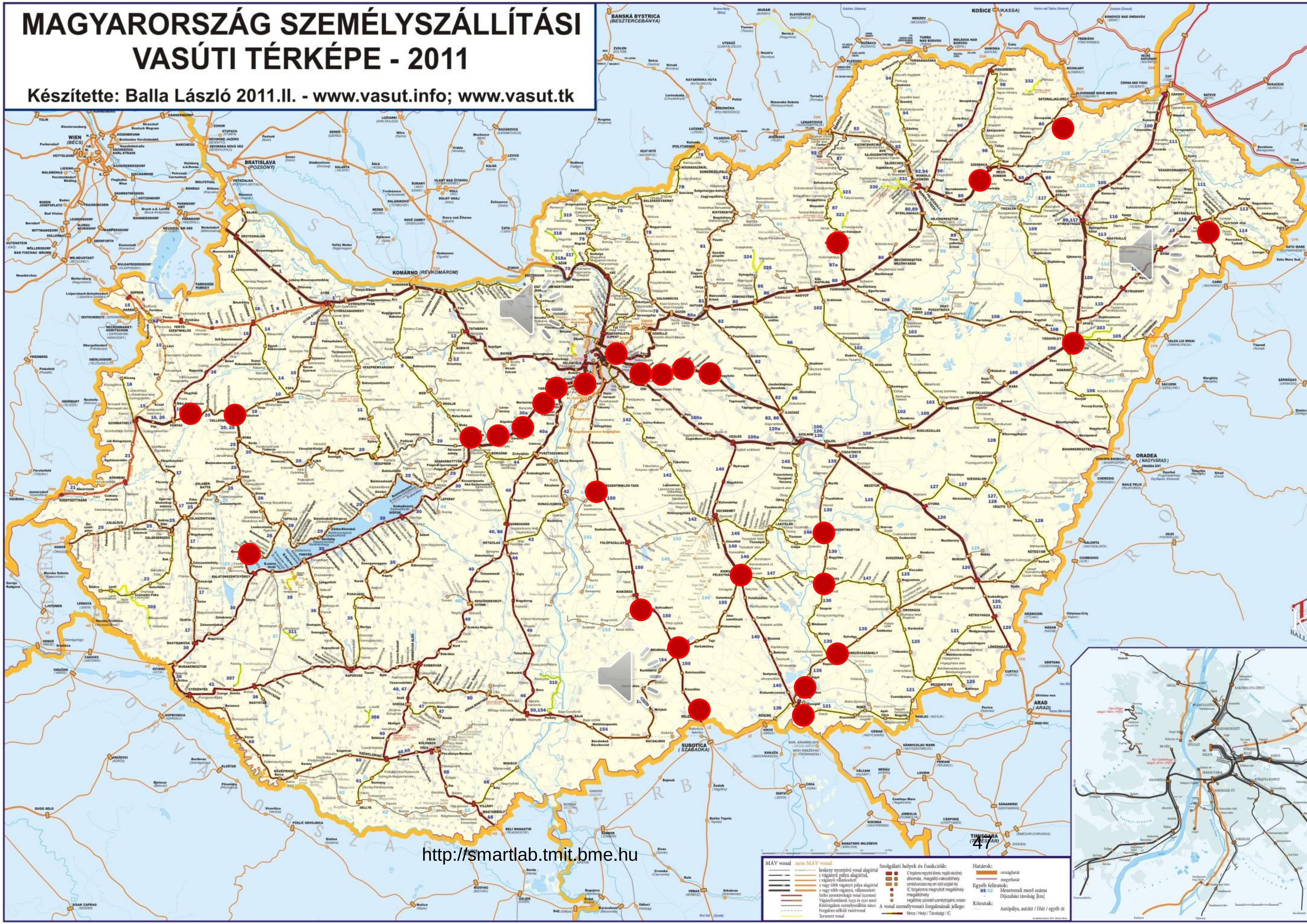

Graph with optimal coverage route
Serial No 1. Unit: 1. Candidate: 1.
Serial no 2. Unit: 1. Candidate 2.
Serial no 3. Unit: 2. Candidate 1.
Serial no 4. Unit: 3. Candidate: 1.
Serial no 5. Unit: 3. Candidate 2.
Serial no 6. Unit: 3. Candidate 3.
Serial no 7. Unit: 4. Candidate: 2.
Serial no 8. Unit: 4. Candidate: 2.
Source
Sink
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
0
12
8
2
4
10

Corpus-based concatenation

Corpus approach
(2003-)

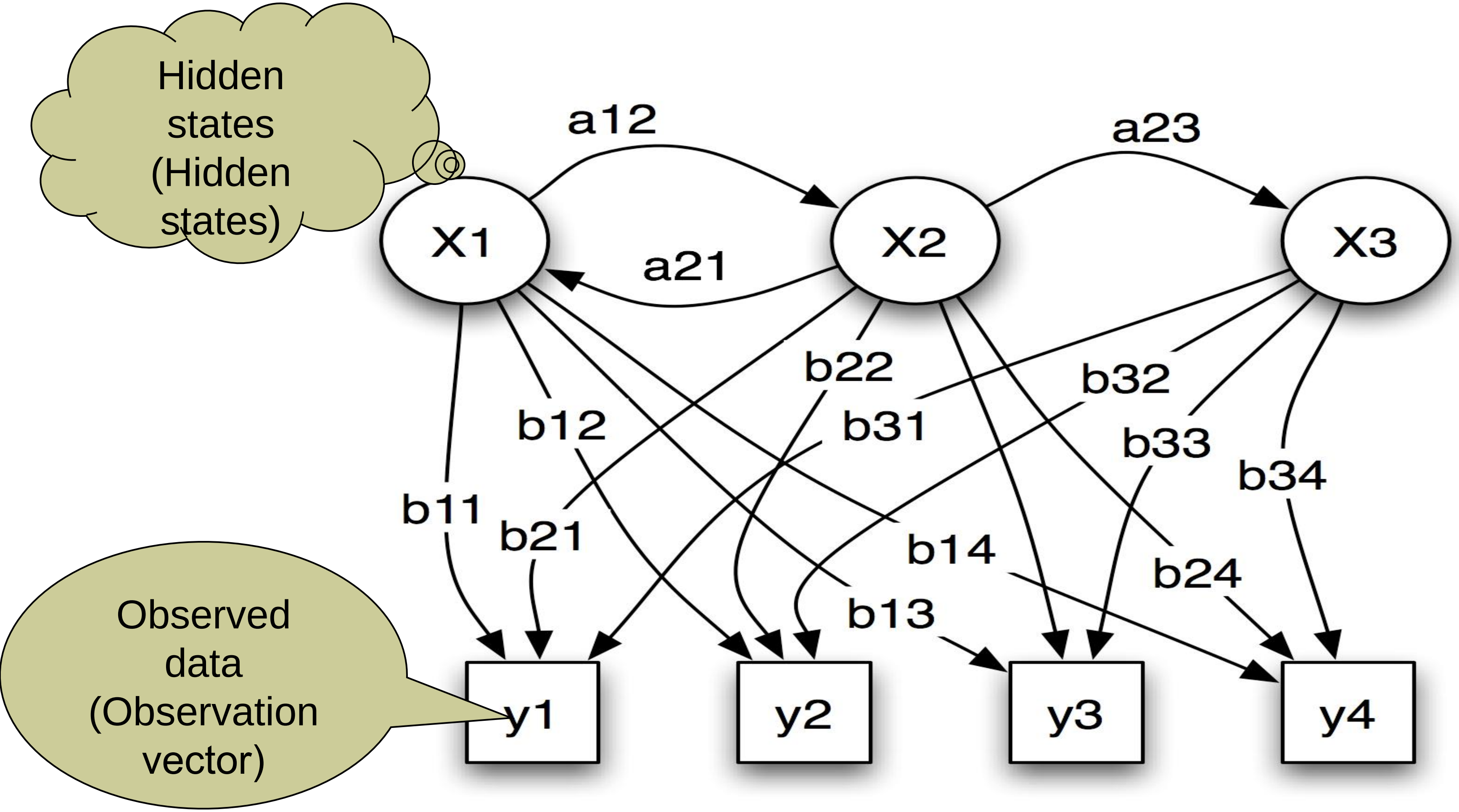
- Resources
  - Digital corpus voice concatenation model
  - Statistical processing
  - Control with computer
- Solution
  - Optimally limited to a topic
  - Practically unlimited vocabulary with slightly worse sound quality
  - Acoustic database n*10Mbyte – n*Gbyte

A new voice at Hungarian train stations


MAGYARORSZÁG SZEMÉLYSZÁLLÍTÁSI VASÚTI TÉRKÉPE - 2011
Készítette: Balla László 2011.II. - www.vasut.info; www.vasut.tk
BANSKÁ BYSTRICA (BESZTERCEBÁNYA)
KOŠICE (KASSA)
BRATISLAVA (POZSONY)
WIEN (BÉCS)
KOMÁRNO (RÉVKOMÁROM)
ORADEA (NAGYVÁRAD)
ARAD (ARAD)
SUBOTICA (SZABADKA)
TIMISOARA (TEMESVÁR)

http://smartlab.tmit.bme.hu

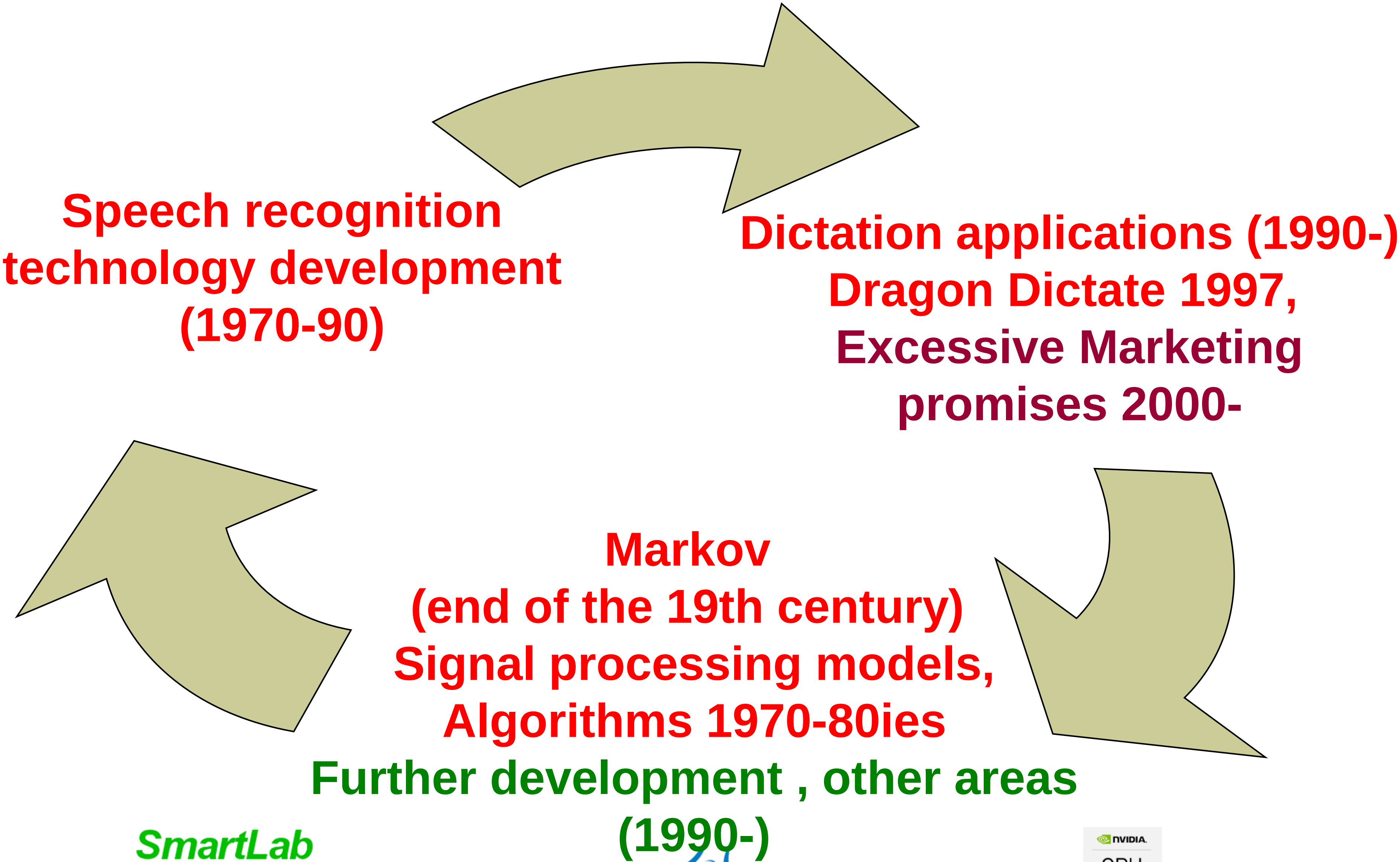
Hidden Markov Model (HMM)



Hidden states (Hidden states)

Observed data (Observation vector)

HMM evolution



Speech recognition technology development (1970-90)

Dictation applications (1990-)
Dragon Dictate 1997,
Excessive Marketing promises 2000-

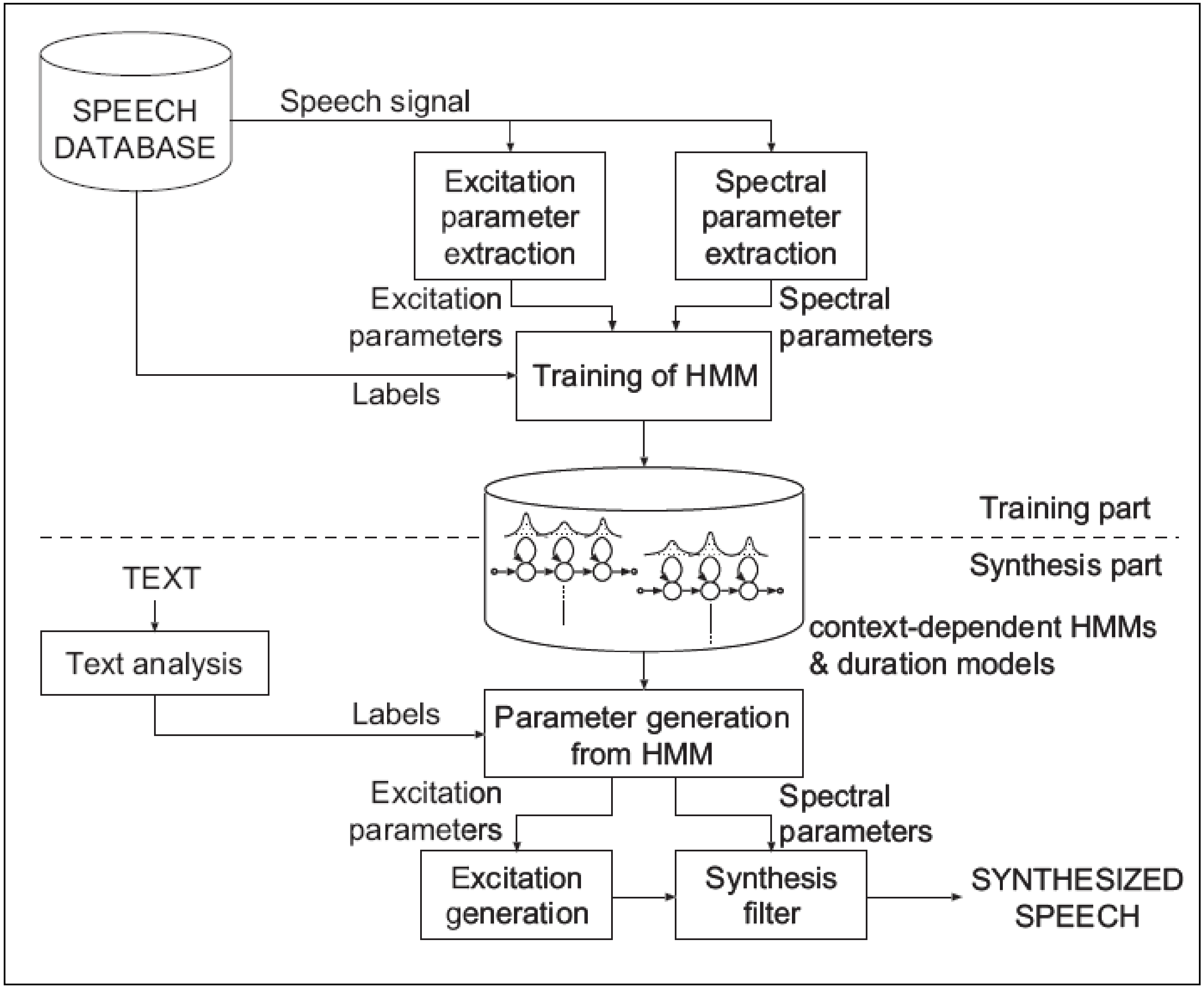
Markov
(end of the 19th century)
Signal processing models,
Algorithms 1970-80ies
Further development , other areas (1990-)

HMM features

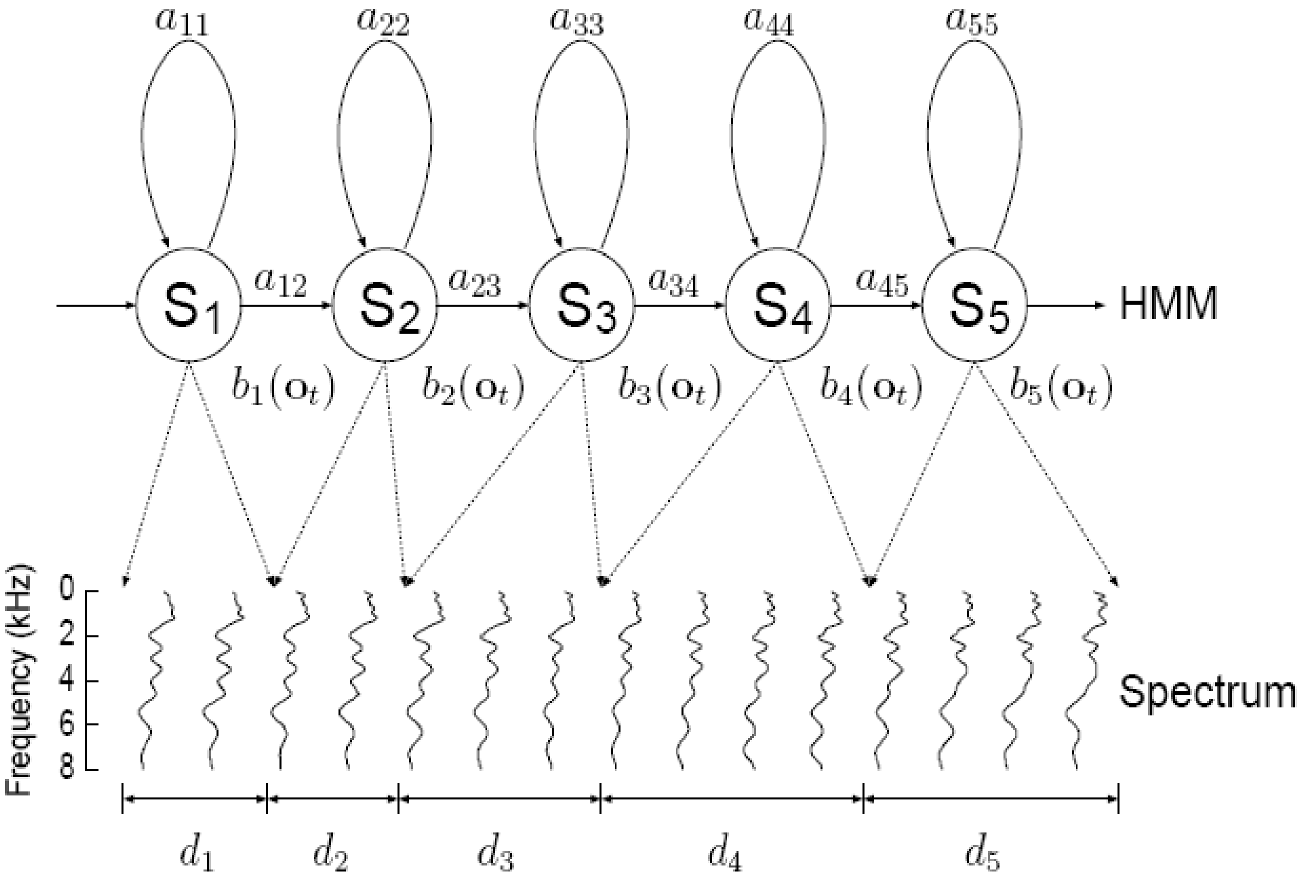
Tokuda et al . (1996-), Tóth, Németh (2008-)

- Resources
  - Digital parametric speech coding model
  - Statistical processing (training database)
  - Computer controlled process
- Advantages
  - Practically unlimited vocabulary
  - Speaker adaptation
  - Slightly poorer speech quality (coding distortions)
  - n*Mbyte acoustic database

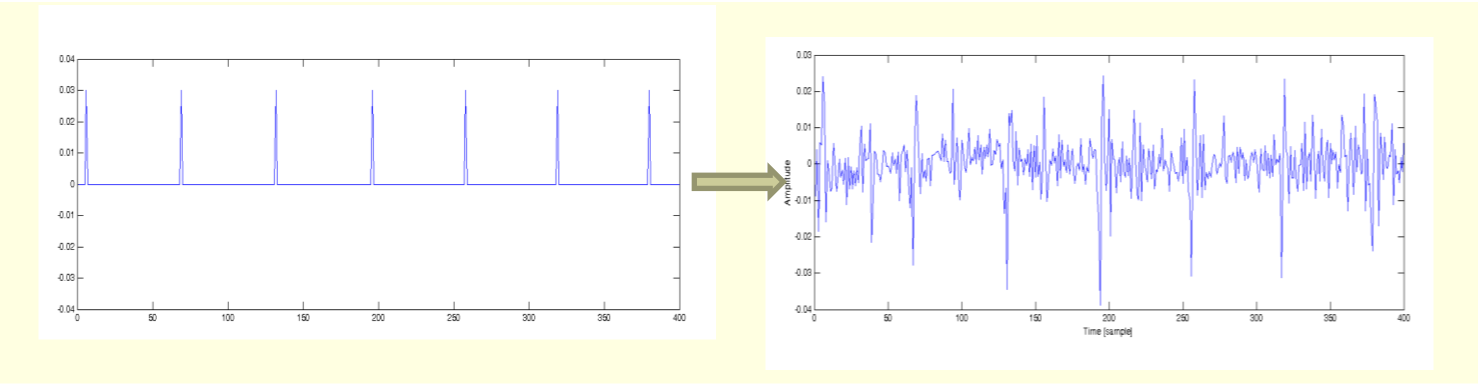
HMM-based TTS

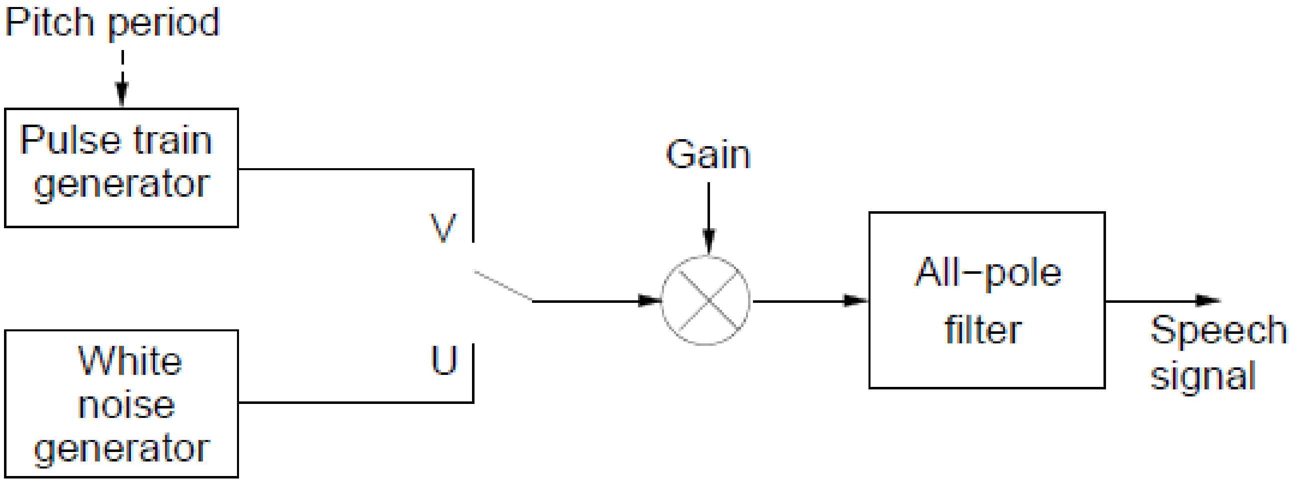


Spectrum modeling



Speech signal excitation models





Pitch period

Pulse train generator

White noise generator

V

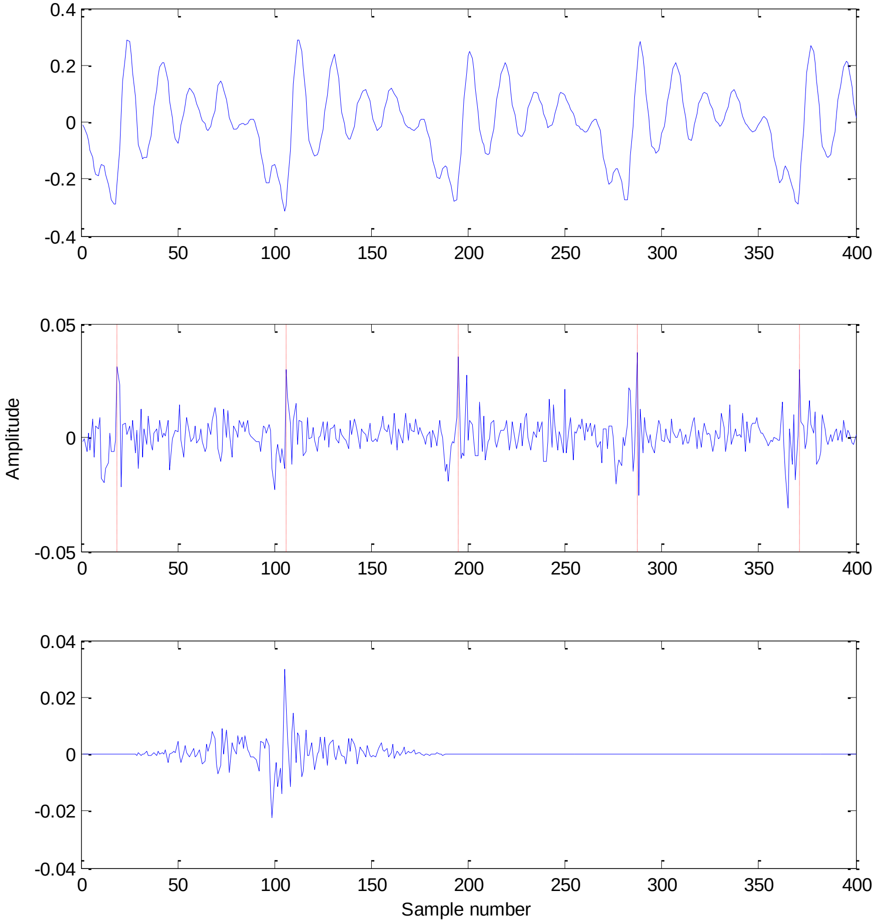
U

Gain

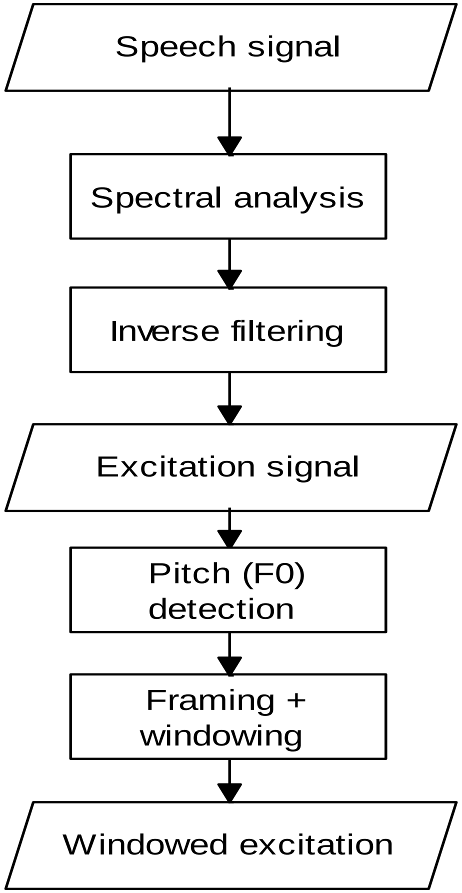
All−pole filter

Speech signal

Excitation signal from natural speech

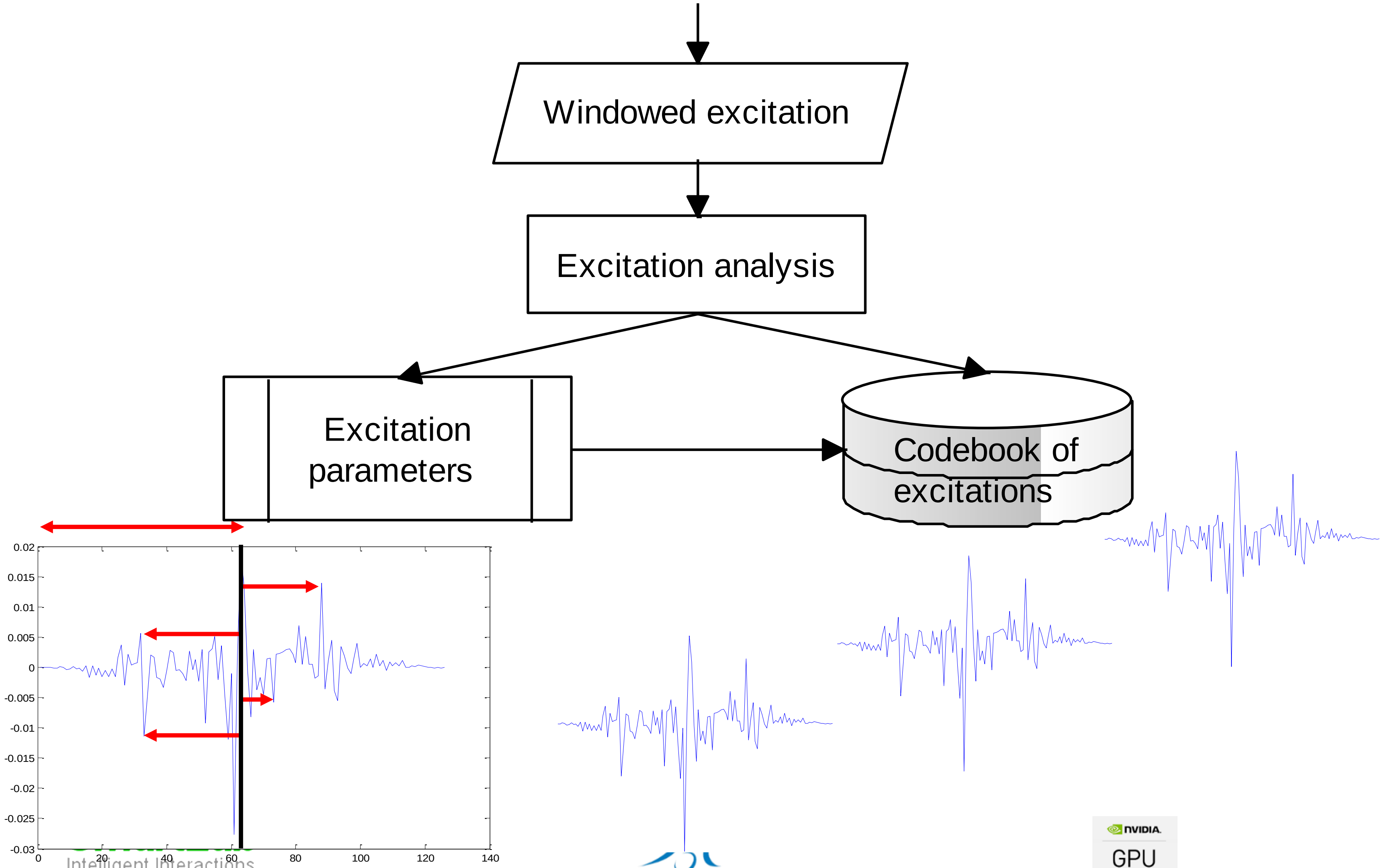


Amplitude



Time [sample number]

Excitation codebook (codebook)



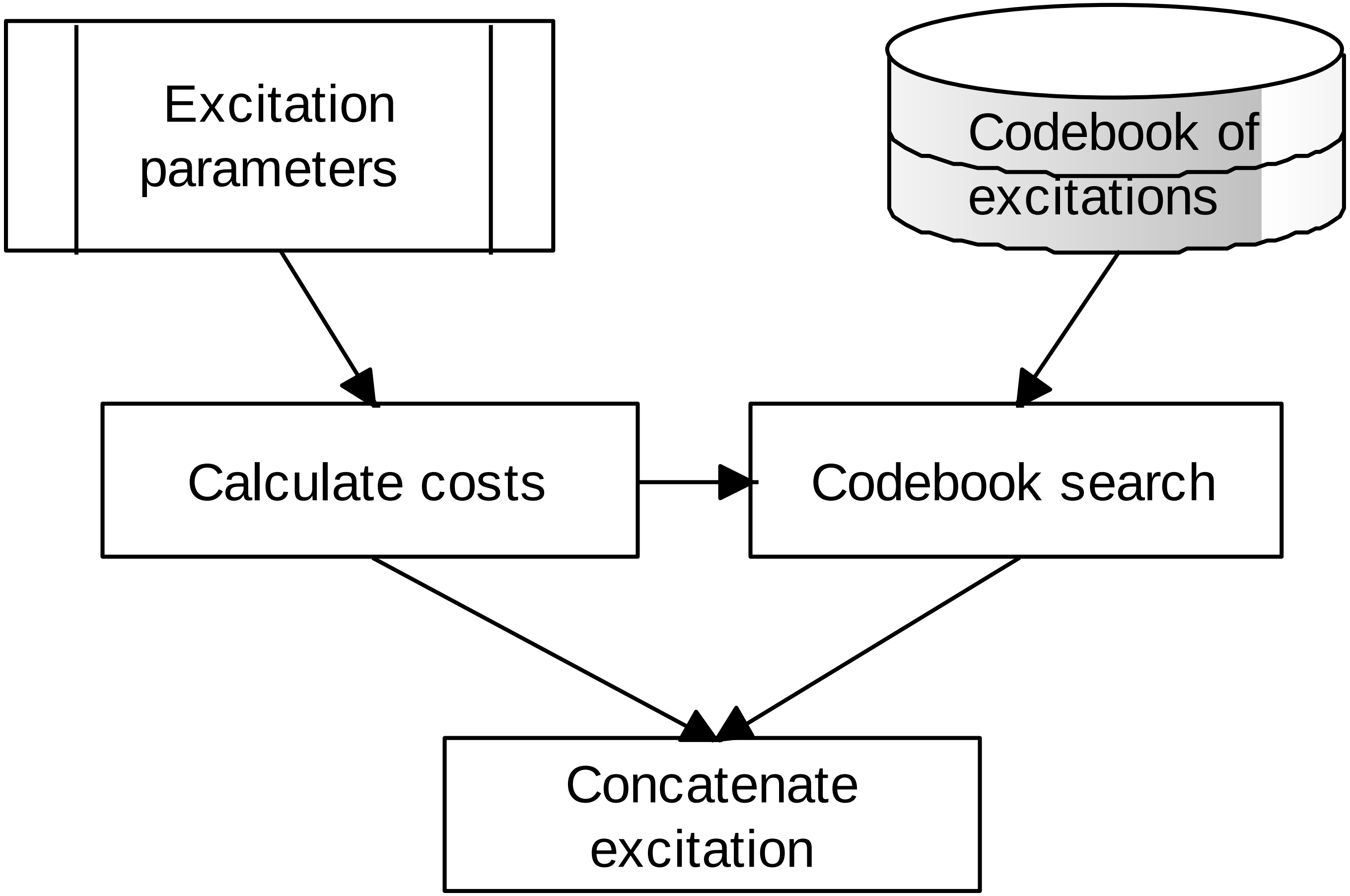
Windowed excitation

Excitation analysis

Excitation parameters

Codebook of excitations

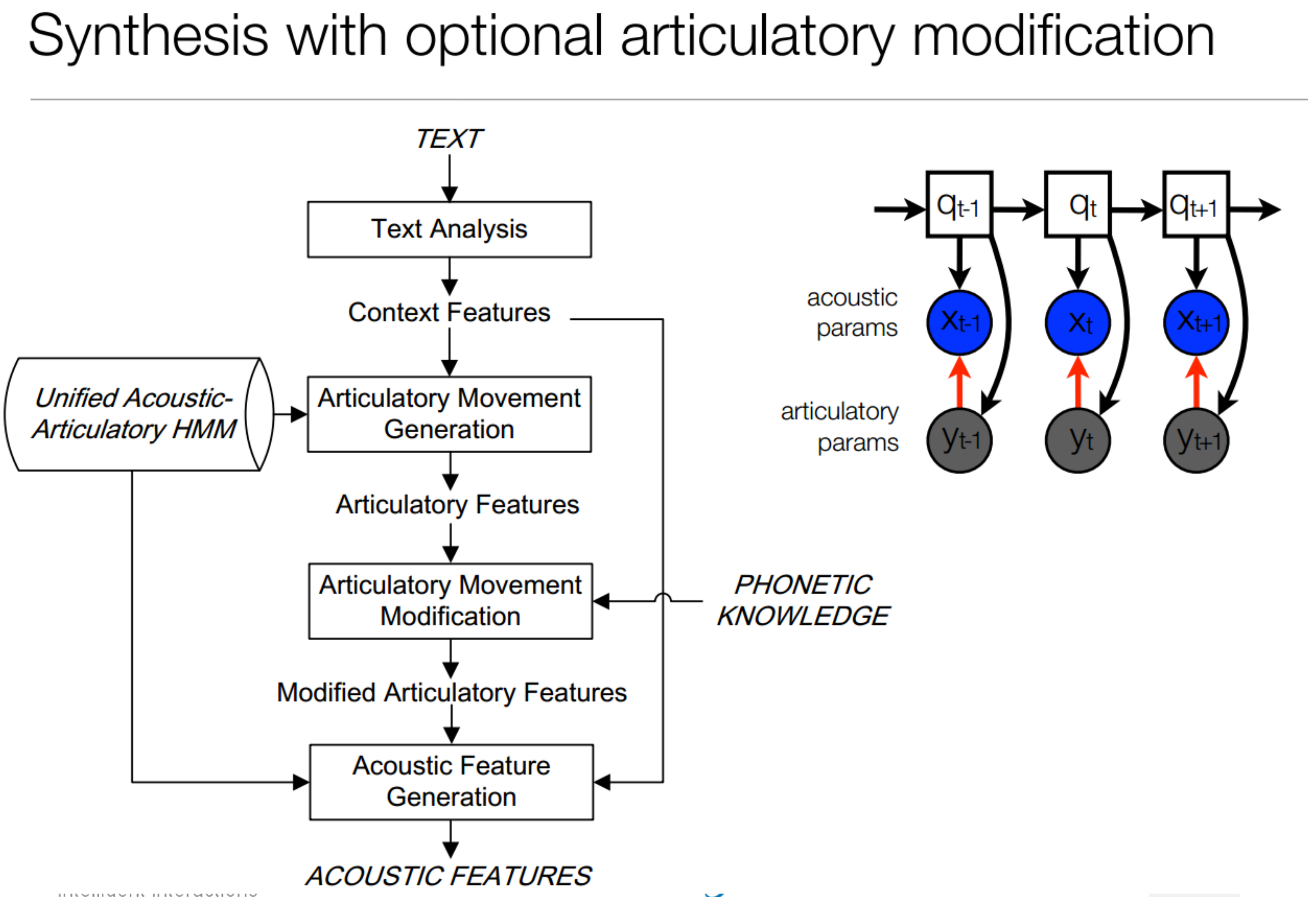
Speech production



Articulatory/formant rebirth (2009-)

- Resources
  - Hybrid approach
  - Digital articulation/formant model
  - Statistical processing
  - Computer controlled
- Advantages
  - Free vocabulary
  - Sound character modification

Articulatory/formant re-Ling et al . (2009-)



Synthesis with optional articulatory modification

TEXT

Text Analysis

Context Features

Unified Acoustic-Articulatory HMM

Articulatory Movement Generation

Articulatory Features

Articulatory Movement Modification

PHONETIC KNOWLEDGE

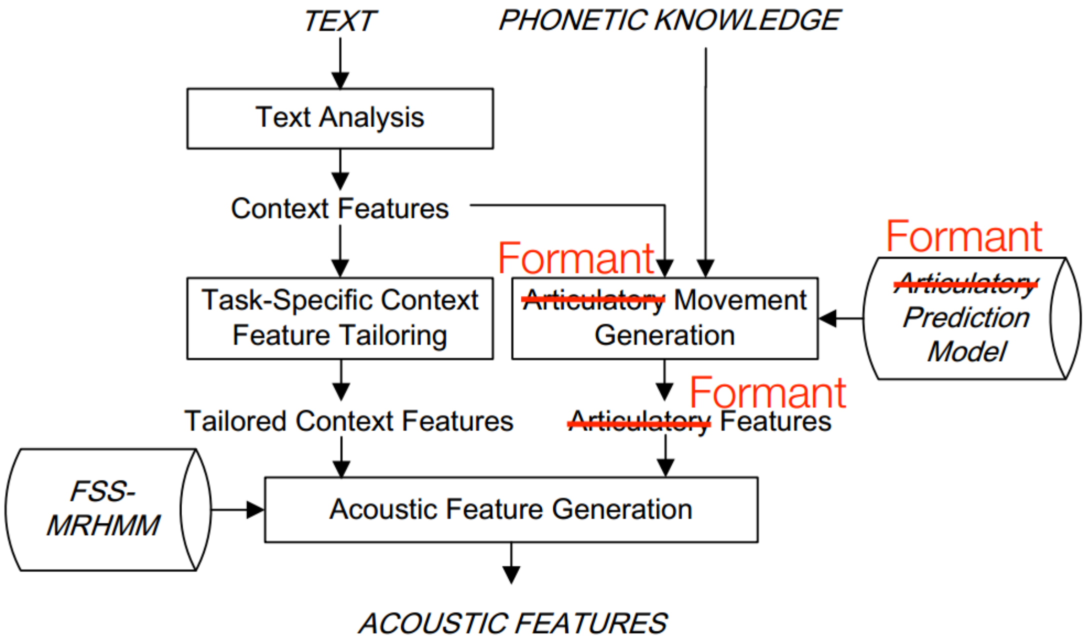
Modified Articulatory Features

Acoustic Feature Generation

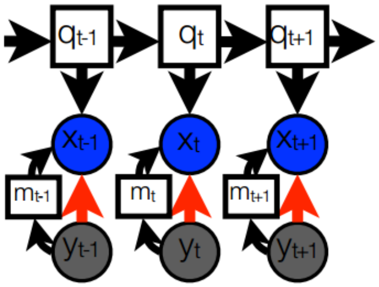
ACOUSTIC FEATURES

Articulatory/formant re- Ling et al . (2009-)

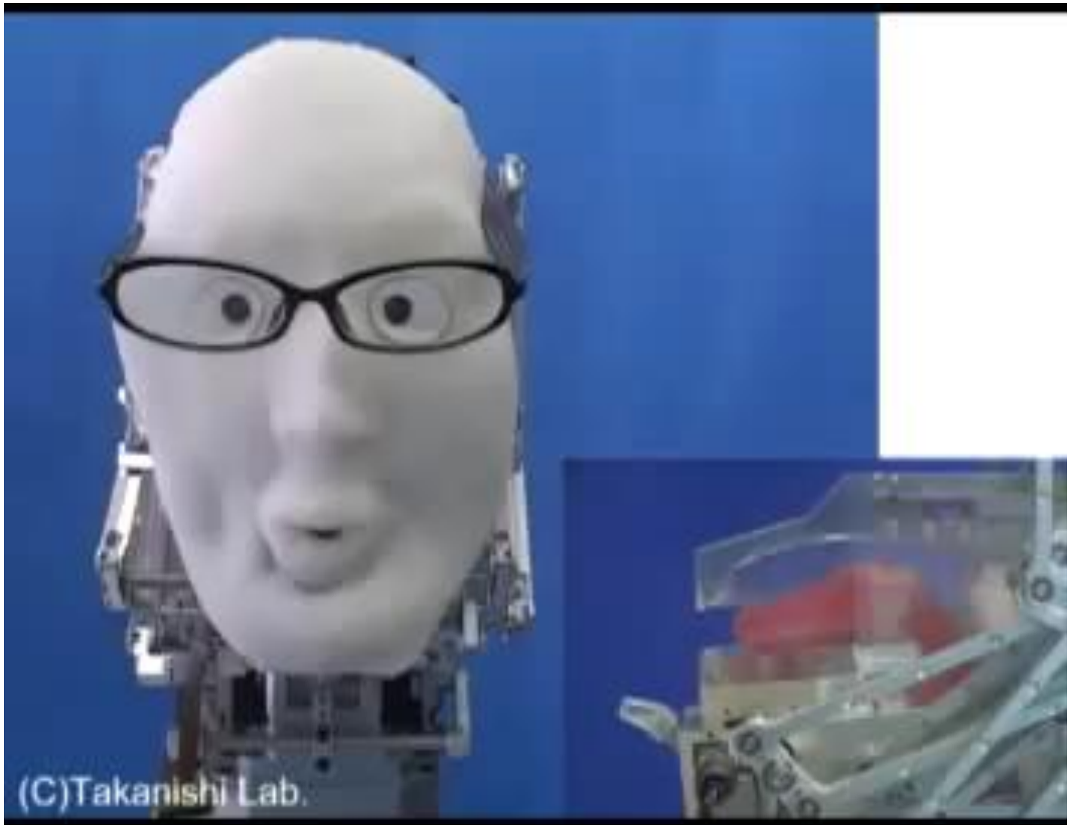
Alternatives to articulation - formants



TEXT
PHONETIC KNOWLEDGE
Text Analysis
Context Features
Formant
~~Articulatory~~ Movement Generation
Formant
~~Articulatory~~ *Prediction Model*
Task-Specific Context Feature Tailoring
Tailored Context Features
Formant
~~Articulatory~~ Features
FSS-MRHMM
Acoustic Feature Generation
ACOUSTIC FEATURES



Articulatory synthesis
Atsuo Takanishi (2008)



Digitally controlled mechanical articulation model

Dyadic synthesis application (2012) Nao robot

Outlook

(even before deep learning)

- Challenges in machine speech generation
  - Adaptation
    - Application scenario (robot <> synthesized audiobook)
    - Speech rate (partner, message type)
    - Person and style (context, emotions, …)
    - Multilingualism (uniform quality, limited resources)
  - Intelligibility in noisy environments
    - Lombard effect
  - High-level control
    - Dialogue systems
  - ……….

Thank you for your attention!