Human-Computer Interaction: Speech Interfaces and e-Inclusion

TRABHCI

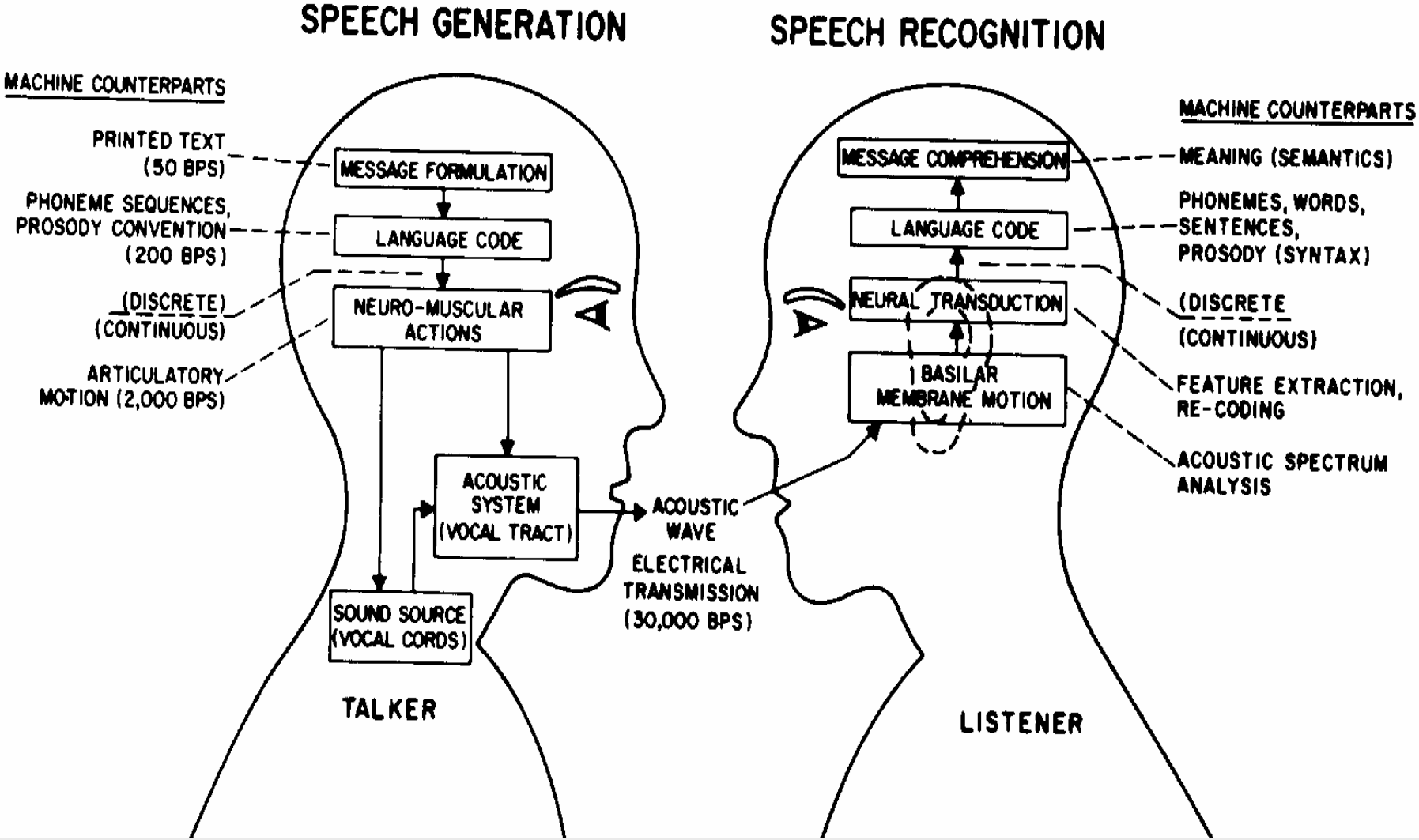
Valencia 2011

Outline

- Human-Computer Interaction
  - Human-Human communication: Speech
  - Human-Computer Interfaces. E-Inclusion.
- Brief Introduction to Speech and Language Technologies
  - The speech signal and its properties.
  - Speech Technologies
  - Automatic Speech Understanding Systems
  - Spoken Dialog Systems
- Speech Technology for e-Inclusion and therapy support
  - Speech Technologies for e-Inclusion
  - Computer-aided Language Learning and Rehabilitation: Pre-linguistic skills.
  - Computer-aided Language Learning and Rehabilitation: Articulatory and Language skills
- Application Development
  - Distributed Speech Recognition
  - Google tools
  - Assistant transcription tools

http://www.youtube.com/watch?v=Y0hl1-06gOo

Human-Computer Interaction



Human-Computer Interaction

- Human-Computer Interaction:
  - Design, evaluation and implementation of interactive computing systems for human use with the study of major phenomena surrounding them.

    [ACM SIGCHI Curricula for Human-Computer Interaction]
  - User Interface is more than a person using an interactive graphics program on a workstation.
    - can be part of spacecraft cockpits or microwave ovens.
  - The design of the HCI must take into account not only the machine or the task but also the human.
  - We will focus here on the more natural way of interaction for the human: Speech.

e-Inclusion

- **Information and Communication Technologies (ICT)** play an essential role in **supporting daily life** in today's digital society.
  - They are used at work, to stay in touch with family, to deal with public services as well as to take part in culture, entertainment, leisure and political dialogues.
- **e-Inclusion** aims to achieve that **"no one is left behind"** in enjoying the benefits of ICT.
  - It focuses on participation of all individuals and communities in all aspects of the information society. e-Inclusion policy, therefore, aims at reducing gaps in ICT usage and promoting the use of ICT to overcome exclusion, and improve economic performance, employment opportunities, quality of life, social participation and cohesion.

Europe´s Information Society Thematic Portal

http://ec.europa.eu/information_society/activities/einclusion/index_en.htm

Human & Environment Interaction

- **Gesture & Speech & Environment**

Speech Gesture Recognition

http://www.youtube.com/watch?v=inlp8qN93Ys

Touch screen based on gesture and speech recognition

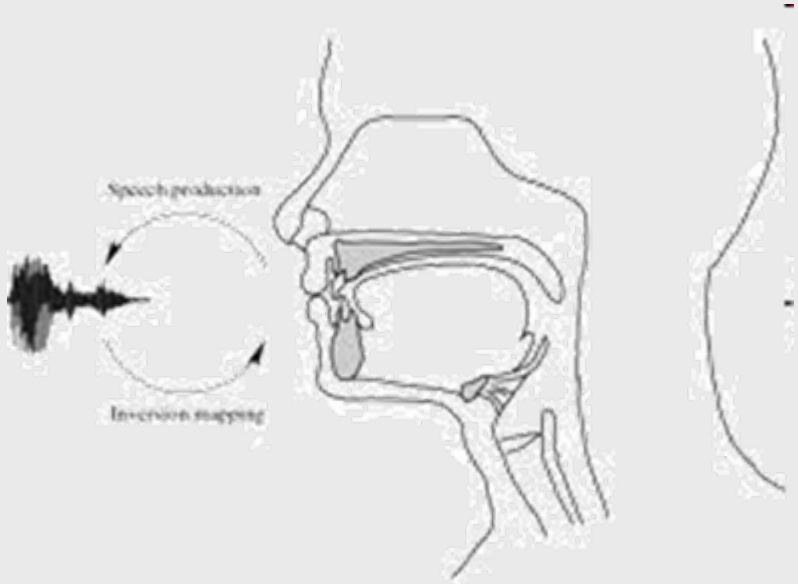
http://video.google.com/videoplay?docid=-8291945120462236123#

Speech Functions and gestures

http://www.youtube.com/watch?v=8BKMoh3RTzk

Multisensorial rooms in special education schools

http://www.youtube.com/watch?v=lCvtIIdQgYk



The Speech Signal and Its Properties

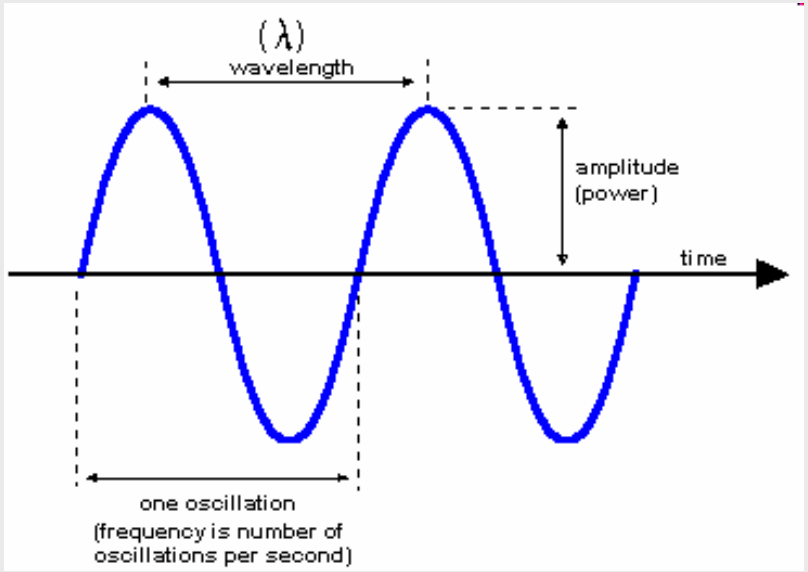
The speech signal and its properties

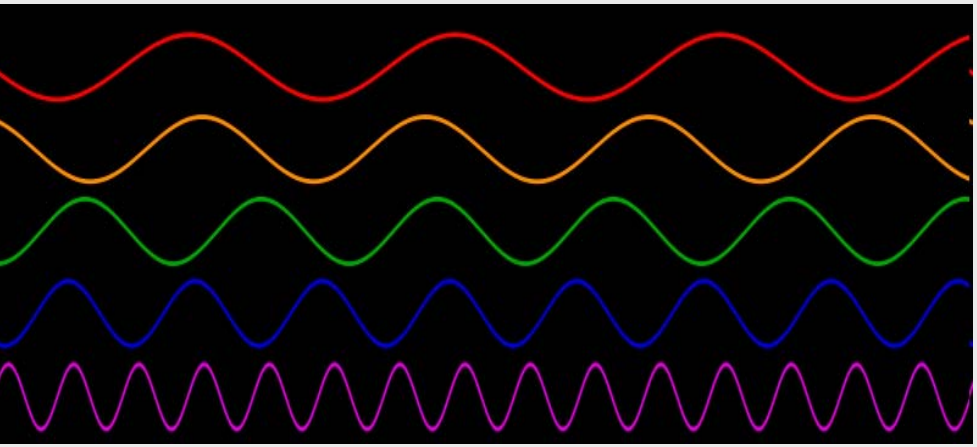
- What is a signal?
  - a signal is a physical magnitude used to convey information from one place to another.
  - a signal is time-dependent variation of a characteristic of a physical phenomenon, used to convey information
- What is speech?
  - The faculty or act of expressing or describing thoughts, feelings, or perceptions by the articulation of words.

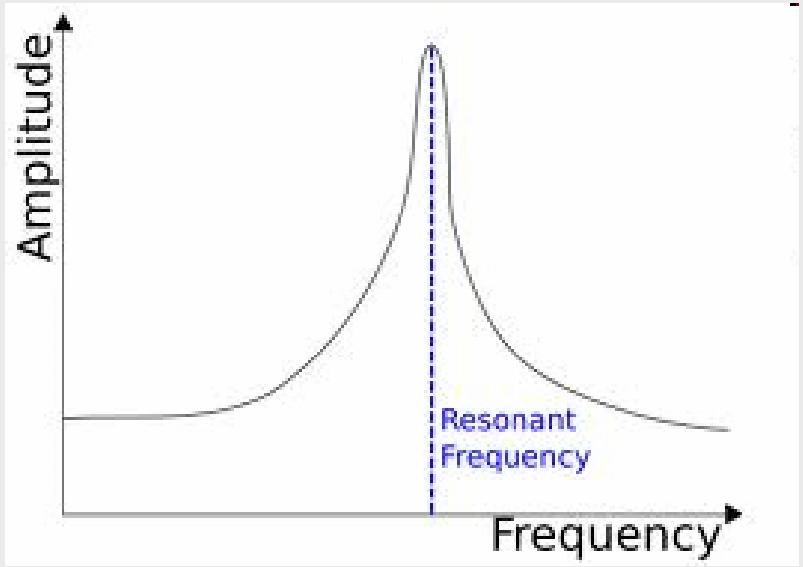
The speech signal and its properties

- How is represented a signal?
  - Time
    - waveform →plot the variation of the physical magnitude with the time (independent variable)
  - Frequency
    - Related with periodic repetition of a physical magnitude. Number of times a physical magnitude is repeated by second
    - Plot the energy distribution of the physical magnitude with the frequency
  - Time-Frequency

The speech signal and its properties

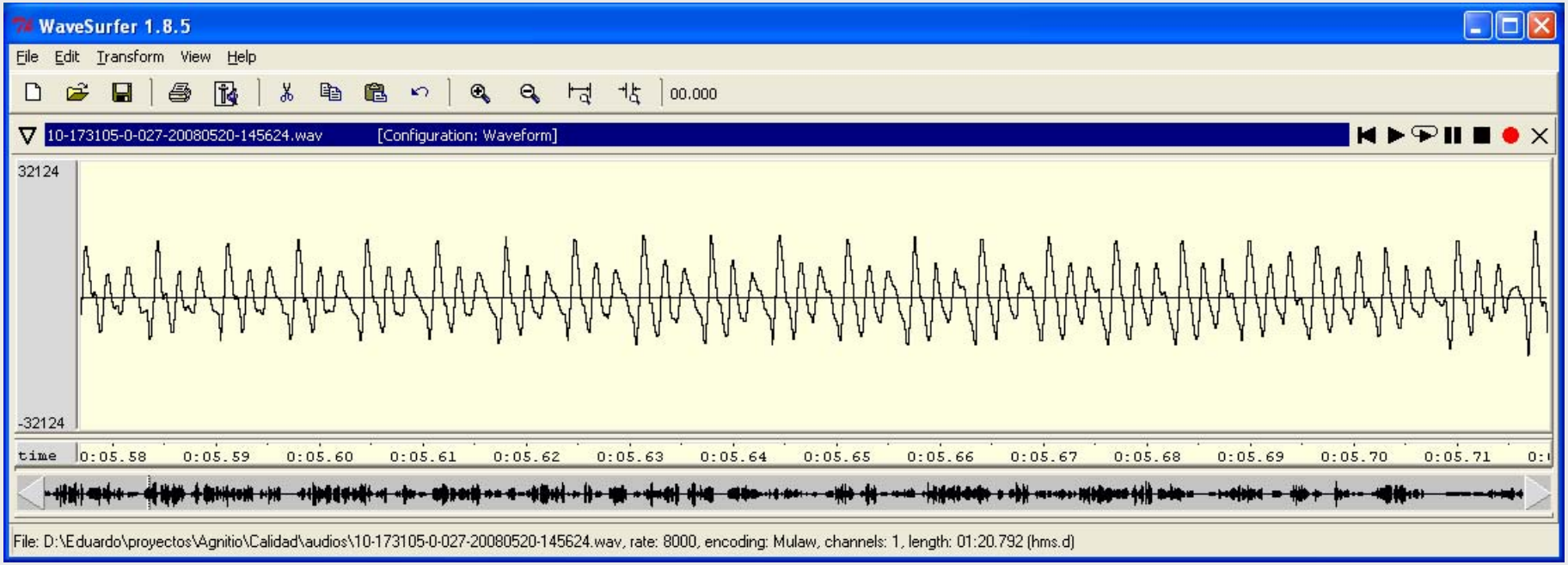






The speech signal and its properties

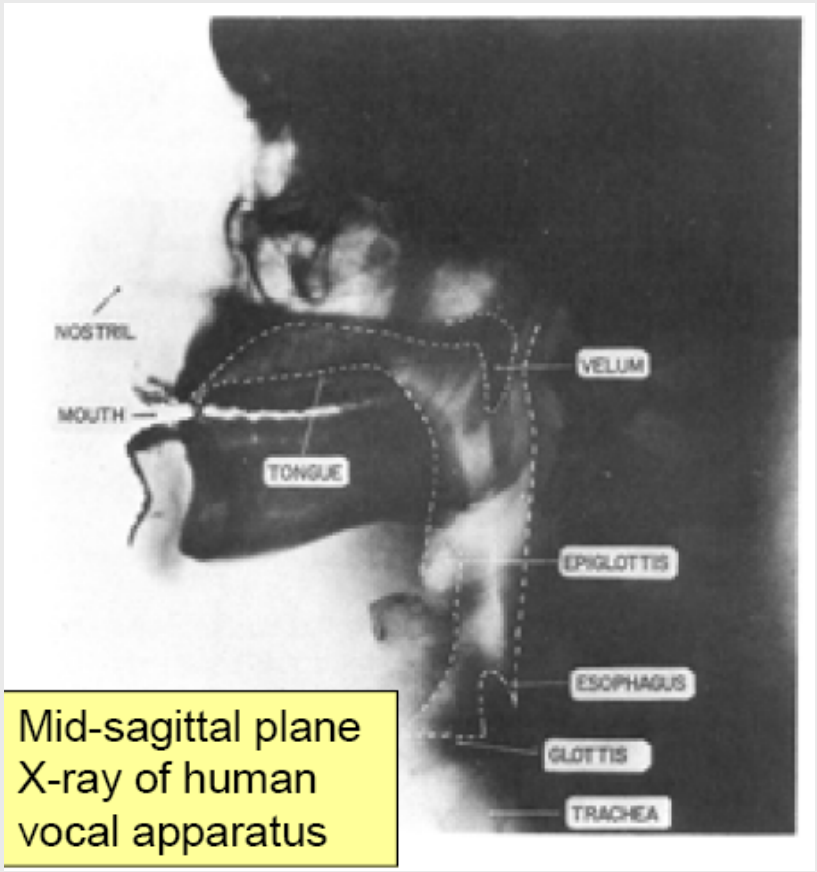
- What is a speech signal?
  - is the physical representation of the speech: a pressure signal converted on a electrical signal by means of a microphone



The speech signal and its properties

- How is produced the speech signal?

Vocal human apparatus



Mid-sagittal plane X-ray of human vocal apparatus

Vocal tract: begins at the glottis (vocal cords) and ends at the lips.

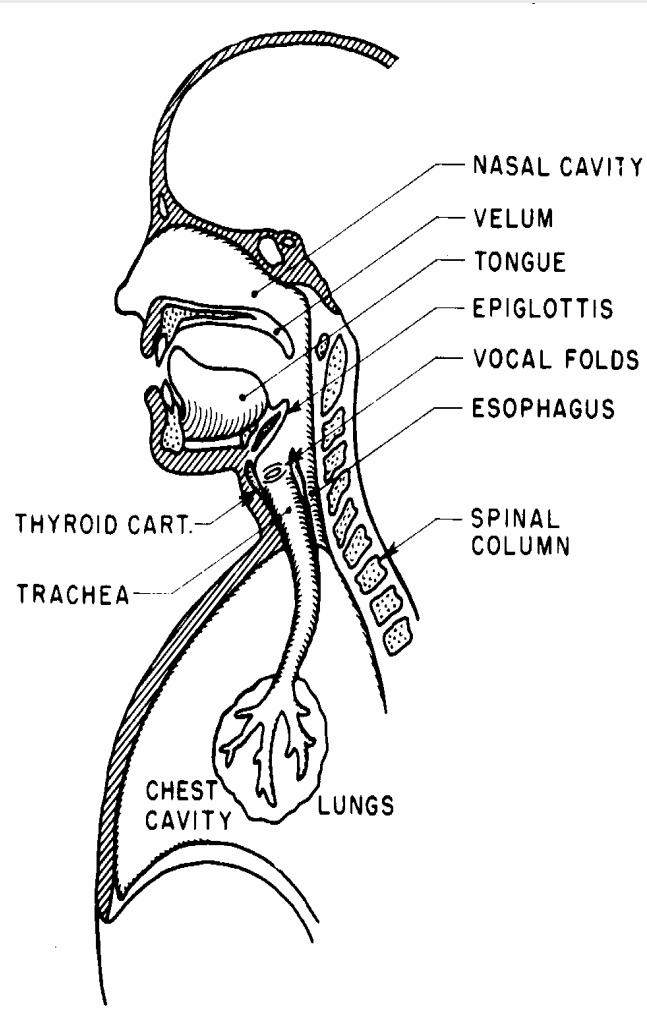
Nasal tract: begins at the velum and ends at the nostrils

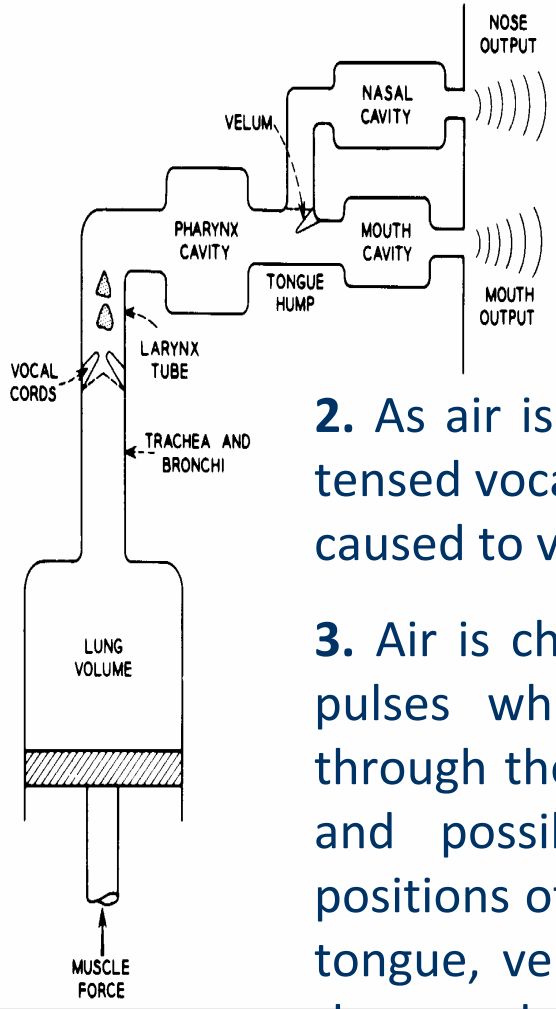
Velum: lowers to couple the nasal tract to the vocal tract to produce the nasal sounds like /m/ (mom), /n/ (night) or /ng/ (sing)

Vocal cords: pair of muscles in the glottis.

The speech signal and its properties

- How is produced the speech signal?





1. Air enters the lungs via normal breathing and no speech is produced (generally) on in-take.

2. As air is expelled from the lung, the tensed vocal cords with in the larynx are caused to vibrate by the air flow

3. Air is chopped up into quai-periodic pulses which are filtered in passing through the pharynx, the mouth cavity, and possibly the nasal cavity; the positions of the various articulator (jaw, tongue, velum, lips, mouth) determine the sound that is produced

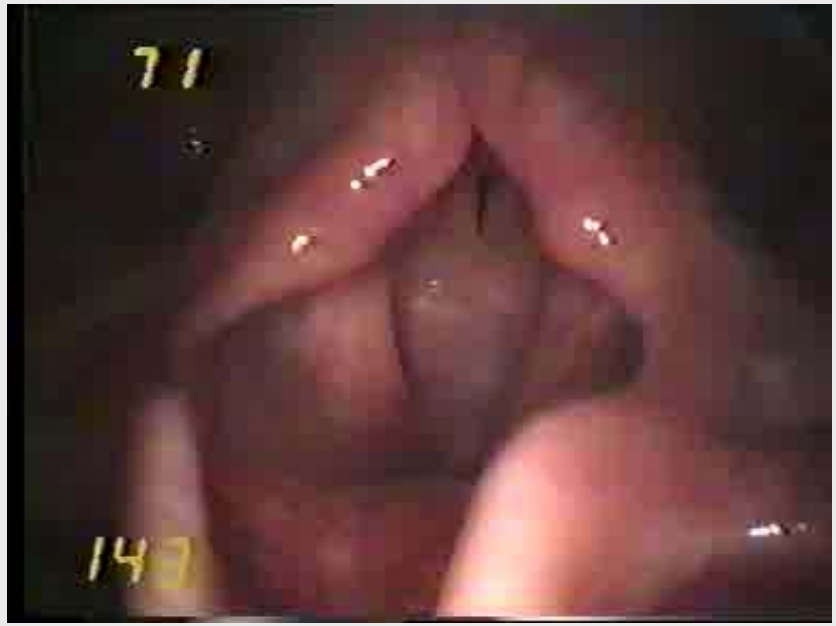
The speech signal and its properties

- The vocal cords
- A pair of elastic structures of tendon, muscle and mucous membrane

  15 mm long in men

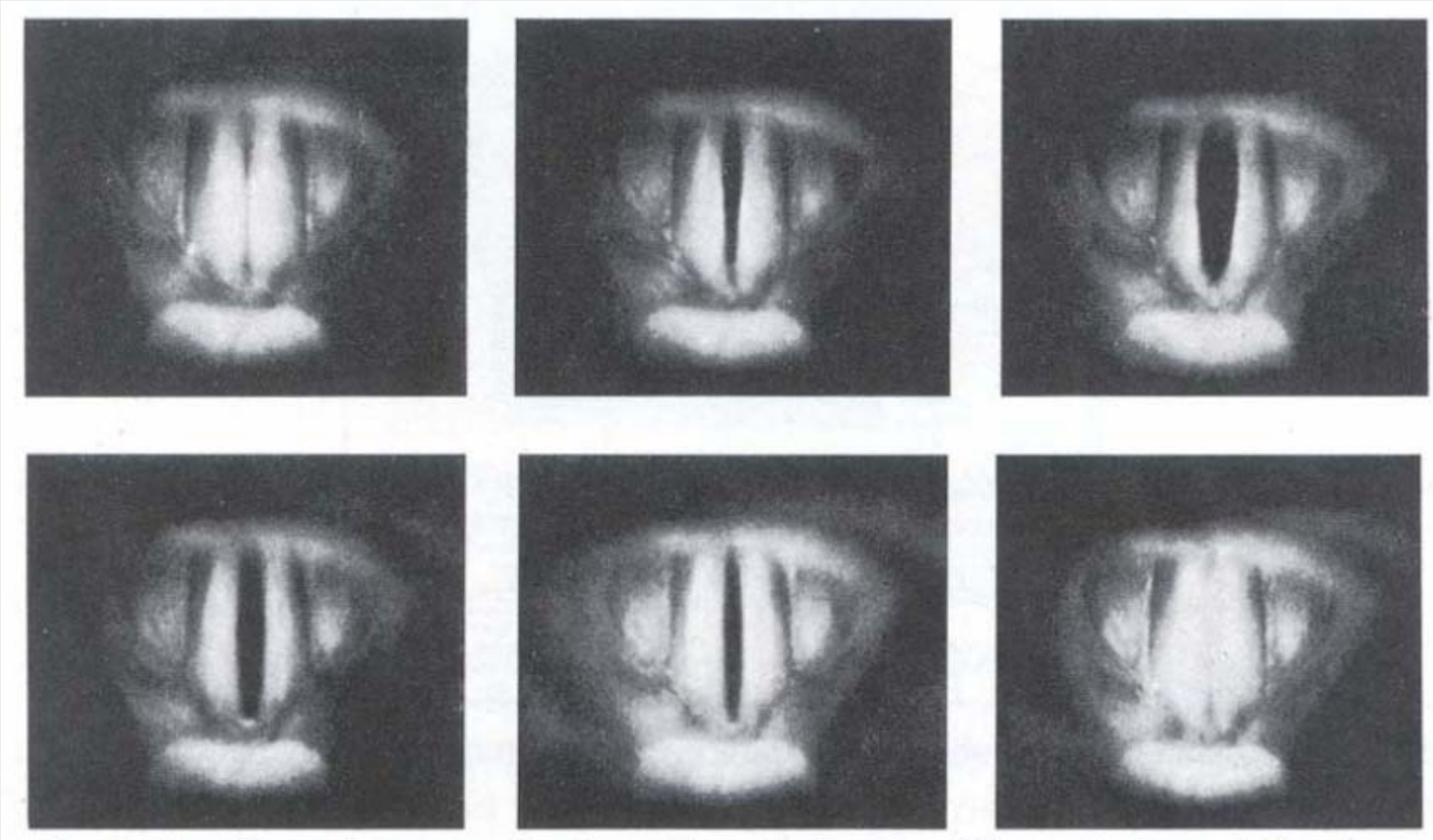
  13 mm long in women
- Can be varied in length and thickness and positioned
- Successive vocal fold openings
  - the fundamental period
  - the fundamental frequency

    or *pitch*
  - -> men: 100-200 Hz
  - -> women: 150-300 Hz



The speech signal and its properties

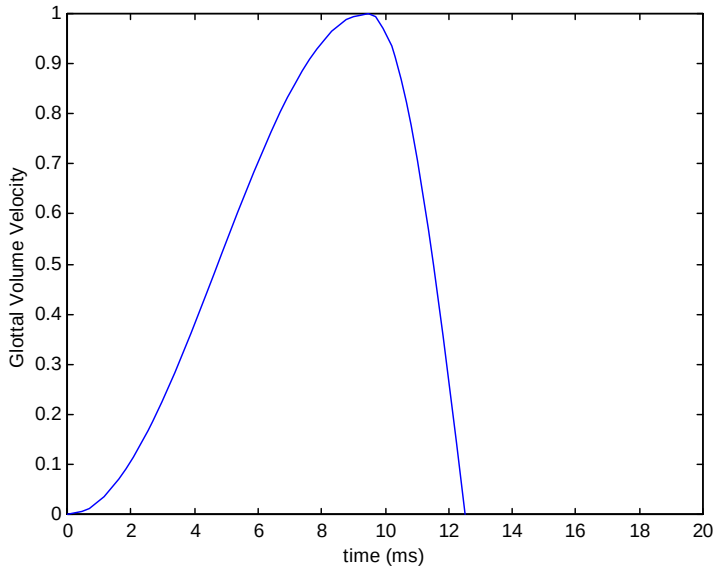
- The vocal cords

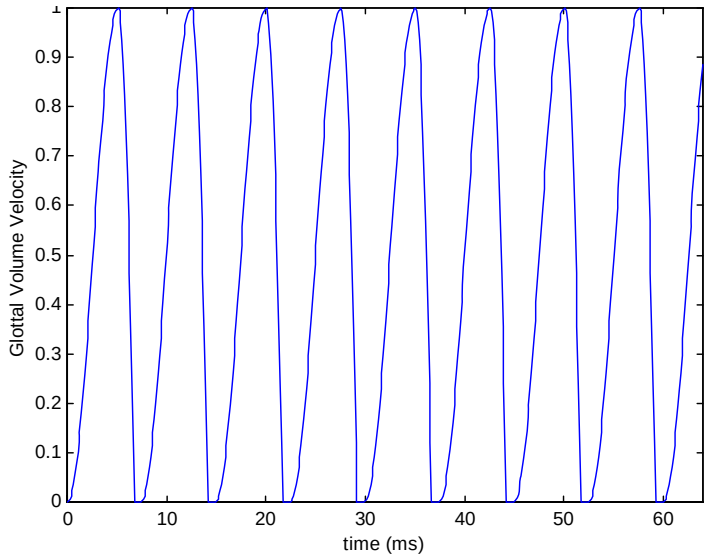


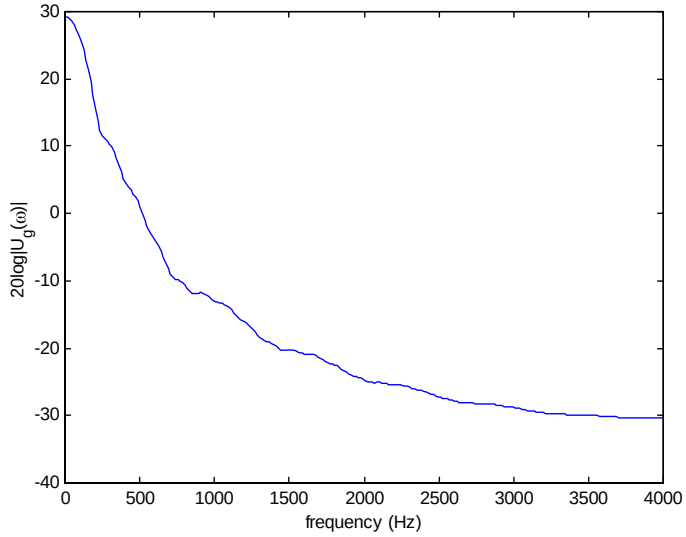
Successive phases in one cycle of vocal cord vibration. The total elapsed time is approximately 8 msec

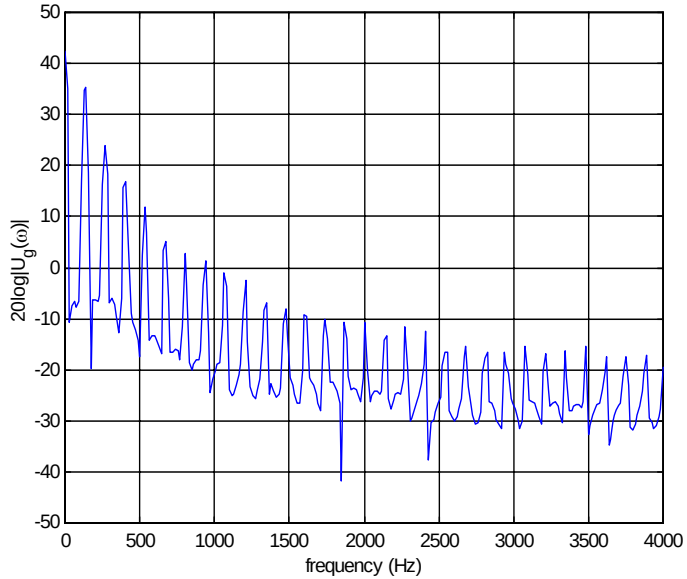
The speech signal and its properties

Vocal cords: frequency Properties



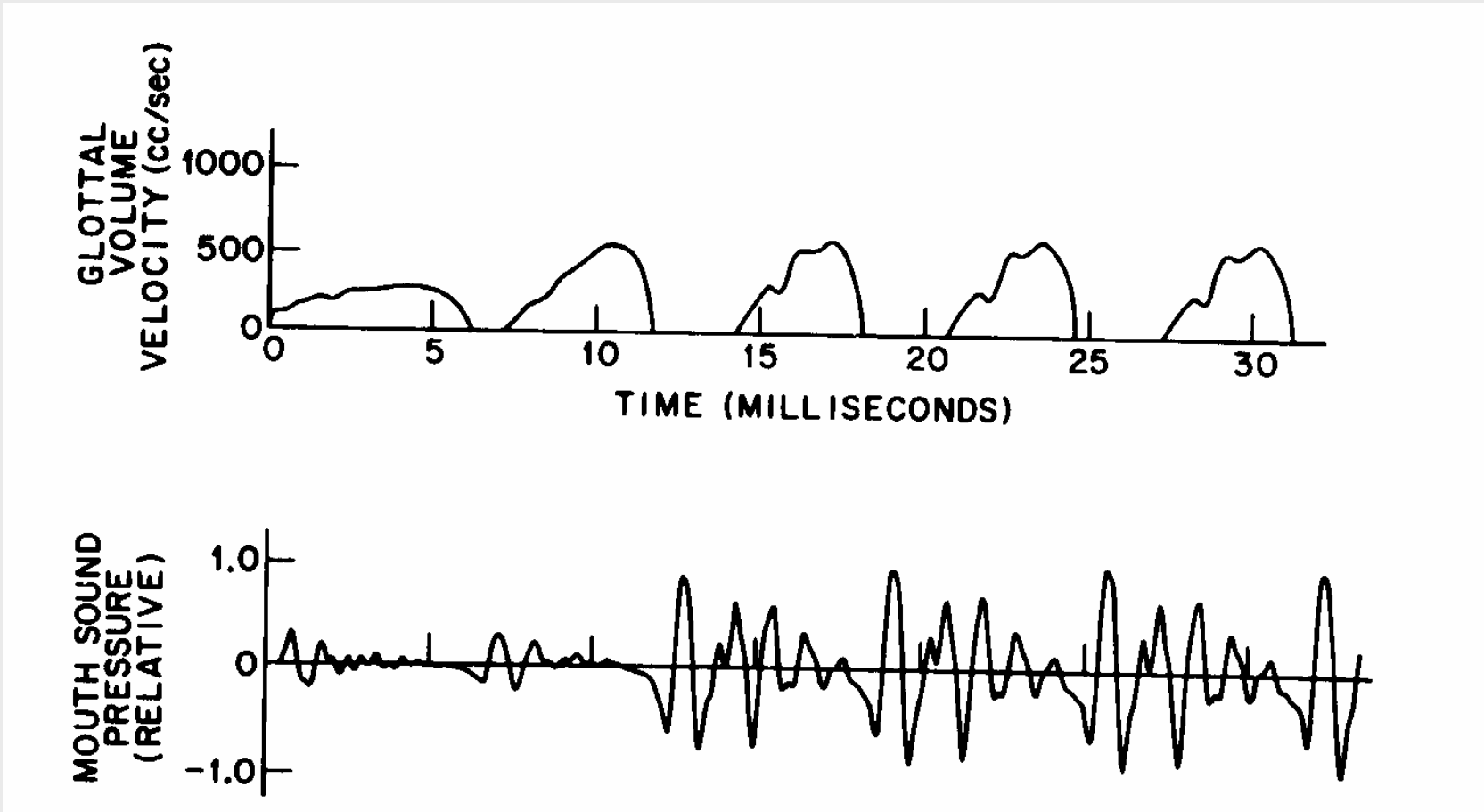






The speech signal and its properties

- From the vocal cords to the lips

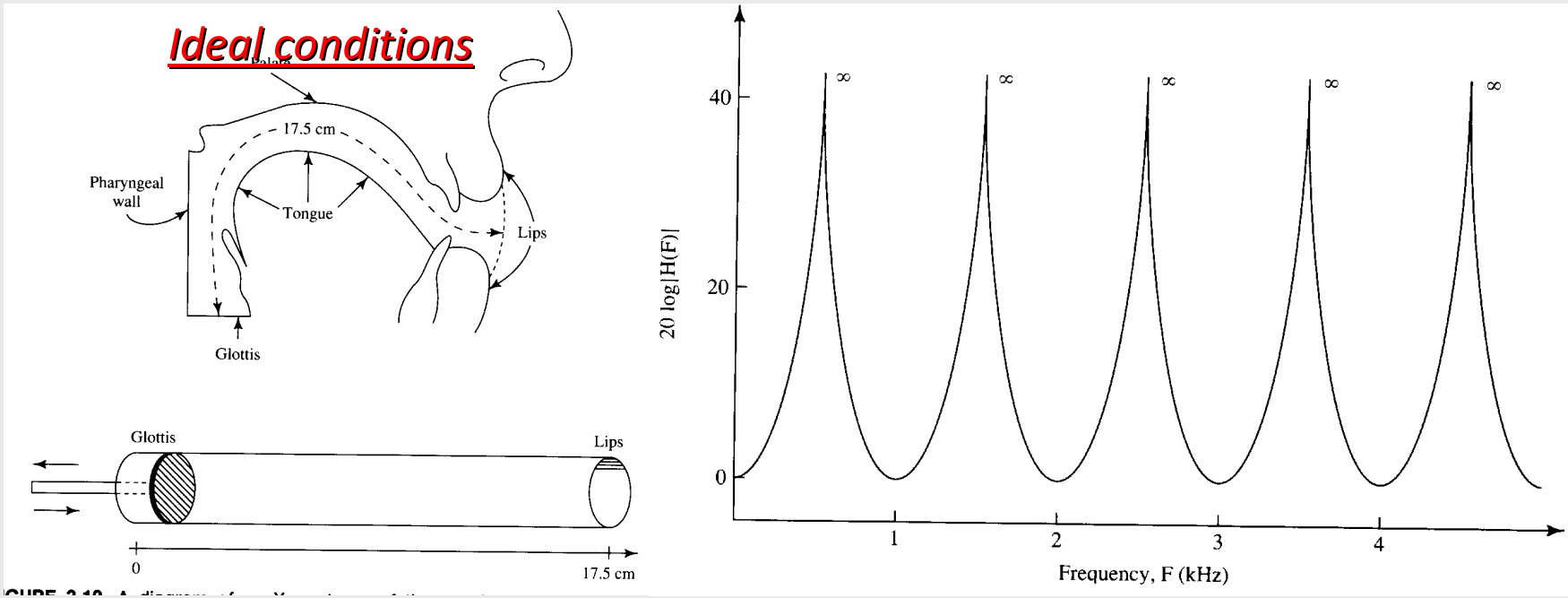


The speech signal and its properties

Vocal Tract: Composed by the Pharyngeal and Oral cavities

Basic functions:

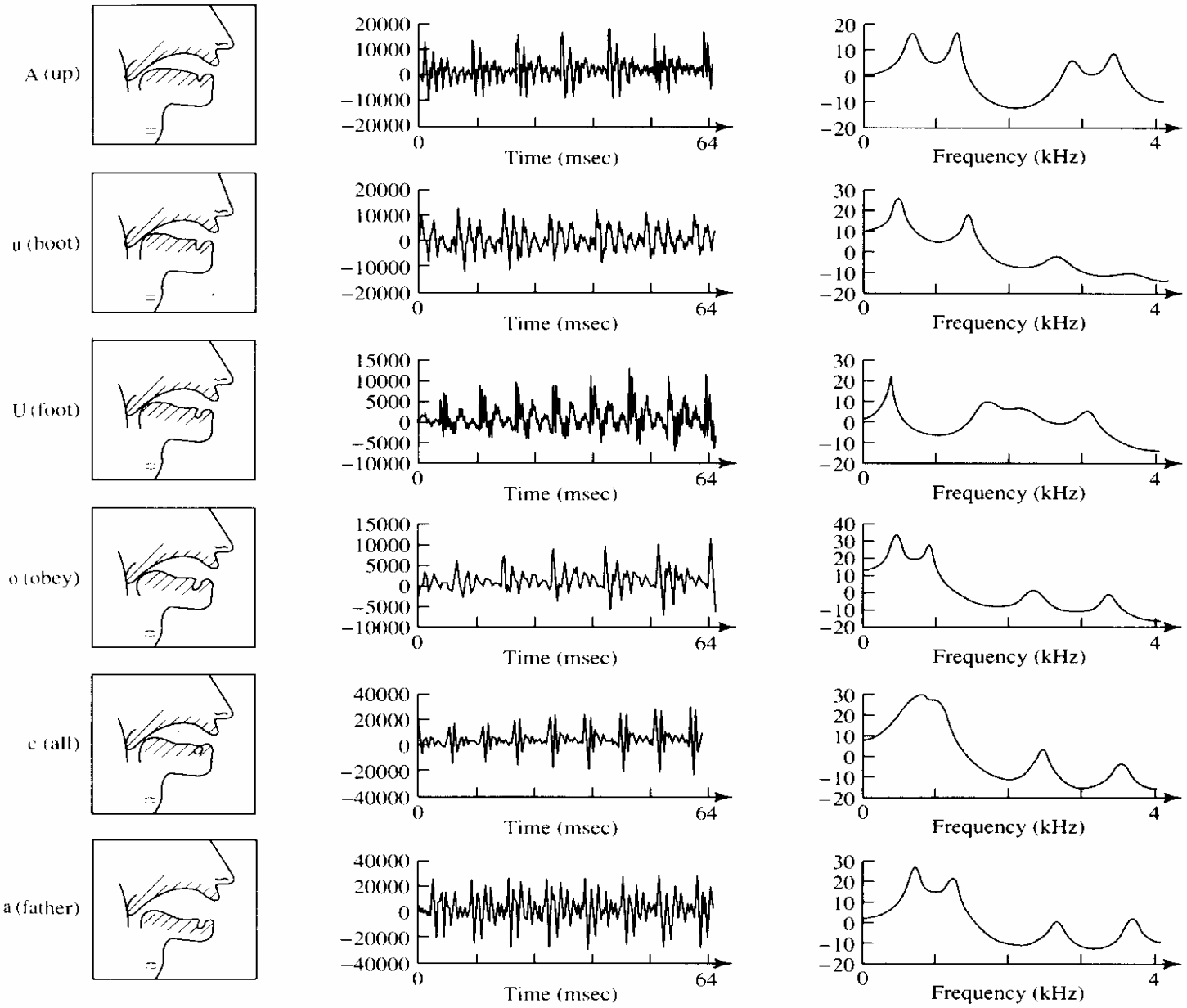
1. Filtering: acoustic filter which modifies the spectral distribution of energy in the glottal sound wave (***formants***)



2. Generation of sounds

A constriction at some point along the vocal tract generates a turbulence exciting a portion of the vocal tract (sound /s/ of six)

The speech signal and its properties



The speech signal and its properties

Types of Excitation

Two elemental excitation:

1. Voiced ..... Vocal cords vibration
2. Unvoiced ... Constriction somewhere along the vocal tract

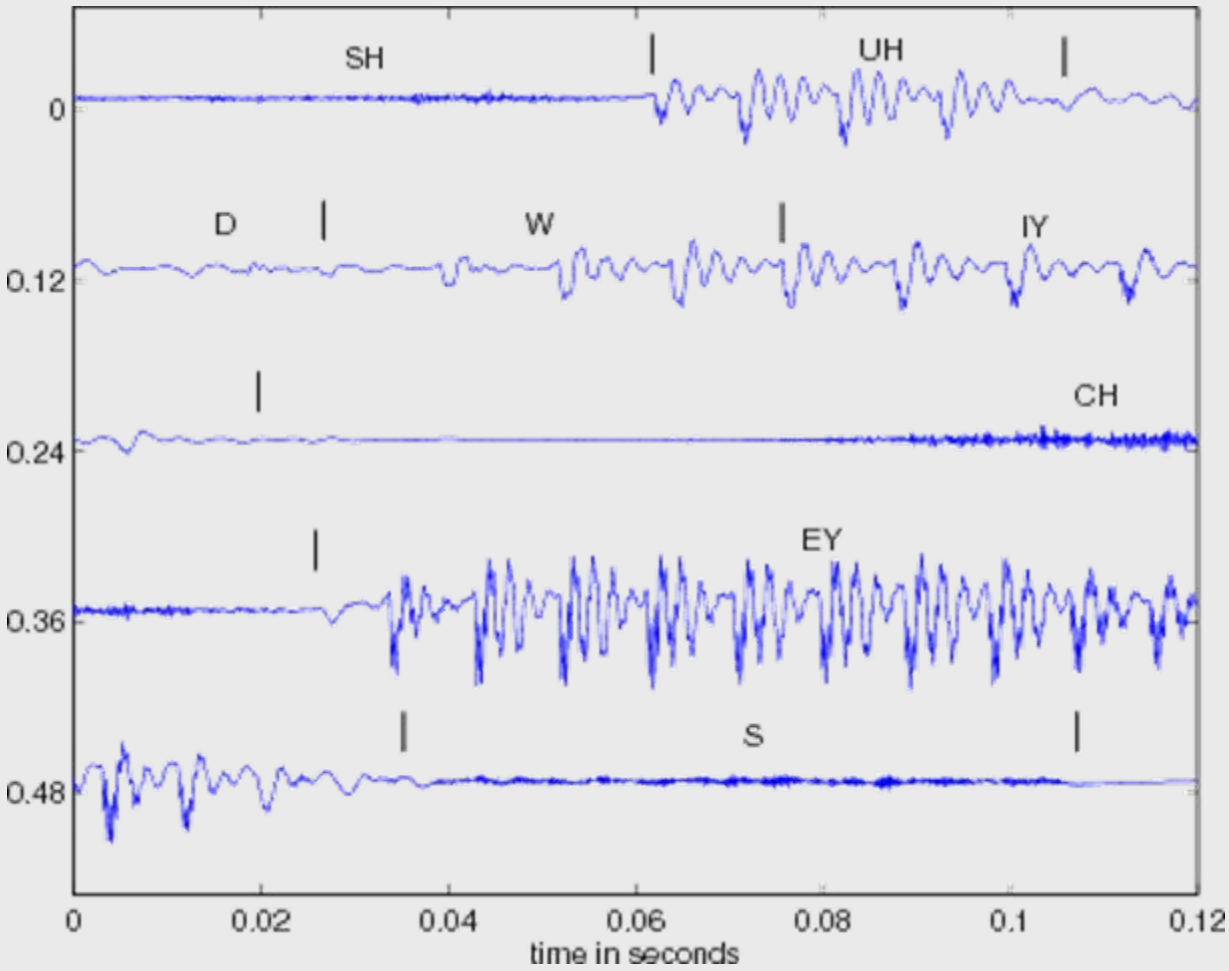
Combinations

3. Mixed ..... Simultaneously voiced and unvoiced

4. Plosive ..... Short region of silent followed by a region of voiced or unvoiced sound
   - /t/ in pat (silence + unvoiced)
   - /b/ in boot (silence + voiced)

5. Whisper ..... Unvoiced excitation generated at the vocal cords

The speech signal and its properties



Should we chase

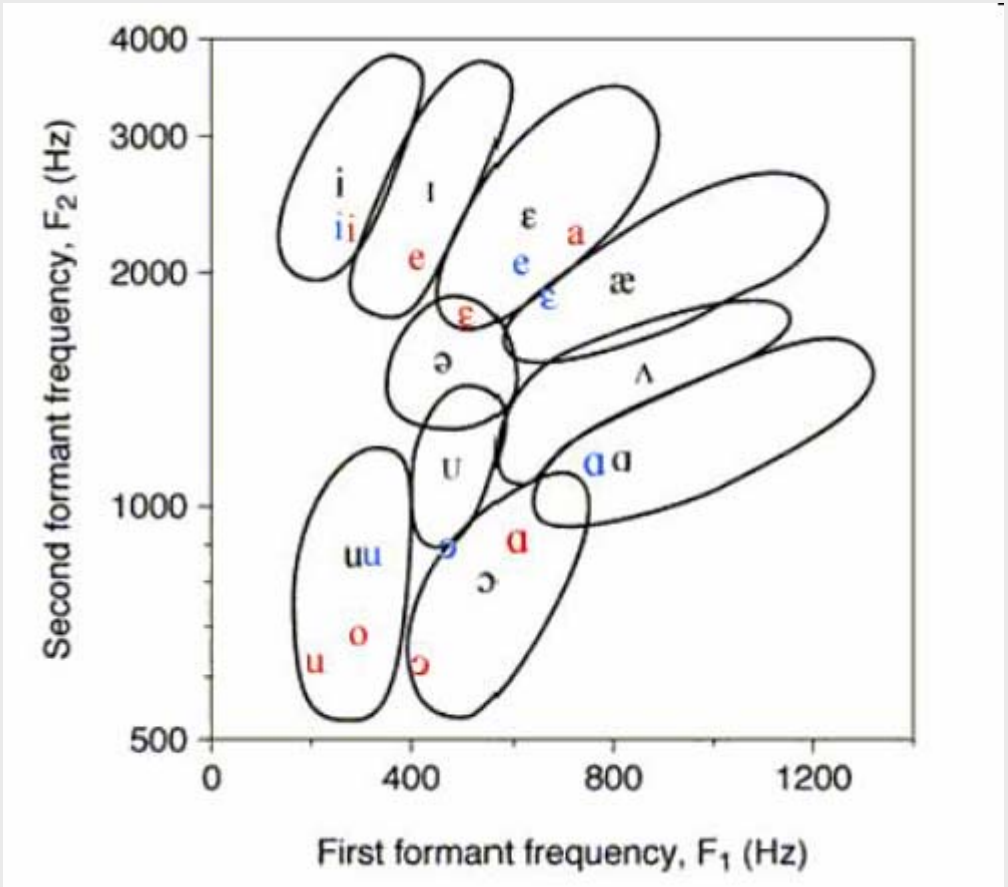
The speech signal and its properties

Speech Main Features

- ✓Pitch (fundamental frecuency)
  From 80 to 400 cicles/sec (Hz)

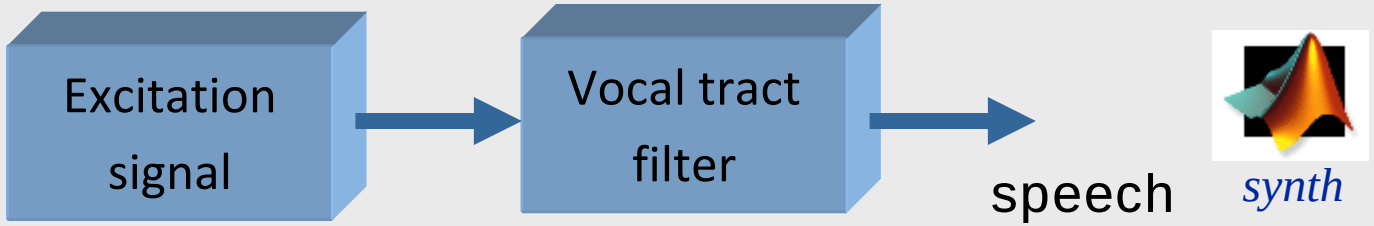
- ✓Formants

	f1	f2
A	700	1150
E	500	1850
I	250	2300
O	400	700
U	300	900



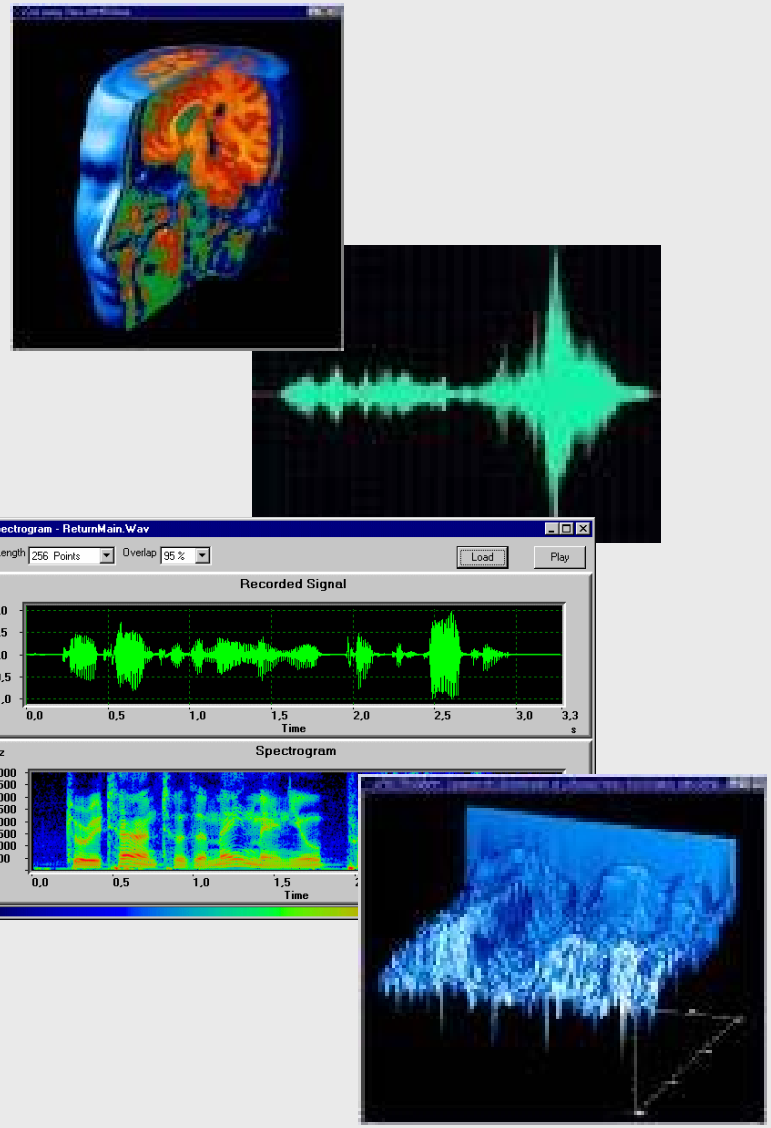
The speech signal and its properties

- Hear the vowels
  http://en.wikipedia.org/wiki/Vowel
- Let's synthesize vowels from scratch



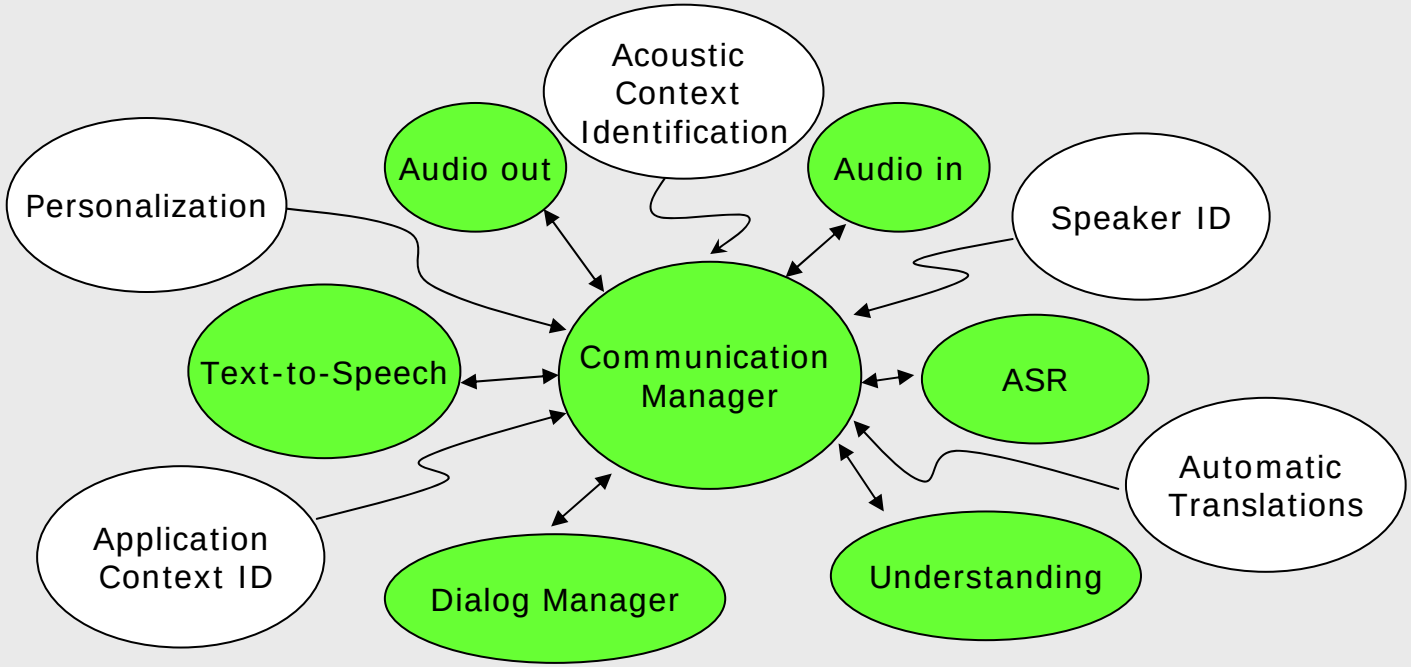
- Let's play with your speech
  download wavesurfer

The Technology



THE TECHNOLOGY

- Spoken Dialog Systems
  - Allow human users to interact with a computer through natural and intelligent conversations as they would do it with human agents.
  - To develop a full system, a wide range of speech and language technologies take part: Automatic Speech Recognition, Speaker Identification, Language recognition, Natural Language Understanding, Spoken Dialog Management, Text-to-Speech conversion.



THE TECHNOLOGY

- Speech Technologies:
  - Speech Enhancement
    - Improve the quality and intelligibility of speech signals distorted by the acoustic environment and transmission channels.
      - Noise, Echo, Reverberation, ...
  - Speech Coding
    - Techniques for compressing the essential information in a speech signal for both, efficient transmission and storage.
  - Speech Synthesis.
    - Process of creating a synthetic replica of a speech signal to transmit a message from a machine to a person.
  - Automatic Speech Recognition.
    - Process of extracting the message information in a speech signal to control the action of a machine by using speech messages.

THE TECHNOLOGY

- Speech Technologies:
  - Speaker Recognition and Identification
    - Process of either identifying or verifying a speaker by his/her voice.
  - Language Identification
    - Process of identifying the language a person is using, given a portion of his/her speech.
  - Automatic Speech Translation.
    - Process of recognizing the speech of a person talking in one language, translating the message content to a second language, and synthesizing an appropriate message in that second language, in order to provide full two-way spoken communication between people who do not speak the same language.

THE TECHNOLOGY

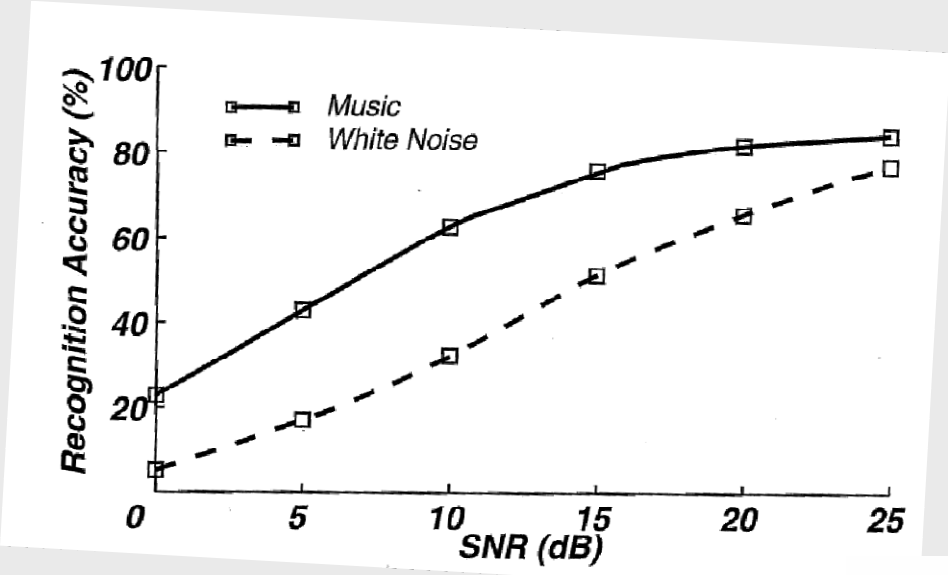
- Natural Language Processing (NLP):
  - Natural Language Understanding
    - Process of extracting the meaning content of a message coming from a human in order to control machines.
  - Spoken Dialog Management:
    - Computer system which must mantain a conversation with humans in order to provide services and perform assigned task in an appropriate way.
    - Is responsible for leading the rest of the modules to collect all the essential information needed to finish successfully the assigned task.
  - Natural Language Generation.
    - Process of constructing a text in a natural way with a predetermined goal.
    - Fundamental stages:
      - Information Selection
      - Information Organization.
      - Natural Language Message Production

THE TECHNOLOGY

- Speech Enhancement:
  - An ASR system rapidly degrades due to acoustic distortion in the input signal
  - Main acoustic degradation:
    - Noise:
      - Access to voice web based application from the car, street, crowded place, industrial plant, etc. can become impossible if acoustic noise is not taken into account
    - Reverberation:
      - Use of distant microphones (hands-free systems) make the performance of the system degrade even in quite environment (like speaking in a bathroom)
    - Acoustic Echo (and electric echo):
      - If microphones and loudspeaker are close together, the signal picked up by system will contain part of the output forcing the ASR to make mistakes
      - The same effect appears in traditional telephone lines due to the limitations of transmitting through a two-wire lind (Hybrid transformer)

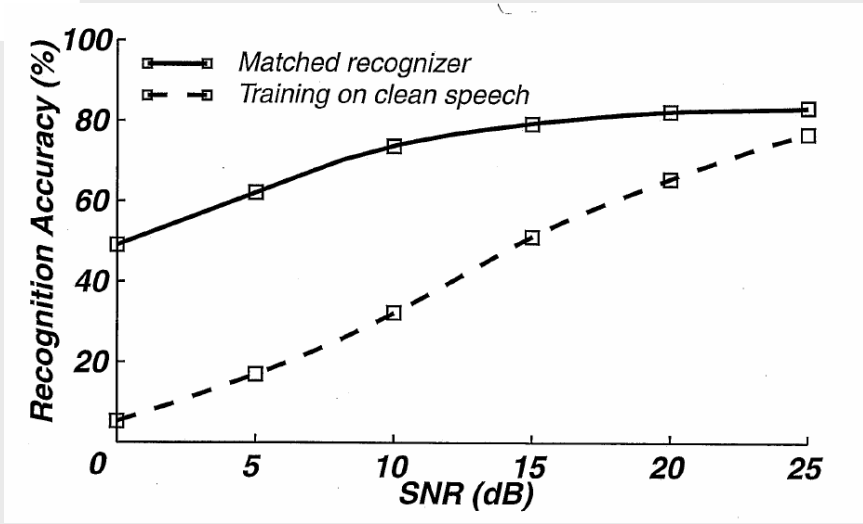
THE TECHNOLOGY

- Ambient Noise Effect



Good ASR systems can perform very well in quiet environment but can become useless when the noise level is high.

Nevertheless, there exists some techniques that allow us to solve this problems, at least partially.



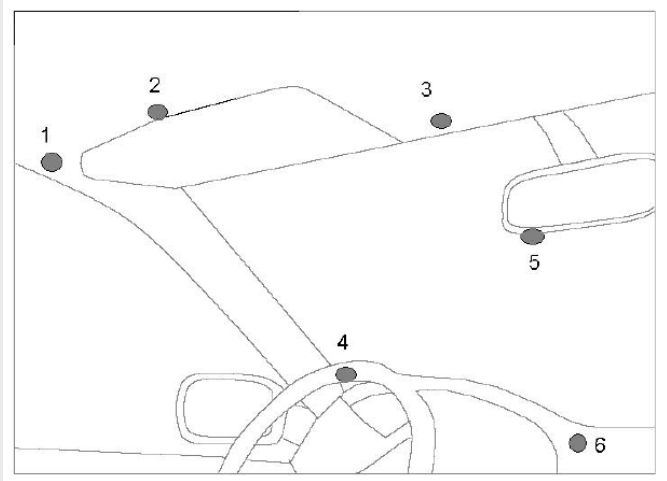
THE TECHNOLOGY

- Reverberation:
  - Use of distant microphones.
  - In-vehicle Speech Recogniton:

When a close-talk microphone is used a state-of-the-art ASR will mistake 9 out of 1000 digits inside a car (considering noise also)

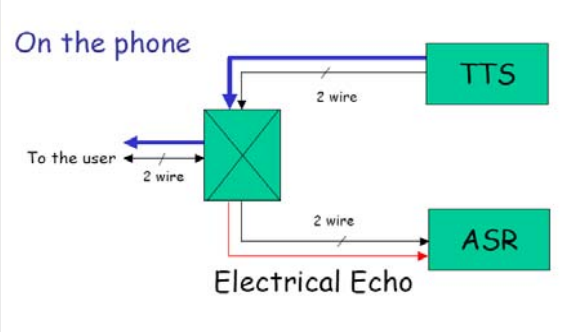
When distant microphones are used (1, 2 or 3 positions) located around 30cm far from the mouth, the error rate increases up to 115 out of 1000.

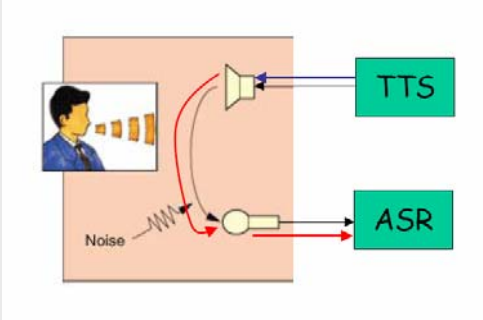
Nevertheless, using appropriate techniques to fight against reverberation and noise the error rate con be reduced up to 1%.



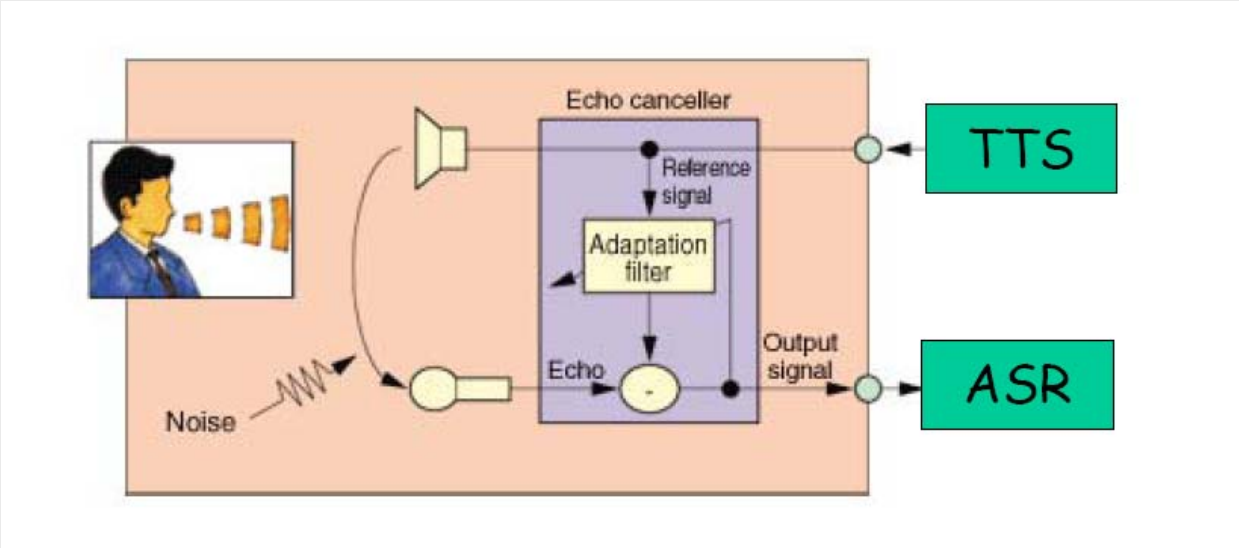
THE TECHNOLOGY

- The Echo:
  - Origin:





  - Solution:

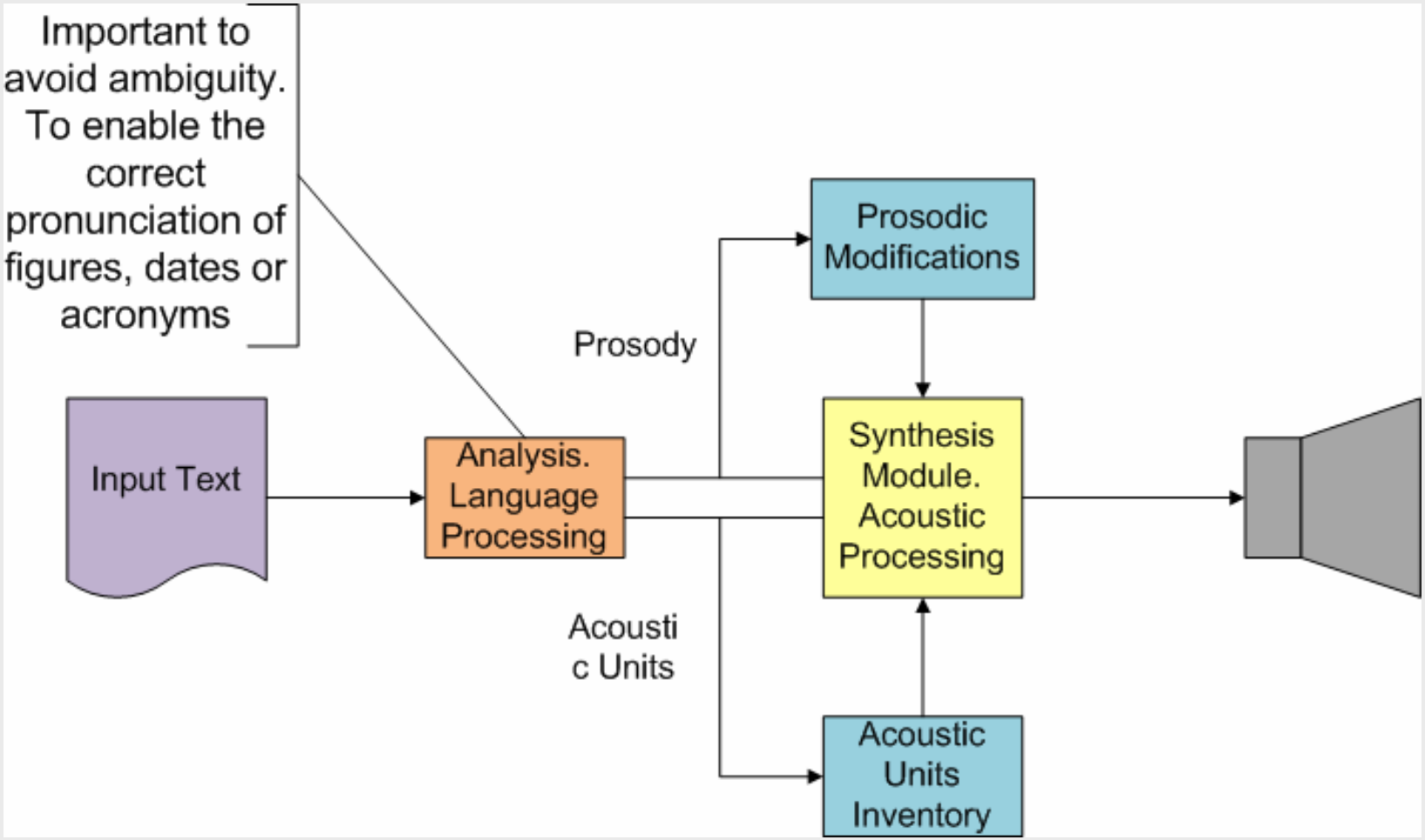


THE TECHNOLOGY

- Text-to-Speech :
  - Speech Synthesis involves the conversion of an input text into speech waveforms.
  - Two basic systems:
    - Voice Response Systems
      - limited vocabulary and syntax
      - pre-recorded units (sentences, words, ...).
    - Text-to-Speech systems (TTS)
      - Unlimited vocabulary and syntax
      - small stored speech units and extensive linguistic processing.

THE TECHNOLOGY

- Typical block diagram of a Text-to-Speech System:

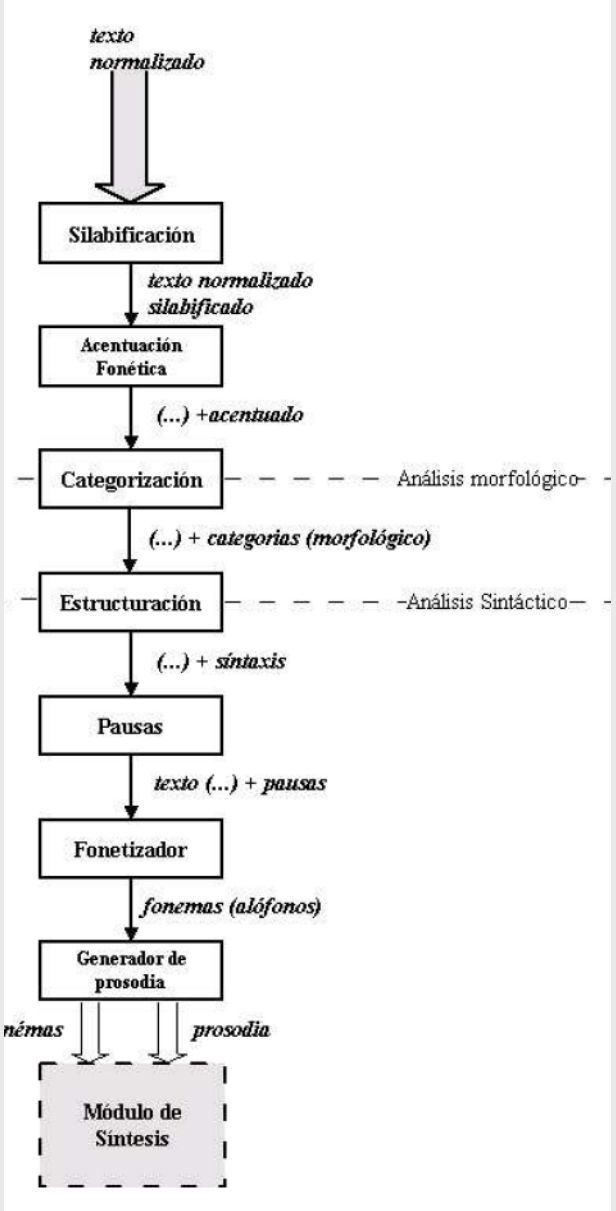


THE TECHNOLOGY

- Linguistic Analysis of the text:
  - The system must know how to pronounce sounds in addition to what sounds it must pronounce.
  - The linguistic analysis module is responsible for deciding which phonemes must be pronounce and which is the correct intonation: Temporal duration, “melody” evolution (pitch), ...
  - It is quite a complex process so it is split into several subtasks.

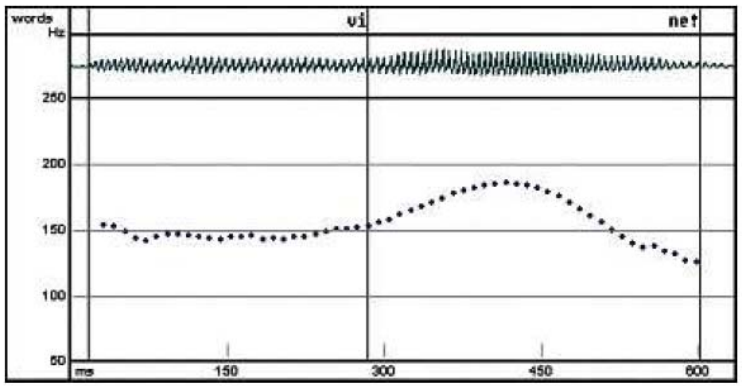
THE TECHNOLOGY

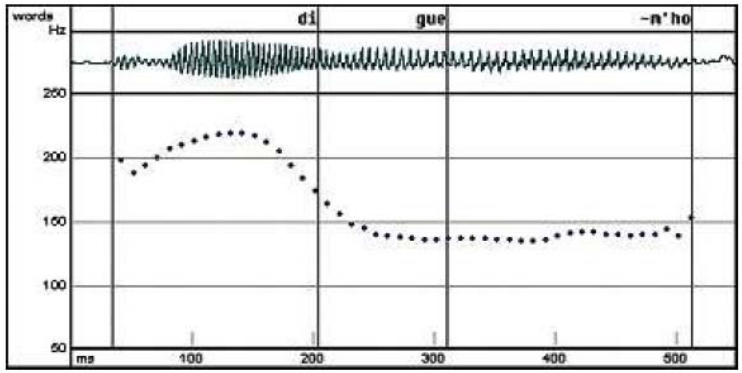
- Linguistic Analysis of the Text:
  - Text Normalization:
    - Split the input text into appropriate work units, sentences.
  - Preprocessing:
    - Ambiguity resolution (acronyms, dates, ...)
  - Syllabifying
  - Phonetic Stress:
    - Important to select and apply the correct prosogy.
  - Categorizer:
    - Assign a tag to every word according to its category (number, name, pause, ...)
  - Structure analyzer:
    - Performs a syntactic analysis of every sentence
  - Pause manager.
  - Grapheme to Phoneme translator:
  - Prosogy Generator



THE TECHNOLOGY

- Prosody Modeling:
  - Key aspect to make the synthetic voice sound natural
    - Rhythm
    - Pauses
    - Intonation
    - Intensity
  - Factors influencing intonation
    - Kind of speech: conversational, read, ...
    - Speaker's attitude..
    - Length of the curve
    - ...





THE TECHNOLOGY

- Some examples:

festvox

AT&T

RealSpeak™

Loquendo
VOCAL TECHNOLOGY AND SERVICES

acapela

Cepstral

Voice Banks

http://www.bbc.co.uk/news/uk-england-manchester-12651740

THE TECHNOLOGY

- Automatic Speech Recognition :
  - Process to convert into text a speech message.
  - Difficulties:
    - Segmentation:
      - There are not clear boundary markers in speech (phoneme/syllable/word/sentence/…)
    - Complexity:
      - 50 phonemes, 5000 sounds, 100000 words.
    - Variability:
      - Anatomy of the vocal tract, speed, loudness, acoustic stress, mood, environment, noise, microphones, dialects, speaking style, context, channel
    - Ambiguity
      - Homophones (two vs. too)
      - Word Boundaries (interface vs. in her face)
      - Semantics (He saw the Grand Canyon flying to N.Y.)
      - Pragmatics (Times flies like an arrow)

THE TECHNOLOGY

- Historic Evolution of ASR systems :
  - 50's .... first attempts
    - Bell Labs, isolated digit recognition, speaker dependent.
    - RCA Labs .... 10 syllable recognition speaker dependent
    - University College in England .... Phonetic recognizer
    - MIT Lincoln Lab .... vowels recognition, speaker independent
  - 60's ... fundamental ideas
    - Dynamic time warping .... Vintsyuk (Soviet Union)
    - CMU ... Continuous Speech Recognition
  - 70's firsts achievements, stochastic approaches
    - LPC, dynamic programming
    - IBM: Large vocabulary project beginings
    - Big budgets in USA: DARPA projects
    - HARPY system (CMU) first successful large vocabulary continuous speech recognition system.

THE TECHNOLOGY

- Historic Evolution of ASR systems :
  - 80's .... Continuous Speech Recognition Expansion
    - Hidden Markov Models: first introduced by IBM, Dragon Systems, popularized by Bell Labs.
    - Introduction of Neural Networks to speech recognition.
  - 90's .... Firsts Commercial Systems
    - Cheap high performance personal computers
    - Dictate systems
    - Integration between speech recognition and natural language processing.
  - 00's .... Systems on the Market, making profits.
    - Phone Integration and Voice Web browsers
    - ASR engines in the operating systems
    - Multimodality, Multilinguality
    - Framework projects EU: Ambient Intelligence

THE TECHNOLOGY

- Forecasts :

Tasks	Machine's error rate today	Human's error rate	Number of years for machines to catch up with humans
Freestyle speech transcription	20 %	4 %	15 years
Connected Digits	0.5 %	0.009 %	30 years
Spelling	5 %	1 %	15 years
Newspaper speech transcription	2 %	0.9 %	5 years

THE TECHNOLOGY

- ASR system categories:
  - Depending on the task or how the user is going to talk to the machine, different ASR strategies must be selected.
  - Depeding on:
    - Task: Isolated commands vs continuous speech, read text speech vs natural speech, …
    - Speaker Attitude: Collaborative, disciplined, familiar with technology
    - Speech Quality: Bandwidth (phone, cellular, Internet, far-field microphone,…), acoustic environment (laboratory conditions, industrial plant, car, street,…), …
    - Interaction: Dialog, one-way comunication, menu browsing, human-human translation,…
    - Speaker dependent vs Speaker Independent: Only one speaker, a reduced group of speakers (profiling), anyone can talk to the system.
    - Vocabulary: Size, similitude among words, Out-of-Vocabulary words (OOV) treatment.
    - Types of tasks:
      - Easy, small devices control (HIFI, oven, ...) .
      - Simple, ticket reservation.
      - Medium, Agenda management.
      - Big, Spoken Document Retrieval.

THE TECHNOLOGY

- Speaker dependent vs. Speaker Independent :
  - Speaker Dependent
    - Trained with only one person speech
    - Low error rate
    - Essential for language or speech pathologies
  - Speaker Independent
    - Trained with huge speech databases recorded with many speakers.
    - Higher error rates.
    - Essential for telephone application
  - Speaker adapted.
    - Initial training with many speakers
    - Retraining or adaptation with only one person's speech.
    - Performance after adaptation is similar to a speaker dependent system

THE TECHNOLOGY

- Sources of Knowledge:
  - **Acoustic**:
    - How sounds are uttered, define the recognition unit (phonemes, words, …)
  - **Lexical**:
    - How words are built from recognition units
  - **Grammatical**:
    - How words are related with each other in a sentence?
    - Speech Recognition Level
  - **Semantic**:
    - What is the meaning of a word?
    - Ambiguity (several meanings for only one word)
    - Essential for a dialog
    - Understanding level
  - **Pragmatic**
    - Relationship among words and their previous uses in the dialog
    - "I like it" ---> It refers to something that appeared previously in the dialog: Ellipsis
    - Dialog level

Human-Computer Interaction

- Voice Input / Voice Output Interfaces:
  - When is Speech considered an appropriate INPUT?
    - When the user is COOPERATIVE
    - Use Speech as INPUT when ...
      - Keyboards or Keypads are not available or they are too small ...
      - Hands-busy situations: Drivers, Industrial Plants Workers,...
      - the user is not a very skilled typist or feels himself uncomfortable using keyboards.
      - the user has some kind of motor disability, specially in his/her hands/arms.
    - DON'T use Speech as INPUT when ...
      - the user must talk to others when performing the task.
      - the task must be performed in a very noisy environment and only distant microphones can be used.
      - as a general rule, when the use of a manual interface is much more easy to use.

Human-Computer Interaction

- Voice Input / Voice Output Interfaces:
  - When is Speech considered an appropriate OUTPUT?
    - When the user is COOPERATIVE
    - Use Speech as OUTPUT when ...
      - Eyes-busy situations: Drivers, Industrial Plants Workers,...
      - the user has some kind of perceptual disability or visual limitation
      - the interface is emulating someone's personality.
      - the situation requires the users full attention.
    - DON'T use Speech as OUTPUT when ...
      - the amount of information to present is high.
      - the user must compare different items.
      - the information to be presented is confidential.

Spoken Dialogue Systems

Speech Technology for e-Inclusion and therapy support

Speech Technologies Applications

- ST can be used for
  - Improve accessibility
  - Control
  - Communication
  - Assessment
  - Treatment
- Most applications focus on
  - Physical disability
  - Speech disorders (dysarthria)

Speech Disorders

- Stuttering:
  - involuntary repetitions and prolongations of sounds
- Speech sound disorders
  - involve difficulty in producing specific speech sounds
    - articulation disorders
      - difficulty learning to physically produce sounds
    - phonemic disorders.
      - difficulty in learning the sound distinctions of a language, so that one sound may be used in place of many.
- Voice disorders
  - impairments, often physical, that involve the function of the larynx or vocal resonance.
- Dysarthria
  - weakness or paralysis of speech muscles caused by damage to the nerves and/or brain. Dysarthria is often caused by strokes, parkinsons disease, head or neck injuries, surgical accident, or cerebral palsy.
- Apraxia
  - involves inconsistent production of speech sounds and rearranging of sounds in a word ("potato" may become "topato" and next "totapo").

Speech disorders

Stuttering

http://www.youtube.com/watch?v=Lj2IsxxCSS8

Some examples of dysarthria

http://www.youtube.com/watch?v=EHNSBo3SsmY

Dysarthria and subtitles

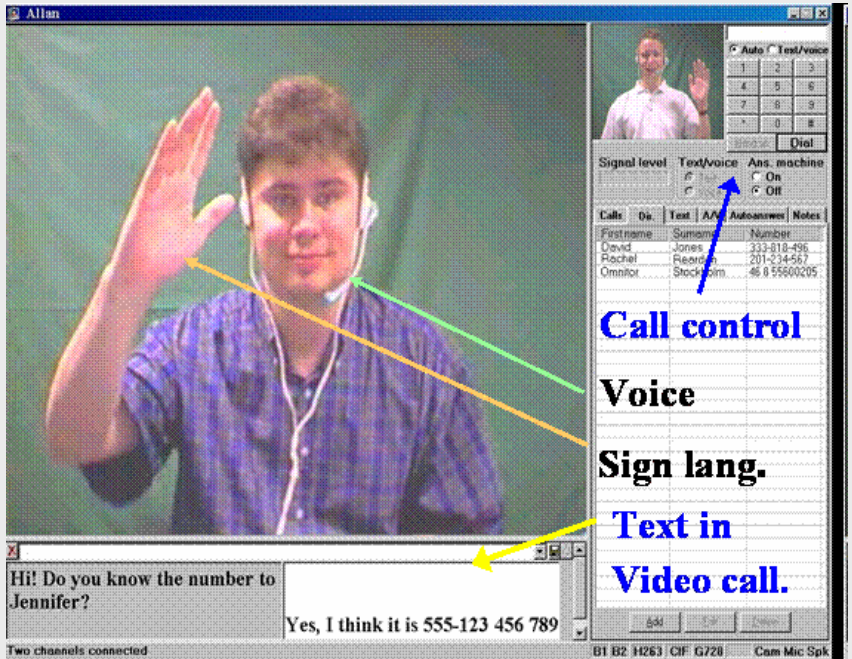
http://www.youtube.com/watch?v=bY95QfUdDSo

Some examples of apraxia

http://www.youtube.com/watch?v=XNB0ihI2srQ

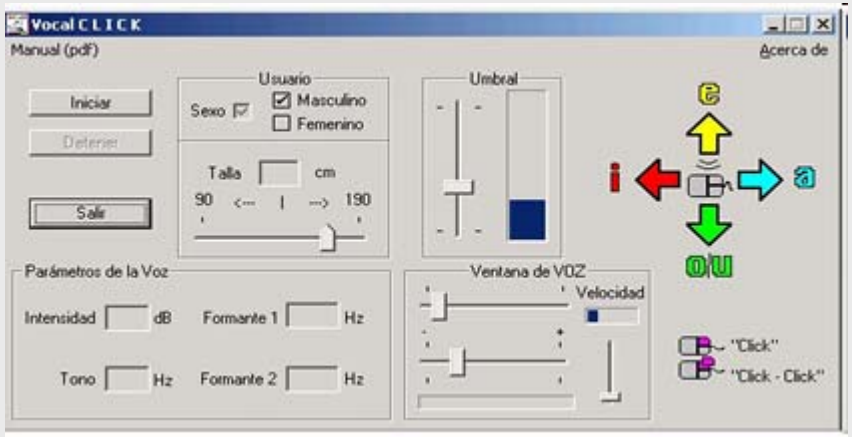
Applications

- Access
  - Speech recognition provides a means of access for some people with physical disability and “normal” speech.
  - Recognition accuracy correlates with intelligibility
    - Works for “normal” speech, mild and moderate dysarthria
    - Does not work for severe dysarthria
  - Personalization of word recognizers for severe dysarthria



Applications

- Control
  - Control of the home environment an essential aspect of independence
  - Home control systems based on personalized speech technologies
  - An example of mouse control
    - VozClick
    - VocalClick http://www.vocaliza.es

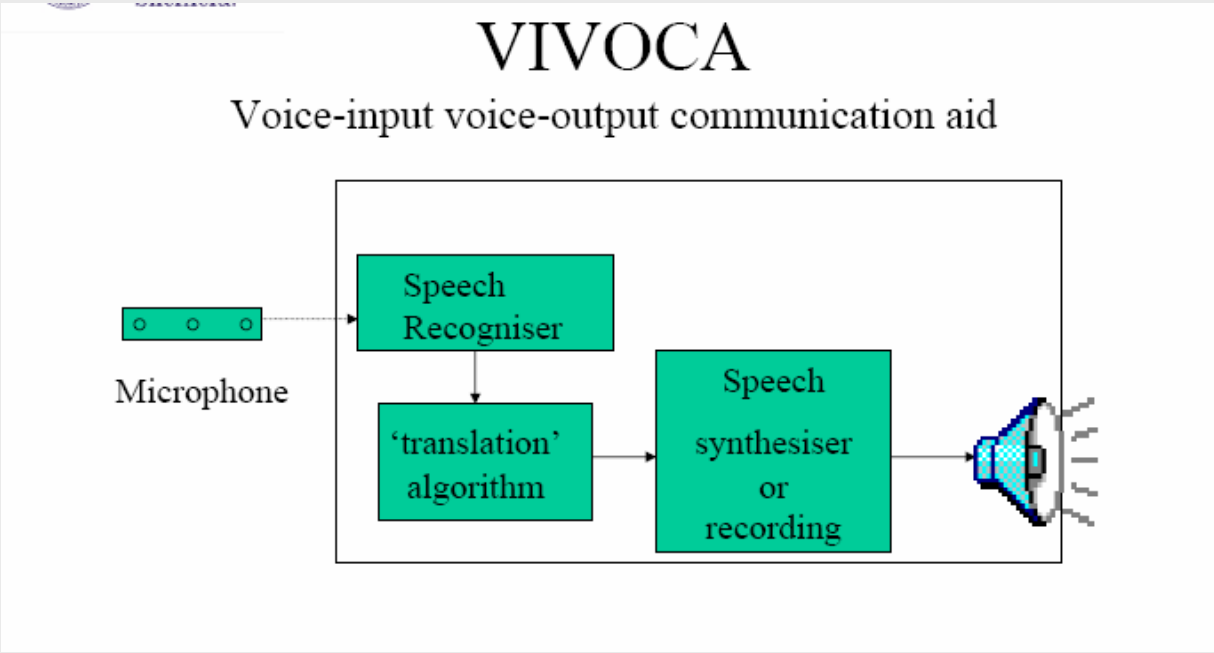


Applications

- Communication
  - Speech synthesis used extensively in assisted communication
  - Personalization of the speech synthesis

Applications

- Communication
  - Voice-Input Voice-Output Communication Aid → VIVOCA
  - Personalization of the speech recognition and synthesis systems



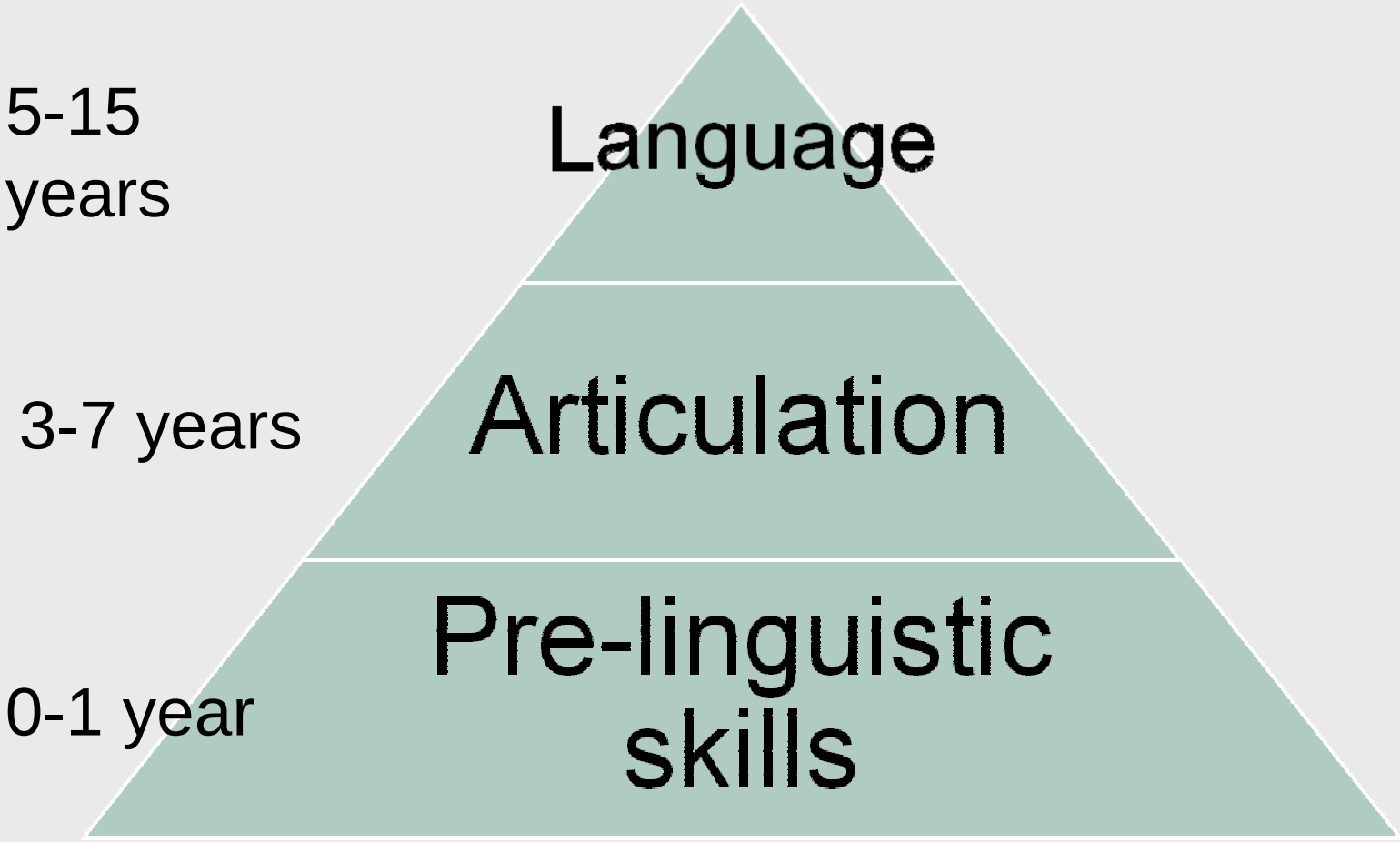
COMPUTER-AIDED LANGUAGE LEARNING AND REHABILITATION: PRELINGUISTIC SKILLS

IP Valencia, 17/03/2010

CALL Systems

- Language Learning Process
- Why?
- Basis
- Examples
  - Pre-linguistic skills
  - Articulation
  - Language

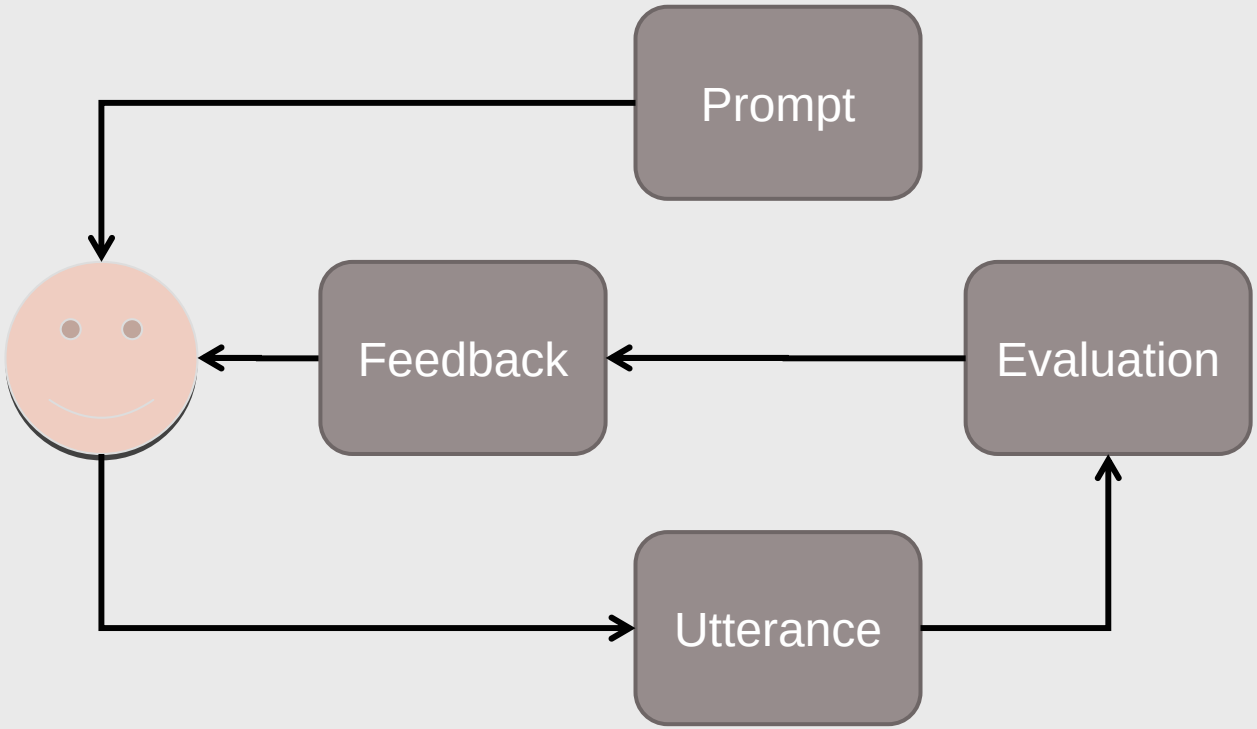
Language Learning Process



Why?

- Emphasis on educational tools based on speech technologies
- Possible users:
  - Impaired users with disordered speech
  - Learners of a new language
- Objective
  - Better communication capabilities

Basis



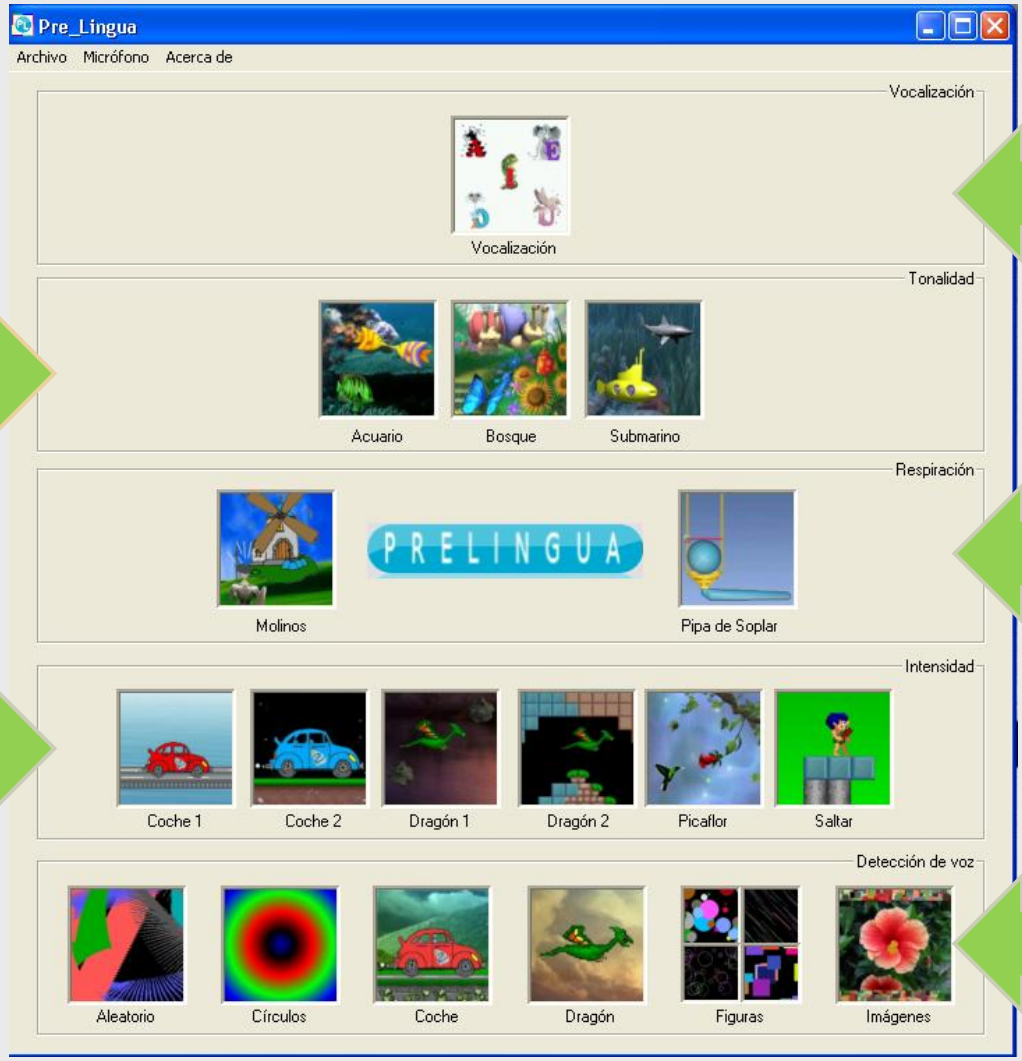
Pre-linguistic skills

- For very small children or with severe disorders
- Graphical feedback!!!
- Control of very basic features
  - Intensity
  - Tone
  - Breathing
  - ...

Voice painter

http://www.youtube.com/watch?v=iP8BvawX8cU

Pre-linguistic skills



Vocalization

Tone

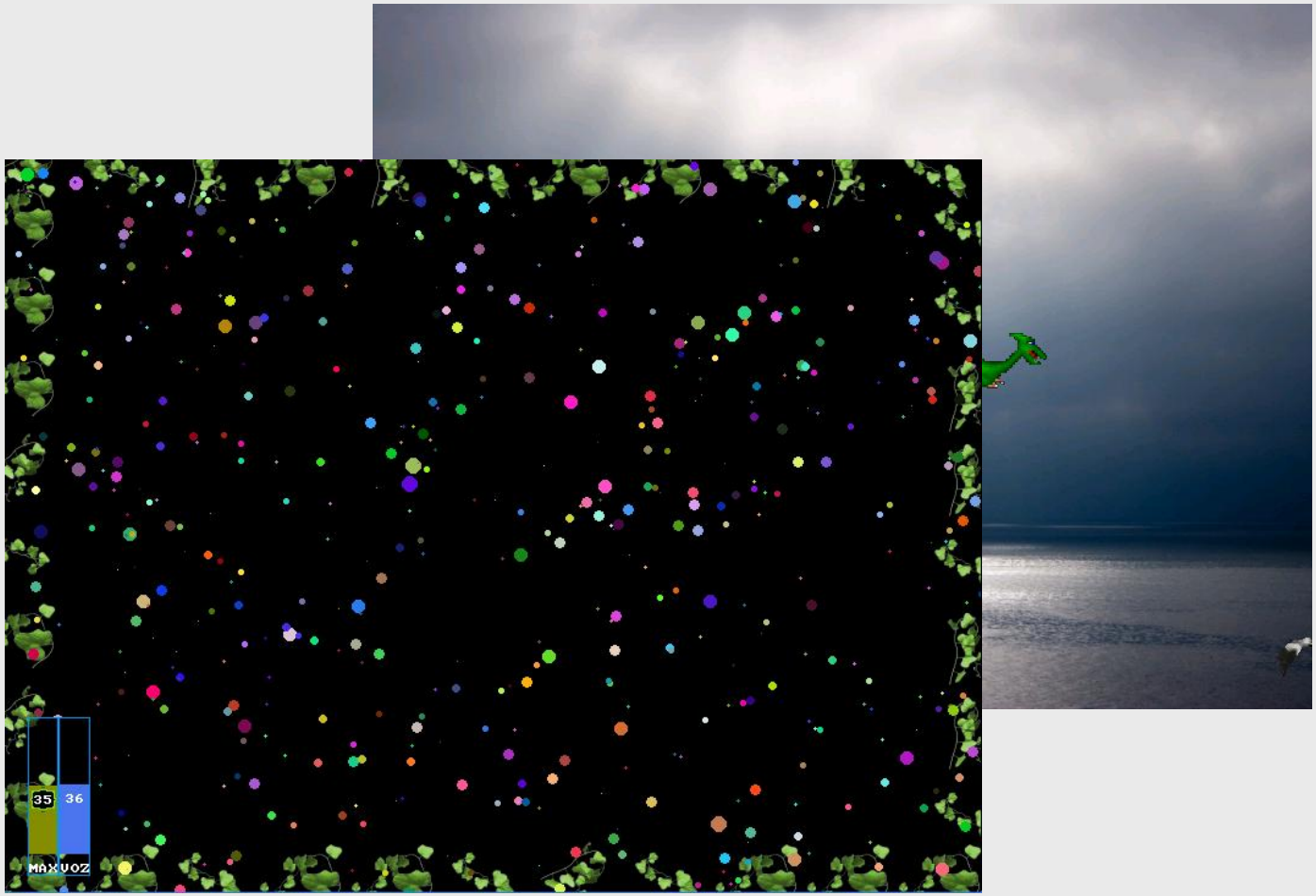
Breathing

Intensity

Voicing

Pre-linguistic skills

- Voicing

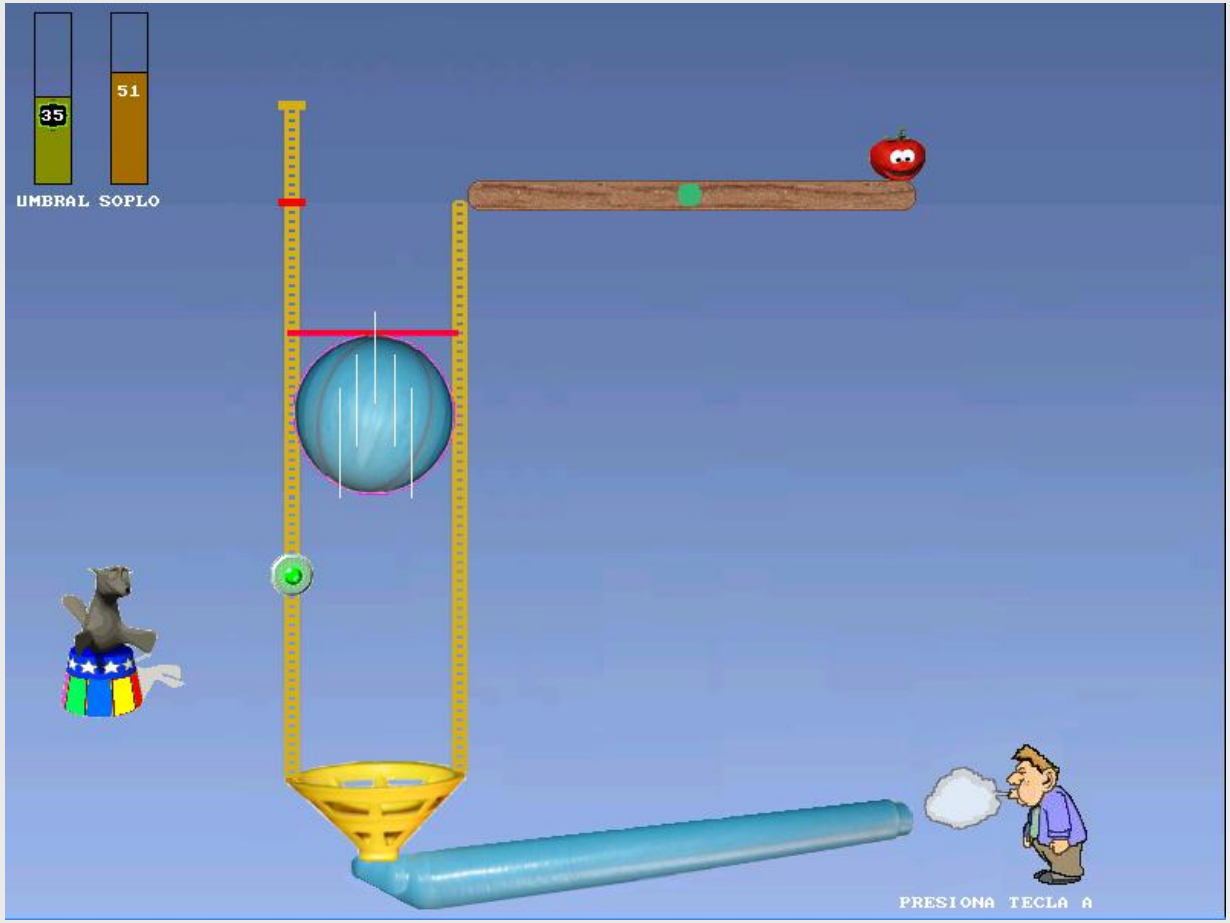


Pre-linguistic skills

- Intensity

Pre-linguistic skills

- Breathe

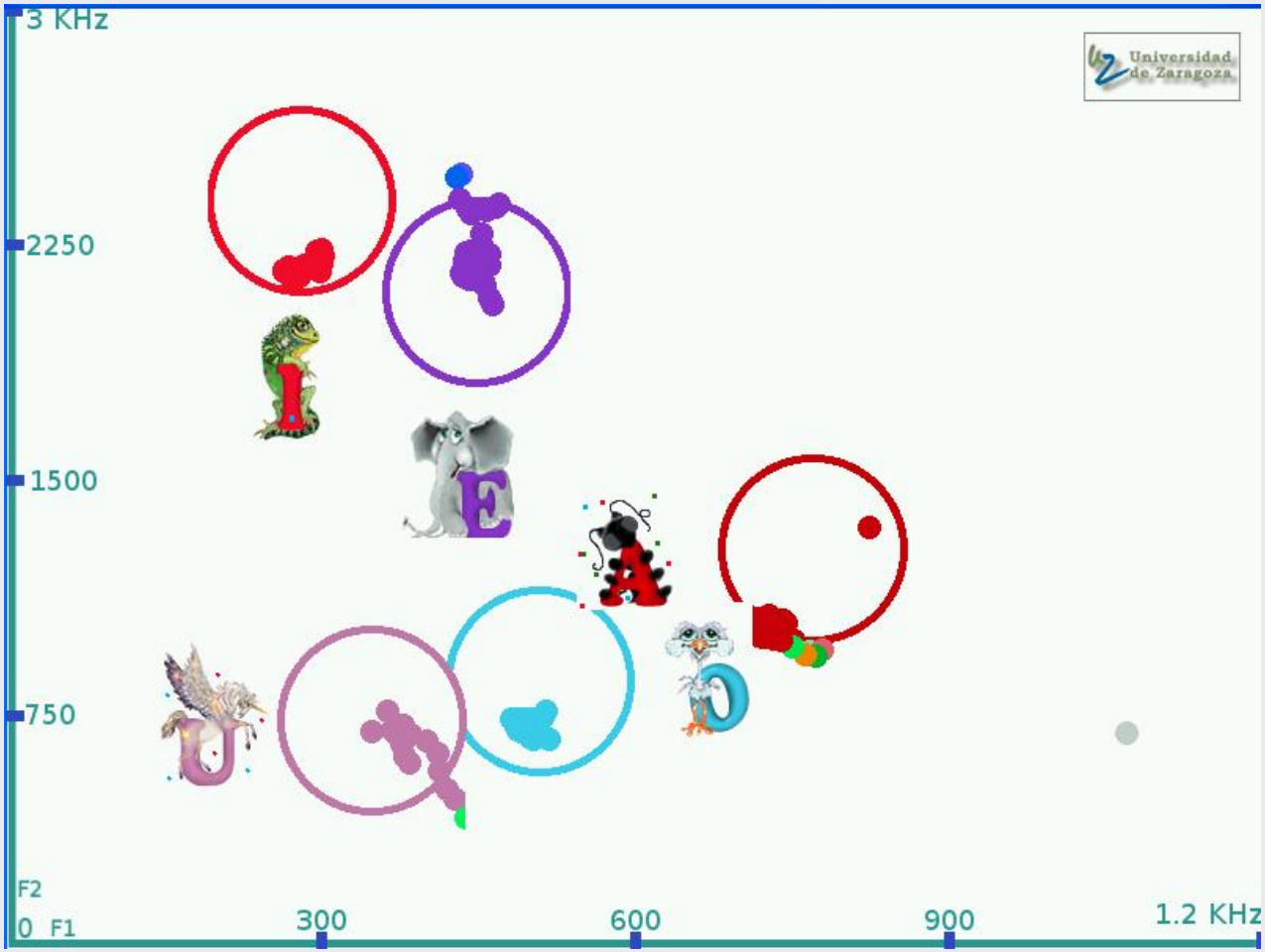


Pre-linguistic skills

- Tone

Pre-linguistic skills

- Vocalization



Examples

- Now, practice

COMPUTER-AIDED LANGUAGE LEARNING AND REHABILITATION: ARTICULATORY AND LANGUAGE SKILLS

IP Valencia, 17/03/2010

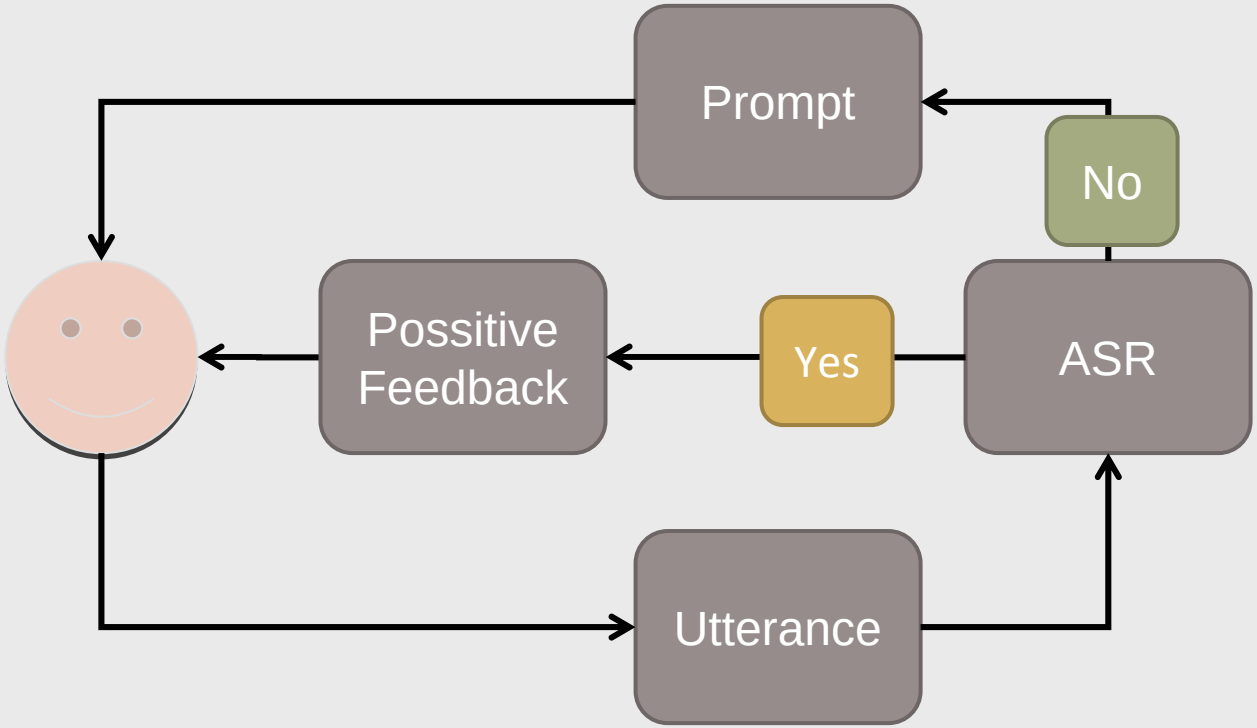
Articulatory skills

- For children-young adults with disorders or
- Learners of a second language

- Word or phoneme based feedback

Evaluation - Alternatives

- Whole word evaluation - ASR

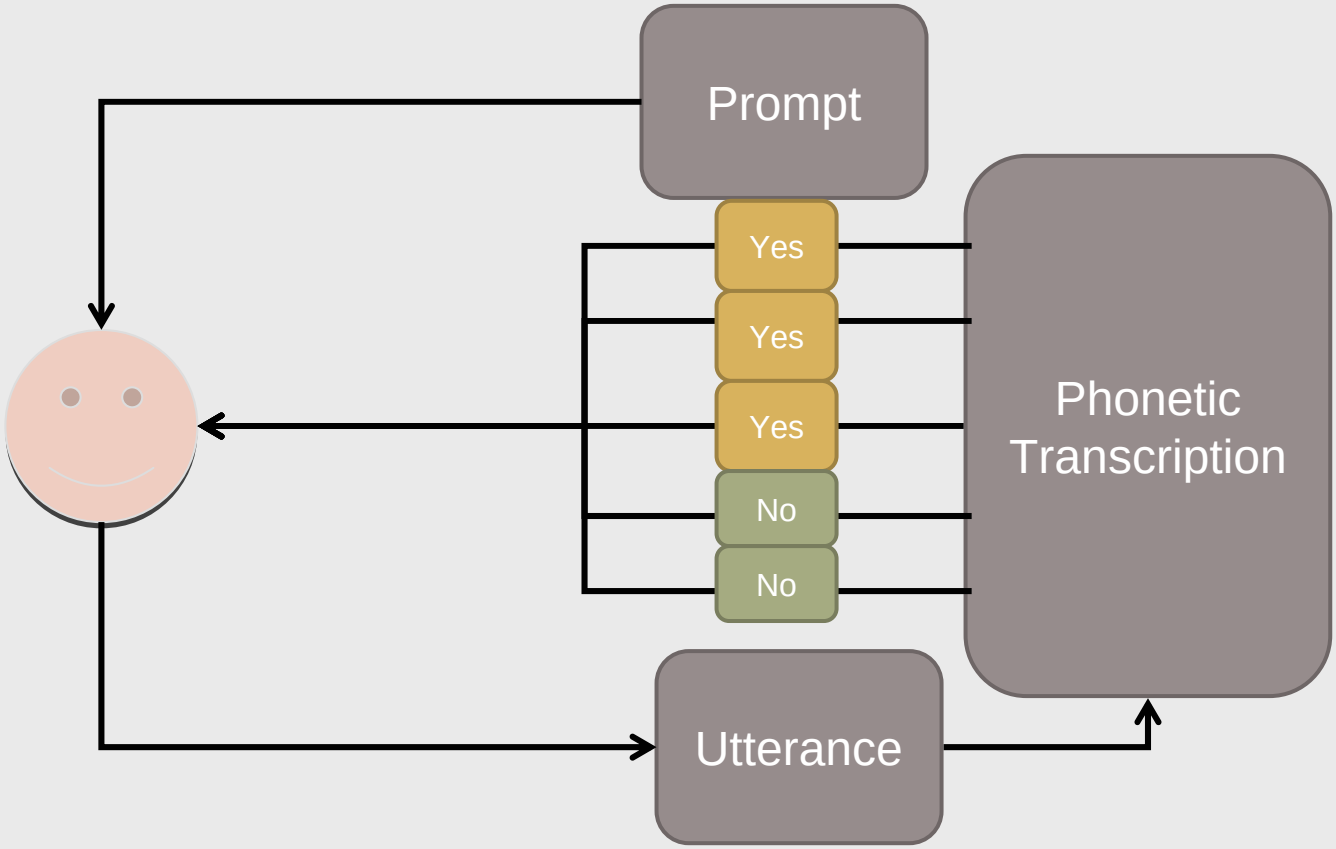


Evaluation - Alternatives

- Whole word evaluation – ASR
- Advantages:
  - Simple: No need to build new blocks
  - Fairly accurate
- Disadvantages:
  - Low correction power when failing

Evaluation - Alternatives

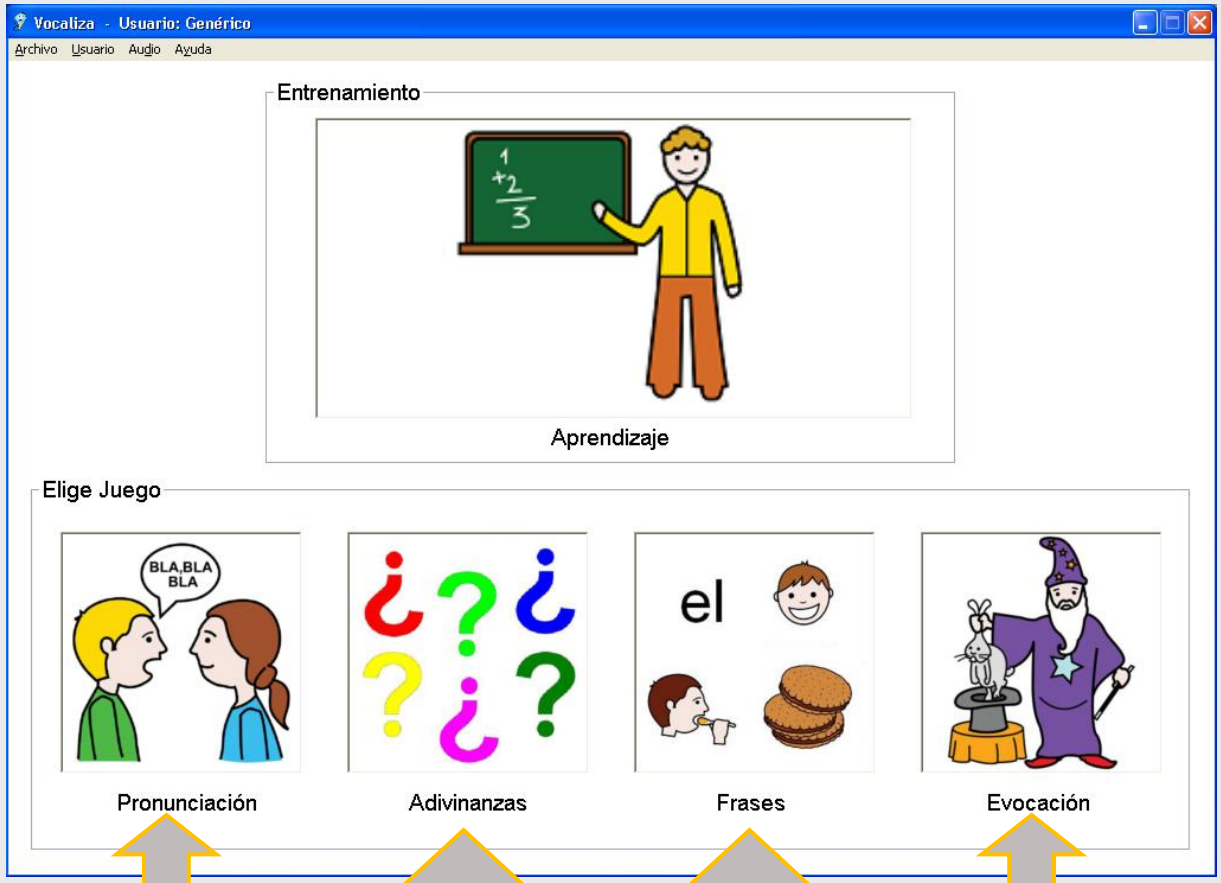
- Phoneme evaluation



Evaluation - Alternatives

- Phoneme evaluation
- Advantages:
  - Great correction power
- Disadvantages:
  - Complex
  - It may lead to different solutions

Articulatory skills



Pronunciation

Riddles

Sentences

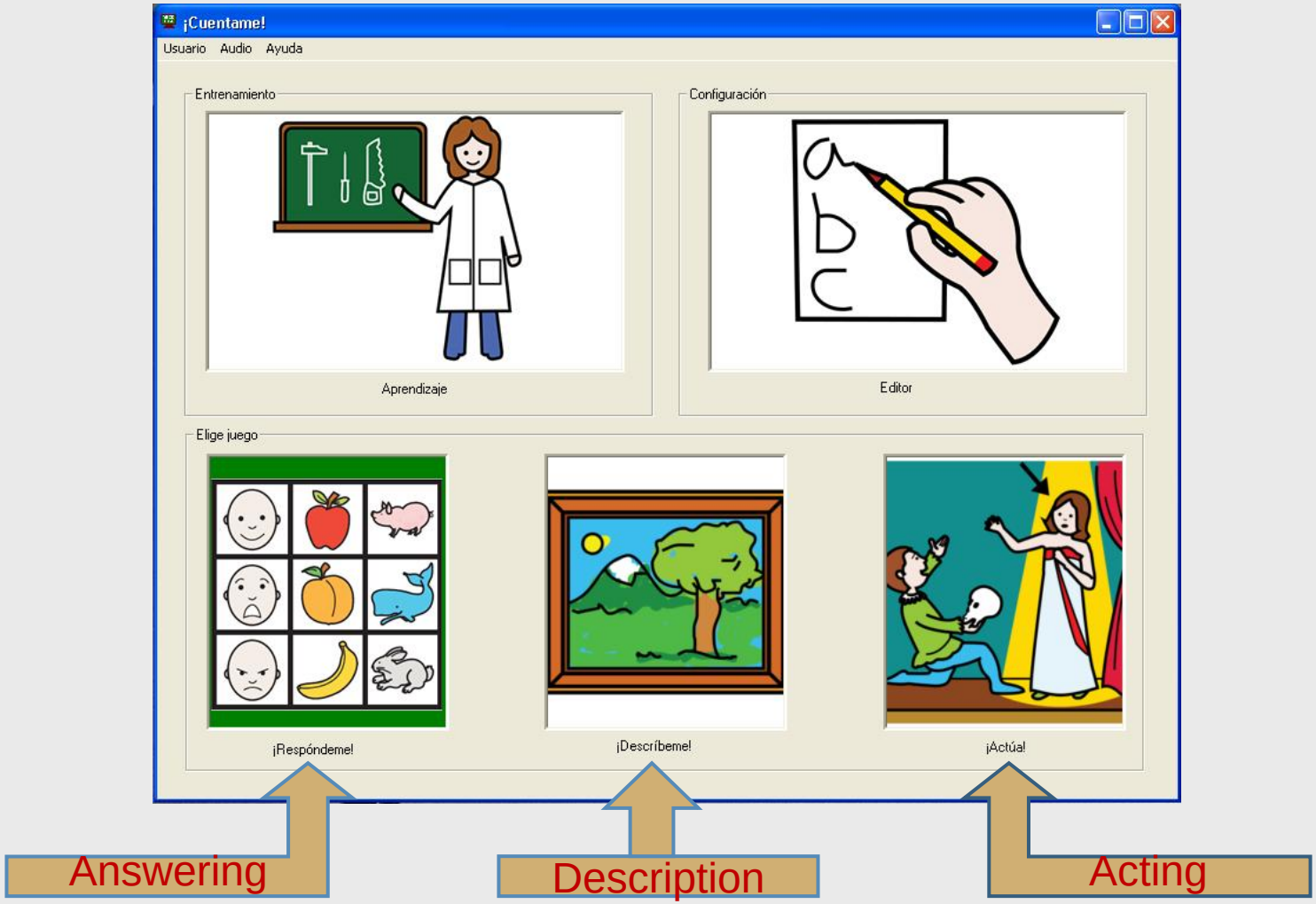
Evocation

Language

- For young adults with disorders or
- Advanced learners of a second language

- Creation of sceneries to be solved by speech

Language


¡Cuentame!
Usuario Audio Ayuda
Entrenamiento
Aprendizaje
Configuración
Editor
Elige juego
¡Respóndeme!
¡Descríbeme!
¡Actúa!
Answering
Description
Acting

Examples

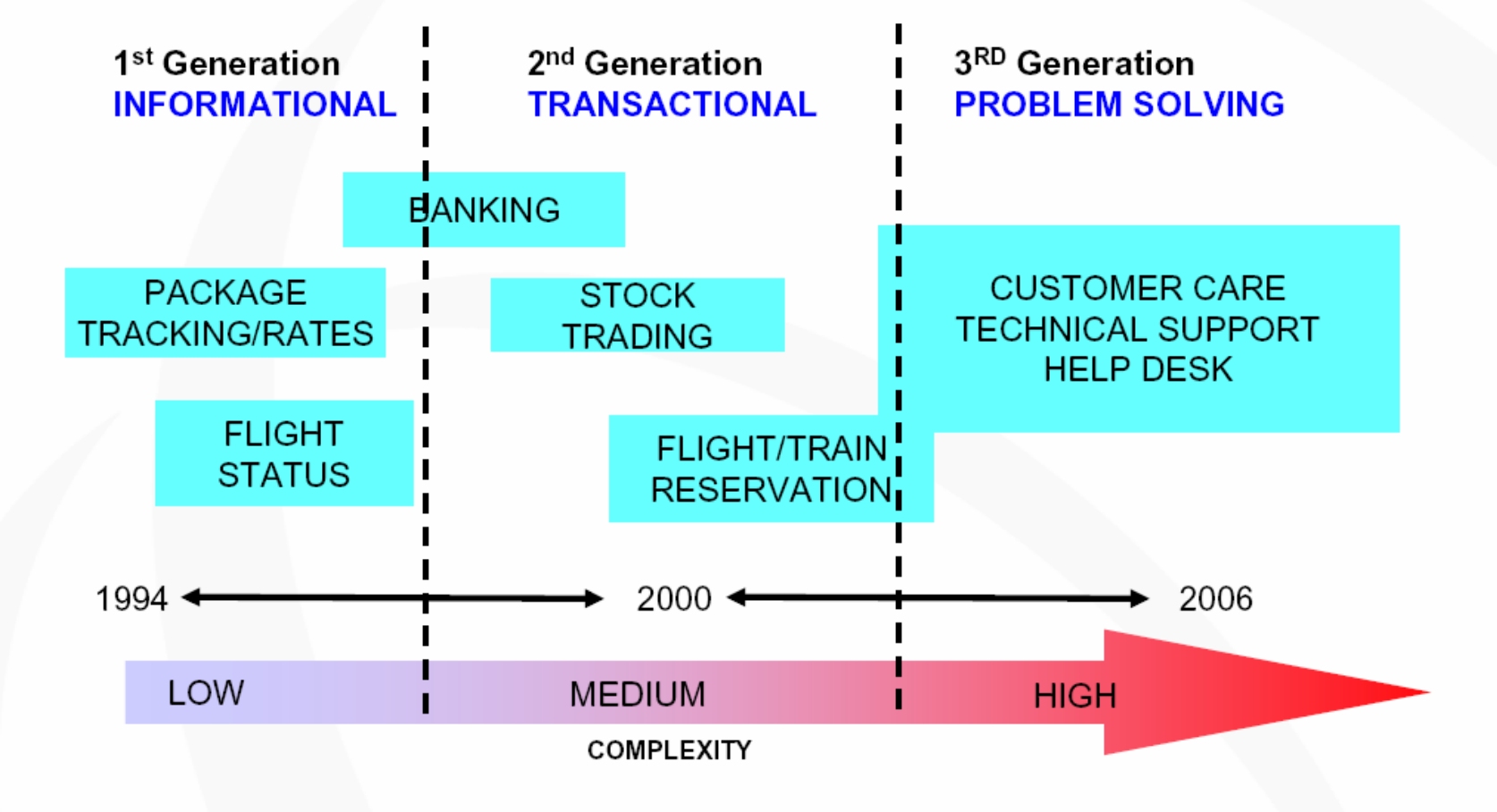
- Now, practice

Business over Voice Portals

- Speech Interfaces: A Case of Success.
  - Business over Voice Portals:
    - Information Access:
      - News, Sports, Traffic, ...
    - Electronic Transactions (e-Trans)
      - Parcel/Shipments Tracking, Electronic Bank, ...
    - Phone Companies customer services:
      - Regiter / Unregister, plans & rates, ...
    - Corporative Portals / Intranets
      - Agenda management, automatic PBX
    - Telematics: Information access and services from the car.
    - Technical Support
      - Cable TV, Internet Services Provider,...
    - Ticket reservation services

Business over Voice Portals

- Business over Voice Portals Evolution



Business over Voice Portals

- What do we mean by Voice Portal?
  - Speech-Enabled Access to Web-based Information and services.
- What can be provided by a Voice Portal?
  - Phone users can have access to a great amount of services available on the Internet
    - 80% - 100% of cellular phone penetration in advanced markets.
    - More than 2 Million cellular users around the World with a clear increasing trend.
    - Only around 50% of homes in Western Europe have an appropriate Internet Access.
  - Access to all the Internet services through the most natural way of communication for the human being: Speech.
    - Fight Against “Digital Divide“
      - Gap between people with effective access to digital and information technology and those with very limited or no access at all.
    - Access to IT technologies for non-skilled people (elder, handicapped, ...)
    - Use of “Natural Spoken Language”

Video

Business over Voice Portals

- What are the advantages of Voice Portals?
  - Global presence.
    - Traditional web-browsers allow the user to access to the Internet only if a computer and a network connetion is available (home, office, ...)
    - Voice Portals allow business to be available 24/7 en from everywhere (vehicle, street, station, airport,...)
  - They do not provide only information but also services:
    - Virtual Personal Agents: Bank, Technical Help Desk, ...
  - Significant Cost Reduction
    - Allow company human resources to provide higher value services
    - They can complement human operators increasing productivity:
      - Human Operator Cost: 3$ per call
      - Automatic System Cost: 20 cents per call. (Goldman Sachs)
  - They open business opportunities thanks to new services with incomes based on advertising of subscription

Voice Web Applications

- What is a Voice Web Application?
  - Web application that provides a speech interface to data available behind the Web Infrastructure.
- Visual Web Applications ..... HTML
  - Graphics, image maps, ...
- Voice Web Applications ..... VoiceXML
  - Speech
- Multimodal Web Application ..... SALT
    - XHTML+VoiceXML (X+V)
  - Speech, Graphics, image maps, ...

Voice Web Applications

- Architecture:

Web browser

HTML

SERVIDOR WEB

Voice XML

Bases de Datos

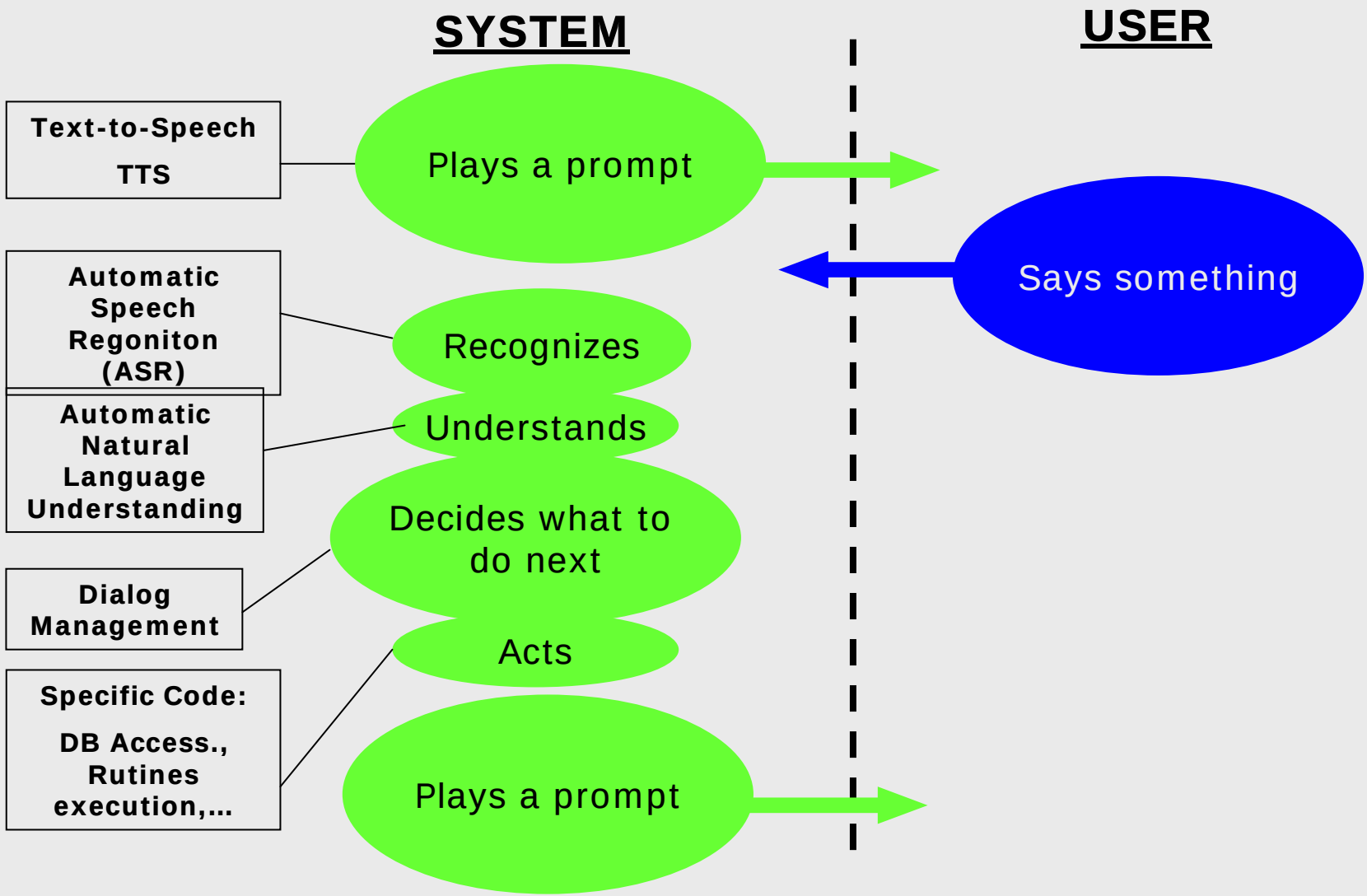
"Voice browser"

Video

Voice Web Applications

- Voice Interface Interaction Model:
  - Basic interactive voice response (IVR)
    - Computer: “For stock quotes, press 1. For trading, press 2, ...”
    - Human: (presses DTMF “1”)
  - Basic speech IVR
    - C: “Say the stock name for a price quote”
    - H: “Nokia”
  - Advanced Speech IVR
    - C: “Stock Services, how may I help you?”
    - H: “Uh, what’s Nokia trading at?”
  - Near-natural language IVR
    - C:”how may I help you?”
    - H: “Uhm, yeah, I’d like to get the current price of Nokia”
    - C: “Nokia is up two at sixty eigth and a half”
    - H: “OK. I want to buy one hundred shares at maker price”.

Behind a Voice Web Application



DEVELOPING VOICE WEB APPLICATIONS

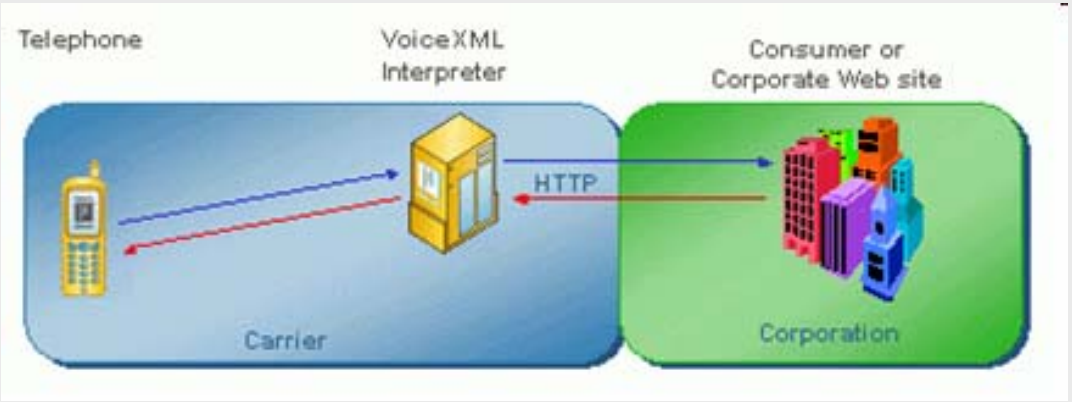
- Some technical aspects:
  - Use of Standards:
    - Developers don't need to be experts on speech technologies.
    - Just think about input/output
    - Two main standards:
      - VoiceXML, XHTML+VoiceXML
        - from w3C
      - SALT (Speech Application Language Tags)
        - Microsoft .NET
    - Challenges: Conversations are different than web-pages
      - Speaking is slower than reading
      - People quickly forget what they have just heard
      - It is NOT very clear what you can not say or what the system is expecting you to say.

VOICE XML

- What is VoiceXML?
  - Markup language for specifying spoken dialogs
    - Voice dialogs use audio prompts and text-to-speech as output.
    - Touch-tone keys (DTMF) and automatic speech recognition as input.
  - Who is developing VoiceXML?
    - W3C Voice Browser working group.
    - VoiceXML Forum (http://www.voicexml.org/)
      - Motorola
      - Lucent
      - IBM
      - VoiceGenie
      - Tellme
      - SpeechWorks
      - PipeBeach
      - BeVocal, Voxeo, Cisco, Nortel, and 10 – 20 more ...

VOICE XML

- Why is VoiceXML important?
  - Standard language enables portability.
  - High-level domain-specific languages simplifies application development.
  - Accessing to the existing web infrastructure by using the telephone.
  - Allows a clean separation between service logic and user interaction.

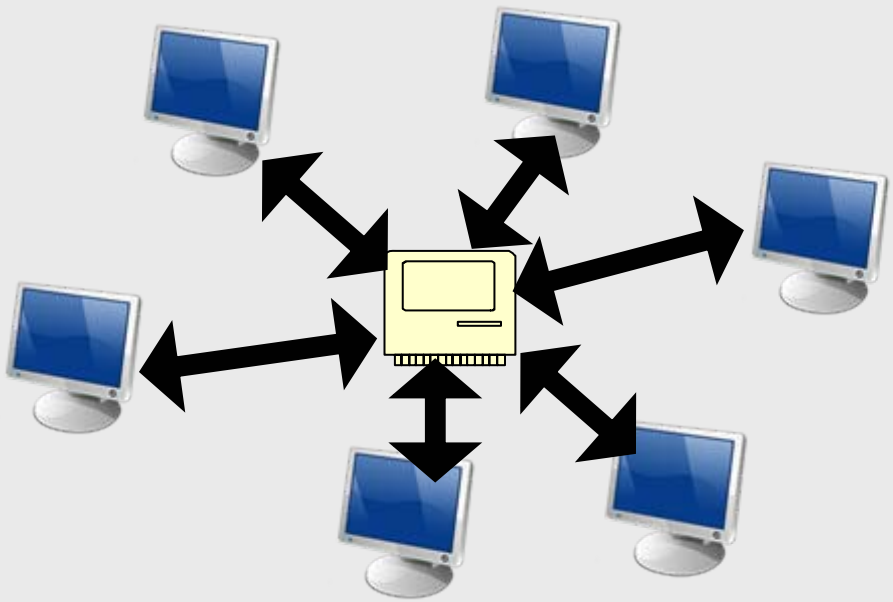


VOICE XML

- How is a Voice XML document?

```
<?xml version="1.0"?>
<vxml version="1.0">
  <form id="hello">
    <block>
          <audio src="http://physionet.cps.unizar.es/~eduardo/wav/alerta.wav"/>
          Hello World!
    </block>
  </form>
</vxml>
```

- More about XML this afternoon with Oscar as guest star.

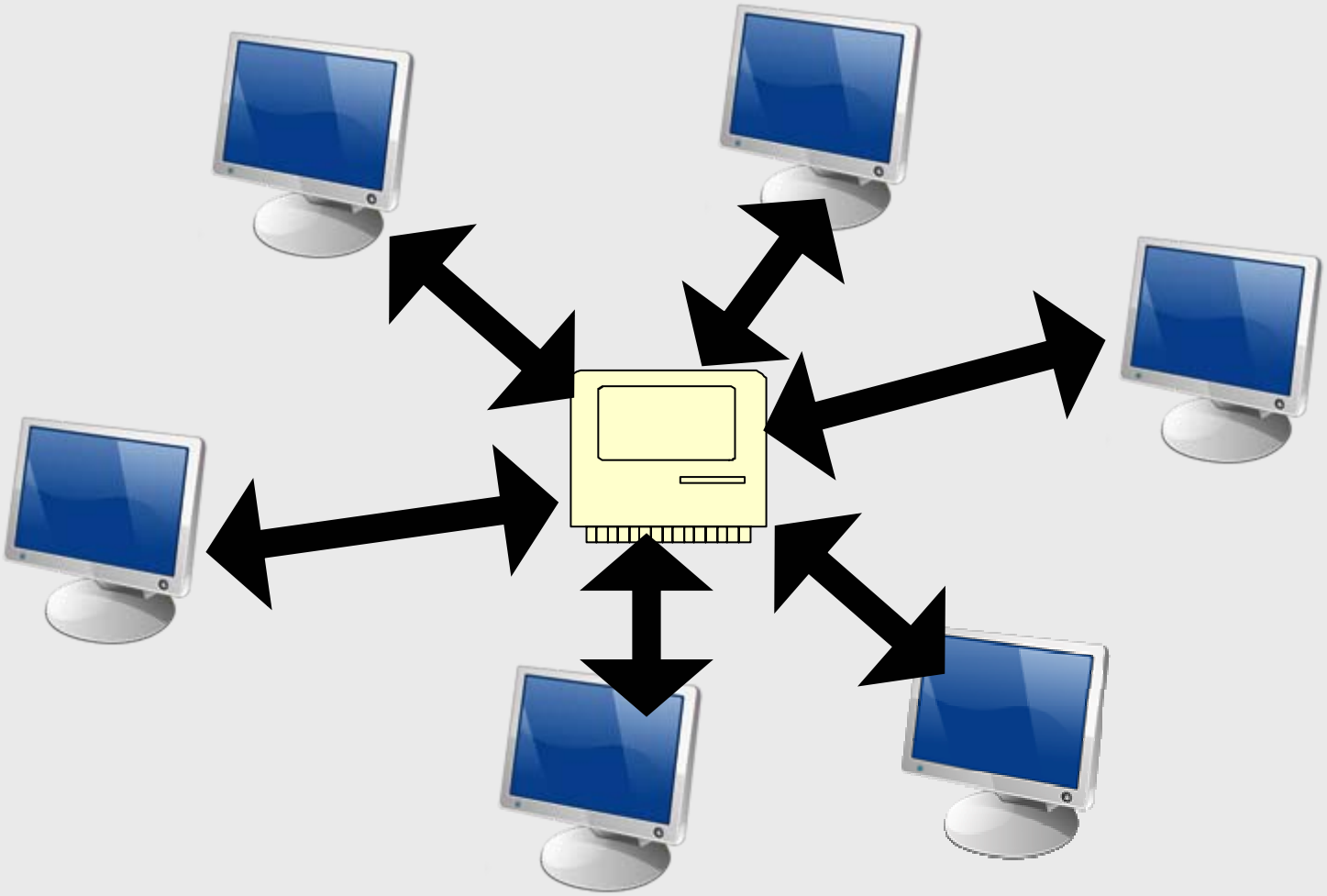


DISTRIBUTED SPEECH TECHNOLOGIES

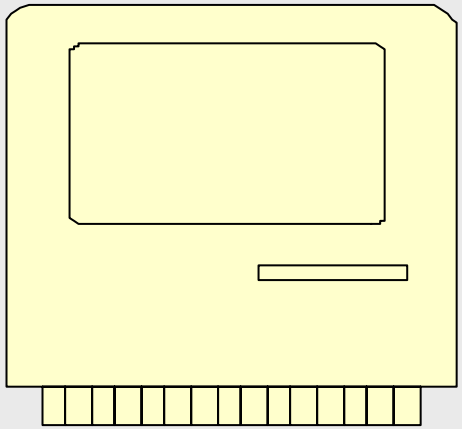
Overview

- Introduction
- Distributed frameworks
- Design
- Applications
  - Language learning
  - Dialog systems
  - Web accessibility

Distributed frameworks



Web-based systems

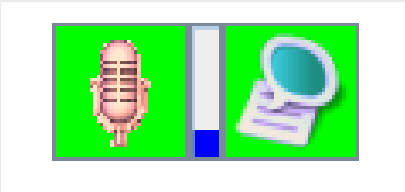


Pros and cons

- Pros:
  - Multi-platform
  - Only requires a Java-enabled web browser
- Cons:
  - Requires a decent Internet connection
  - Careful to cover all browsers

UZ Java Applet

- Simple Applet for ASR and TTS
- Graphic version



- HTML Code

```
<applet name="vivoreco" id="vivoreco" code="RecoFrontEndsimpleJApplet.class"
                                                    width="90" height="40"  archive=" sRecoFrontEnd_simple.jar">
                        <param name="host" value="gtc3pc23.cps.unizar.es">
        <param name="port" value="22229">
        <param name="sinte_speaker" value="Jorge" >
        <param name="sinte_service“ value="http://gtc3pc23.cps.unizar.es:8080/tts_servlet_unizar_cache_codec/sinte">
        <param name="sinte_codec" value="3">
        <param name="sinte_INI"   value="on">
</applet>
```

Javascript code

- TTS

```
void document.vivoreco.UZSinte(String sentence, String spk);
void document.vivoreco.UZSinteStop();
```

- ASR

```
void document.vivoreco.UZStopReco();
void document.vivoreco.UZStartReco();
void document.vivoreco.UZStartRecoGrammar(String
url_grammar);
void recopushini();
void recoend();
void recoerror();
```

Language learning

- Replication of language learning tools in web-based systems
- Cross-platform
- Relies on Internet connection

http://web.vocaliza.es

Dialog systems

- Development of oral dialog systems via web-based interfaces
- Aim: Language training tools, for instance

http://web.vocaliza.es

Web accessibility

- Blind people can’t access web content due to the visual nature of the web
- Speech can enable web-reading via TTS
- UZ Applet can provide a whole TTS-ASR experience

Web accessibility

- Development of accessible webs
  - HTML tags to indicate “readable” parts of the site (headlines, texts, links…)
  - Keyboard control of the reading process: Advance forward and back with keys
  - Enable the recognition of simple commands to speed up common processes

Web accessibility

- HTML tagging

```
<span class=”headings-sinte” title=”Synthesize this”></span>
<p class=”headings-sinte”> Synthesize this </p>
<a class=”headings-sinte” title=” Synthesize this”>But not this</a>
<a class=”headings-sinte”> Synthesize this </a>
```

- Elements and sub-elements

```
<span class=”headings-sinte” title=”I have sub-elements”>
<p class=”subheadings-sinte”>Synthesize this</p>
<a class=”subheadings-sinte”>Synthesize this</a>
</span>
```

Web accessibility

- Page control

"Ctrl+(right arrow)" or "Tab": Synthesizes next element on the list.
"Ctrl+(down arrow)": Re-synthesizes last element.
"Ctrl+(left arrow)": Synthesizes previous element.
"Ctrl+(flecha arriba)": Synthesizes first element.

- 3 levels of control
  - Main elements
  - Sub-elements
  - Long texts

Web accessibility

- Extra elements
  - Inclusion of UZ Applet
  - Inclusion of Javascript files
  - Definition of body onload()

```
<script type="text/javascript" charset="iso-8859-1" src="Uzaccess_vars.js"></script>
<script type="text/javascript" charset="iso-8859-1" src="Uzaccess_sinte.js"></script>
<body onload="mensaje_bienvenida('Welcome')" onunload="salida()">
<a id="ghost-link"></a>
```

Web accessibility

- Example

http://www.vocaliza.es/ar2/ar2_frames.htm

HTML5

- What is html5?
  - HTML5 is the last version of markup language for the web. Adds many new syntactical features. These include the <video>, <audio>, and <canvas> elements, as well as the integration of SVG content

http://en.wikipedia.org/wiki/HTML5

http://www.w3.org/TR/html5/

Google speechwebkit

- Browser Chrome beta 11
  - Integrates html5+ajax+javascript

```
<form>
<input id="speech" size="100" type="text" x-webkit-
speech speech onwebkitspeechchange="inputChange();">
</form>
```

Download Google Chrome beta 11

http://www.google.com/intl/en/landing/chrome/beta/

http://www.w3.org/2005/Incubator/htmlspeech/

http://chrome.blogspot.com/2011/03/talking-to-your-computer-with-html5.html