Seminar on Language Technology

Data-Driven Speech Synthesis

Konstantin Tretjakov
kt@ut.ee

11.12.07

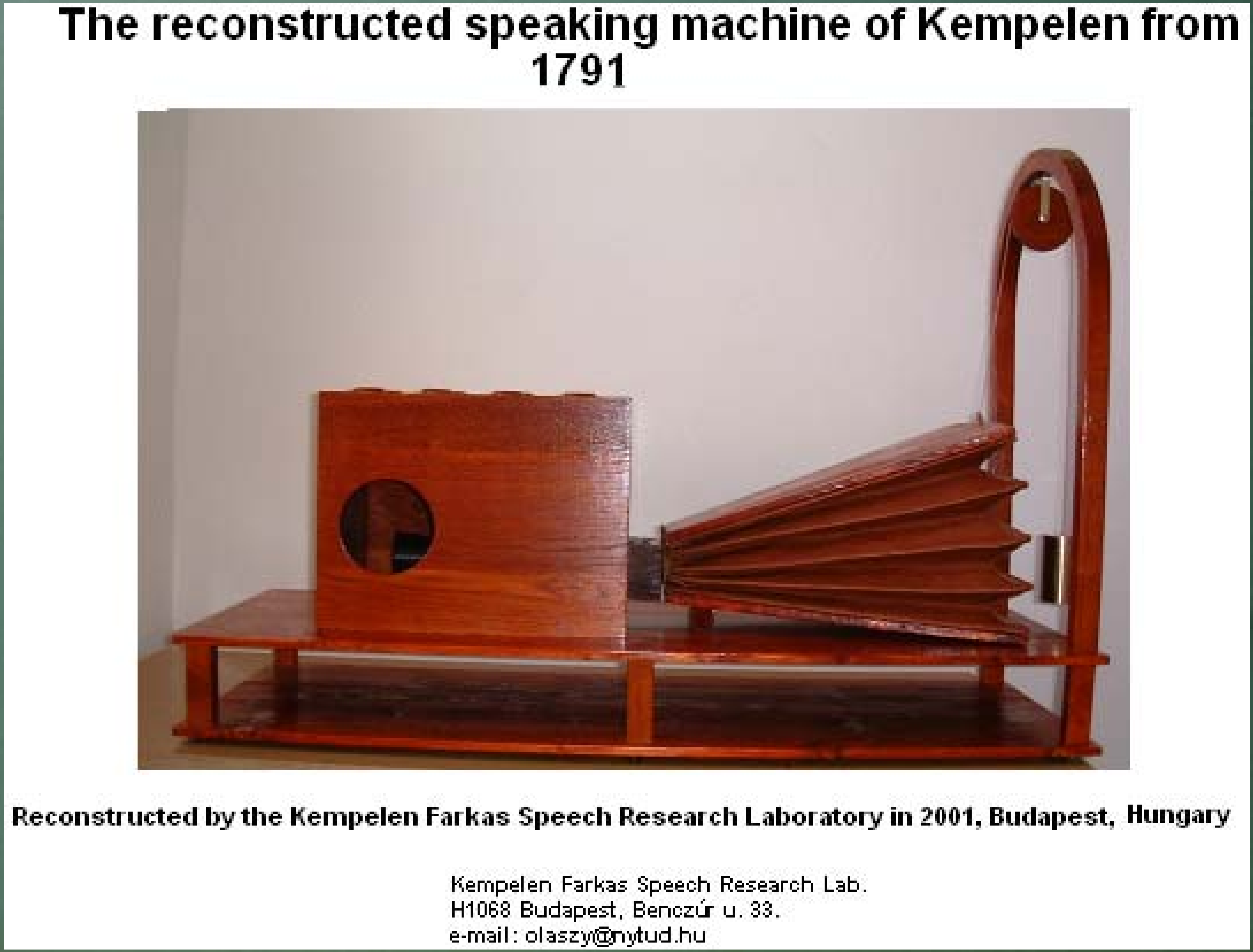
Speech Synthesis

“Computers are getting smarter all the time. Scientists tell us that soon they will be able to talk with us.

(By “they”, I mean computers. I doubt scientists will ever be able to talk to us.)

- Dave Barry

Speech Synthesis in year 1791

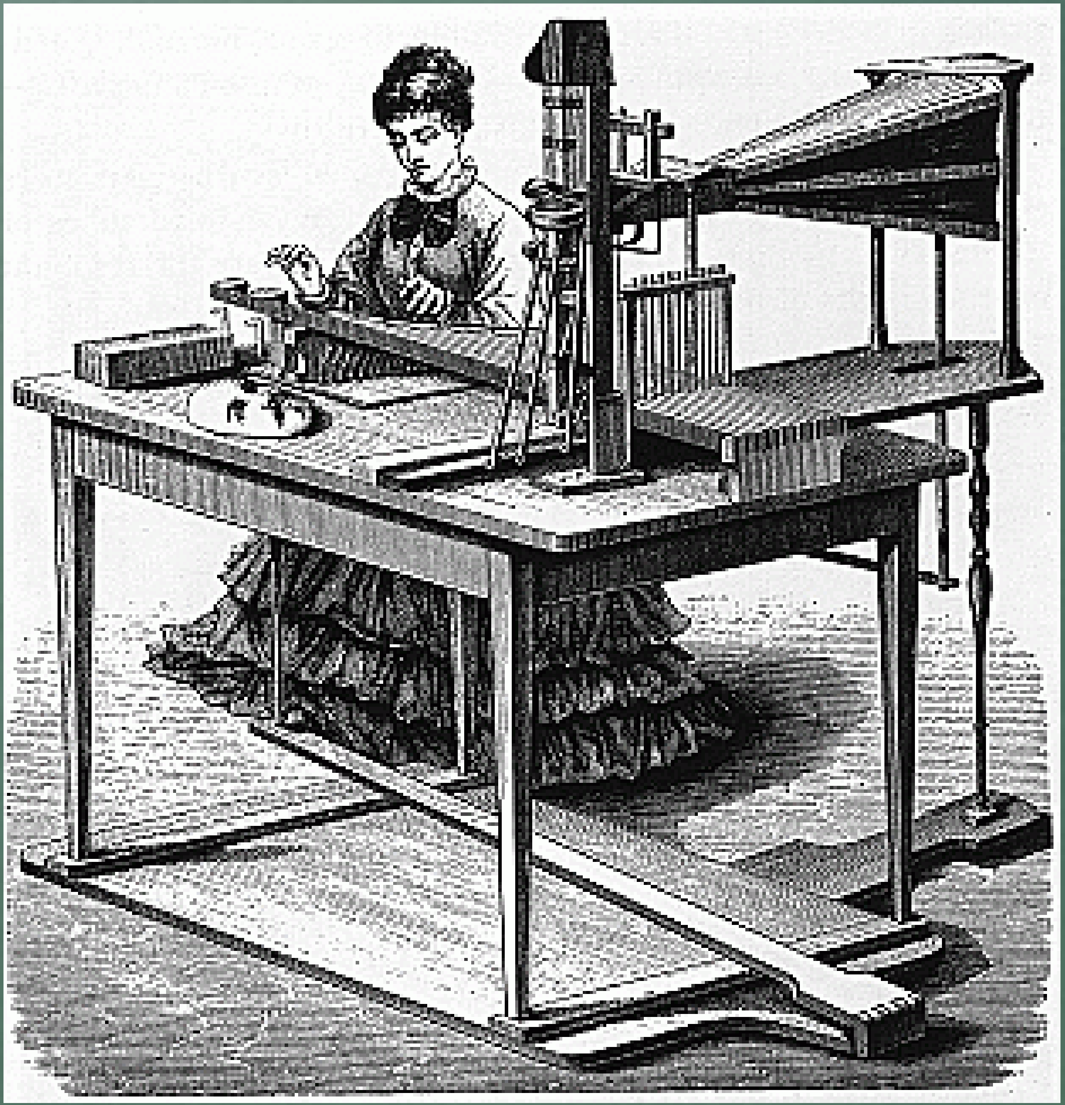


The reconstructed speaking machine of Kempelen from 1791

Reconstructed by the Kempelen Farkas Speech Research Laboratory in 2001, Budapest, Hungary

Kempelen Farkas Speech Research Lab.
H1068 Budapest, Benczúr u. 33.
e-mail: olaszy@nytud.hu

Speech Synthesis in year 1835

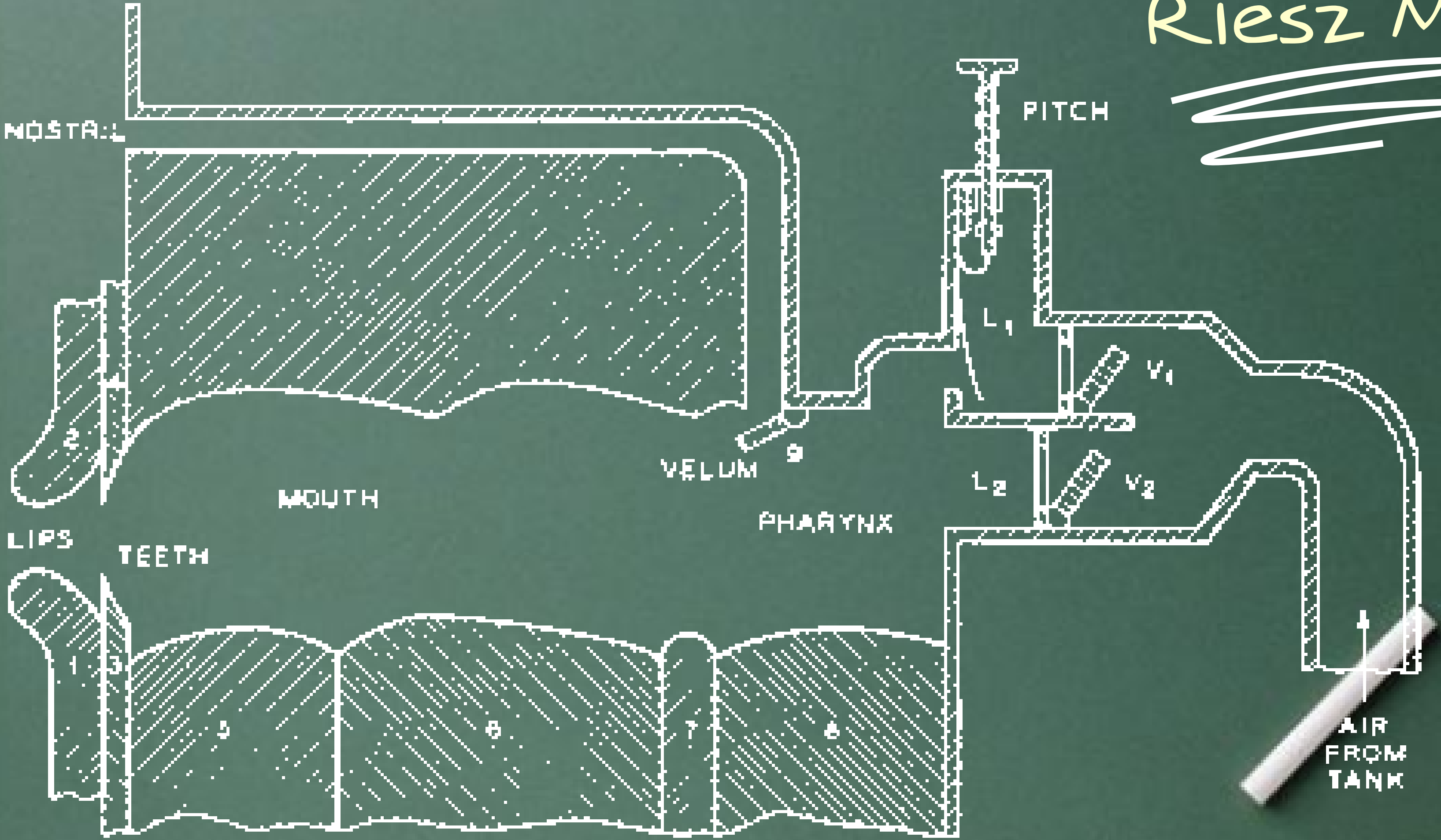


J. Faber
"Euphonia"

http://www.ling.su.se/staff/hartmut/kemplne.htm

Speech Synthesis in year 1937

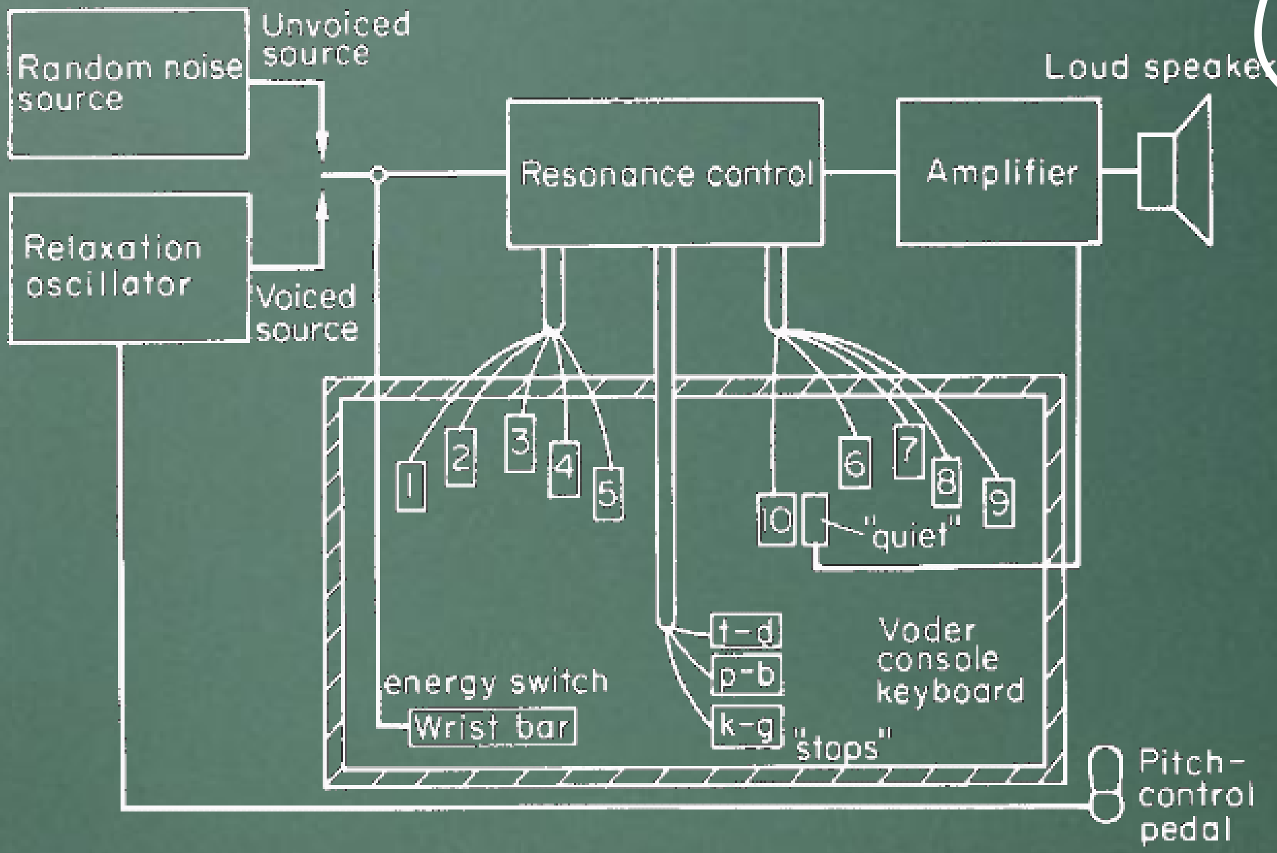
Riesz Model



http://www.ling.su.se/staff/hartmut/kemplne.htm

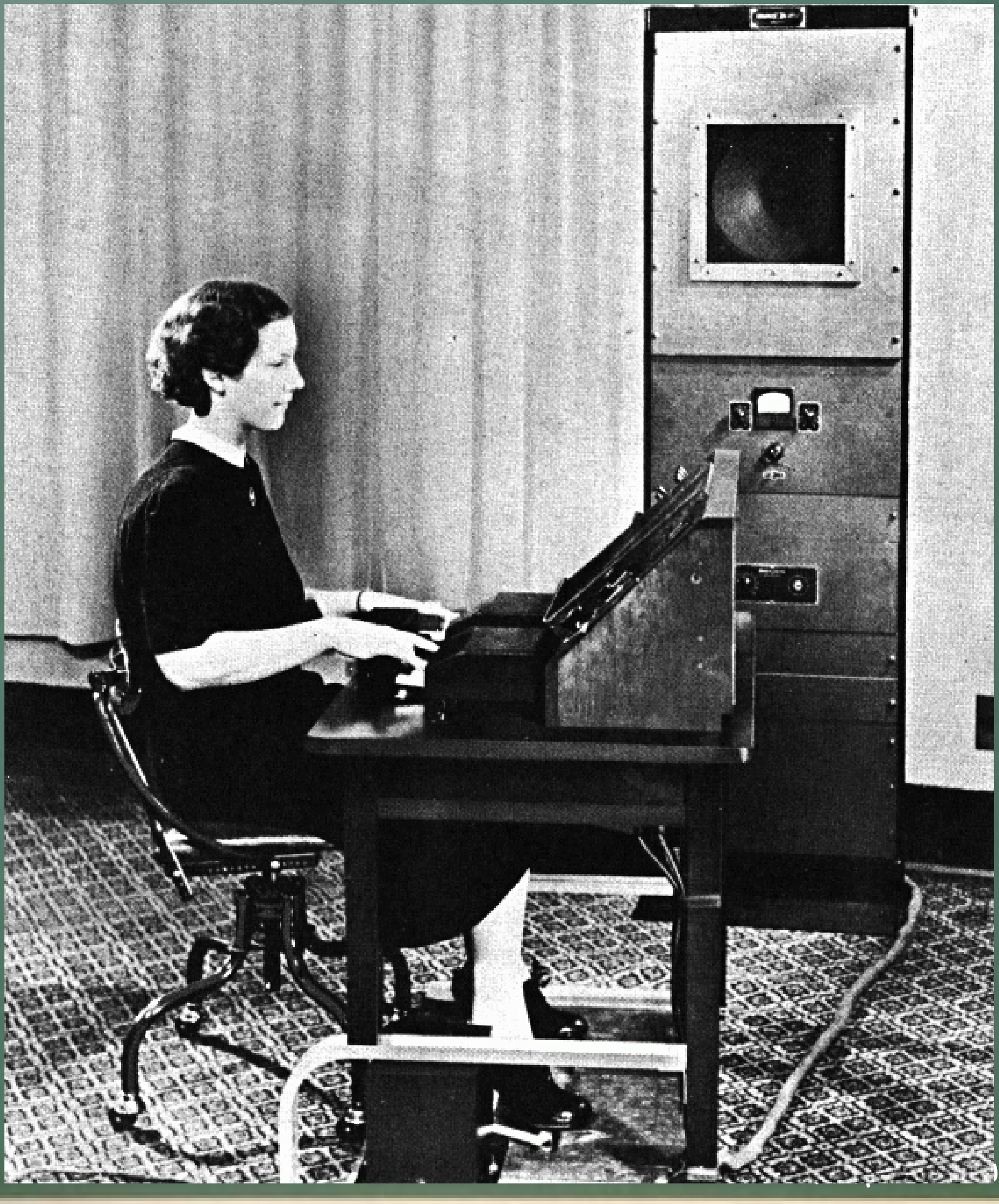
Speech Synthesis in year 1939

H.Dudley
"VODER"



http://www.ling.su.se/staff/hartmut/kemplne.htm

Speech Synthesis in year 1939



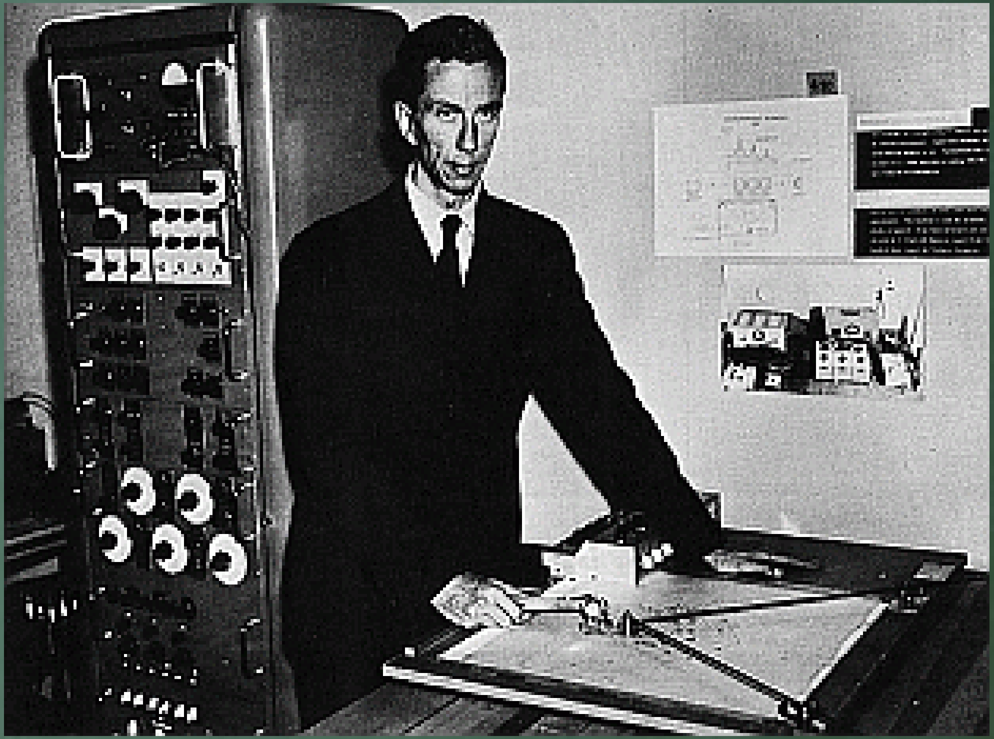
H.Dudley
"VODER"

w.ling.su.se/staff/hartmut/kemplne.htm

Speech Synthesis in year 1953

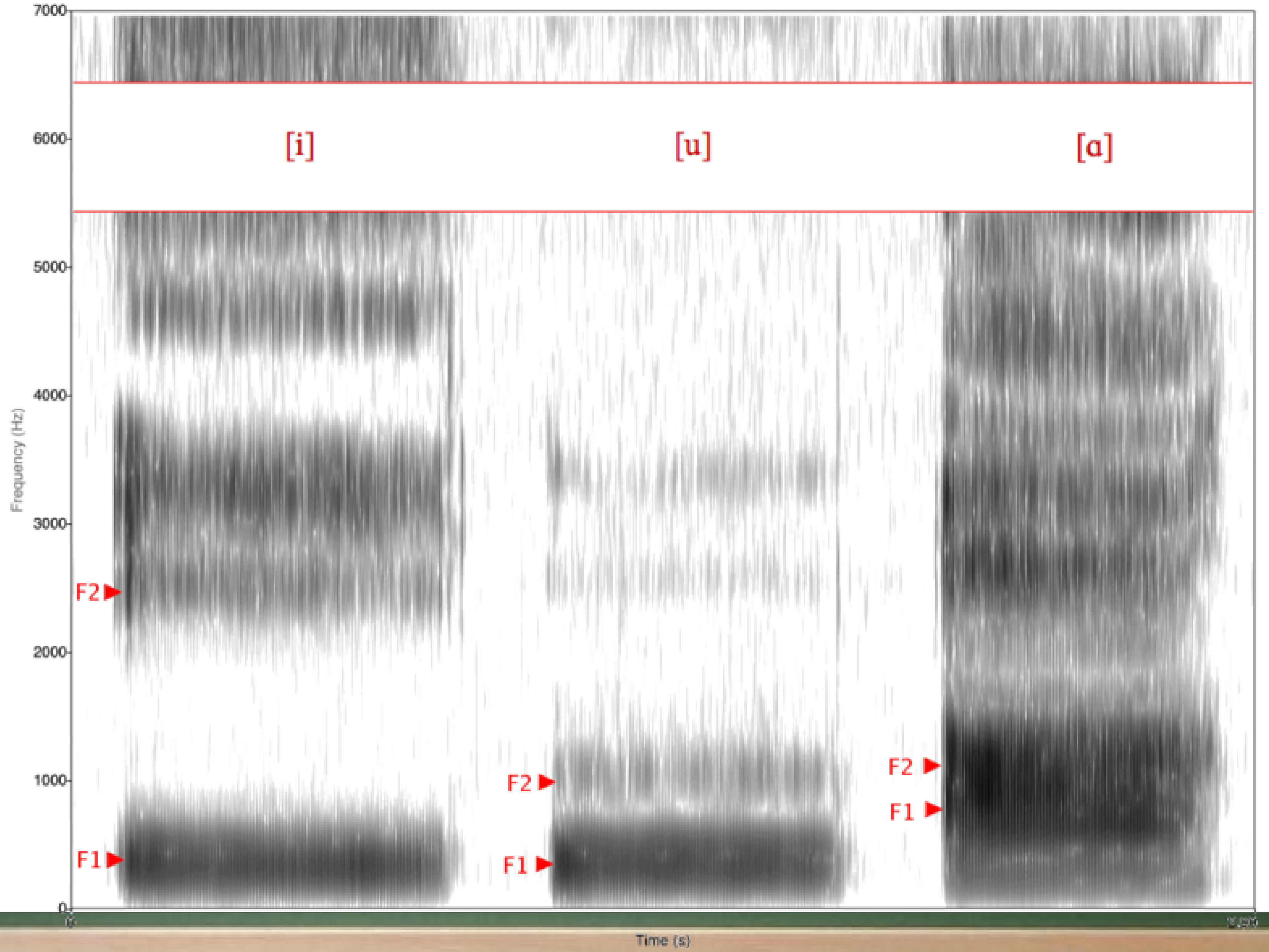
Gunnar Fant's "OVE" (Orator Verbis Electris)

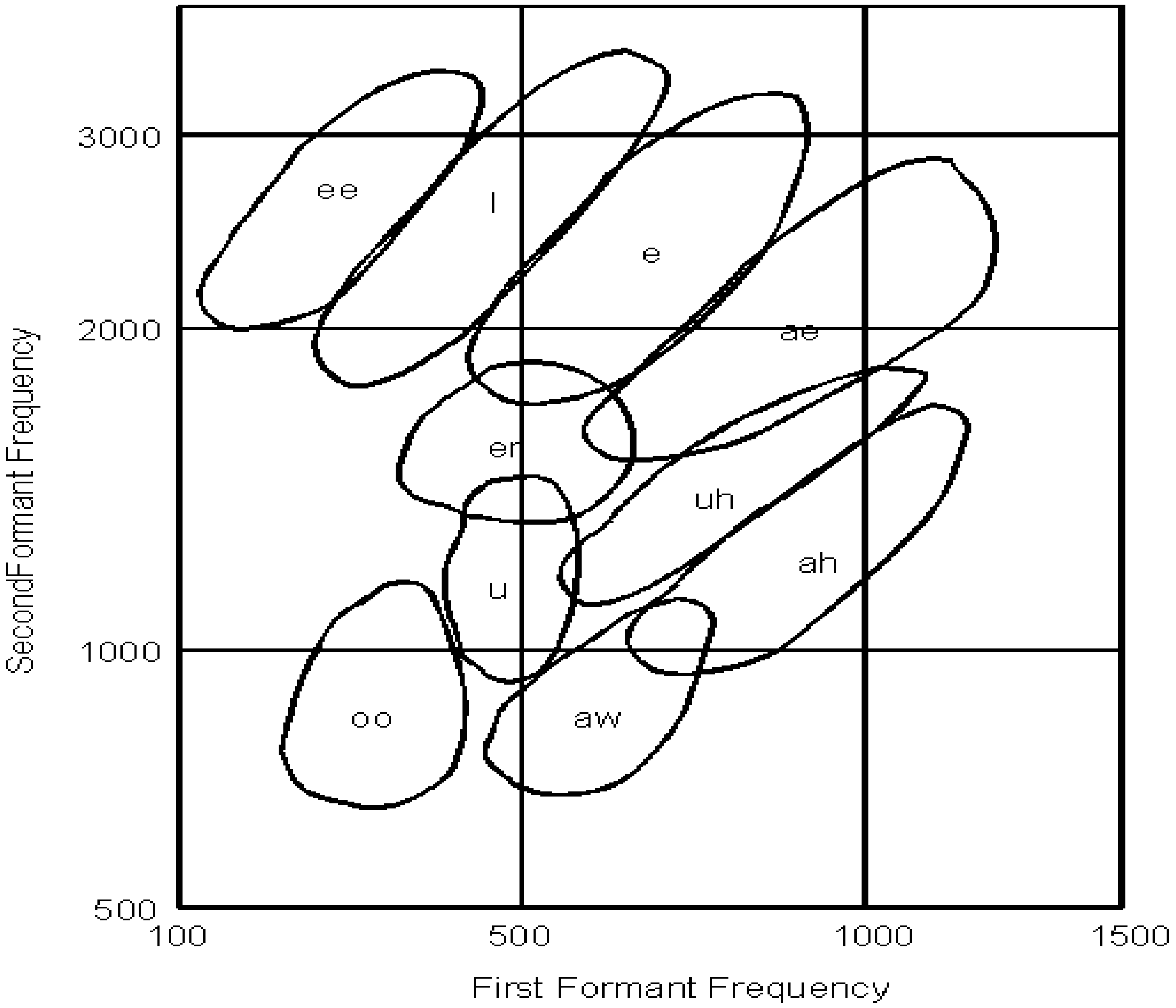
Formant Synthesizer for vowels



http://www.ling.su.se/staff/hartmut/kempIne.htm

Formant Synthesis





Second versus First Formant Frequency for Some Common Vowels

http://www.geofex.com/Article_Folders/wahpedl/voicewah.htm

Modern Speech Synthesis

- 1968 - First full TTS (Umeda et al.)
- 1977 – Diphone concat. (J. Olive)
- 1979 – MITTalk (Allen et al)
- 1984 – DECTalk (Klatt, DEC)
- 1995 – Eurovocs
- 200? - IBM

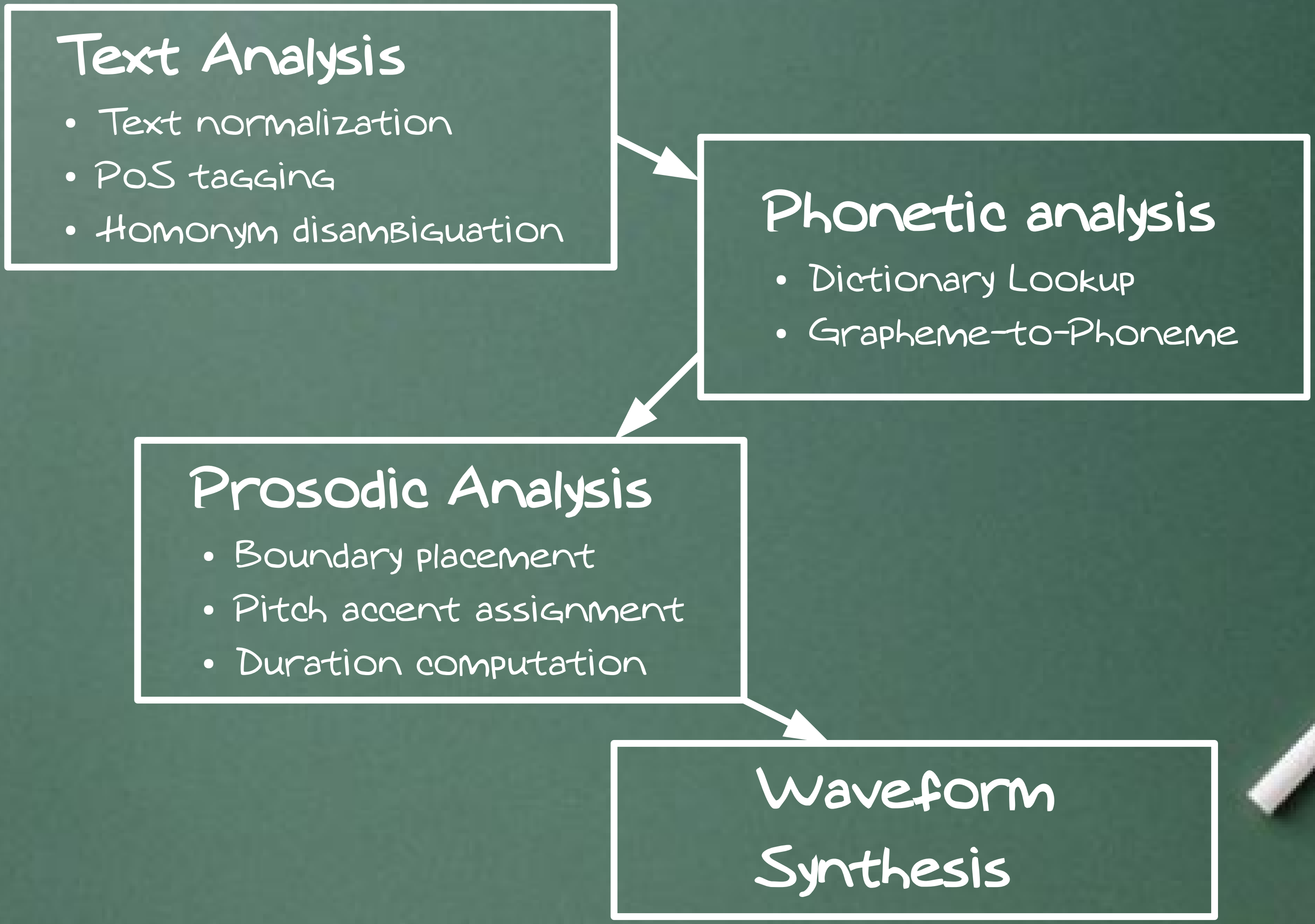
Modern Speech Synthesis

- 1968 - First full TTS (Umeda et al.)
- 1977 – Diphone concat. (J. Olive)
- 1979 – MITTalk (Allen et al)
- 1984 – DECTalk (Klatt, DEC)
- 1995 – Eurovocs Rule-Based
- 200? - IBM Data-driven

Outline

- ~~History of Speech Synthesis~~
- Text-To-Speech System Architecture

Text-to-Speech System



Text Analysis

- Text normalization
- PoS tagging
- Homonym disambiguation

Phonetic analysis

- Dictionary Lookup
- Grapheme-to-Phoneme

Prosodic Analysis

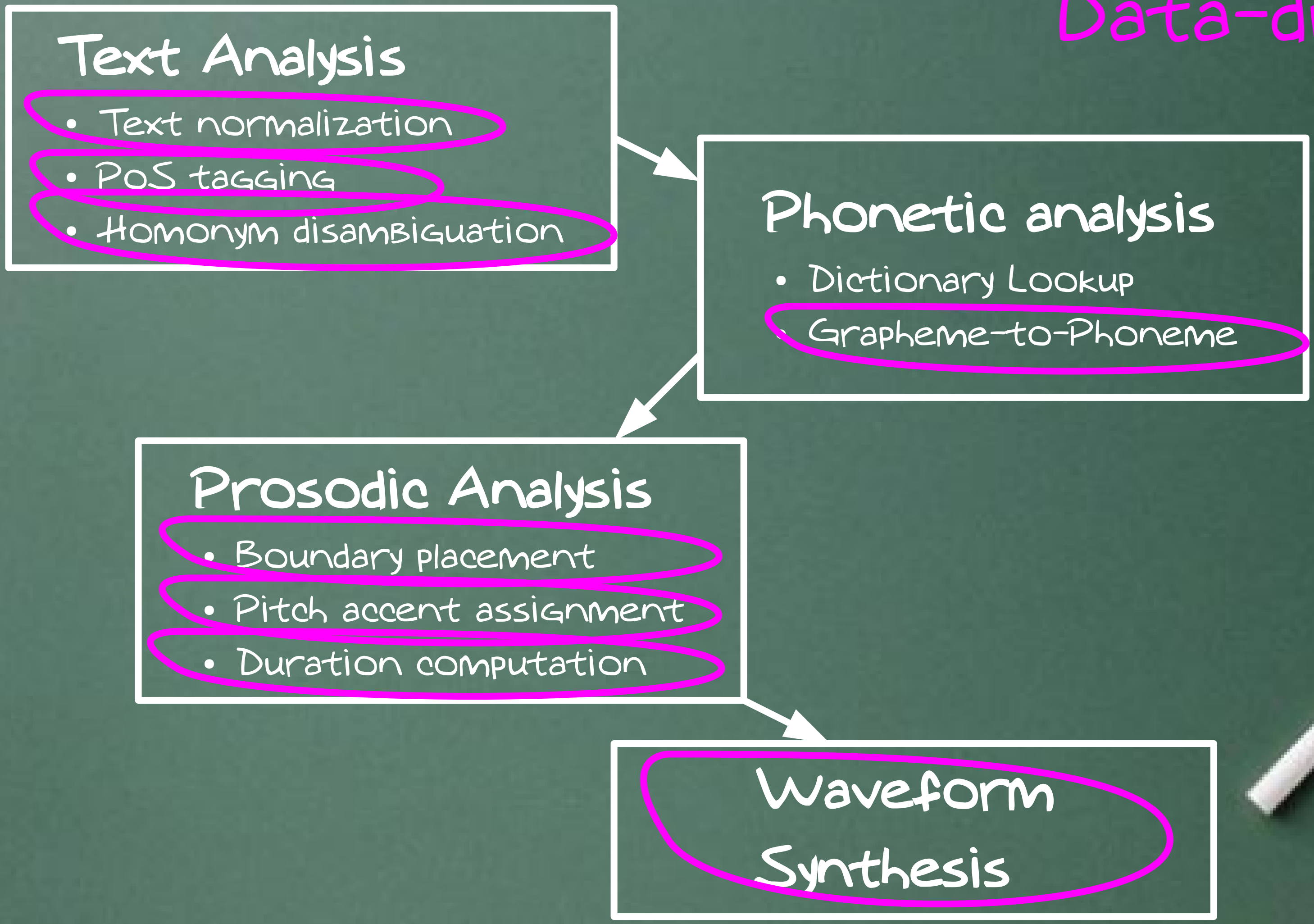
- Boundary placement
- Pitch accent assignment
- Duration computation

Waveform Synthesis

http://www.stanford.edu/class/linguist236/

Text-to-Speech System

Data-driven?



Text Analysis

- Text normalization
- PoS tagging
- Homonym disambiguation

Phonetic analysis

- Dictionary Lookup
- Grapheme-to-Phoneme

Prosodic Analysis

- Boundary placement
- Pitch accent assignment
- Duration computation

Waveform Synthesis

1) Text Normalization

- He stole $100 million from the Bank.
- It's 13 St. Andrews St.
- The home page is http://www.ut.ee.

Method:

- Split to tokens.
- Map tokens to words.
- Identify types for words.

2) Phonetic Analysis

- My latest project is to learn how to better project my voice.
- On May 5 1996, the university bought 1996 computers.
- Yesterday it rained 3 in. Take 1 out, then put 3 in.

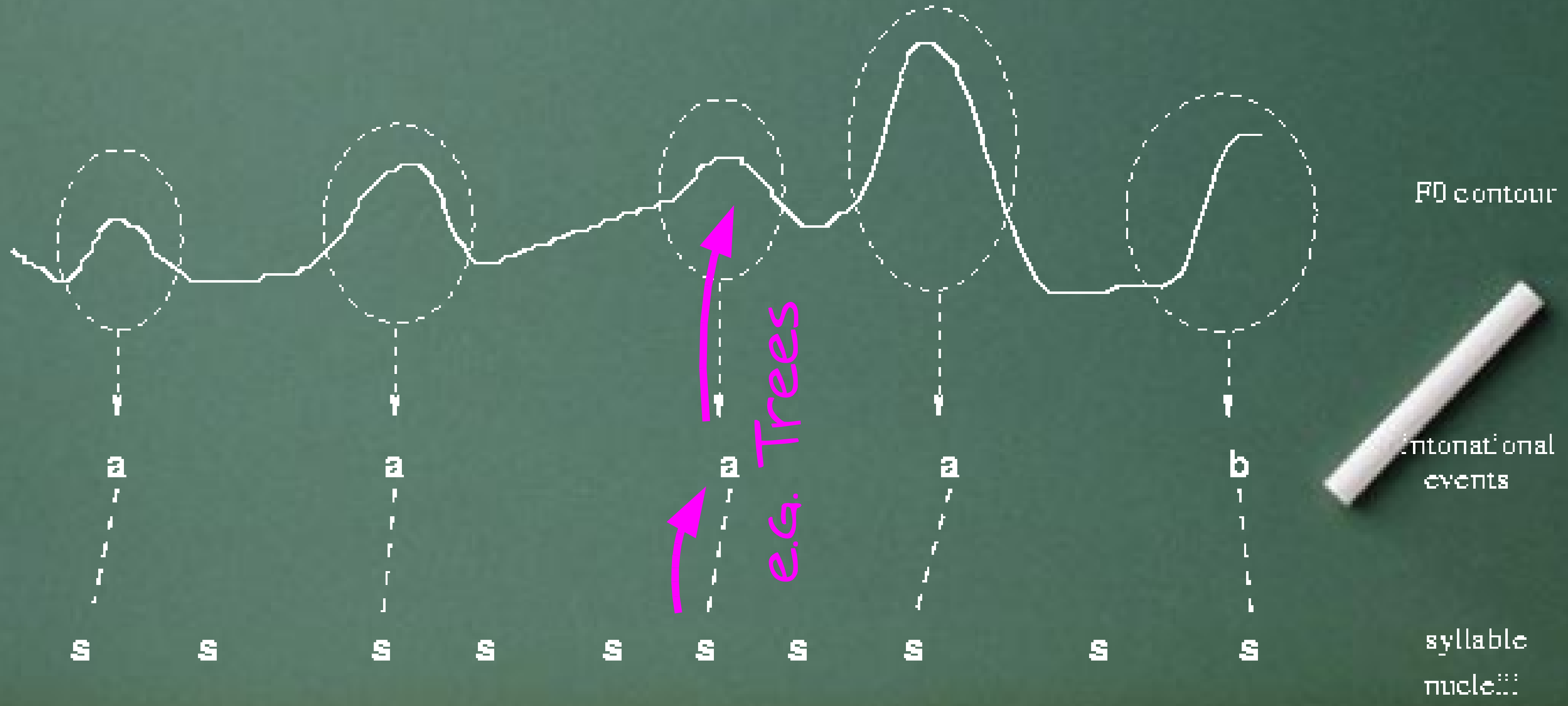
2) Phonetic Analysis

- How to pronounce a word?
  - Look in the dictionary!
    - But what about unknown words and names?
    - Complex languages: German/French/Turkish
  - Letter to sound rules
    - .. also neural networks (NETTalk)
    - .. pr. by analogy (PRONOUNCE)
    - .. case-based (MBRTalk)
    - … and **much** more.

more later

3) Prosodic Analysis

- Prosody: phrases, accents, F0 contour, duration
- The Tilt Intonation Model



4) Waveform synthesis

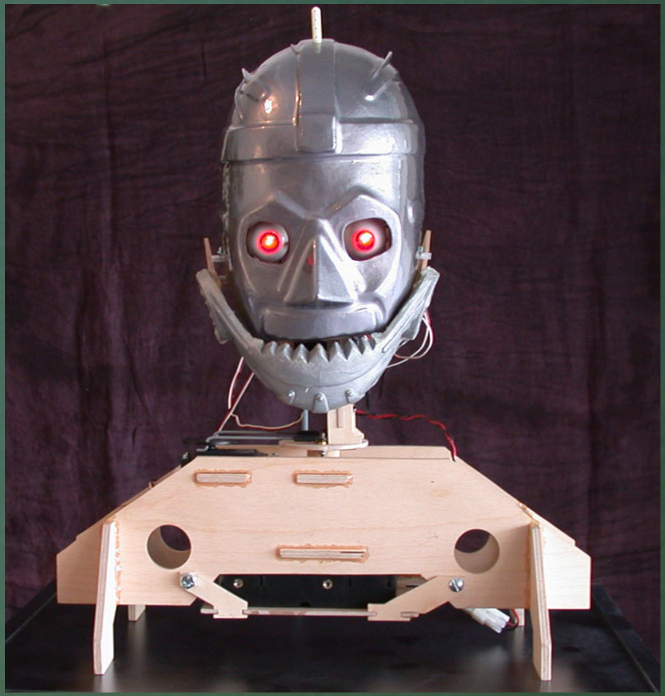
- Articulatory synthesis (a-la VODER)
- Formant (a-la OVE)
- Concatenative synthesis
  - Domain-specific ("talking clock", "weather")
  - Diphones (PSOLA, MBROLA)
  - Unit selection

4) Waveform synthesis

- Domain-specific synthesis is easy:

```
#!/bin/bash
hours=`date +"%-l"`
mins=`date +"%-M"`
ampm=`date +"%-P"`

play $hours.wav
play $mins.wav
play $ampm.wav
```



4) Waveform synthesis

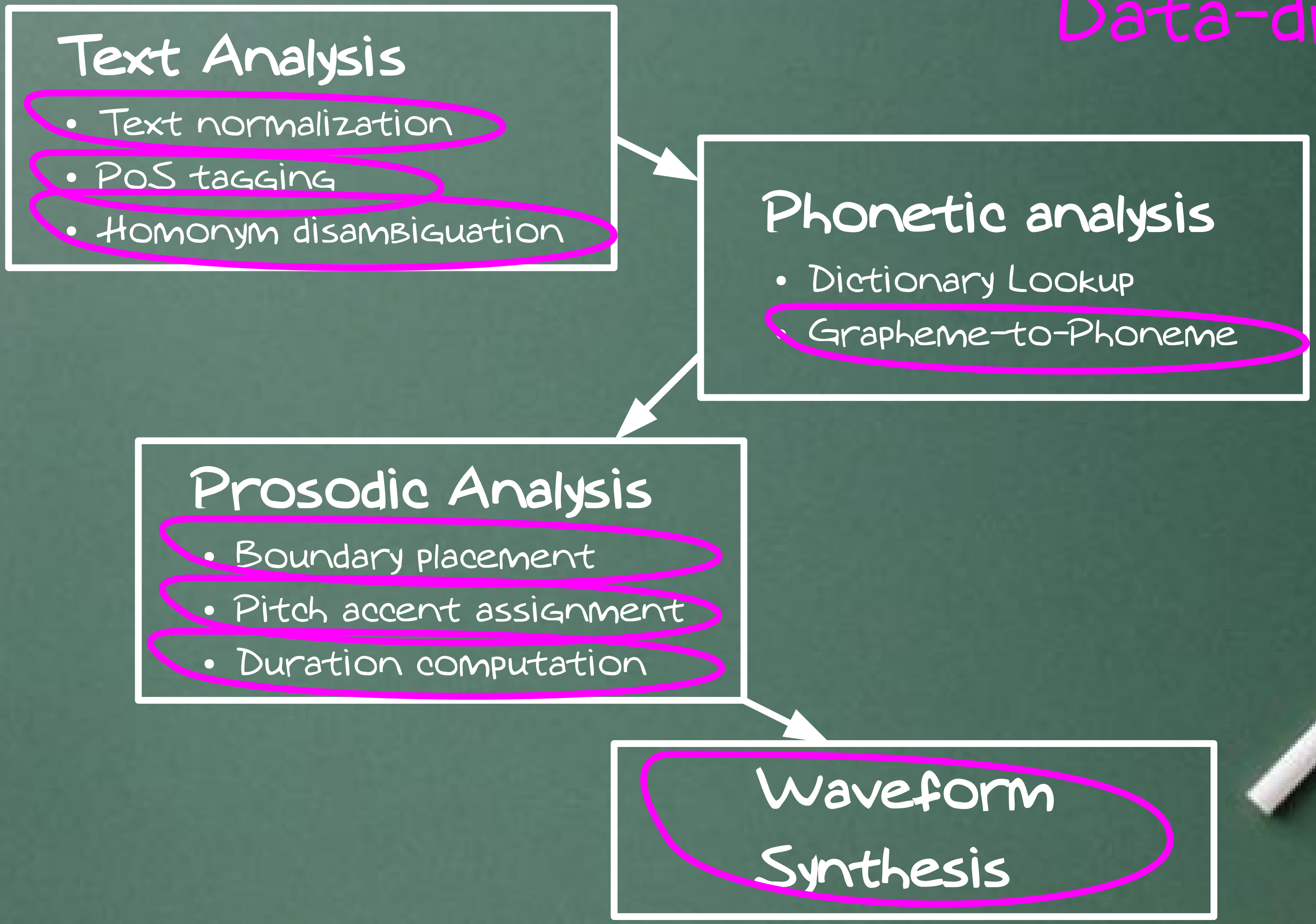
- Diphone synthesis
  - Use diphones: middle of one phone to middle of next.
  - Just a bit of DSP to connect diphones.
    - PSOLA
    - MBROLA ☆☆

4) Waveform synthesis

- Unit selection
  - Use the entire speech corpus as the acoustic inventory.
  - Select at runtime the longest available string of phonetic segments.
  - Minimize number of concatenations.
  - Reduce DSP.

Text-to-Speech System

Data-driven?



Text Analysis

- Text normalization
- PoS tagging
- Homonym disambiguation

Phonetic analysis

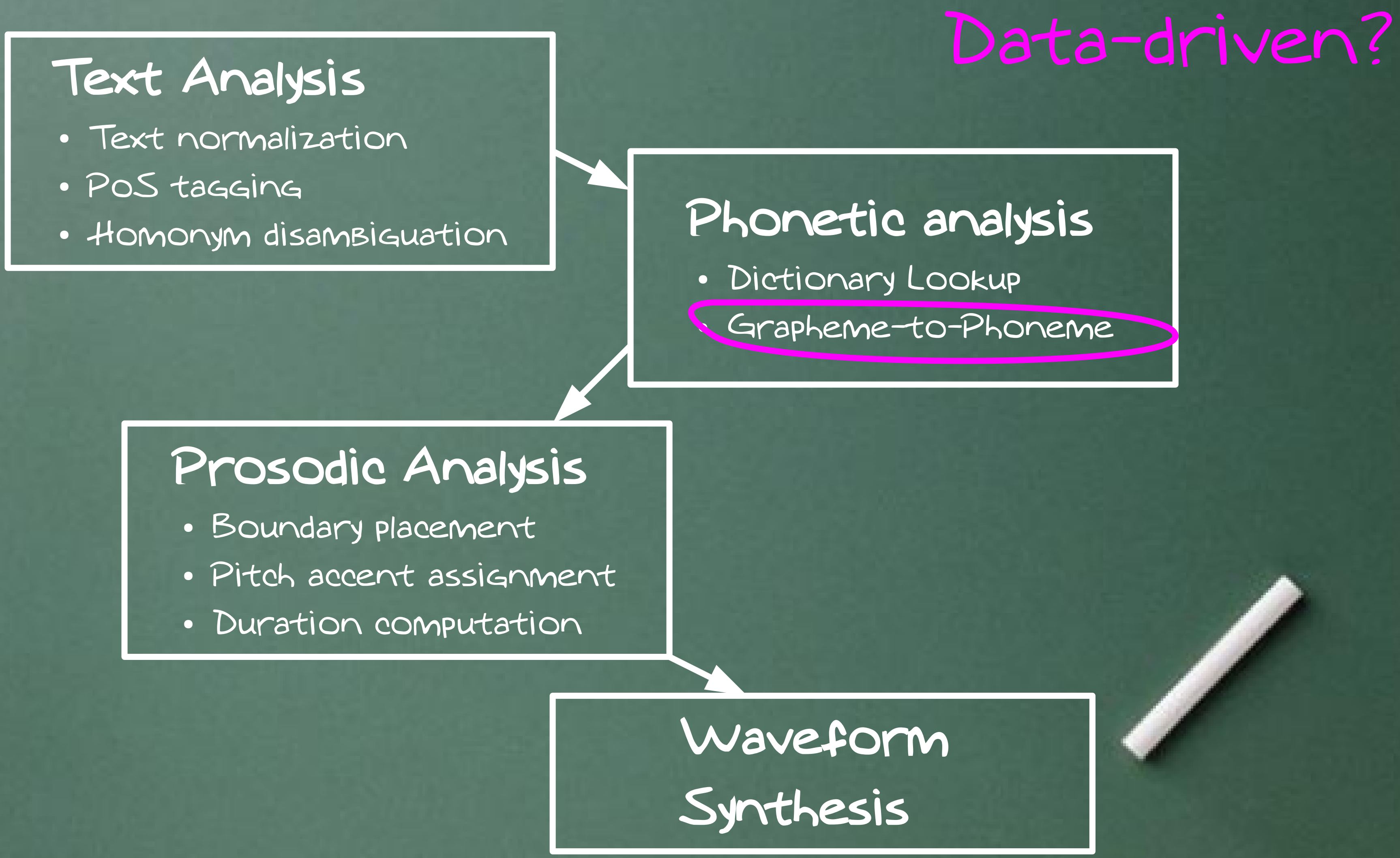
- Dictionary Lookup
- Grapheme-to-Phoneme

Prosodic Analysis

- Boundary placement
- Pitch accent assignment
- Duration computation

Waveform Synthesis

Text-to-Speech System



Data-driven?

Text Analysis

- Text normalization
- PoS tagging
- Homonym disambiguation

Phonetic analysis

- Dictionary Lookup
- Grapheme-to-Phoneme

Prosodic Analysis

- Boundary placement
- Pitch accent assignment
- Duration computation

Waveform Synthesis

Outline

- ~~History of Speech Synthesis~~
- ~~Text-To-Speech System Architecture~~
- Grapheme-to-Phoneme transcription

GTP transcription

- Lexicon:
  - “cepstra” -> (k eh p)' (s t r aa)
  - What about unknown words?
  - Commercial systems have 3-part system:
    - Big dictionary
    - Special code for names/acronyms/etc
    - Machine-learned letter-to-sound (LTS) system for other unknown words

Learning LTS rules

- Induce LTS from a dictionary of the language (Black et al. 1998)
- Two steps:
  - Alignment
  - Decision tree-based rule-induction

Alignment

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

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

- Black et al. propose 2 methods:
  - Expectation-Maximization
  - Estimate p(letter | phone) from valid alignments, take best.
- Devil in the details

Decision trees for LTS

- Now that aligned data is available, train a decision tree:
  - ###chek -> ch
  - checked -> _
- 92-96% letter acc. (58-75% word acc.) for English

GTP transcription

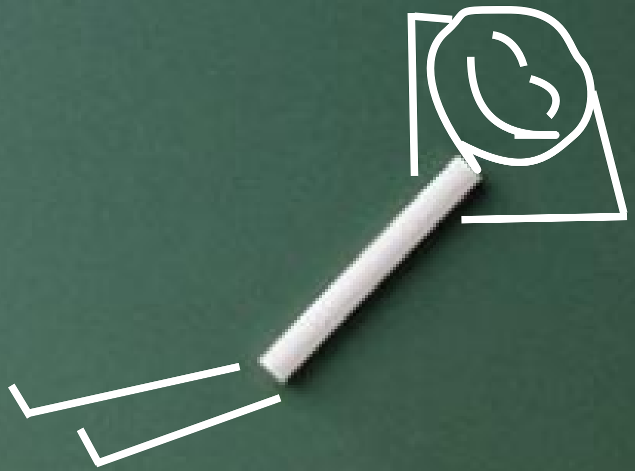
- ~~Decision-tree based (Black et al.)~~
- ANN-based (NETTalk, Sejnowski et al.)
- Pronunciation-by-Analogy (Damper et al.)
- Memory-based (MBRTalk, Stanfill)
- Transducer-based (I. Bulyko)
- Non-segmental (A. Cohen)

GTP transcription

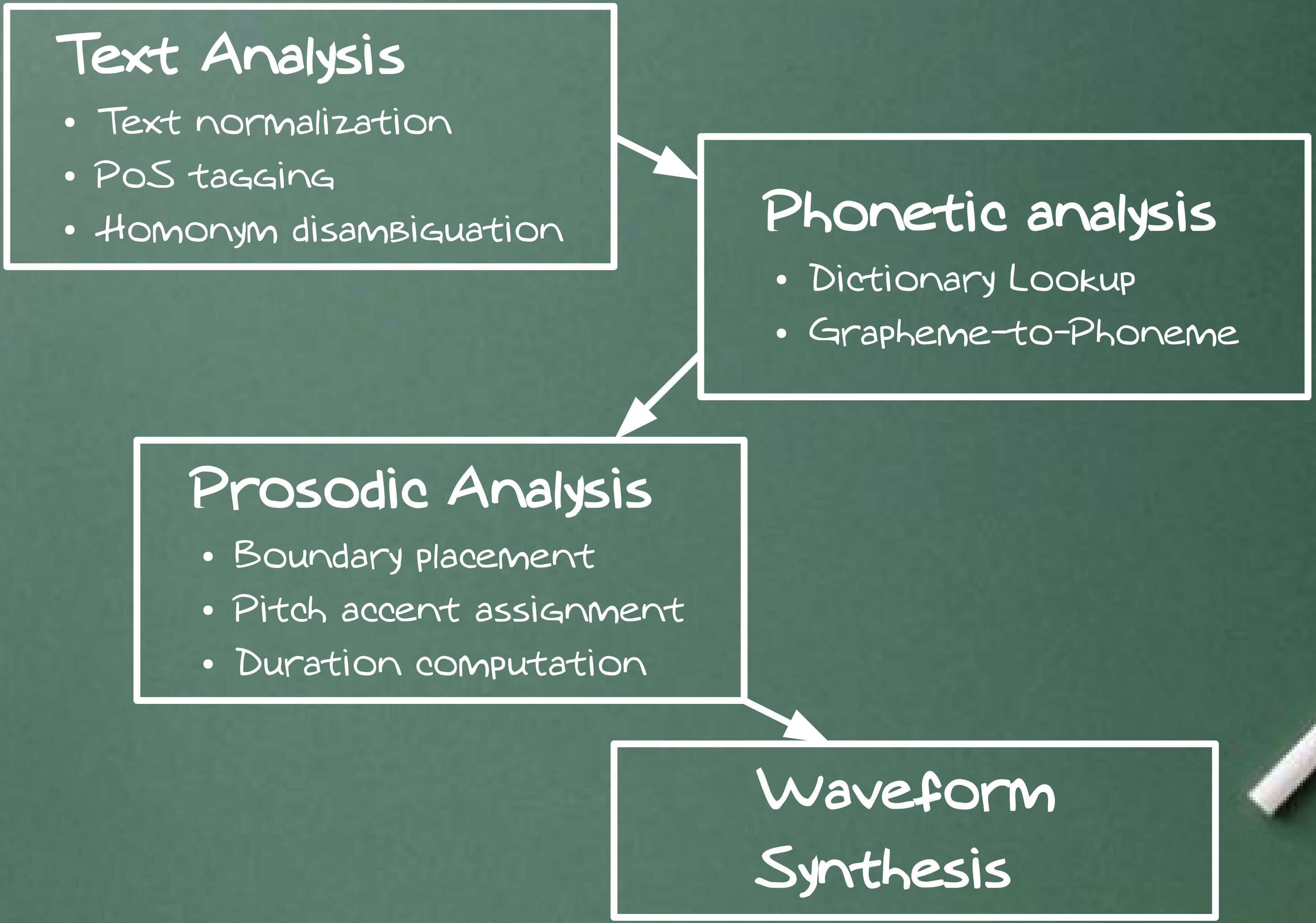
- ~~Decision-tree based (Black et al.)~~
- ~~ANN-based (NETTalk, Sejnowski et al.)~~
- ~~Pronunciation-by-Analogy (Damper et al.)~~
- ~~Memory based (MBR Talk, Stanfill)~~
- ~~Transducer-based (I. Bulyko)~~
- ~~Non segmental (A. Cohen)~~

Outline

- ~~History of Speech Synthesis~~
- ~~Text-To-Speech System Architecture~~
- ~~Grapheme to Phoneme transcription~~
- Conclusion



Text-to-Speech System



Text Analysis

- Text normalization
- PoS tagging
- Homonym disambiguation

Phonetic analysis

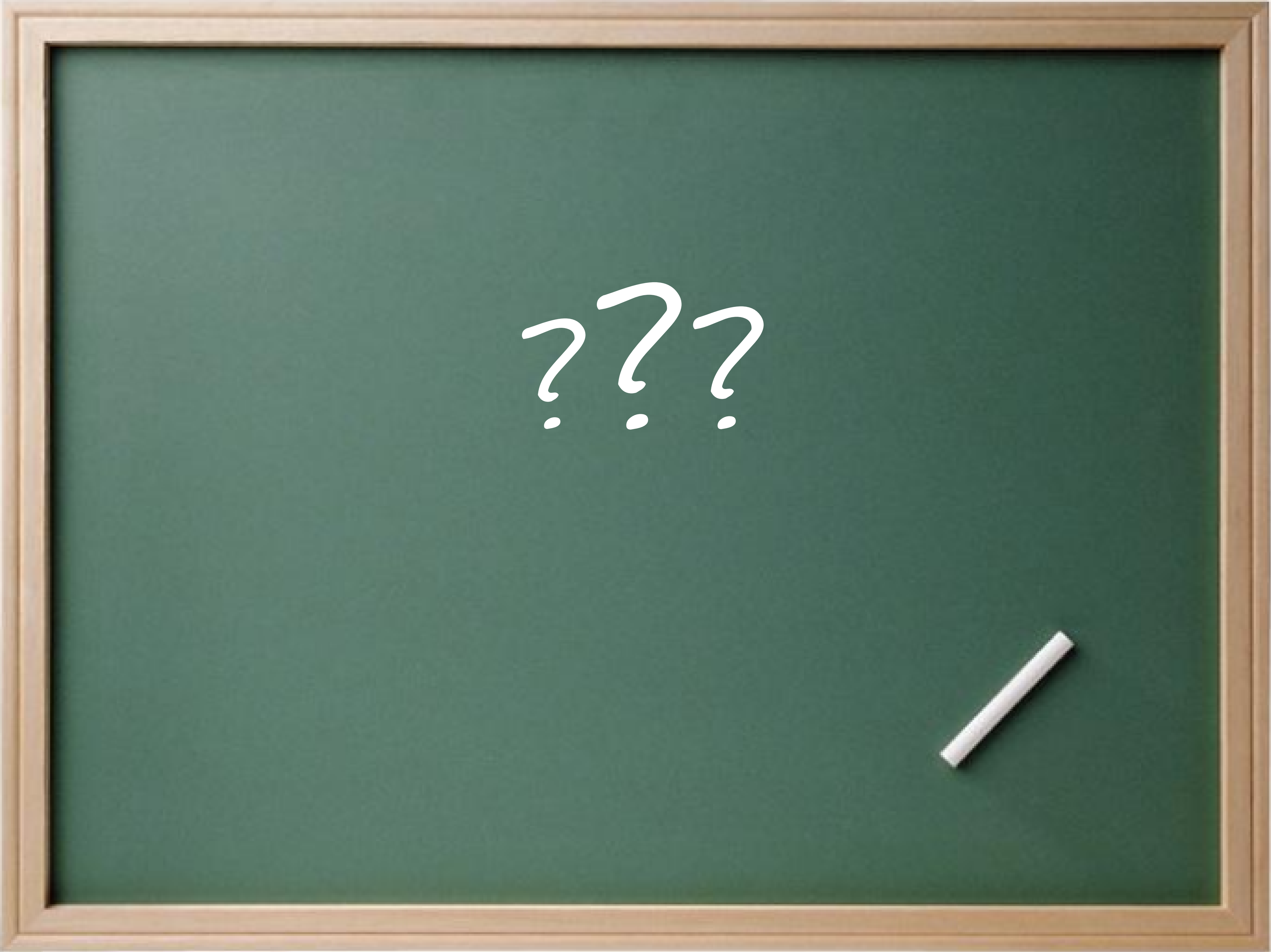
- Dictionary Lookup
- Grapheme-to-Phoneme

Prosodic Analysis

- Boundary placement
- Pitch accent assignment
- Duration computation

Waveform Synthesis

http://www.stanford.edu/class/linguist236/



???