Overview of the subject
Speech Information Systems

- Course data sheet:
- https://portal.vik.bme.hu/kepzes/targyak/VITMAD02/en/

- Information Systems
- Language, Hearing, Speech Chain
- Speech Processing
- Speech Coding, Compression
- Machine Speech Generation
- Text-to-Speech
- Speech Recognition
- Speaker Identification

Contributors (BME TMIT SmartLabs)

Dr. Németh Géza

university professor
laboratory head

I.B.151
nemeth@tmit.bme.hu

Bartalis Mátyás

departmental engineer

Dr. Kiss Gábor

research associate

Fegyó Tibor

research assistant

Jenei Attila

departmental engineer

Dr. Mihajlik Péter

Senior Research Fellow

Dr. Olaszy Gábor

professzor emeritus

Dr. Sztahó Dávid

research associate

Dr. Zainkó Csaba

assistant professor

https://smartlab.tmit.bme.hu/

Timetable

Date	Lecture	Lab (IB210 and IB211)
2/10/2025	1	Odd week
2/13/2025		VITMAD02 L1 group1
2/17/2025	2	Even week
2/20/2025		VITMAD02 L1 group2
2/24/2025	3	Odd week
2/27/2025		VITMAD02 L2 group1
3/3/2025	4	Even week
3/6/2025		VITMAD02 L2 group2
3/10/2025	5	Odd week
3/13/2025		VITMAD02 L3 group1
3/17/2025	6	Even week
3/20/2025		VITMAD02 L3 group2
3/24/2025	7	Odd week
3/27/2025		VITMAD02 L4 group1
3/31/2025	8	Even week
4/3/2025		VITMAD02 L4 group2
4/7/2025	9	Odd week
4/10/2025		VITMAD02 L5 group1
4/14/2025	10	Even week
~~4/17/2025~~	Spring break	~~VITMAD02 L5~~
~~4/21/2025~~	Spring break	~~Odd week~~
~~4/24/2025~~	OTDK	
4/28/2025	11	Even week
~~5/1/2025~~	Labour day	
5/5/2025	12	Odd week
5/8/2025		VITMAD02 L5 group2
5/12/2025	13	Even week
5/15/2025		VITMAD02 L6 group1
5/19/2025	14	Odd week
5/22/2025		VITMAD02 L6 group2

Evaluation(s)

All exercises/labs KZH (online test)

Scan the QR code to vote
or open the link

https://forms.office.com/e/v0kUb4FxBp

May 23, 14.00-16 CET : Test

May 28, Room defined later: Retake test

Min. 5 exercises/labs have to be completed

Useful websites

The Hungarian Speech website:

https://www.magyarbeszed.hu/

The Hungarian Speech book:

https://www.magyarbeszed.hu/download/A-magyar-beszed.pdf

Speech processing software:

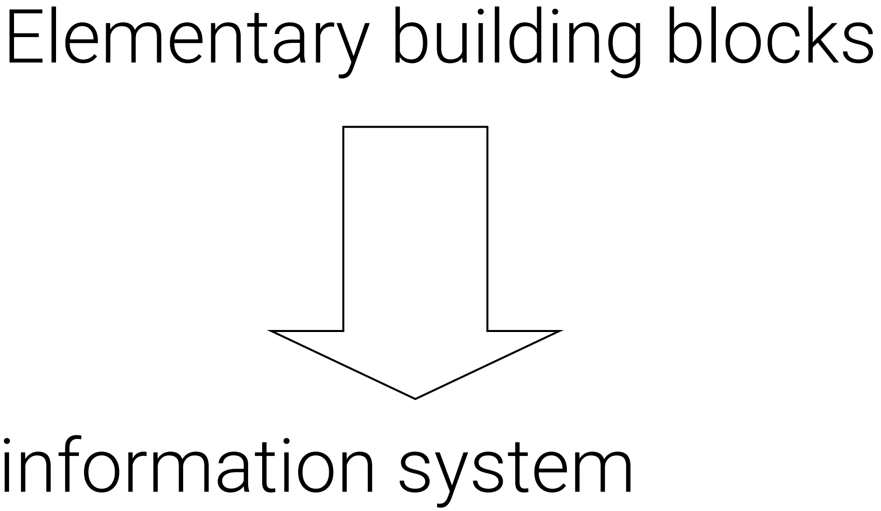
- Audacity: https://www.audacityteam.org/
- Praat: https://www.fon.hum.uva.nl/praat/

Copyright

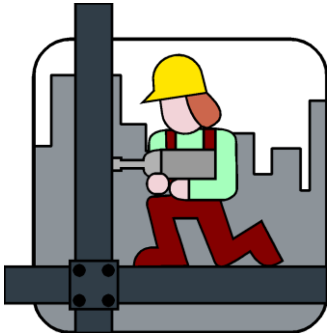
Lecture

- Building blocks of speech information systems
- Structure of speech information system
- Classification of dialogue systems
- Control of dialogue systems
- Description of dialogue systems

Speech information systems


Elementary building blocks

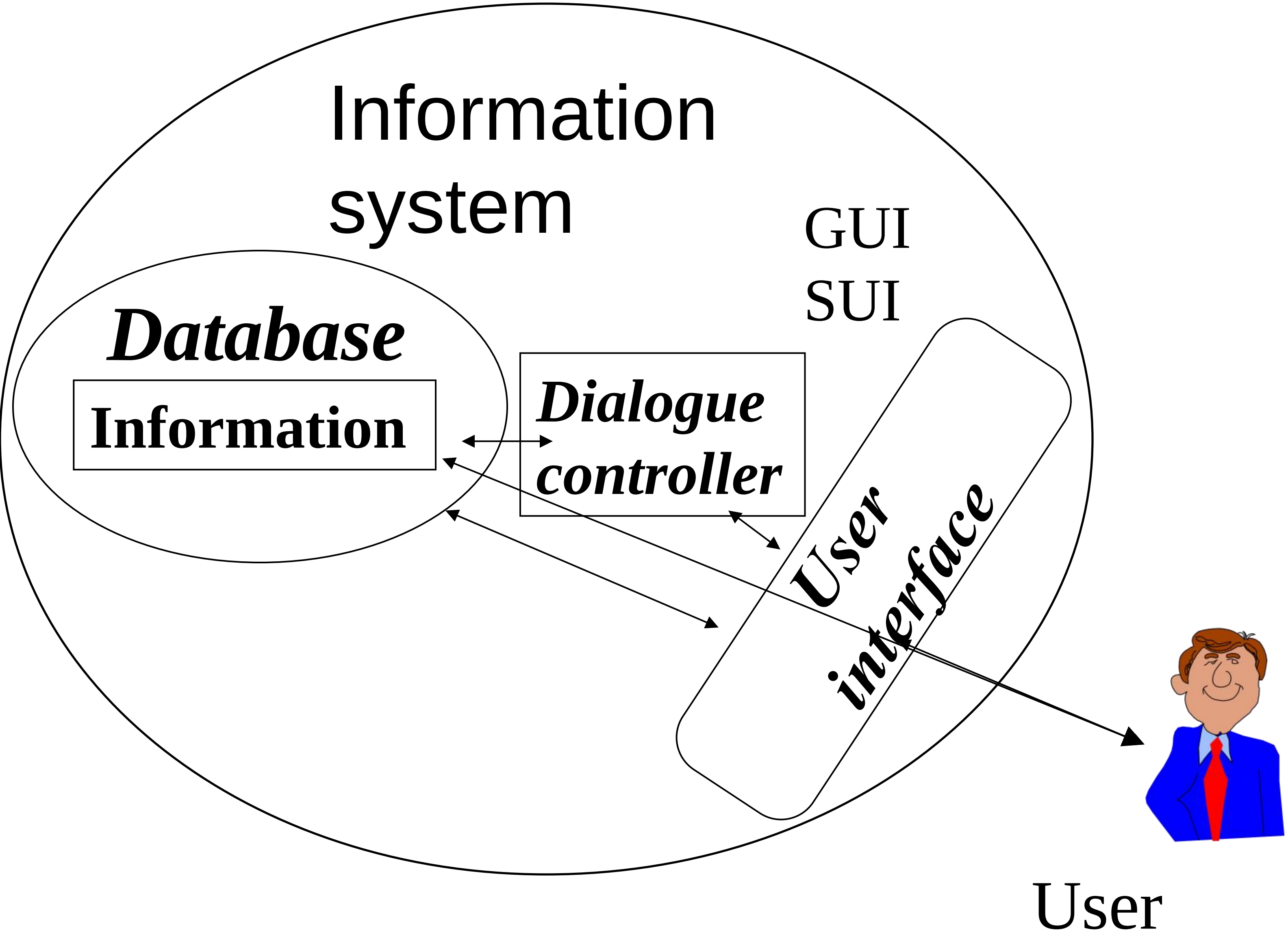
Speech information system



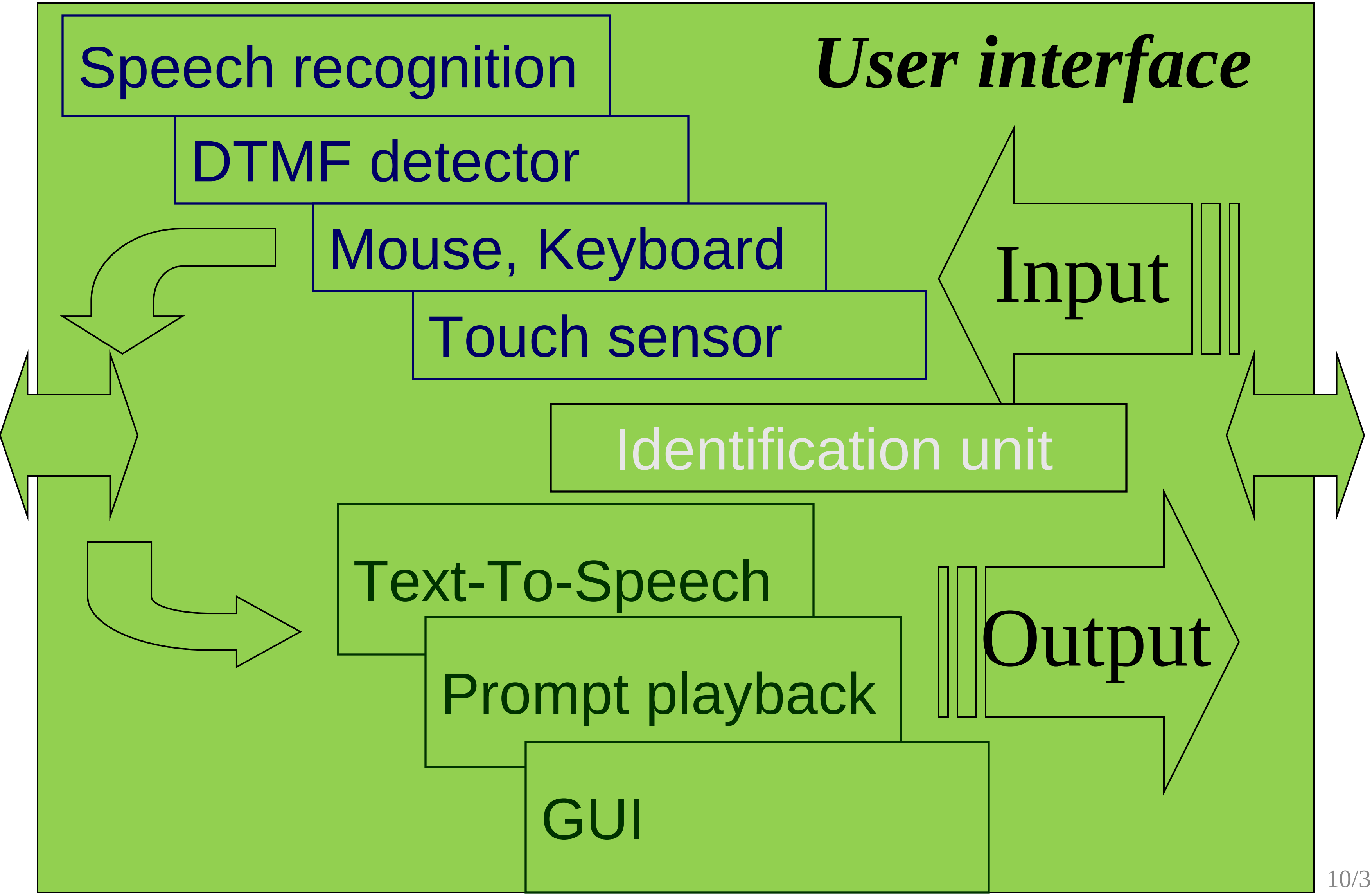
„What are we putting together?”

„How do we put it together?”

Speech information system



Speech information system



User interface

Speech recognition

DTMF detector

Mouse, Keyboard

Touch sensor

Input

Identification unit

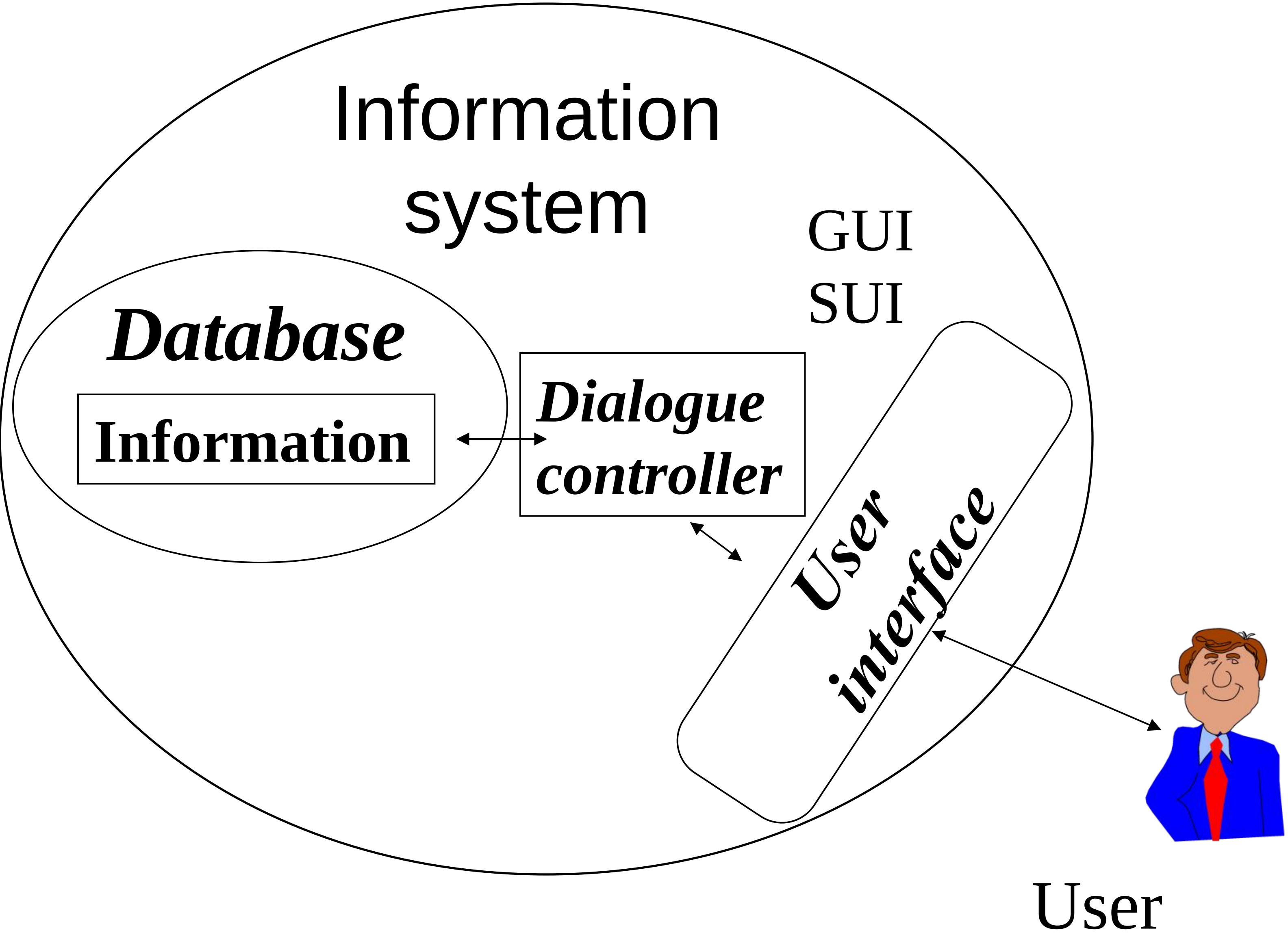
Text-To-Speech

Prompt playback

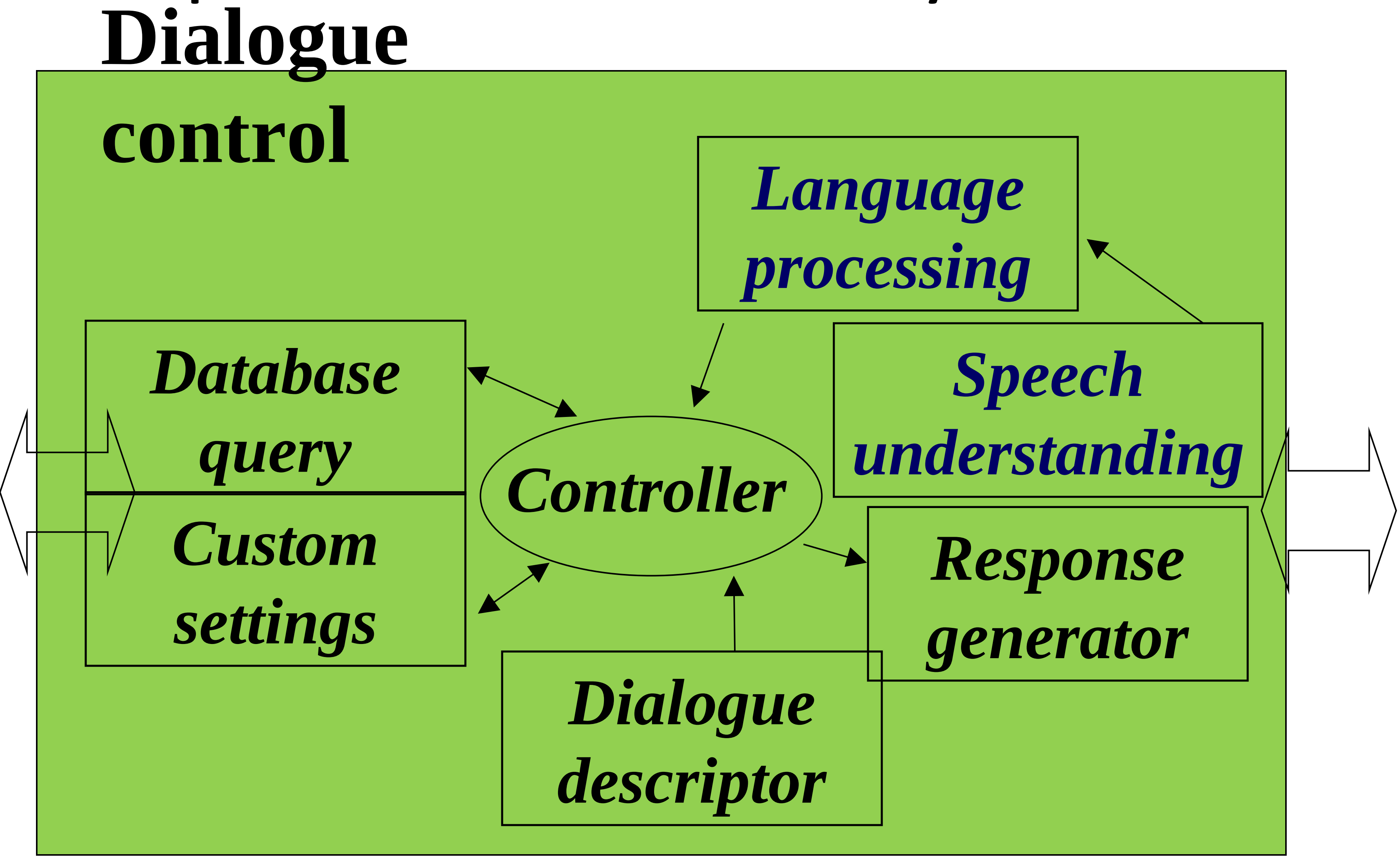
GUI

Output

Speech information system



Speech information system



System modality

- Which of our senses does it affect?

PC

- Mouse
- Keyboard
- GUI
- Speech

Information desk

- GUI
- Touch screen
- Pointing detection
- Gesture

Phone application

- Wired (classic)
- Mobile (old)
- Mobile
    - parallel or alternating sound and image
    - SMS
    - MMS, ...
- IP phone
- Videophone
- .....

Classification of dialogue systems

- According to the nature of the control
  - System controlled
  - User controlled
  - Mixed initiative
- According to the control method
  - DTMF
  - Speech recognition
  - Network information identification
  - Other

System v. user-controlled

- The system determines the navigation

  Menu system, with offered choices

- User defines navigation

  No fixed route

- Mixed initiative

  Possibility to modify navigation

Menu system design considerations

- Using building blocks
- 4 options
- Maximum depth: 4-5 levels
- Preferably same instruction -> same function
- User-dependent menu system
- Only provide information relevant to the topic
- Appropriate level of detail (little, a lot)
- Highlighting new features and important elements

System-independent custom options

- User level
  - Length and detail of questions and explanations
  - Number of choices
  - Number of choices offered
- Reading parameters
  - Speaker selection (male, female, …)
  - Setting the speech rate
  - Duration of pauses (e.g. between sentences)
- Adaptive change/user-driven

DTMF control

- Dual Tone Multi-Frequency (4x4 frequency)
- Data entry on the phone keypad
- Advantages:
  - Very reliable, practically 100%
  - proven technology
  - cheap
- Disadvantages:
  - The configurable menu structure is not user friendly
  - Difficult to use if the keyboard is not available
  - A human operator may also be required.

Voice control

- Advantages:
  - Talking on the phone is natural
  - Applicable to a wider range of topics (not just numbers)

- Disadvantages:
  - More unreliable
  - In some cases much slower than DTMF
  - Limited vocabulary

Voice control II.

- Yes/No systems
  - Slow, unnatural
  - Most reliable in speech recognition systems
  - You have to choose the Hungarian/non-English equivalent of Yes/No carefully.
    - The distance between "Yes" and "No" may be smaller than in English
  - A 2-word dictionary is not enough (yes, good, ok, okay, okay, yeah, aha, okay,...)
  - Data entry in a tree structure

Voice control III.

- Small dictionary
  - Recognize specific (10-20) words
  - More convenient, but menu-based
  - Speaker-independent/adaptive
  - Fast access even in deep structures.
  - Can be mixed with DTMF control “fall -back ”
  - Data entry is difficult

Voice control IV.

- Any information on a given topic can be said or queried.
- User-driven
- Human use
- E.g.: Ticket purchasing system
- Dictation systems
- Language analysis (artificial intelligence ??) required

Confirmation (Verification)

- Required (incorrect input)
- Expected (user's sense of security)
- Confirmation can be
  - Explicit (e.g. with direct feedback)
  - Implicit (e.g.: Hidden in the next question)

Explicit confirmation for each data

- Query for each data item
  - Question to be decided
  - Simple structure,
  - Inconvenient dialogue process for the user

H: I would like to travel to Szeged.

R: Would you like to travel to Szeged?

H:No.

R: Where would you like to go?

H: Szöged.

R: Do you want to travel to Szöged?

H: Yes.

…

Explicit confirmation with correction

- Query for each data item
  - Same as before.
  - In addition to a Yes/No answer, corrected data can also be provided.
  - Faster dialogue progress
  - Less stumbling

H: I would like to travel to Szeged.
R: Would you like to travel to Szeged?
H: No, to Szöged.
R: Do you want to travel to Szöged?
H: Yes.
…

Explicit confirmation for multiple data

- Query all data at once
  - Fewer questions
  - In addition to the Yes/No answer, corrected data can/cannot be provided.
  - More natural

R: From Budapest To Szeged would like to travel ?

H:No, to Szöged .

R: Do you want to travel from Budapest to Szöged?

H: Yes .

- The check is only done at the end of the data entry.

Implicit Confirmation

- Next data request embedded validation
- Closer to natural dialogue
- The length of the question increases
- Repair is more difficult
- The system is more complicated

Mon: To Szeged I would like to travel .
R: Where would you like to travel to Szeged from?
H: No , to Szöged .
R: Where would you like to travel to Szöged from?

Incorrect confirmation

- Recognition problem

R: Would you like to travel to Szeged?
H: Yes, to Szeged. (Recognized: not to Szeged)
R: Where would you like to travel?

R: Do you want to travel to Szeged?
H: Szeged.. (Recognized: Szeged)
R: The ticket will be ... Forint...

- Multiple confirmation for critical

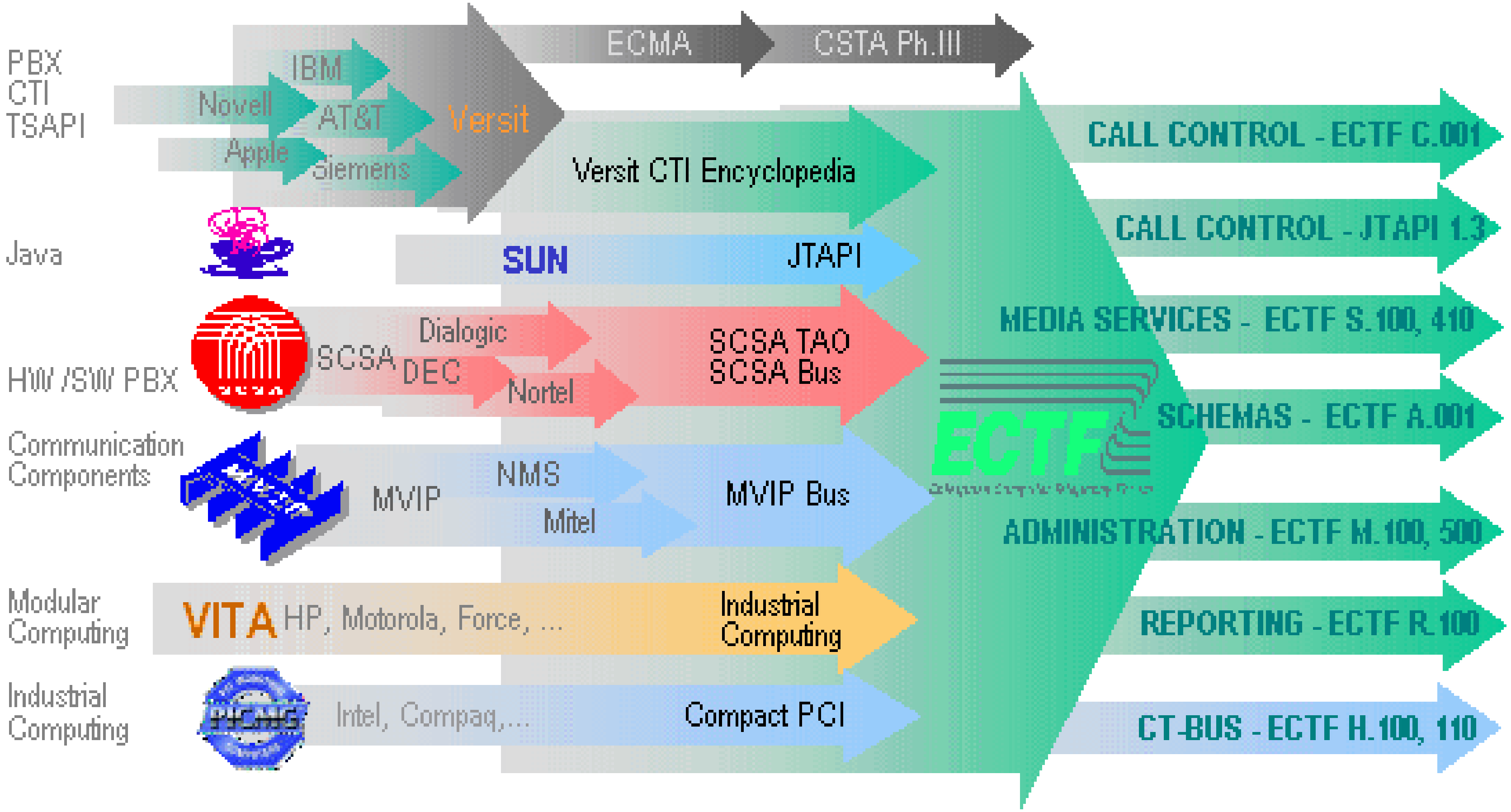
R: When do you want to travel to Szeged?
H: No, Szeged. (Recognizing: ...)
A: I don't understand. Repeat when you want to travel.!

Application development tools

- SAPI (http://www.microsoft.com/speech/)
  - SAPI 4.0 <> 5.0 <> .NET
  - TAPI (http://msdn.microsoft.com, search TAPI)
- JSAPI (http://java.sun.com/products/java-media/speech/)
  - experimental phase (e.g. FreeSpeech TTS)
  - JTAPI (http://java.sun.com/products/jtapi/)
- VocApi (Philips, Bosch, Siemens, Opel, Sony, Volkswagen..)
  - www.speech.philips.com/vc/Pages/vc_322_u.htm
  - Small appliances, telephone, washing machine, photocopier
- MRCP (http://tools.ietf.org/wg/speechsc/)
  - Media Resource Control Protocol
  - standard IP interface

Application development tools II.

- ECTF (http://www.ectf.org/)

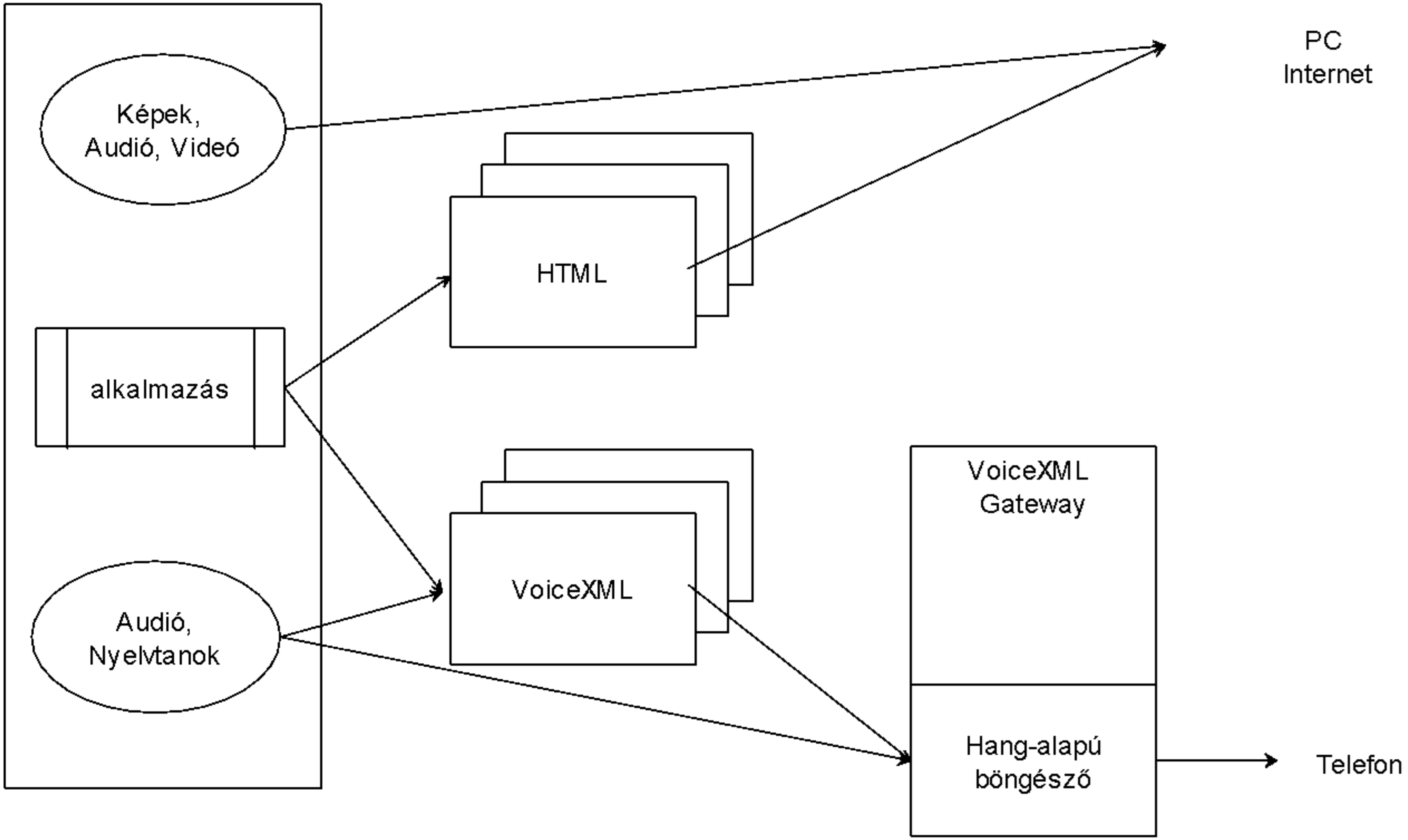


VoiceXML (http://www.voicexml.org/)

- Voice eXtensible Markup Language
  - Version 1.0 March 17 , 2000 .
  - W3C Recommendation for VoiceXML 2.0 , 2004.Mar.16
  - W3C v3.0 Working Draft Dec 2008 19 (http://www.w3.org/Voice/)
- example:

```
<?xml version="1.0"?>
<vxml version="1.0">
    <form>
        <field name="drink">
            <prompt>Would you like coffee, tea, milk, or
            nothing?</prompt>
            <grammar src="drink.gram" type="application/x-jsgf"/>
        </field>
    <block>
        <submit next="http://www.drink.example/drink2.asp"/>
    </block>
    </form>
</vxml>
```

VoiceXML (http://www.w3.org/Voice/Activity.html)

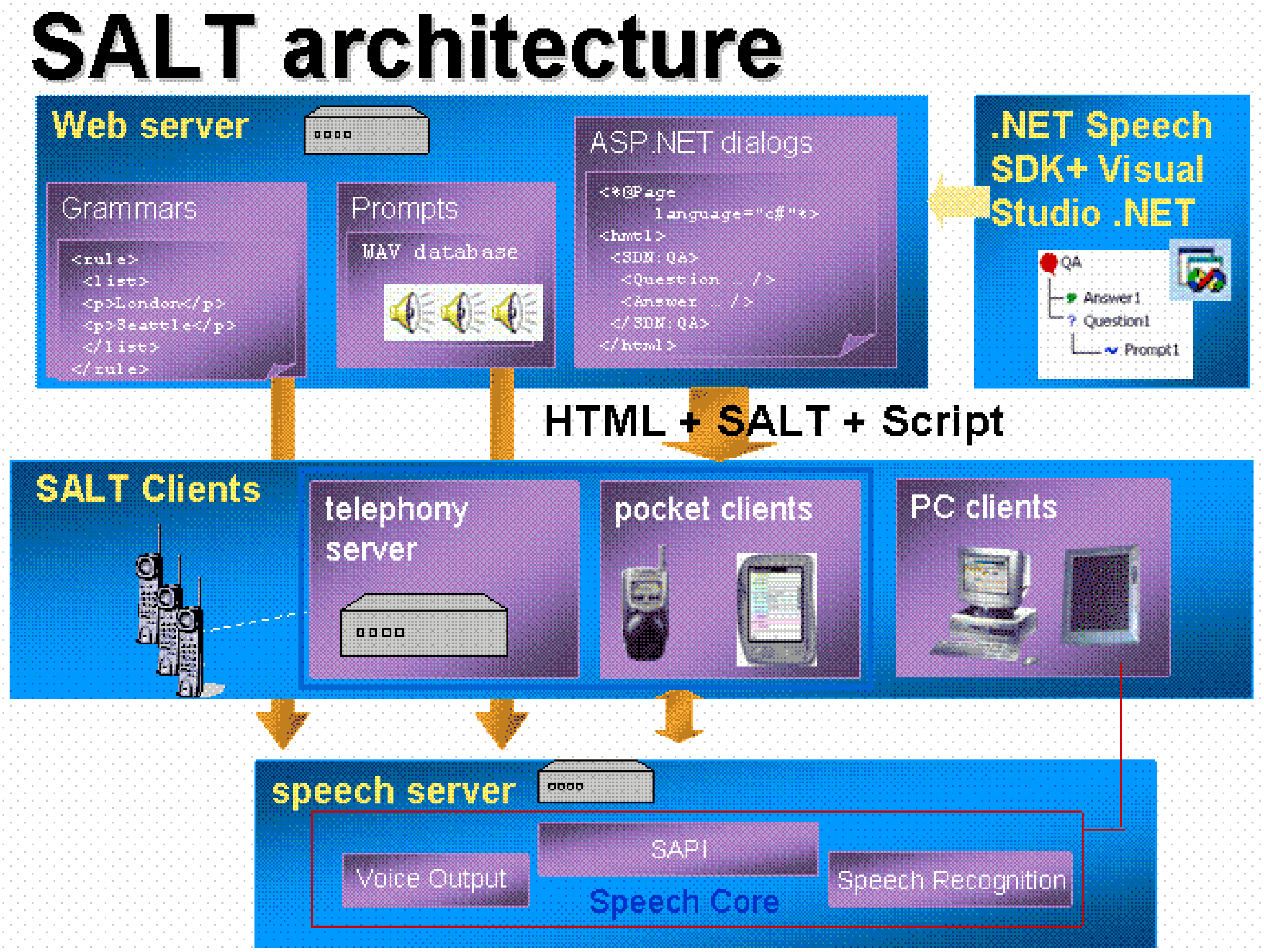


SALT (http://www.saltforum.org/)

- Speech Application Language Tags
  - Version 1.0 July 15, 2002
  - W3C submission, 13 Aug 2002
- Multimodal and telephone HTML (cHTML , XHTML, WML, etc.) extension with strong .NET orientation
- Example:

```
<!-- HTML -->
<html xmlns:salt="http://www.saltforum.org/2002/SALT">
        ...
        <input name="txtBoxCity" type="text" />
        <input name="buttonCityListen" type="button" onClick="listenCity.Start();" />
        ...
        <!-- SALT -->
        <salt:listen id="listenCity">
        <salt:grammar name="g_city" src="./city.grxml" />
        <salt:bind targetelement="txtBoxCity"
        value="//city" />
        </salt:listen>
</html>
```

SALT (http://www.saltforum.org/)



```
<rule>
 <list>
 <p>London</p>
 <p>Seattle</p>
 </list>
</rule>
```

```
<*@Page
     language="c#"*>
<hmtl>
 <SDN:QA>
  <Question ... />
  <Answer ... />
 </SDN:QA>
</html>
```

Speech API-s of large companies

Google Speech API:

- https://www.cloudskillsboost.google/course_templates/7

Microsot Aure AI Speech:

- https://azure.microsoft.com/en-us/products/ai-services/ai-speech?msockid=122c022bb722642b208217a3b6fc651b

Amazon Connect:

- https://docs.aws.amazon.com/connect/latest/adminguide/connect-conversational-ai-bots.html

META LLAMA:

- https://ai.meta.com/blog/future-of-ai-built-with-llama/

IBM speech services:

- https://www.ibm.com/search?lang=en&cc=us&q=Speech%20services

………..