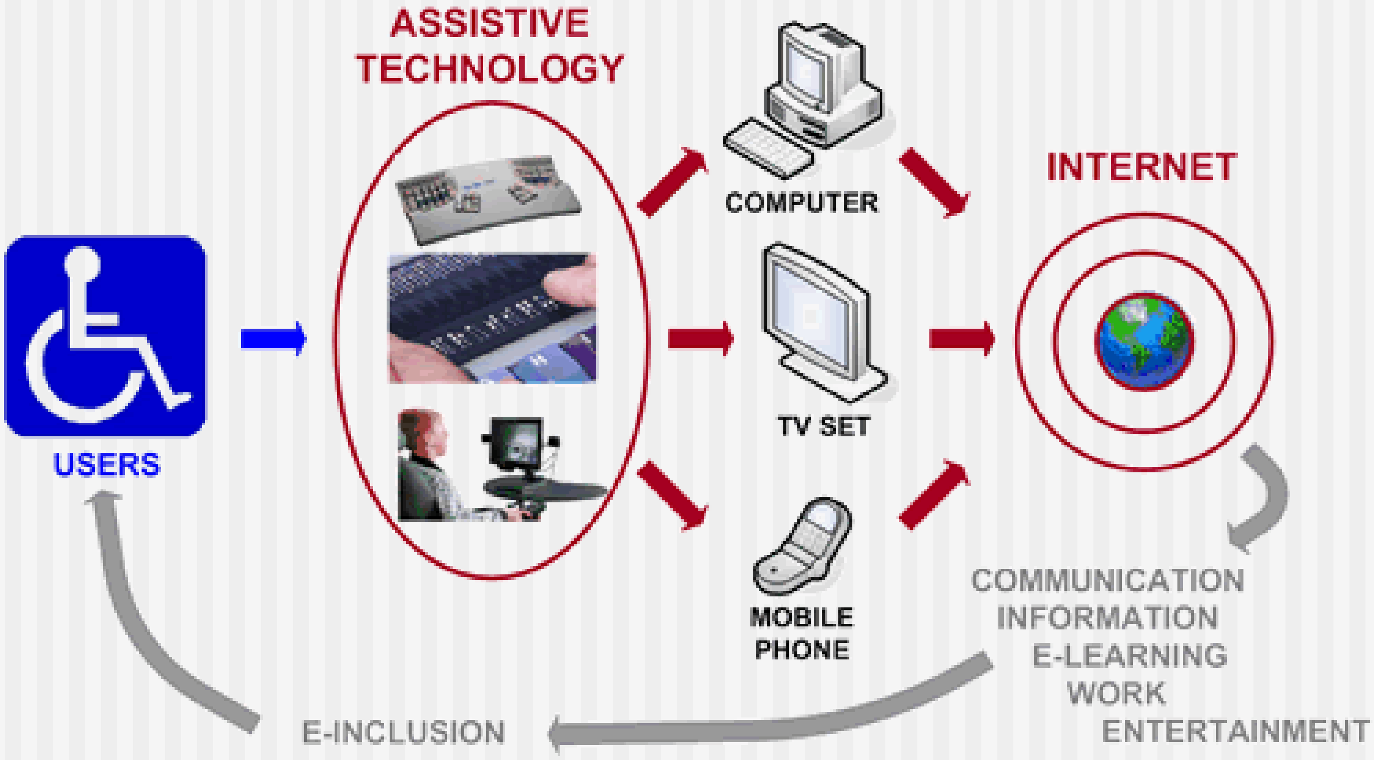
Speech Interfaces and e-Inclusion



TRABHCI
Rome 2013

Outline

- Spoken Language Communication and 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
    - Computer-aided Language Learning and Rehabilitation: Pre-linguistic skills.
    - Computer-aided Language Learning and Rehabilitation: Articulatory and Language skills
- Application Development
    - Kinect
    - Google tools
    - Assistant transcription tools

Spoken Language and E-inclusion

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

Spoken Language and E-inclusion

Speech Technologies

The aim of speech technology is to make communication between humans and humans, and humans and machines more efficient and easy.

Speech technologies includes several technologies as:

Speech analysis

Speech synthesis

Speech recognition

Speaker recognition

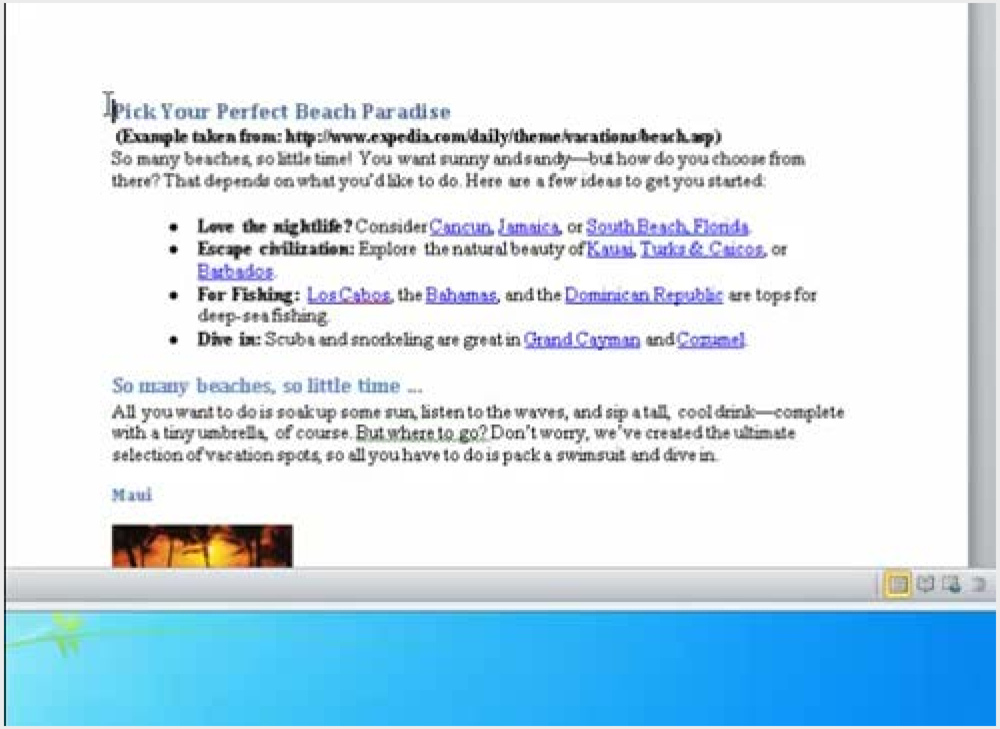
....

Spoken Language and E-inclusion

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

Spoken Language and E-inclusion

- speech technology can help people with disabilities and elderly people.
  - Blind and non-speaking people were amongst the first to be provided with commercially available speech synthesis systems
  - Screen-readers
  - Communication boards



Pick Your Perfect Beach Paradise

(Example taken from: http://www.expedia.com/daily/theme/vacations/beach.asp)

So many beaches, so little time! You want sunny and sandy—but how do you choose from there? That depends on what you'd like to do. Here are a few ideas to get you started:

- **Love the nightlife?** Consider Cancun, Jamaica, or South Beach, Florida.
- **Escape civilization:** Explore the natural beauty of Kauai, Turks & Caicos, or Barbados.
- **For Fishing:** Los Cabos, the Bahamas, and the Dominican Republic are tops for deep-sea fishing.
- **Dive in:** Scuba and snorkeling are great in Grand Cayman and Cozumel.

So many beaches, so little time ...

All you want to do is soak up some sun, listen to the waves, and sip a tall, cool drink—complete with a tiny umbrella, of course. But where to go? Don't worry, we've created the ultimate selection of vacation spots, so all you have to do is pack a swimsuit and dive in.

Maui

Spoken Language and E-inclusion

- Speech disorders, individuals lose the ability to produce their own speech
  - Use of alternative augmentative communication (AAC) devices
  - “Voice Output Communication Aid” or VOCA.
  - Voice banking for people who are at risk for losing their *voice*
    - Restoration of disordered speech
    - VOCA personalization

      original 🔈 adaptada 🔈

    - Traslation systems

      Speech2Speech

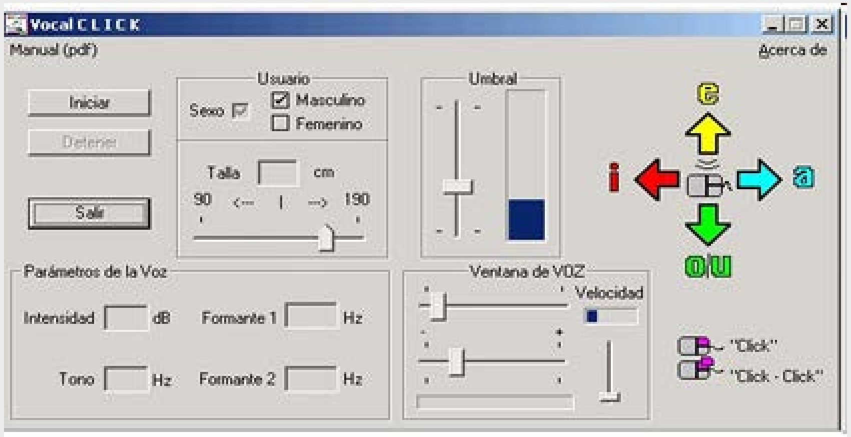
      Speech2Text: subtitling

      Speech2SignLanguage

Prof. Stephen Hawking, Amyotrophic *lateral* sclerosis (ALS), makes use of a VOZ device

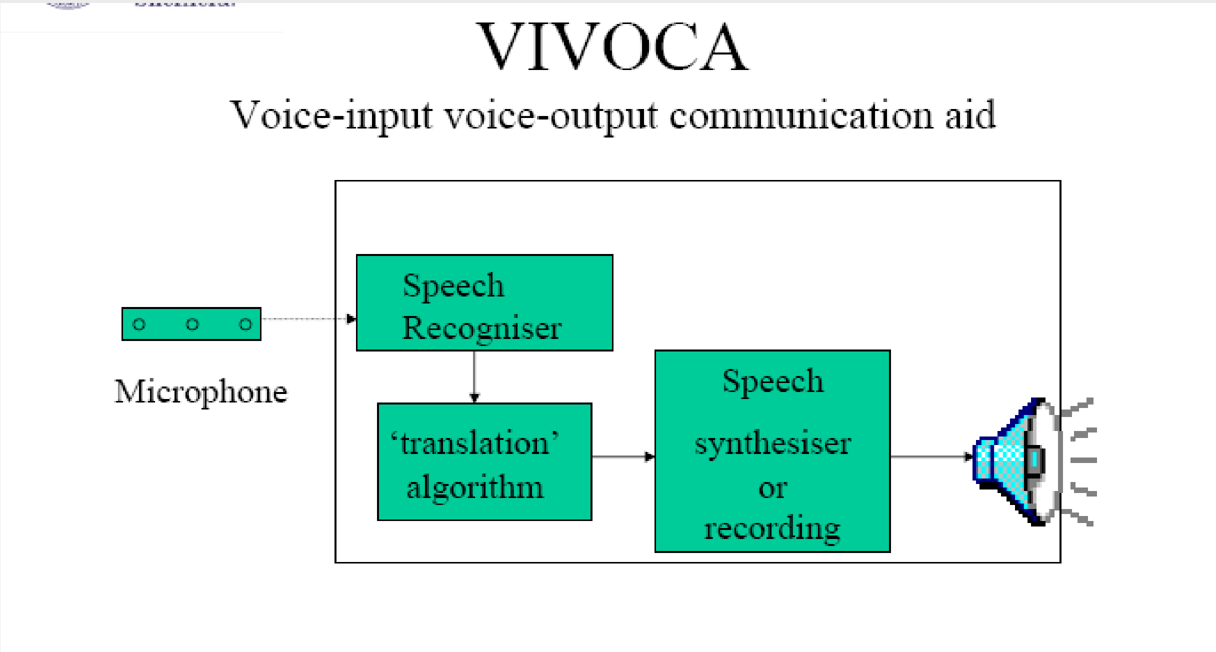
Spoken Language and E-inclusion

- Control of your environment
  - 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



Spoken Language and E-inclusion

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



VIVOCA

Voice-input voice-output communication aid

Microphone

Speech Recogniser

'translation' algorithm

Speech synthesiser or recording

The speech

The Speech is the particular and individual use of a language made by a speaker.

The speech is an individual act, opposed to a the language, which is social.

It is a vehicle of communication: Sender, Channel, Receiver

Two layers:

Physical support: voice signal

Sounds, Prosody, Emotion

Linguistic structure: Message

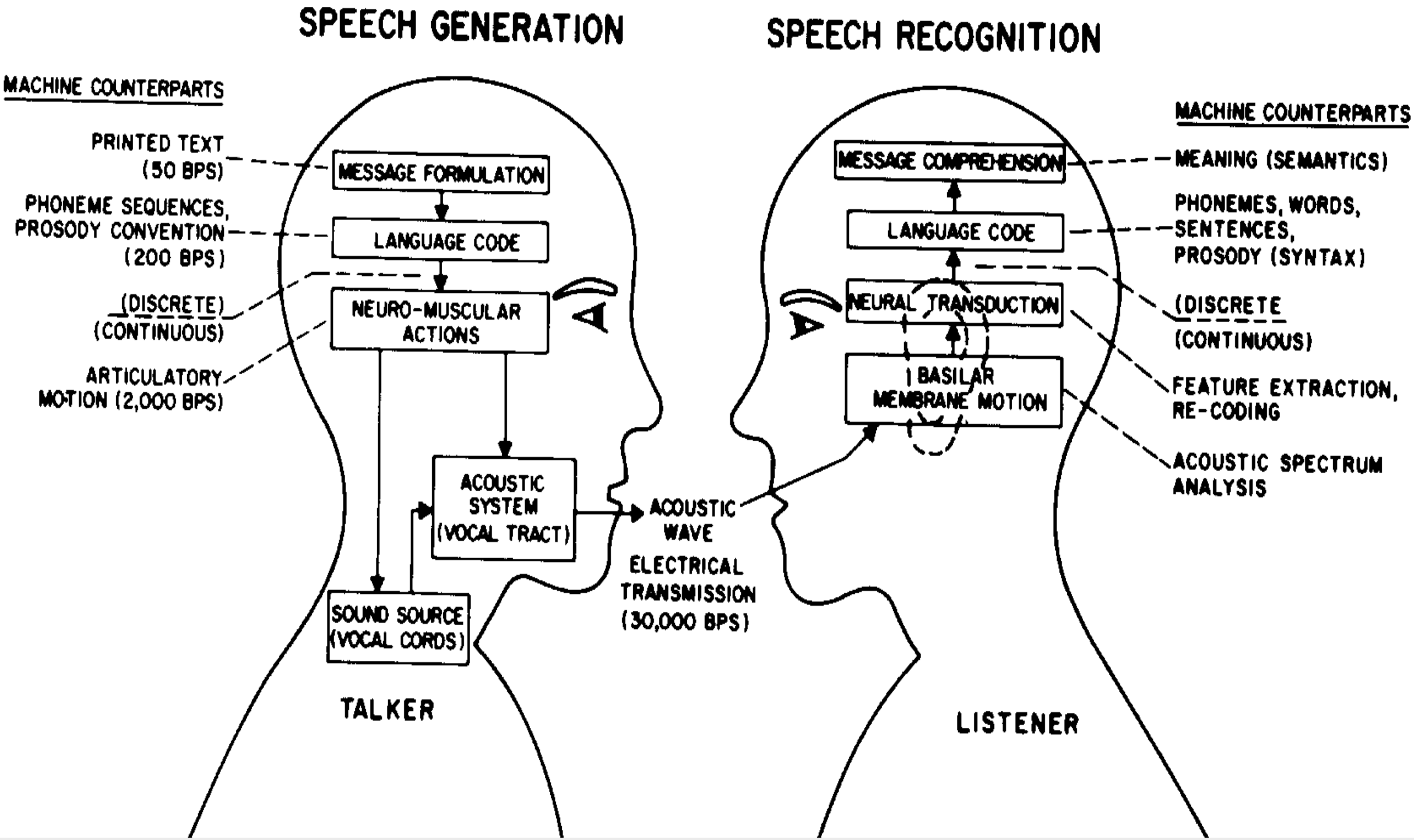
Lexicon, Syntax

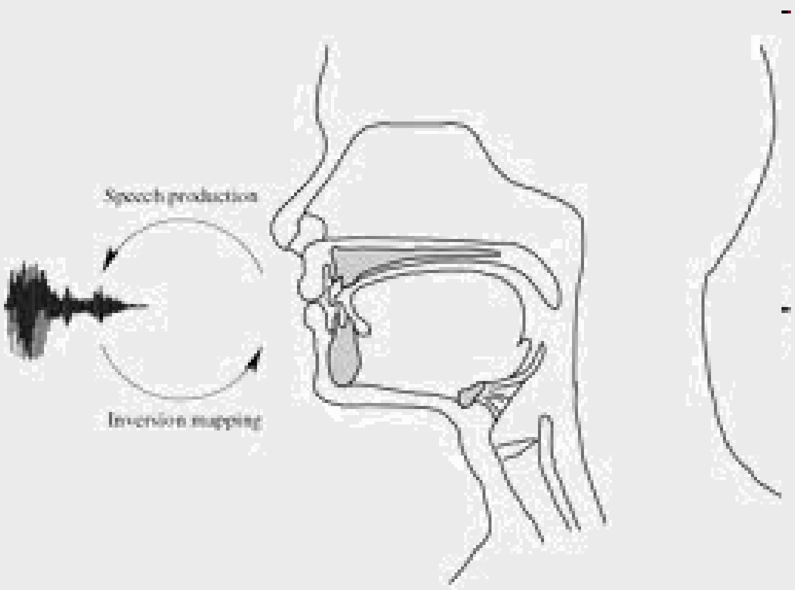
help can disabilities technology people speech with .

Semantics, Pragmatics

Time flies like an arrow

Human-Computer Interaction





The Speech Signal

and

Its Properties

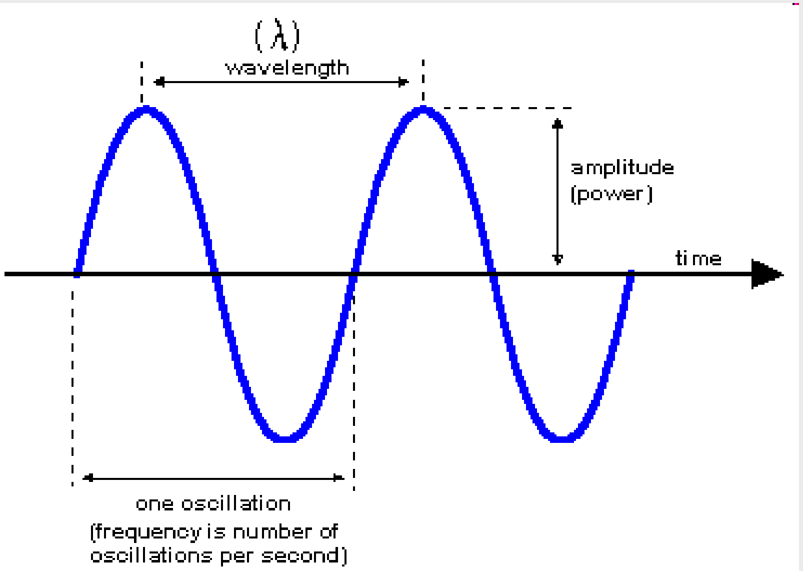
The speech signal and its properties

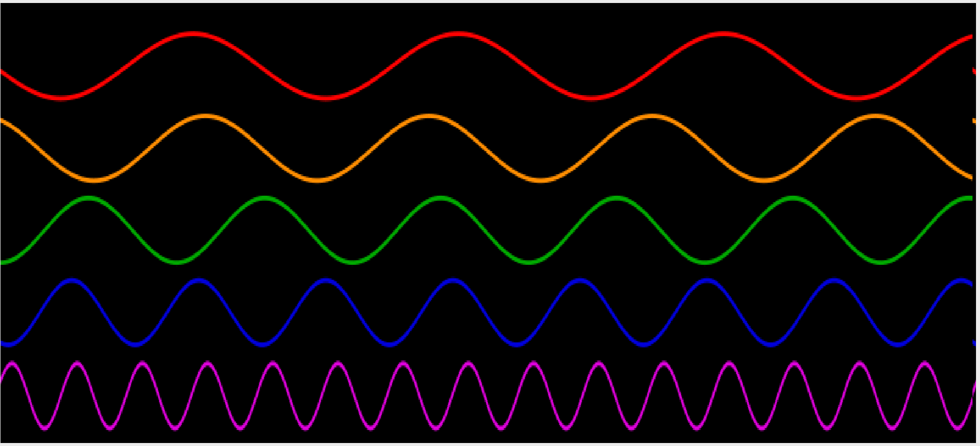
- What is a signal?
  - a time-dependent variation of a physical magnitude (voltage, current, EM field, pressure, ...) used to convey information from one place to another.
- What is speech?
  - The faculty or act of expressing or describing thoughts, feelings, or perceptions by the articulation of words.

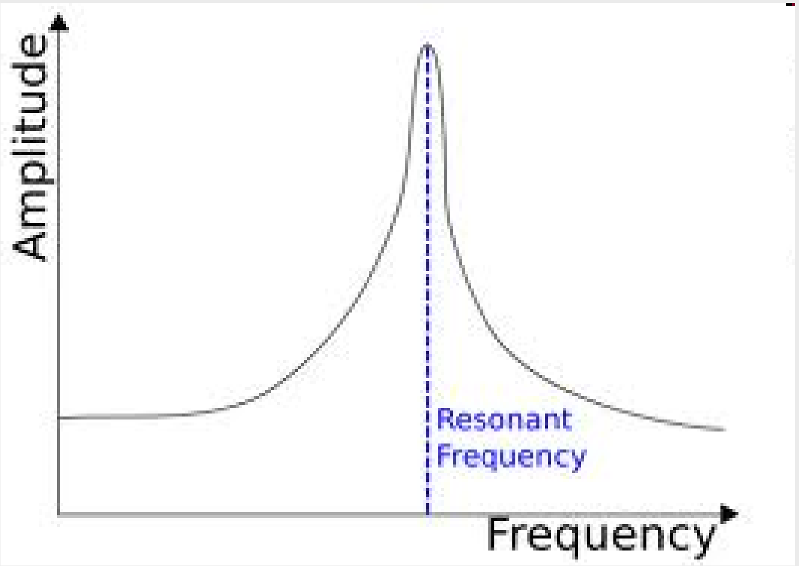
The speech signal and its properties

- How is represented a signal?
  - Time
    - waveform → represents the variation over time of the physical magnitude over time (independent variable)
  - Frequency
    - Related with periodic repetition of a physical magnitude.
      - Number of repetition of a phenomenon per time unit.
    - Represents the energy distribution of the physical magnitude over frequency
  - Time-Frequency

The speech signal and its properties

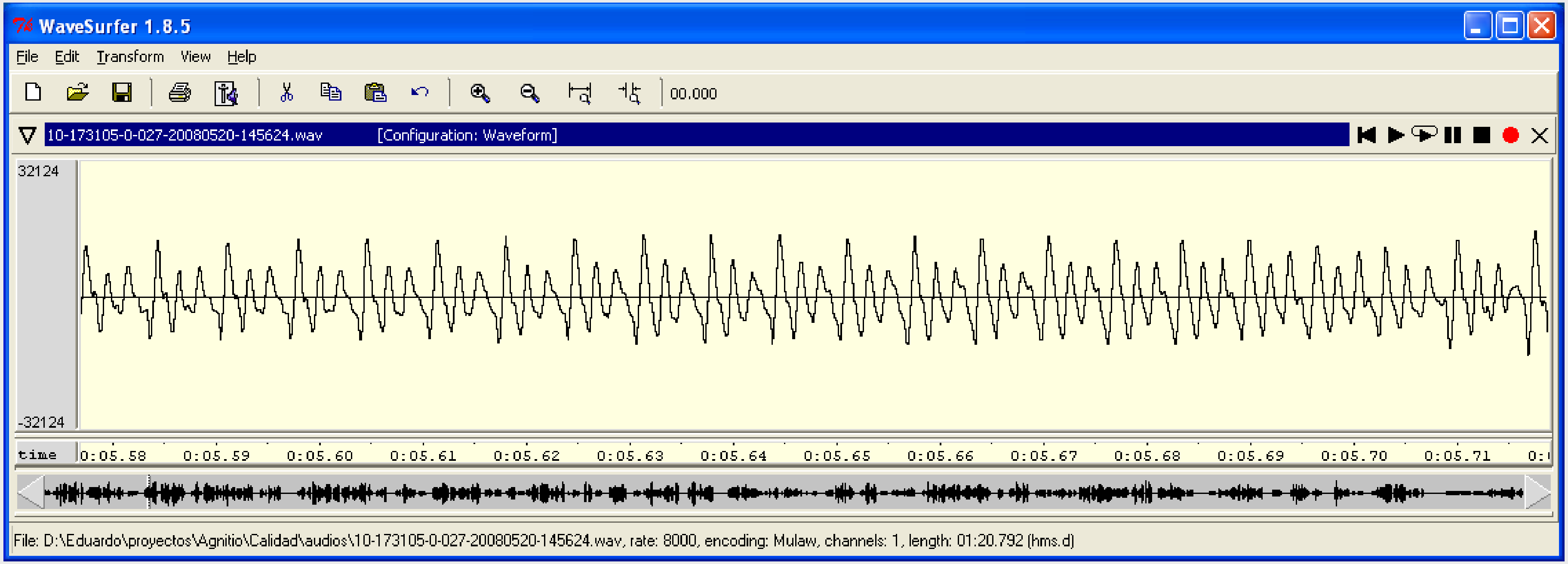






The speech signal and its properties

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



File: D:\Eduardo\proyectos\Agnitio\Calidad\audios\10-173105-0-027-20080520-145624.wav, rate: 8000, encoding: Mulaw, channels: 1, length: 01:20.792 (hms.d)

The speech signal and its properties

- How is produced the speech signal?

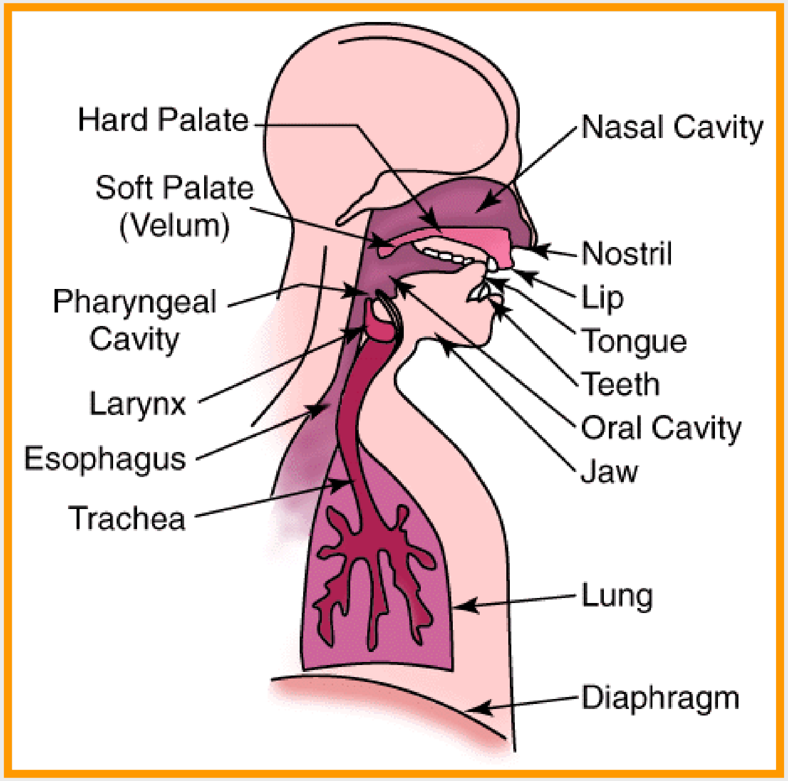
Vocal human apparatus

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

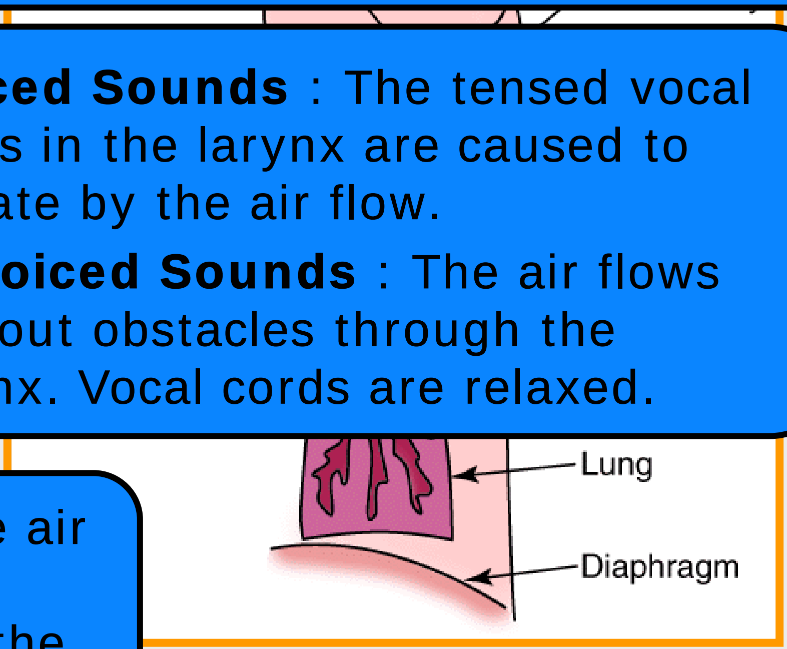
Vocal human appa

Voiced Sounds : The positions of several articulators (jaw, tongue, velum, lips, mouth) determine the sound that is produced.

Unvoiced Sounds : The air finds some obstacles in some point of the vocal tract.

Voiced Sounds : The tensed vocal cords in the larynx are caused to vibrate by the air flow.

Unvoiced Sounds : The air flows without obstacles through the larynx. Vocal cords are relaxed.



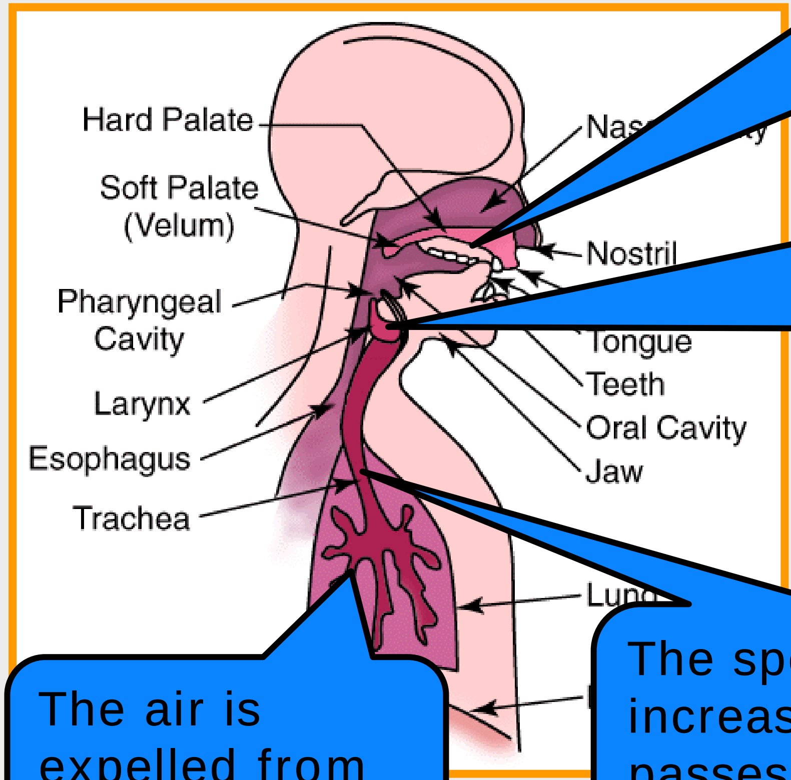
The air is expelled from the lung

The speed of the air increases as in passes through the Trachea

The speech signal and its properties

- How is produced

Vocal human appa



Voiced Sounds : The positions of several articulators (jaw, tongue, velum, lips, mouth) determine the sound that is produced.

Unvoiced Sounds : The air finds some obstacles in some point of the vocal tract.

Voiced Sounds : The tensed vocal cords in the larynx are caused to vibrate by the air flow.

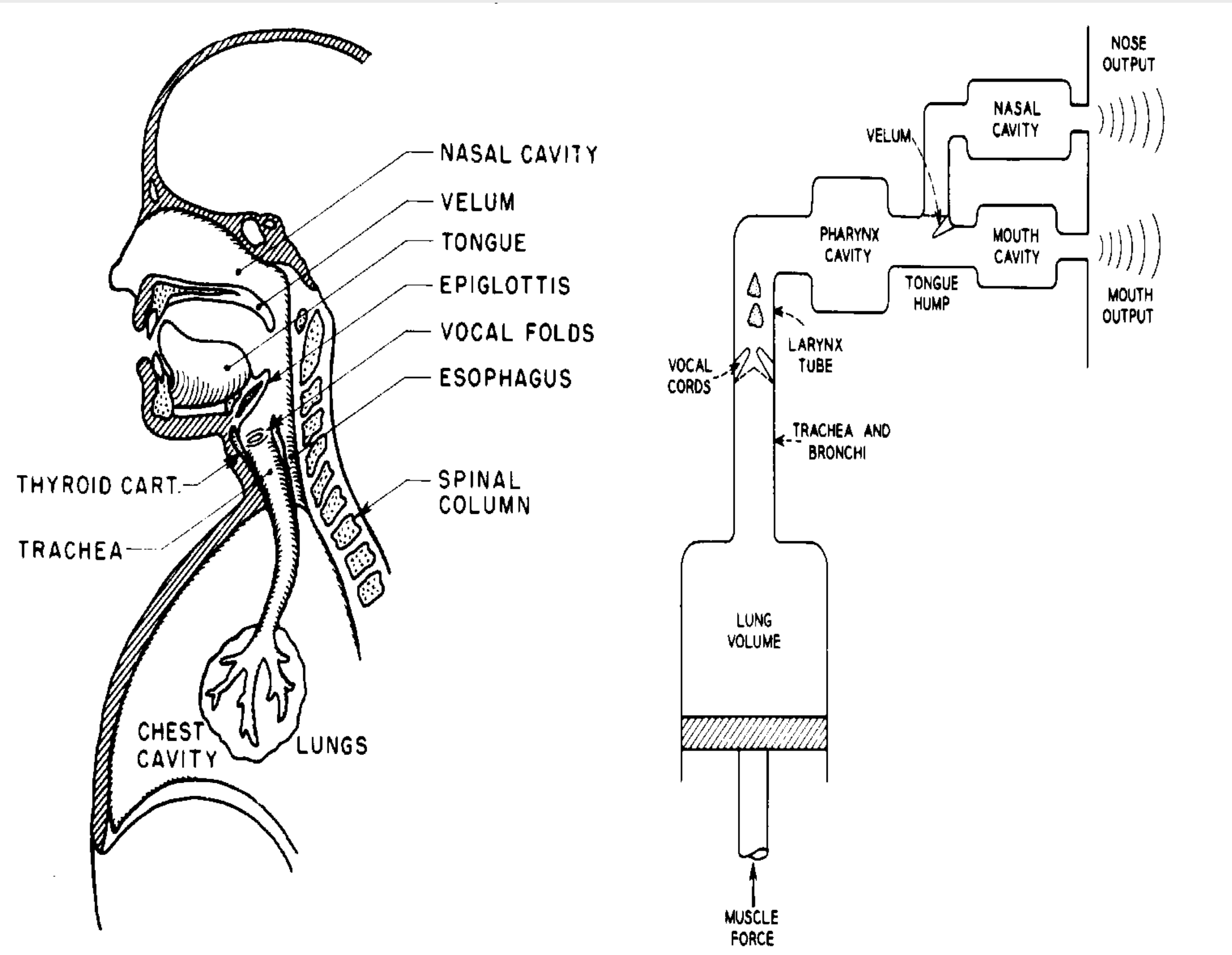
Unvoiced Sounds : The air flows without obstacles through the larynx. Vocal cords are relaxed.

The speed of the air increases as in passes through the Trachea

The air is expelled from the lung

The speech signal and its properties

- How is produced the speech signal?



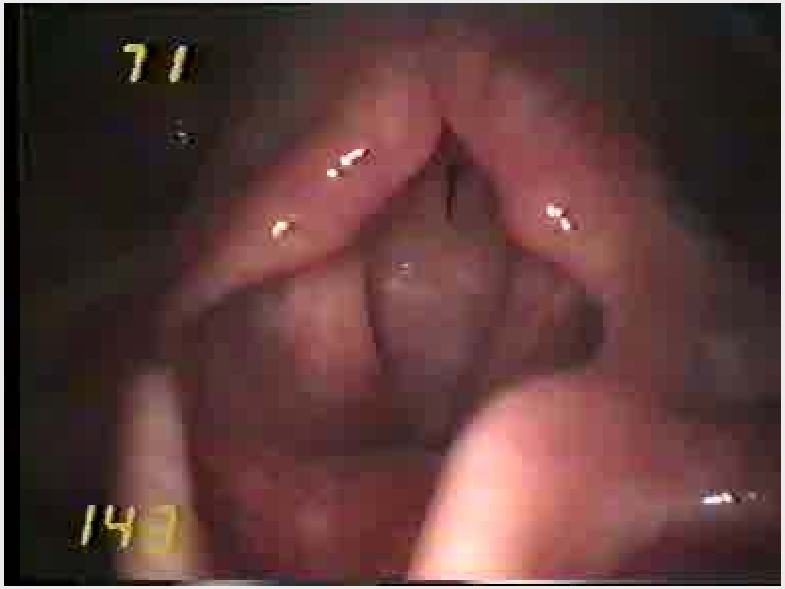
The speech signal and its properties

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

  15 mm long in men

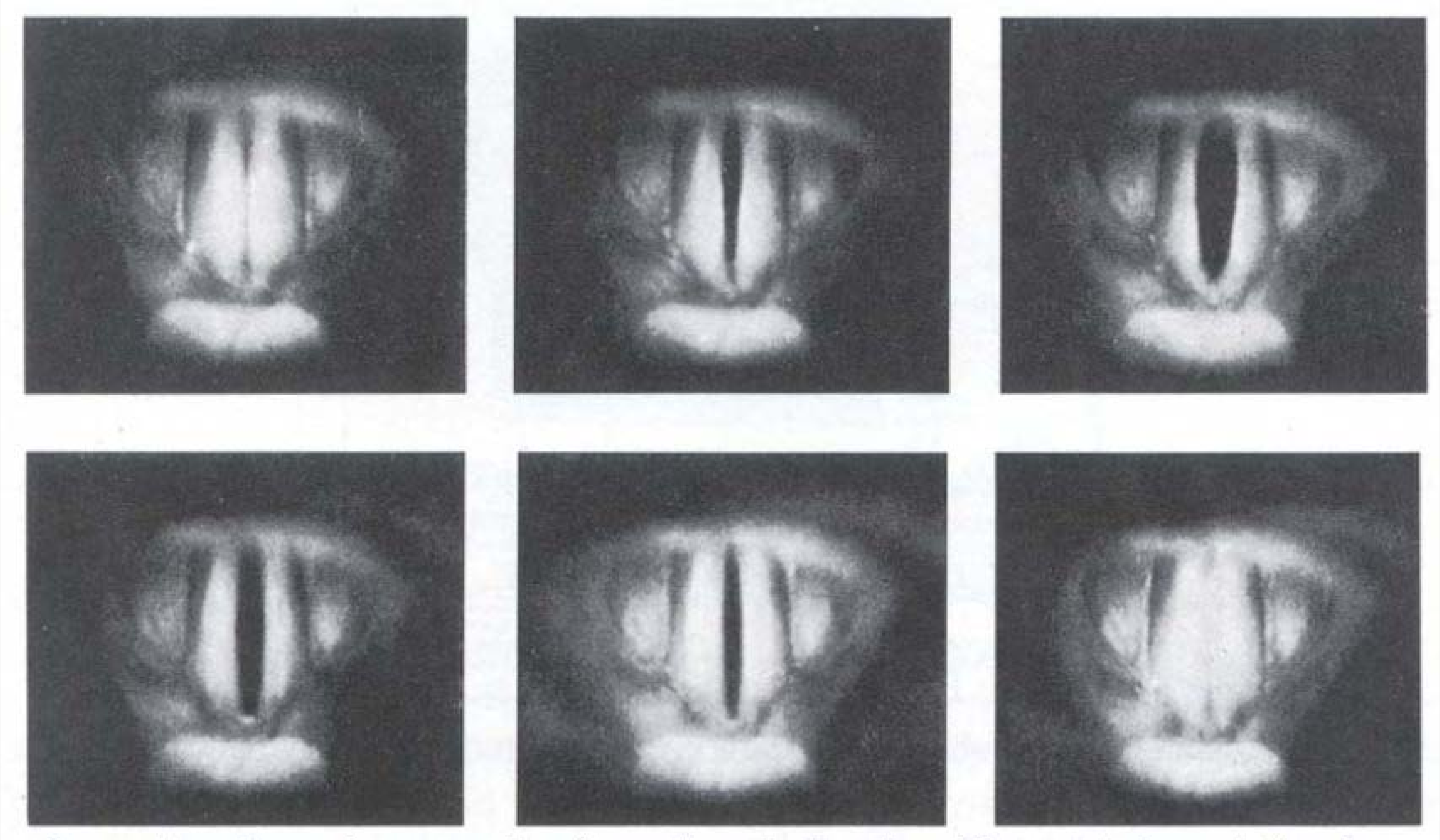
  13 mm long in women
- Can be varied in length and thickness and positioned
- Successive vocal fold openings
  - the fundamental period
  - the fundamental frequency

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



The speech signal and its properties

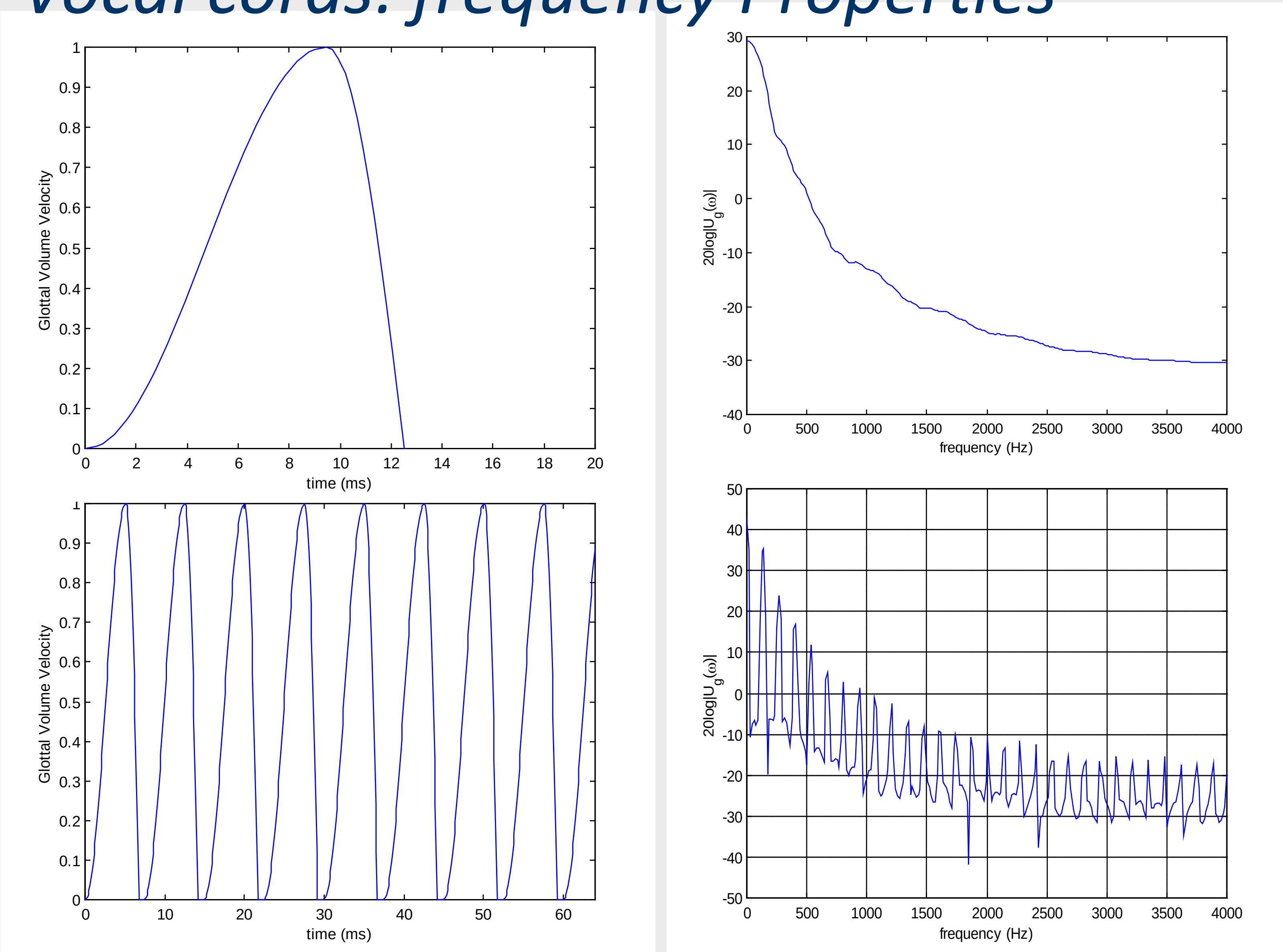
- The vocal cords



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

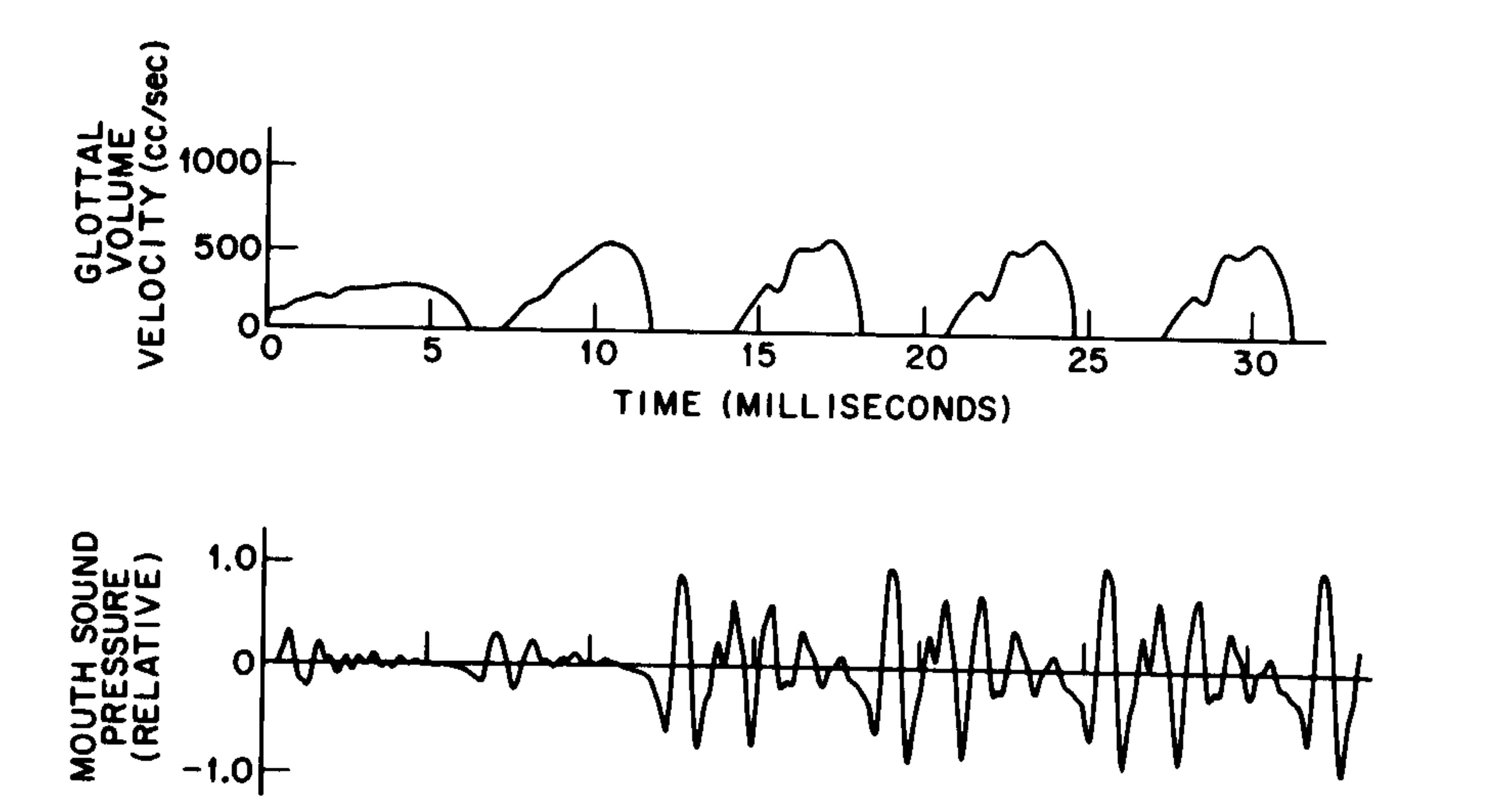
The speech signal and its properties

Vocal cords: frequency Properties



The speech signal and its properties

- From the vocal cords to the lips



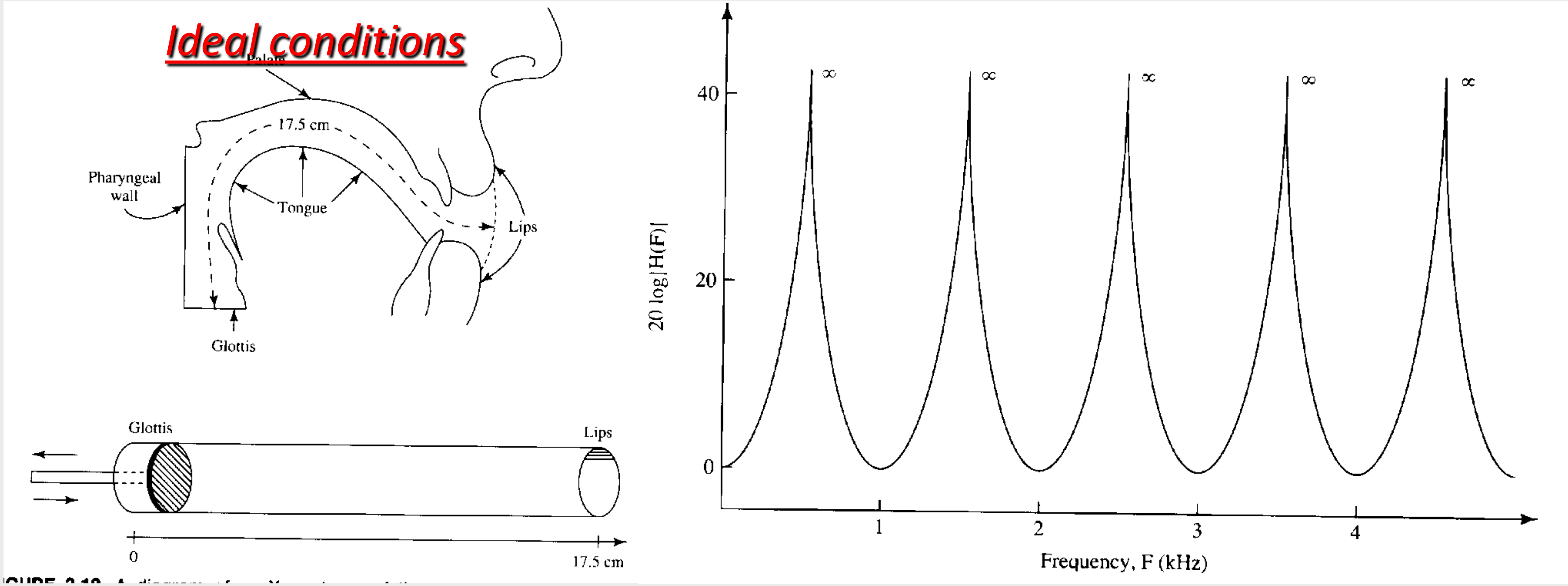
The speech signal and its properties

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***)

Ideal conditions

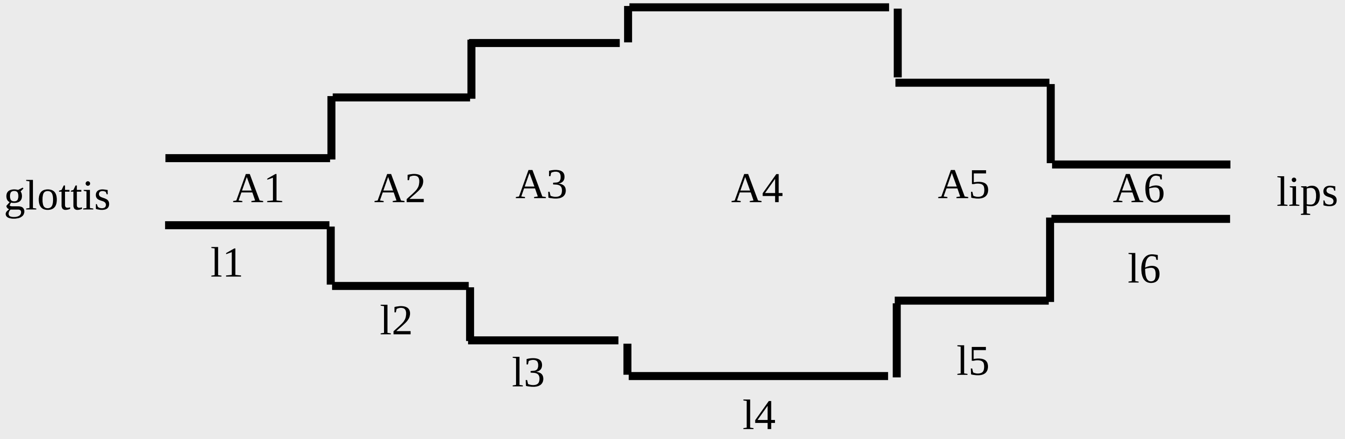


2. Generation of sounds

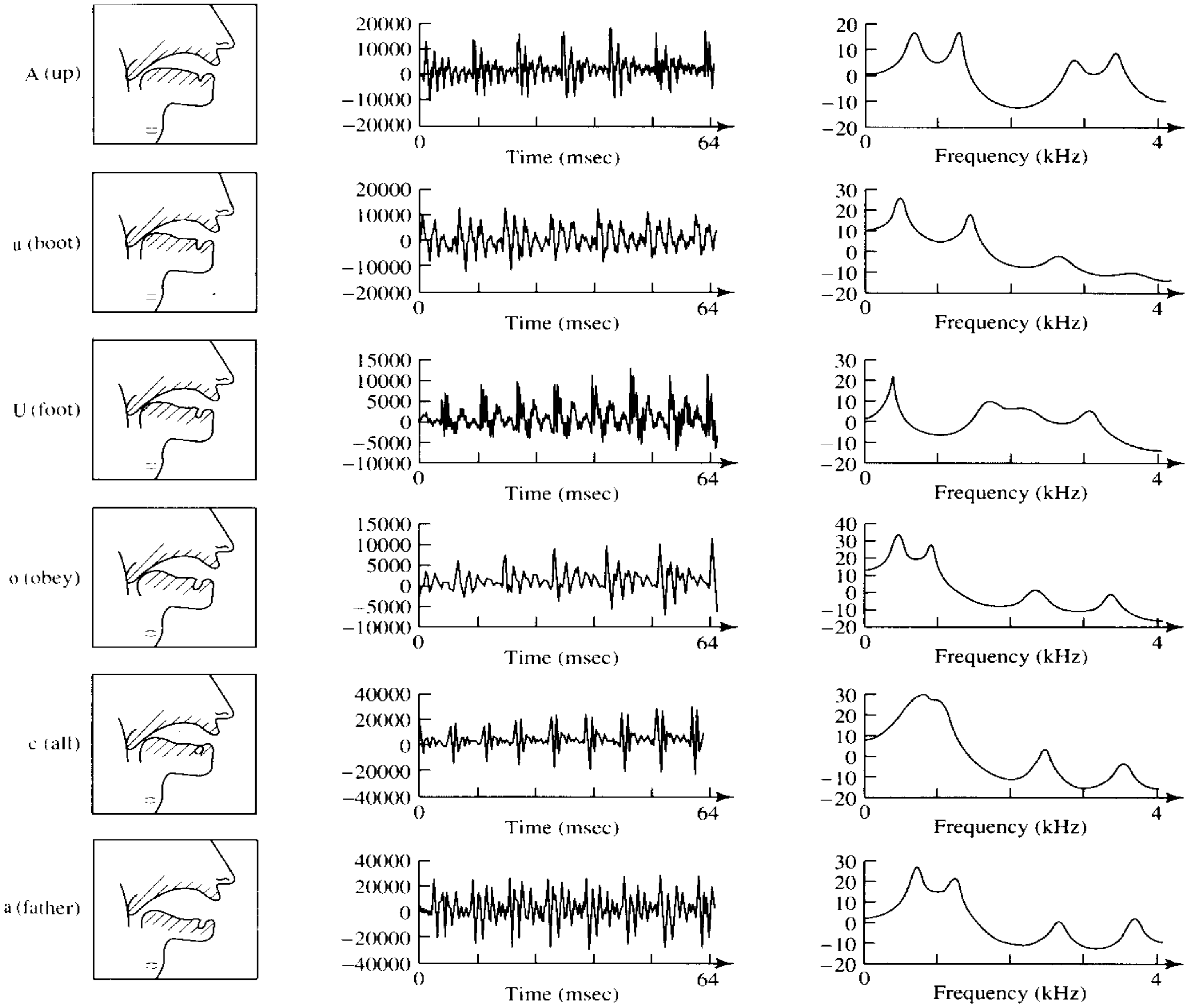
   A constriction at some point along the vocal tract generates a turbulence exciting a portion of the vocal tract (sound /s/ of six)

The speech signal and its properties

- The vocal Tract Tube Model
  - Describes how the vocal tract modifies the spectrum of the excitation signal to produce every sound
  - Articulators vary the shape of the vocal tract and thus the frequency response.



The speech signal and its properties



The speech signal and its properties

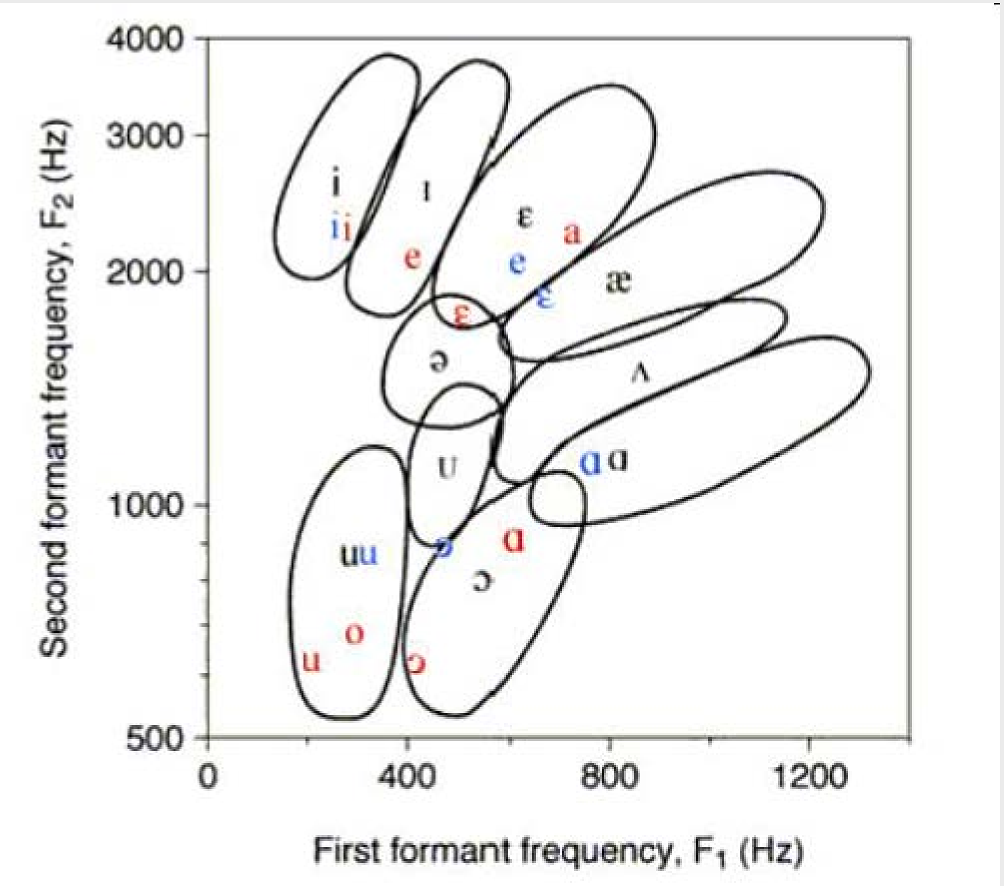
Speech Main Features

✓Pitch (fundamental frecuency)

From 80 to 400 cicles/sec (Hz)

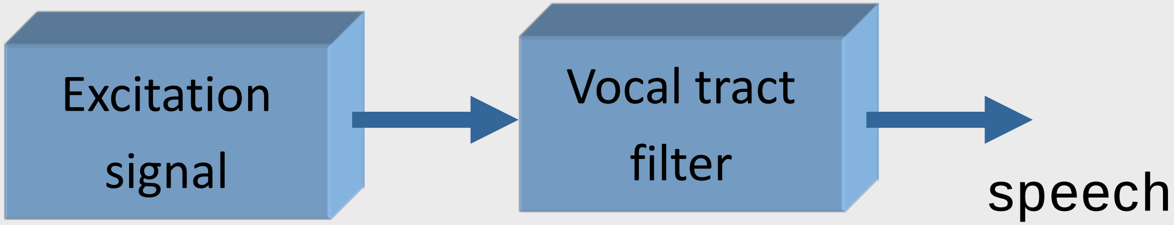
✓Formants

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



The speech signal and its properties

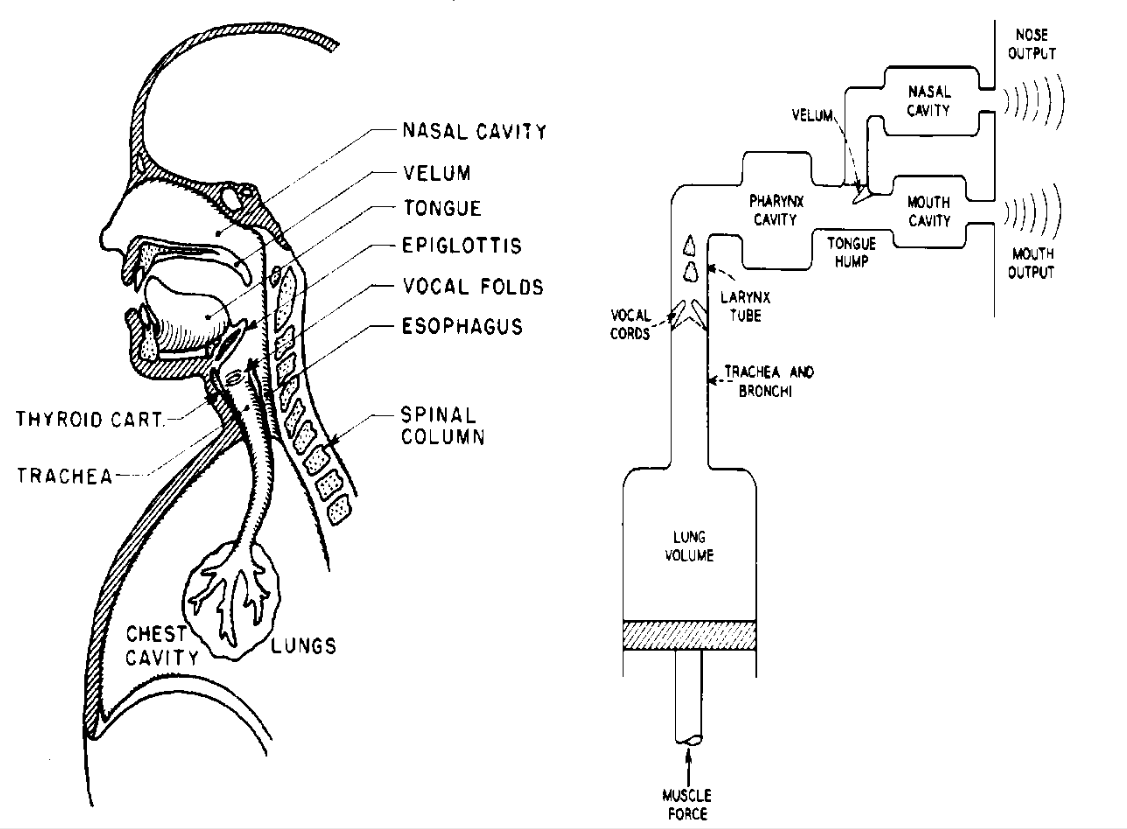
- Hear the vowels
  http://en.wikipedia.org/wiki/Vowel
- Let’s synthesize vowels from scratch

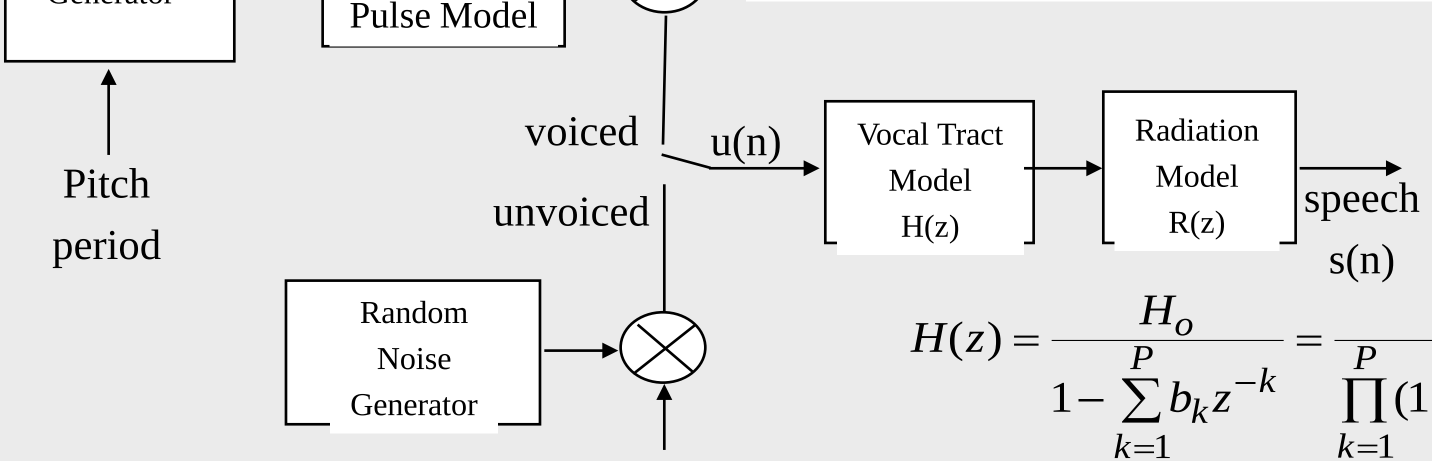


synth

- Let’s play with your speech
  download wavesurfer

Source-Filter Model





Voiced sounds Gain

Impulse Generator

Glottal Pulse Model

Pitch period

voiced

unvoiced

u(n)

Vocal Tract Model H(z)

Radiation Model R(z)

speech s(n)

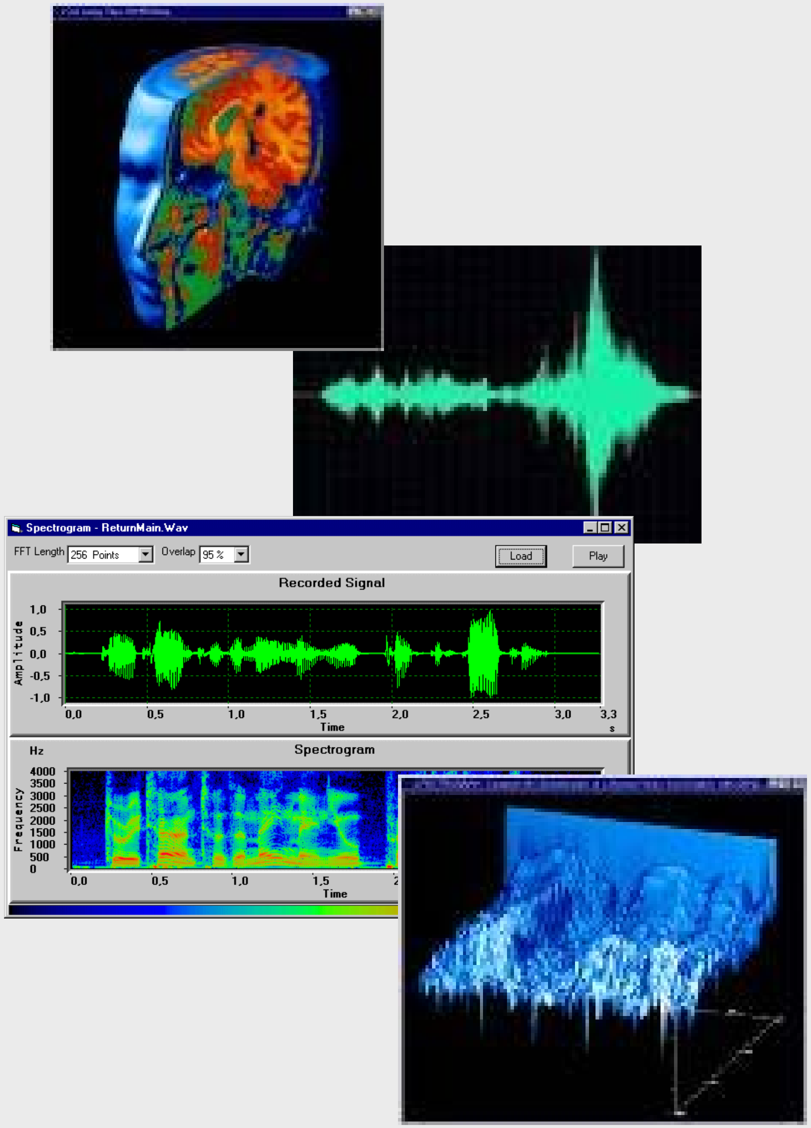
Random Noise Generator

Unvoiced sounds Gain

$$H(z) = \frac{H_o}{1 - \sum_{k=1}^{P} b_k z^{-k}} = \frac{H_o}{\prod_{k=1}^{P} (1 - p_k z^{-1})}$$

$$R(z) = 1 - z_o z^{-1} \qquad z_o \approx 1,\ z_o < 1$$

- The Technology



Spectrogram - ReturnMain.Wav

FFT Length 256 Points

Overlap 95 %

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

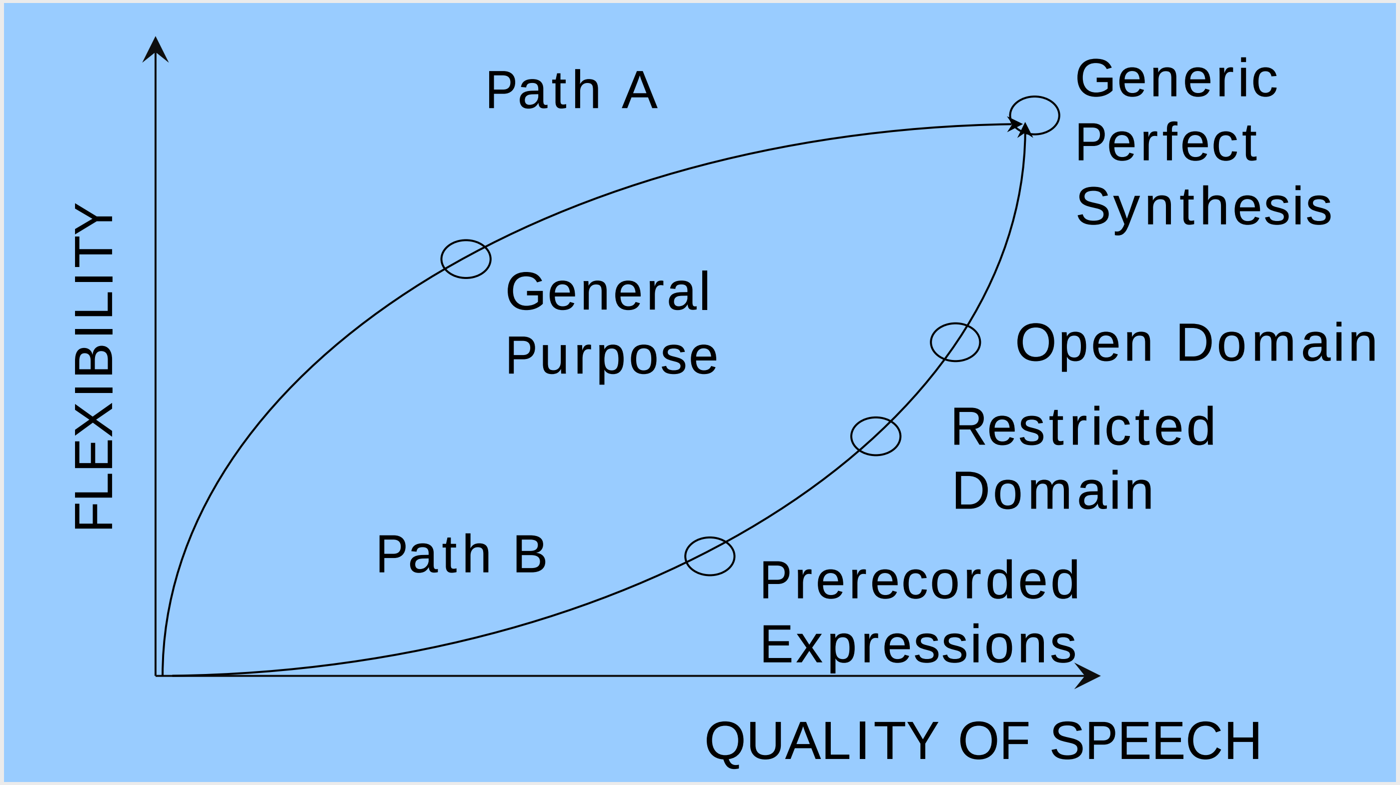
- Speech Enhancement:
  - An ASR system rapidly degrades due to acoustic distortion in the input signal
  - Main acoustic degradation:
    - Noise:
      - Access to voice web based application from the car, street, crowded place, industrial plant, etc. can become impossible if acoustic noise is not 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

- 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

- Text-to-Speech :
    - Trade-off between FLEXIBILITY and QUALITY OF SPEECH.

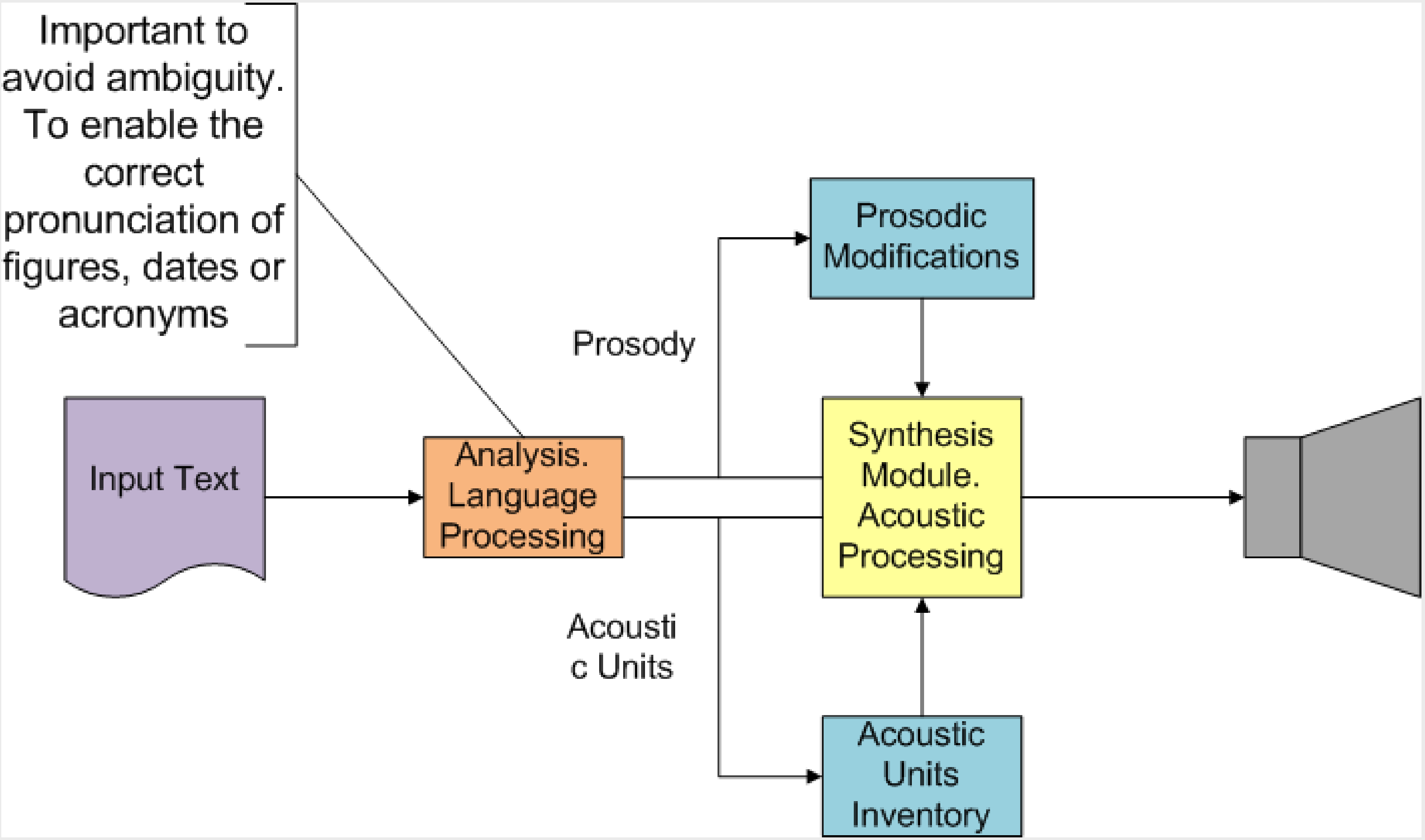


THE TECHNOLOGY

- Text-to-Speech :
  - Formant Synthesizer:
    - Parametric model: Vocal Tract model using formants
  - Concatenative Synthesizer:
    - Diphone synthesizer: Concatenation of prerecorded short segments plus signal processing.
    - Unit Selection: Concatenation of prerecorded segments (short and long) plus some signal processing and algorithms to select the best sequence of units.
  - HMM-based synthesis.
    - Parametric model: vocoder based on source-filter theory plus statistical model (HMM)

THE TECHNOLOGY

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

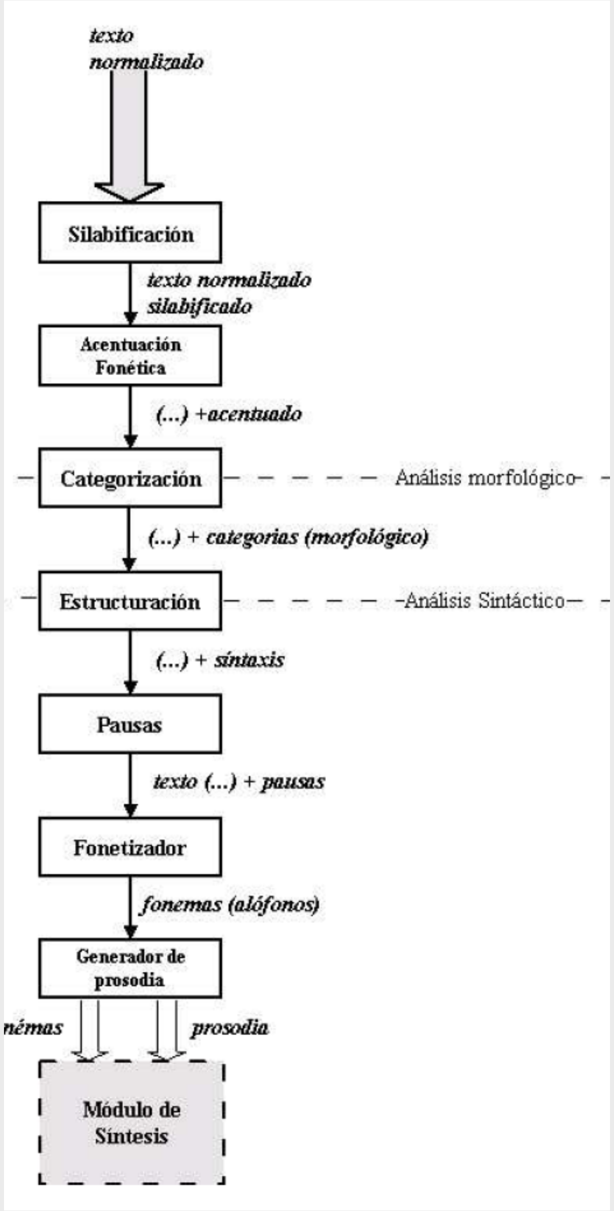

Important to avoid ambiguity. To enable the correct pronunciation of figures, dates or acronyms
Input Text
Analysis. Language Processing
Prosody
Acousti c Units
Prosodic Modifications
Synthesis Module. Acoustic Processing
Acoustic Units Inventory

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.

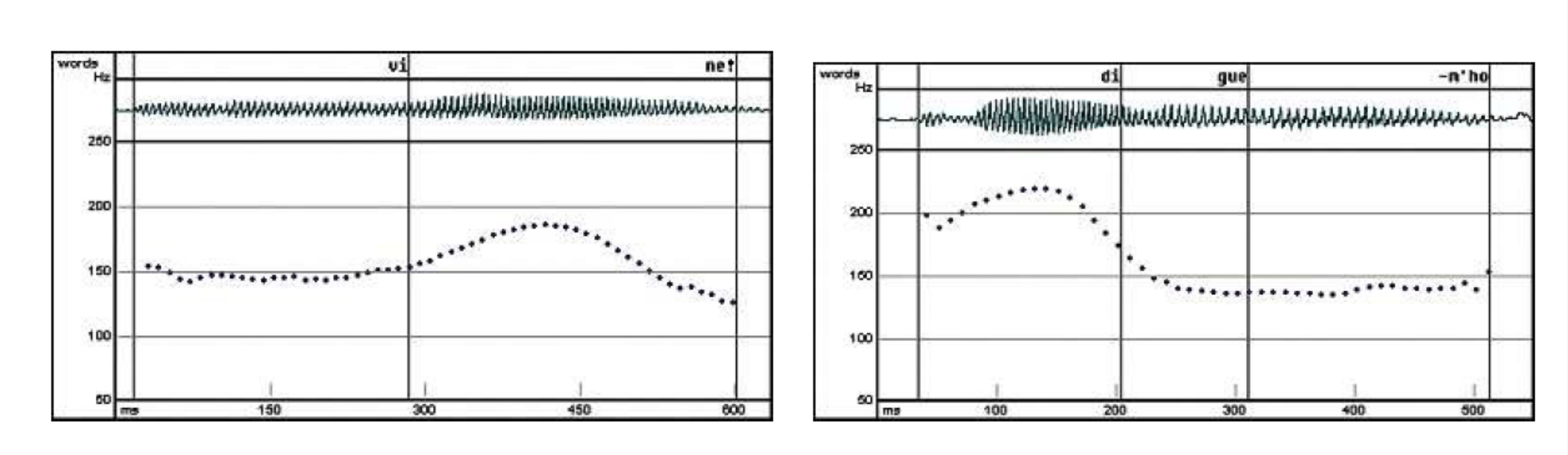
THE TECHNOLOGY

- 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 prosody.
  - 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:
  - Prosody 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:

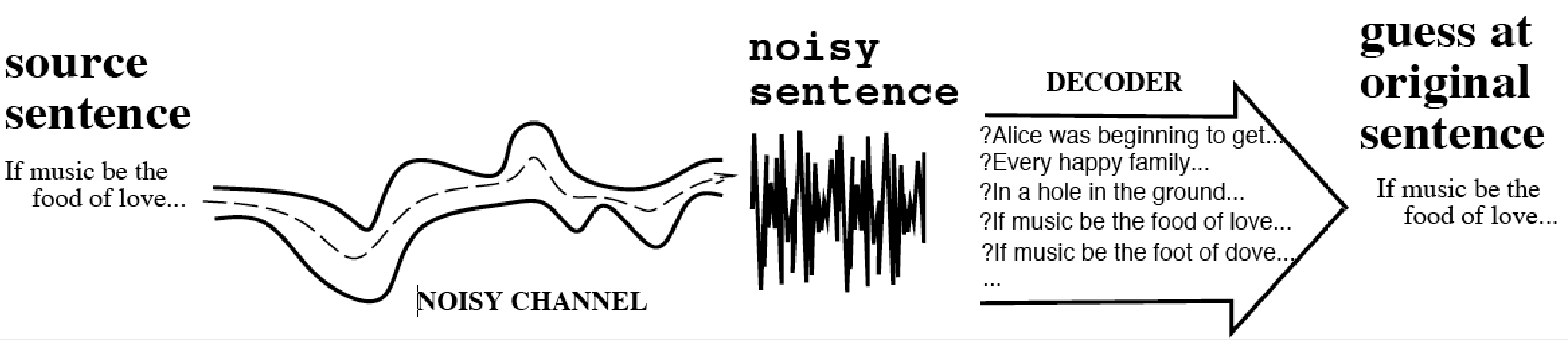
THE TECHNOLOGY

- 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)

Building an ASR System

- Build a statistical model of the speech-to-words process
  - Collect lots of speech and transcribe all the words
  - Train the model on the labeled speech
- Paradigm: The Noisy Channel Model
  - Automatic speech recognition (ASR) is a process by which an acoustic speech signal is converted into a set of words
  - Acoustic input considered a noisy version of a source sentence

Building an ASR System

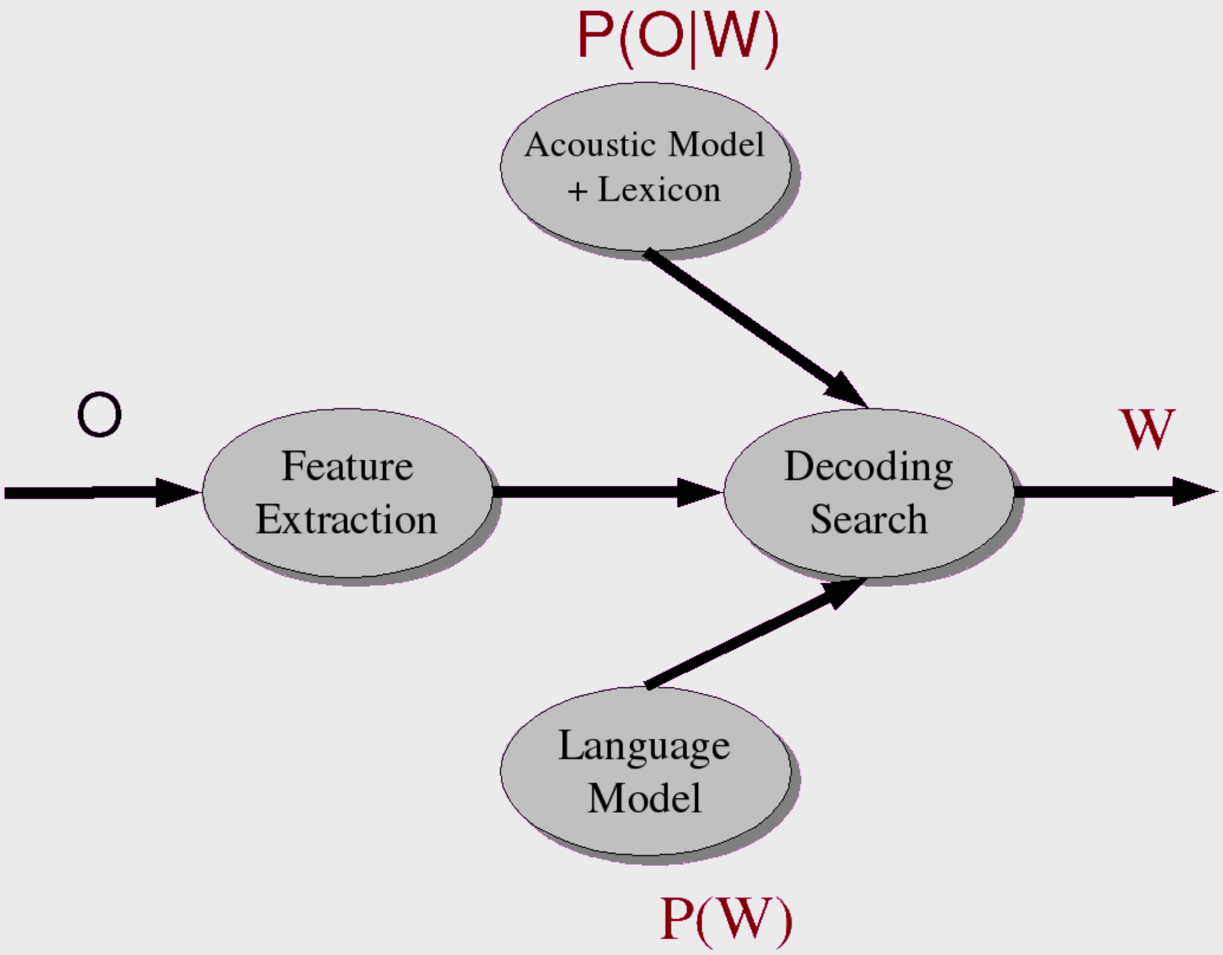

source sentence
If music be the food of love...
NOISY CHANNEL
noisy sentence
DECODER
?Alice was beginning to get...
?Every happy family...
?In a hole in the ground...
?If music be the food of love...
?If music be the foot of dove...
...
guess at original sentence
If music be the food of love...

What is the most likely sentence out of all sentences in the language L, given some acoustic input O?

Search through space of all possible sentences.

Pick the one that is most probable given the waveform

Building an ASR System



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

THE TECHNOLOGY

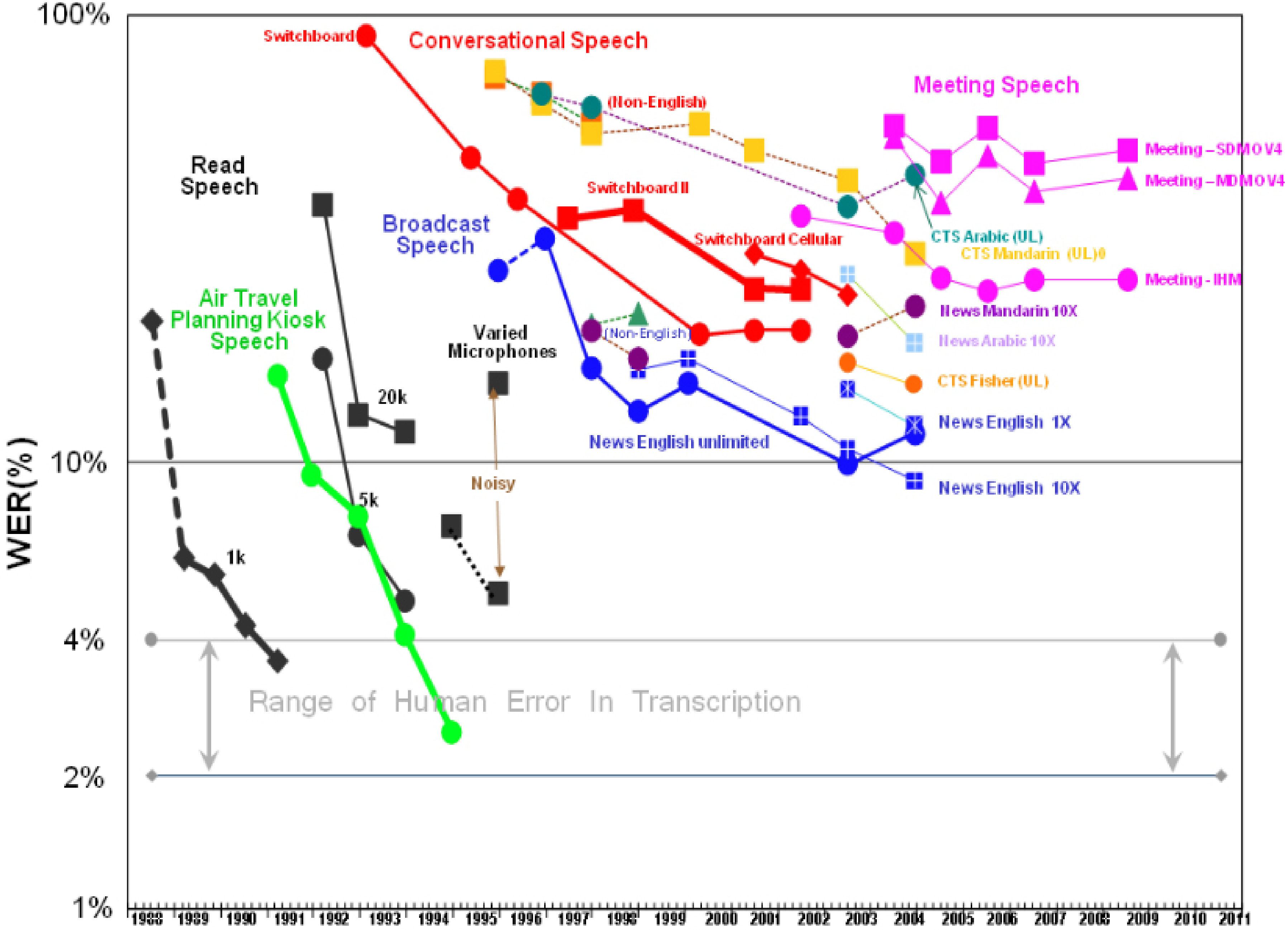
- Errors in a ASR system.
  - Deletions:
    - The speaker says something but nothing is the returned by the systems
  - Substitutions:
    - The output of the system is a different word than the one uttered by the speaker.
  - Insertions:
    - The user said nothing but a word is the output of the systems (acoustic artifacts leaded the system)

THE TECHNOLOGY

- Sources of errors：

Problem	Cause
Deletion or Substitution	The user said something out-of-vocabulary
	The uttered word does not belong to the active grammar
	The user started speaking before the system was ready to listen.
	Confused words sound alike.
	Too long pauses between sentences.
	Disfluencies (false start, “uhmmm”, “eeehh”, ...)
	The user has an accent or cold
	The user has a voice substantially different than the model.
	The microphone is not properly
Insertion	Non-speech sound (e.g.. Cough, laugh,...)
	Background speech triggers recognition
	The user is talking to another person.

NIST STT Benchmark Test History – May. '09



- 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 easier to use.

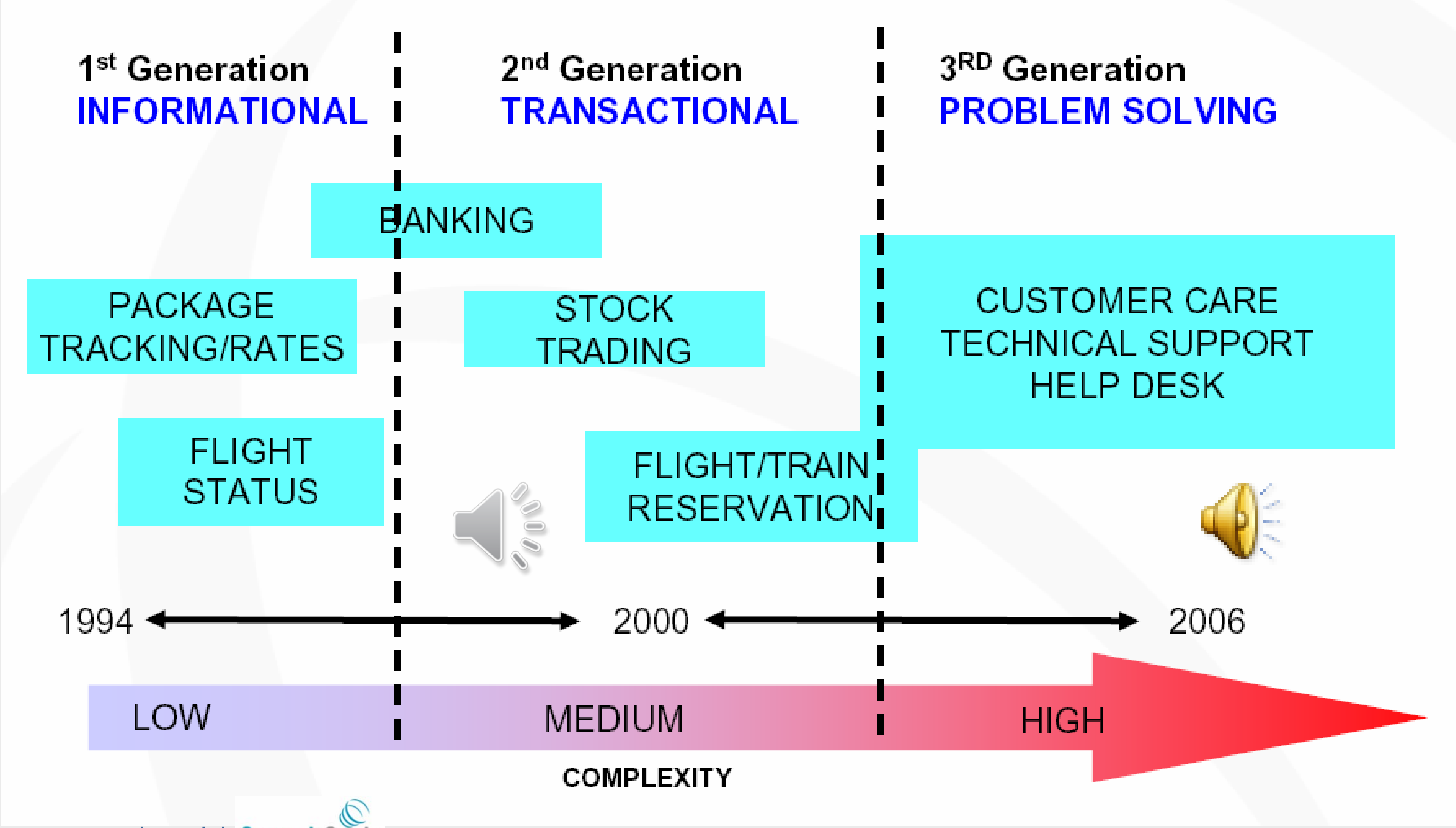
- Voice Input / Voice Output Interfaces:
  - When is Speech considered an appropriate OUTPUT?
    - When the user is COOPERATIVE
    - Use Speech as OUTPUT when …
      - Eyes-busy situations: Drivers, Industrial Plants Workers,…
      - the user has some kind of perceptual disability or visual limitation
      - 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

Spoken dialogue systems

- Application that enables the communication between the human and the machine, in the most natural way.
- Speech is the most natural way for humans to communicate:
  - SPOKEN DIALOGUE SYSTEMS
- Functional requisites
  - Understand instructions uttered by the user
  - Report the user any event that takes place during the execution of the requested actions.

Spoken Dialogue System Generation


1st Generation INFORMATIONAL

2nd Generation TRANSACTIONAL

3RD Generation PROBLEM SOLVING

BANKING

PACKAGE TRACKING/RATES

STOCK TRADING

CUSTOMER CARE TECHNICAL SUPPORT HELP DESK

FLIGHT STATUS

FLIGHT/TRAIN RESERVATION

1994 2000 2006

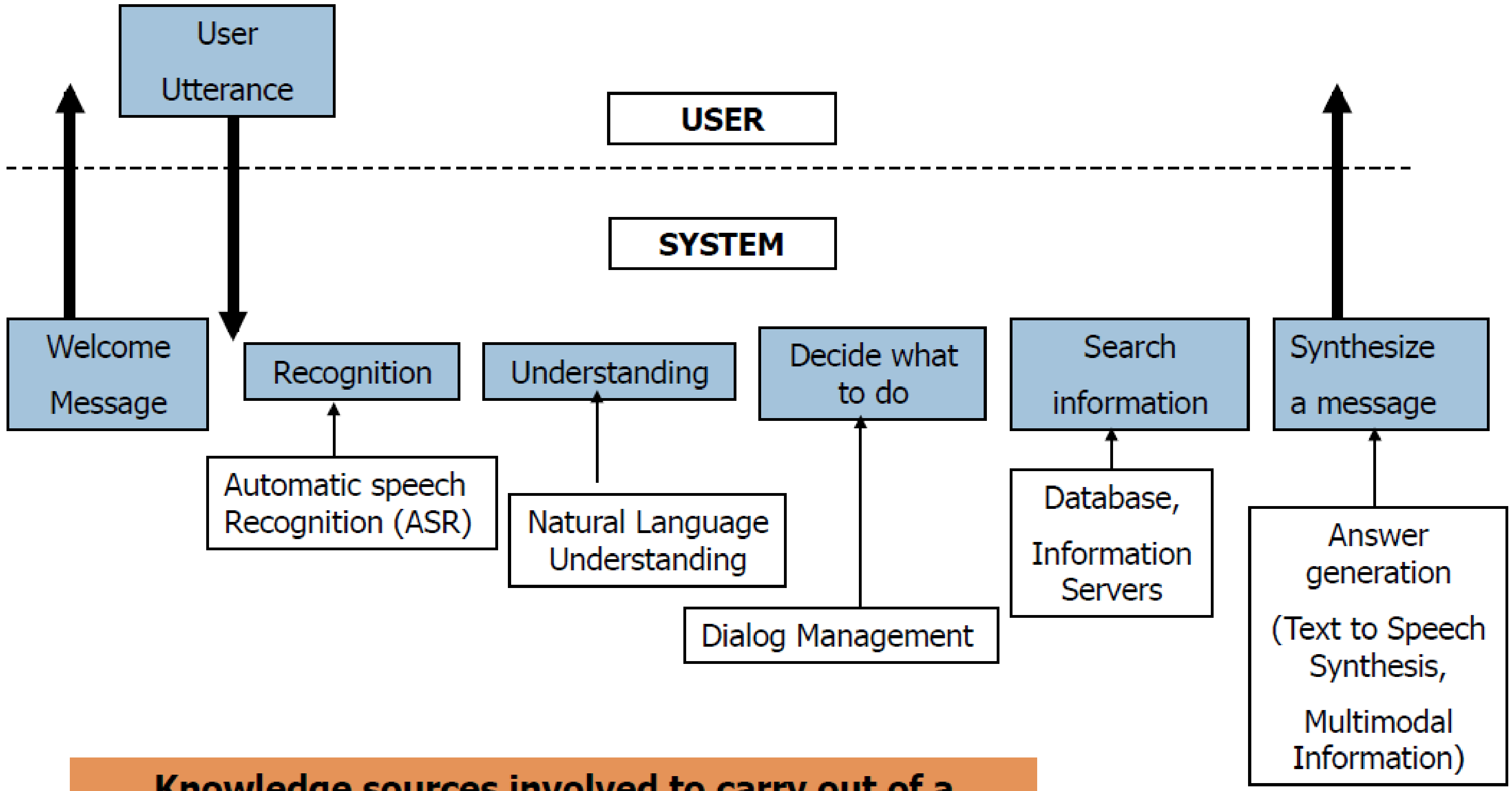
LOW MEDIUM HIGH

COMPLEXITY

Fuente R. Pieraccini, SpeechCycle

TV
14
D

Spoken Dialog System Scheme



User Utterance

USER

SYSTEM

Welcome Message

Recognition

Understanding

Decide what to do

Search information

Synthesize a message

Automatic speech Recognition (ASR)

Natural Language Understanding

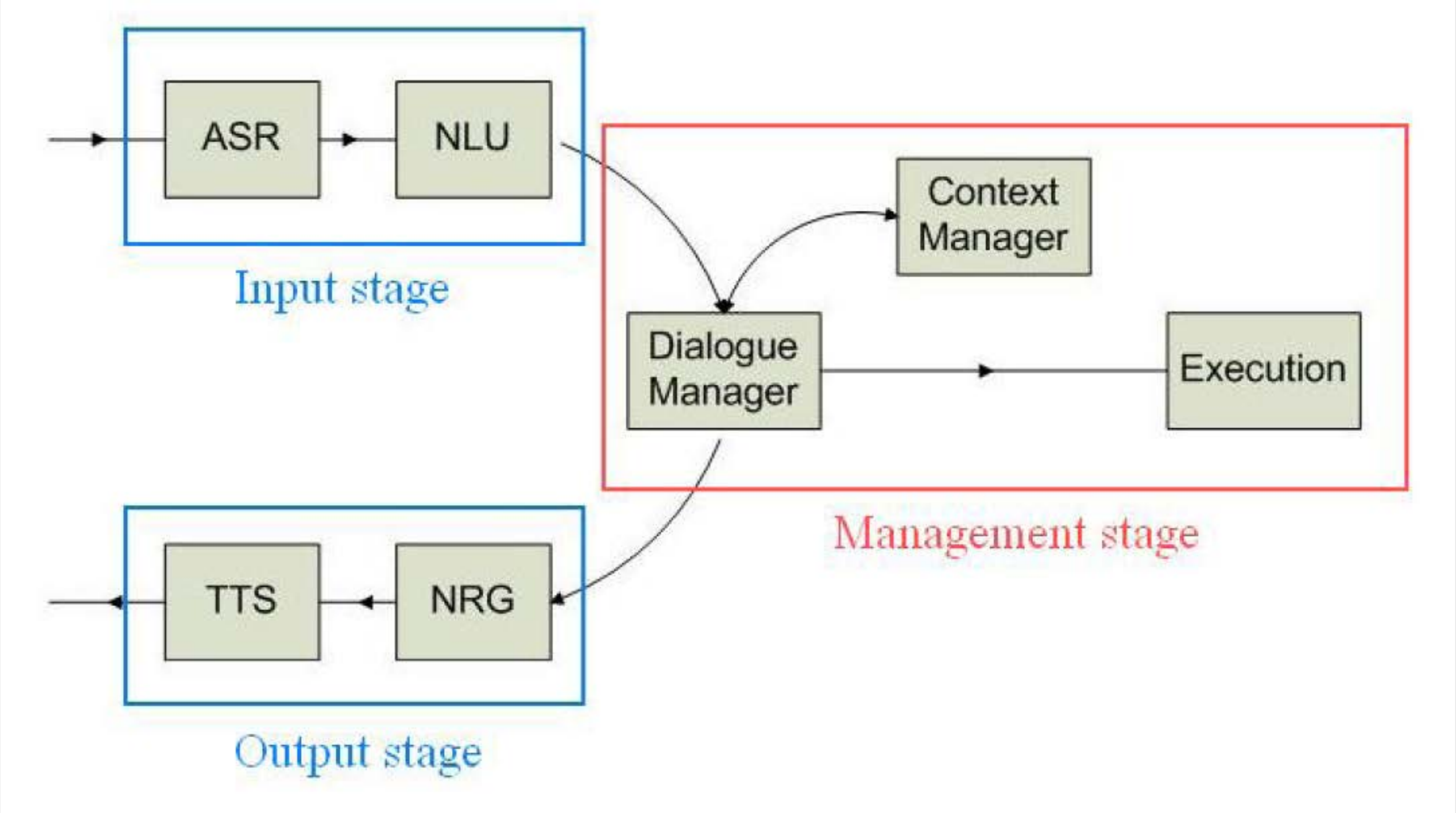
Dialog Management

Database, Information Servers

Answer generation (Text to Speech Synthesis, Multimodal Information)

Knowledge sources involved to carry out of a spoken dialog system

Spoken dialogue system modules



Input Stage

- Automatic Speech Recognition System.
  - Process input speech signal and returns the hypothesized transcription.
  - The returned sequence of words may contain errors.
  - Along with the transcription, confidence values can be delivered.
- Natural Language Understanding Module:
  - Extract the semantic content of the utterance
  - Usually, the output semantics are delivered as a set of dialogue concepts that model different aspects of the application domain.
  - The sequence of dialogue concepts may also contain errors and confidence values can be also delivered.

Management Stage

- Dialogue Manager.
  - Determines which actions the user wants to carry out as a function of the sequence of dialogue concepts and the dialogue context (history, user profile, ...)
  - If some information is missing, the user must be asked for fulfilling it.
  - Generates the semantics to communicate the corresponding message to the user.
- Dialogue Context:
  - Contains all the useful information to carry out the proposed actions: User info, past user utterances, ...
- Execution Module:
  - Carries out the actions proposed by the user and determined by the dialogue manager: Device control, Database queries, ...

Output Stage

- Natural Language Generation:
  - Builds a lexically and syntactically appropriate sentence that conveys the concepts that must be presented to the user.
- Text-to-Speech Synthesizer:
  - Generates an acoustic signal that synthetizes the message generated in the previous module.

Other Mudules

- Language ID:
  - In multilingual environments, it can be useful to identify automatically the language used by the user.
- Speaker ID:
  - To determine the identity of the user in order to customize or personalized the application.
- ...

Dialogue Management

- Dialogue Initiative:
  - System Initiative:
    - The Most Basic ones. The system ask the user at each time to execute a certain action, or explicitly confirm that the info provided is correct
    - The user freedom is reduced and they have an important lack of naturalness.
  - User Initiative:
    - The user decides how to carry on the dialogue without a predefined structure.
    - Linguistic constructions are more complex and the system must be more flexible and advanced.
  - Mixed Initiative:
    - Tradeoff between both approaches. The initiative belongs to the user or the system depending on the dialogue situation

Dialogue Management

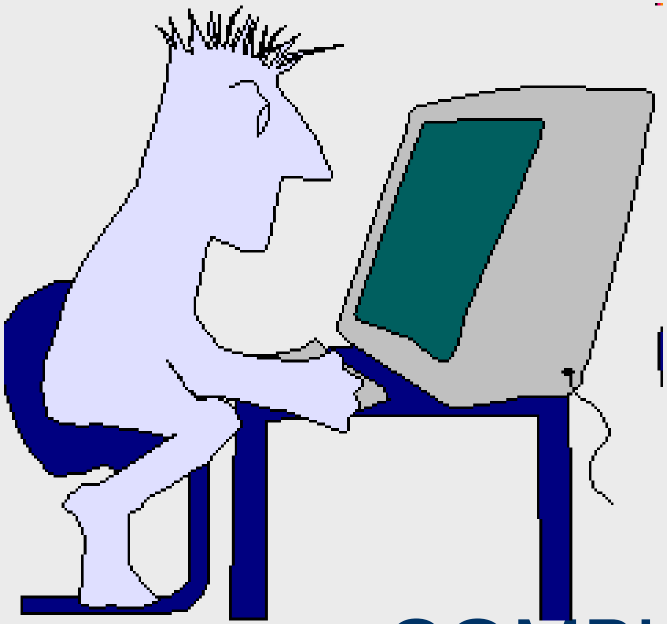
- Techniques for managing the dialogue:
  - Finite State Machines (FSM):
    - Set of states in which the system performs specific actions.
    - The dialogue flow goes from one state to the next one according to a predefined order:
      - High level of control by the designer. Low flexibility and naturalness.
  - Frames:
    - Each action is represented by a frame:
      - Data structure with several fields that must be fulfilled prior to carry out the corresponding action.
    - More flexible and natural. The user can fulfilled more than one field per turn.

Dialogue Management

- Techniques for managing the dialogue:
  - Plan-Based:
    - The Agent set a goal to accomplish
    - The goal is reached through several dialog acts.
  - Stochastic:
    - Statistical modeling of the dialogue from a training database (with lots of dialogue examples)
      - Supervised Learning
      - Reinforcement Learning
    - More flexibility and naturalness.

Applications

- Information Retrieval, Services and Transactions:
  - Search and Retrieval of information.
    - Movies, flights, trains, restaurants,...
  - Control of Devices and Applications
    - Robots, multimedia centers, ...
- Problem Solving:
  - Technical support (Cable TV, Modems, Logistics, ...)
- Education:
  - CALL (Computer-Aided Language Learning)
  - ITS (Intelligent Tutoring System)
- Games and Entertainment
  - People with special needs, rehabilitation (serious games).

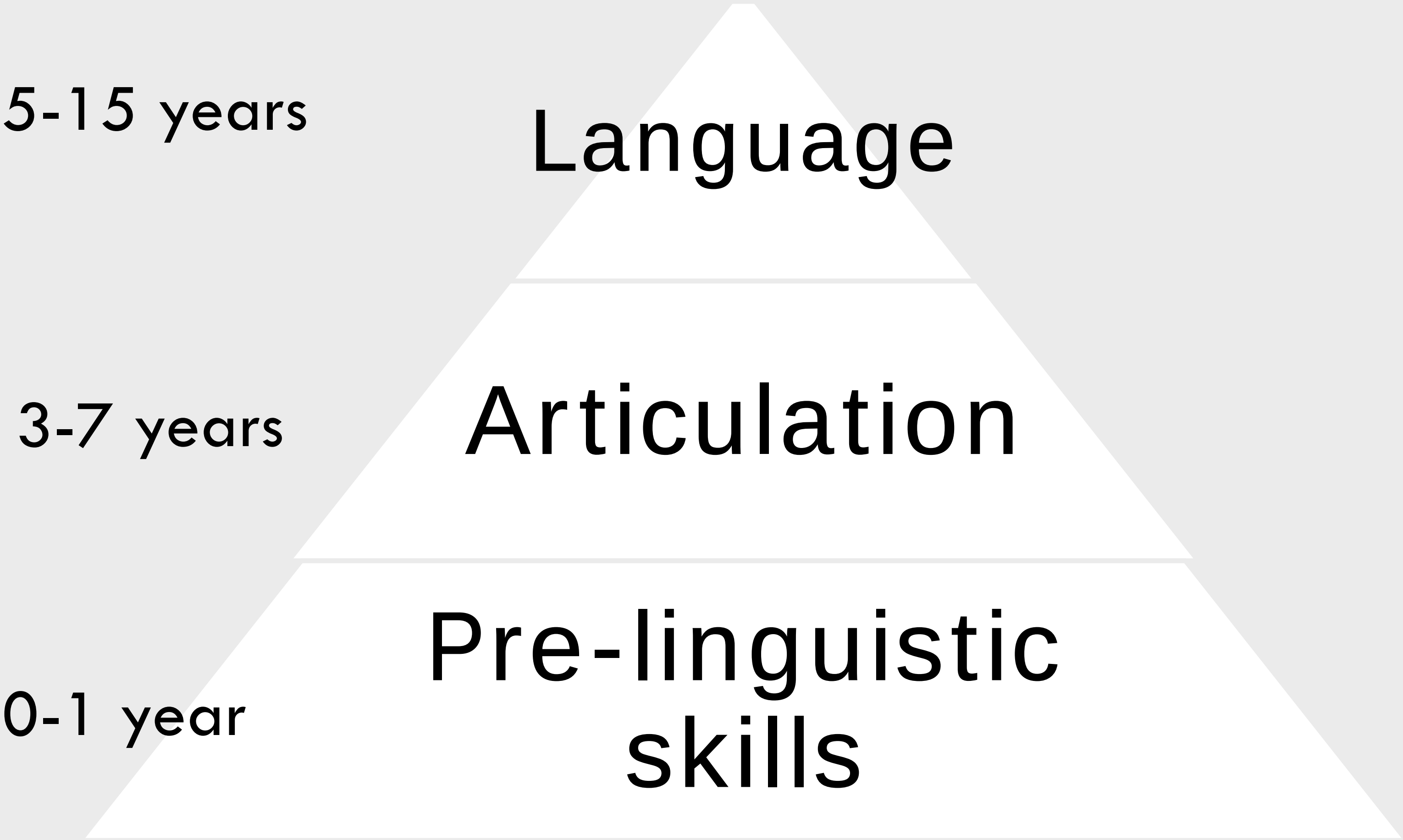


COMPUTER-AIDED LANGUAGE LEARNING AND REHABILITATION: PRELINGUISTIC SKILLS

CALL Systems

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

Language Learning Process



5-15 years

Language

3-7 years

Articulation

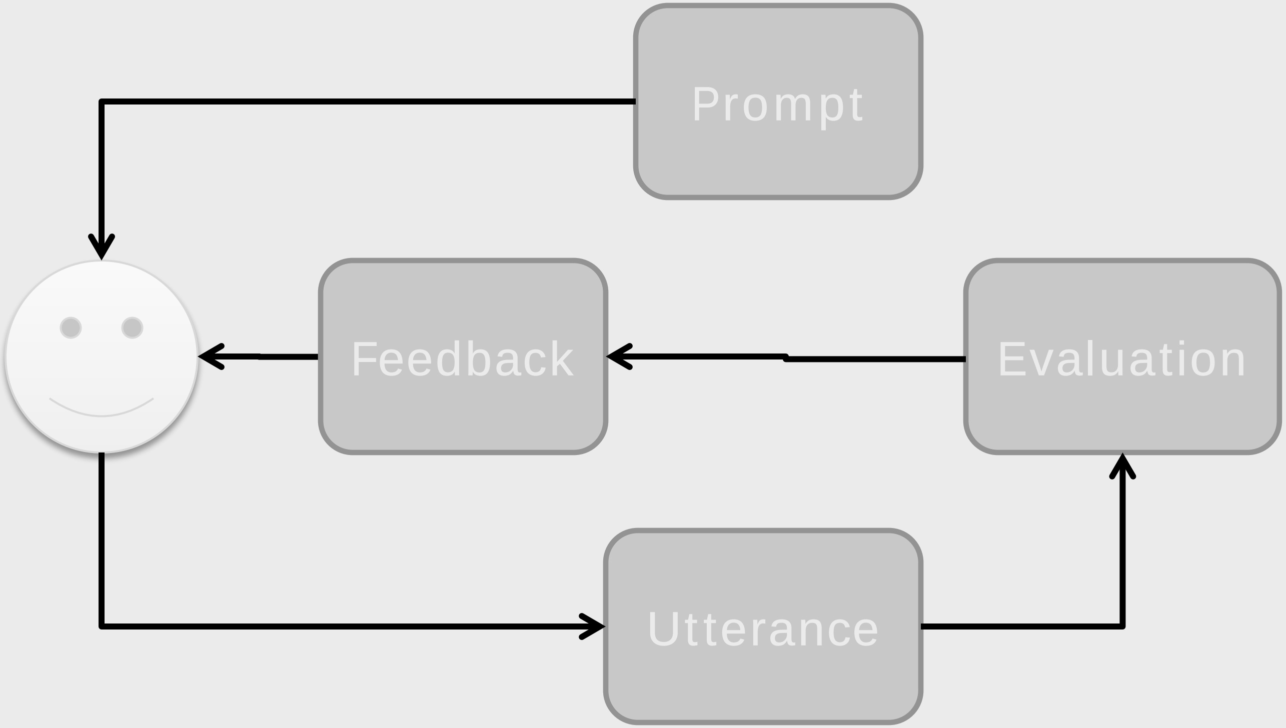
0-1 year

Pre-linguistic skills

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



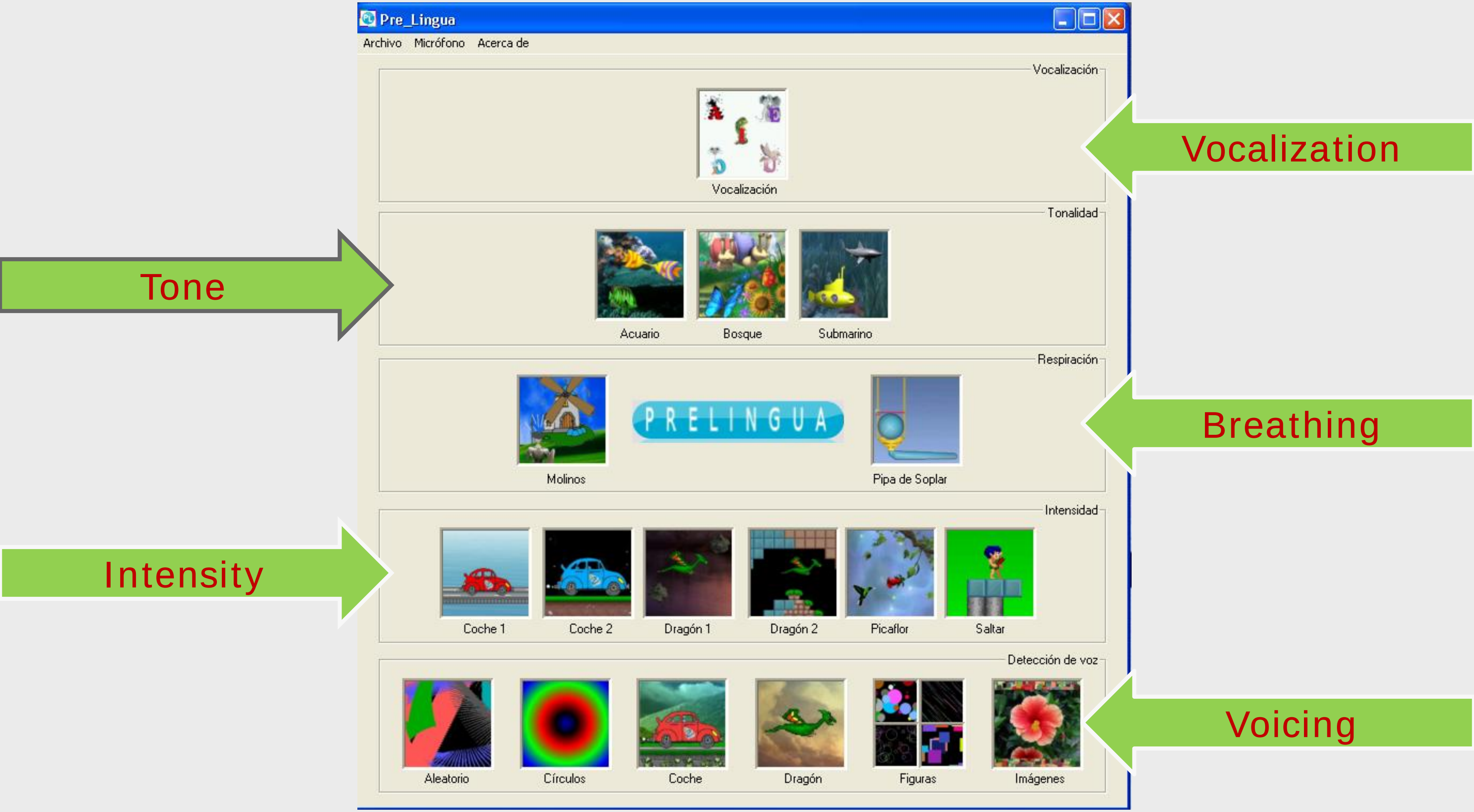
Pre-linguistic skills

- 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

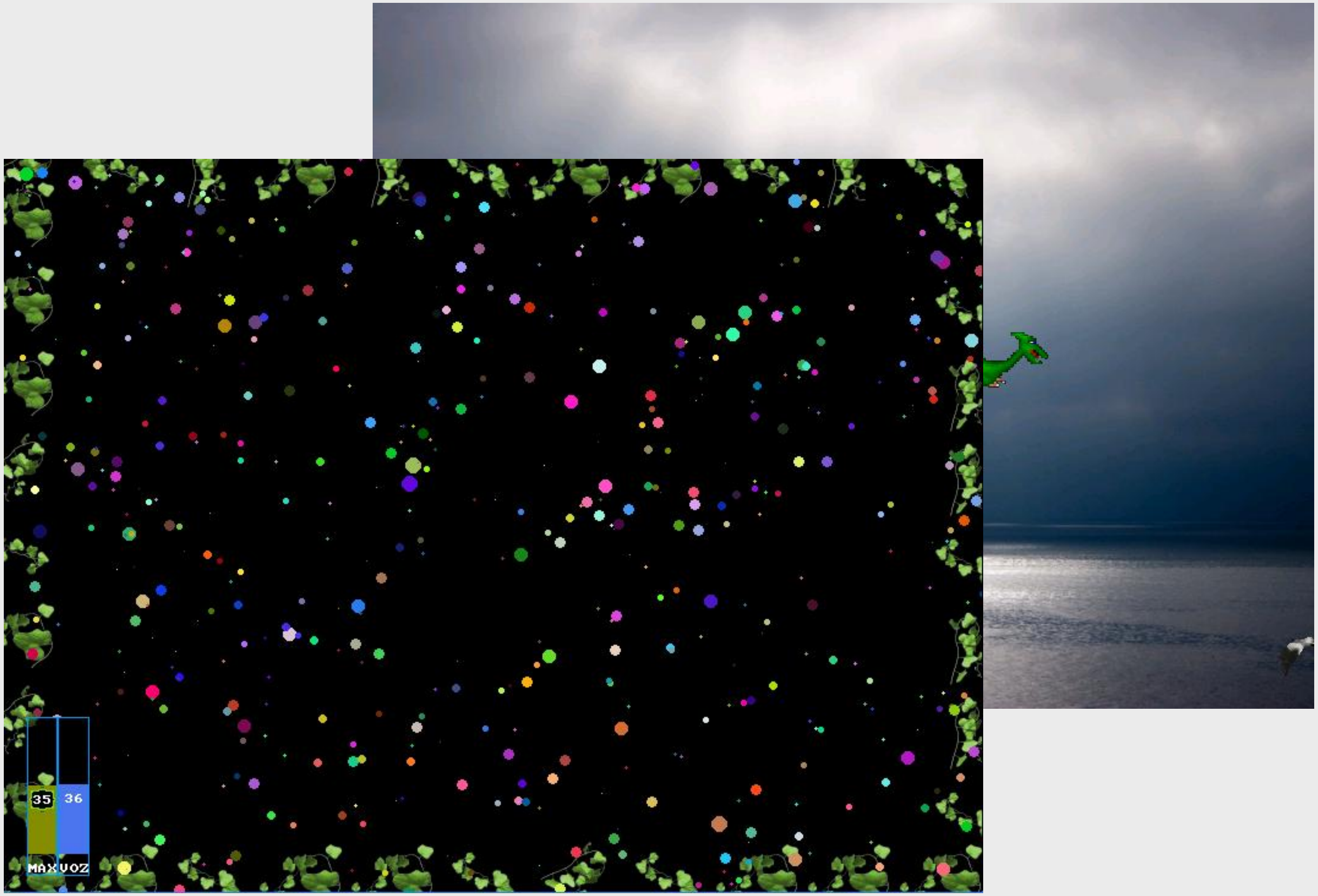
Breathing

Intensity

Voicing

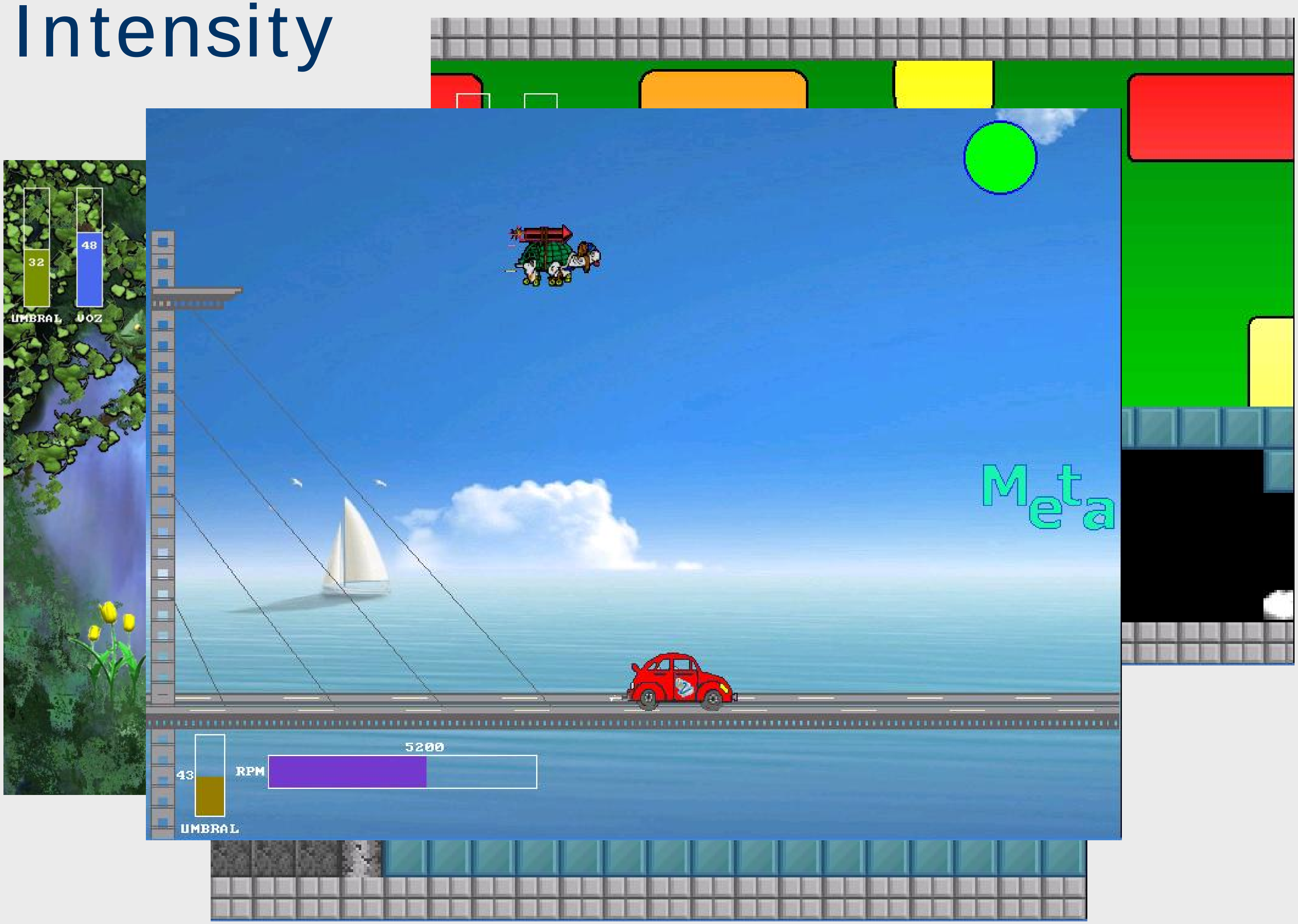
Pre-linguistic skills

- Voicing



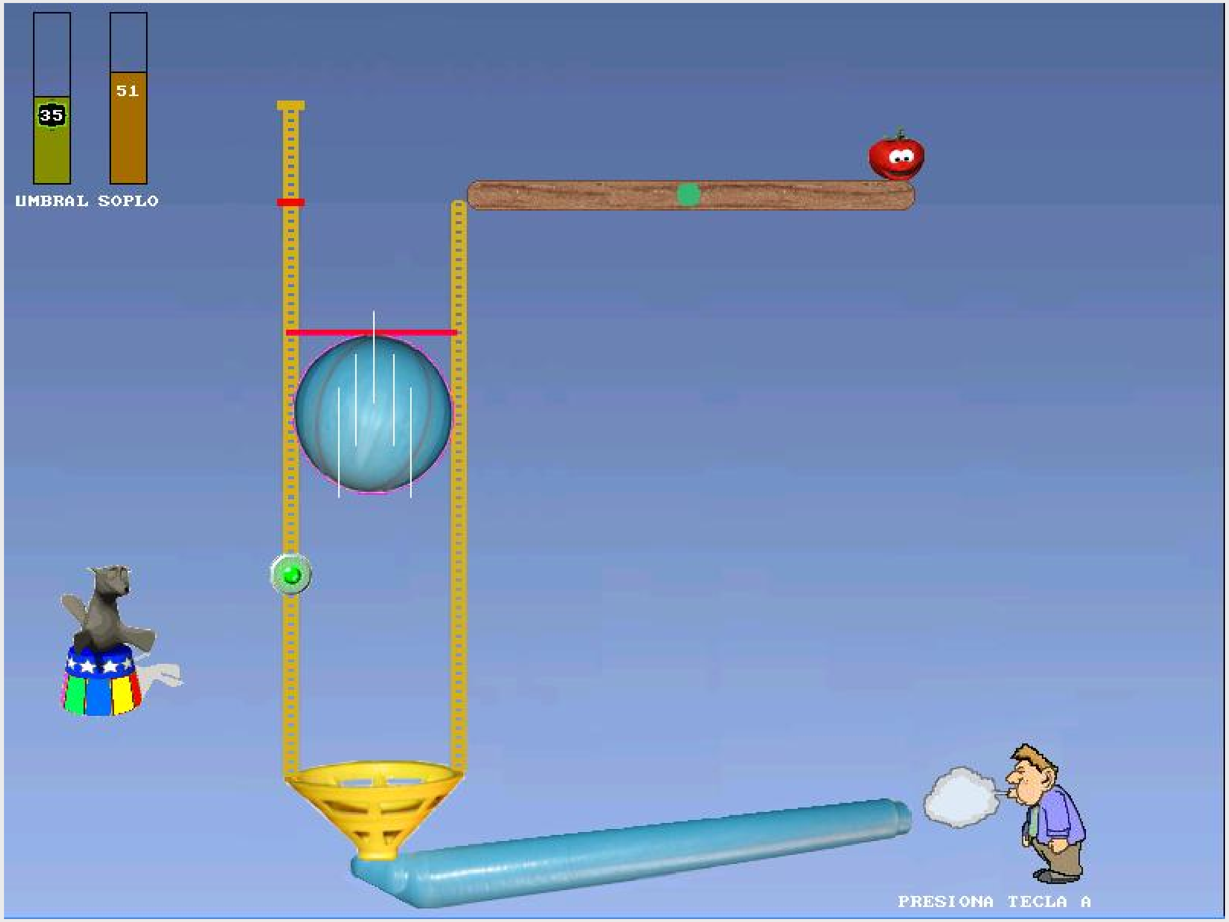
Pre-linguistic skills

- Intensity



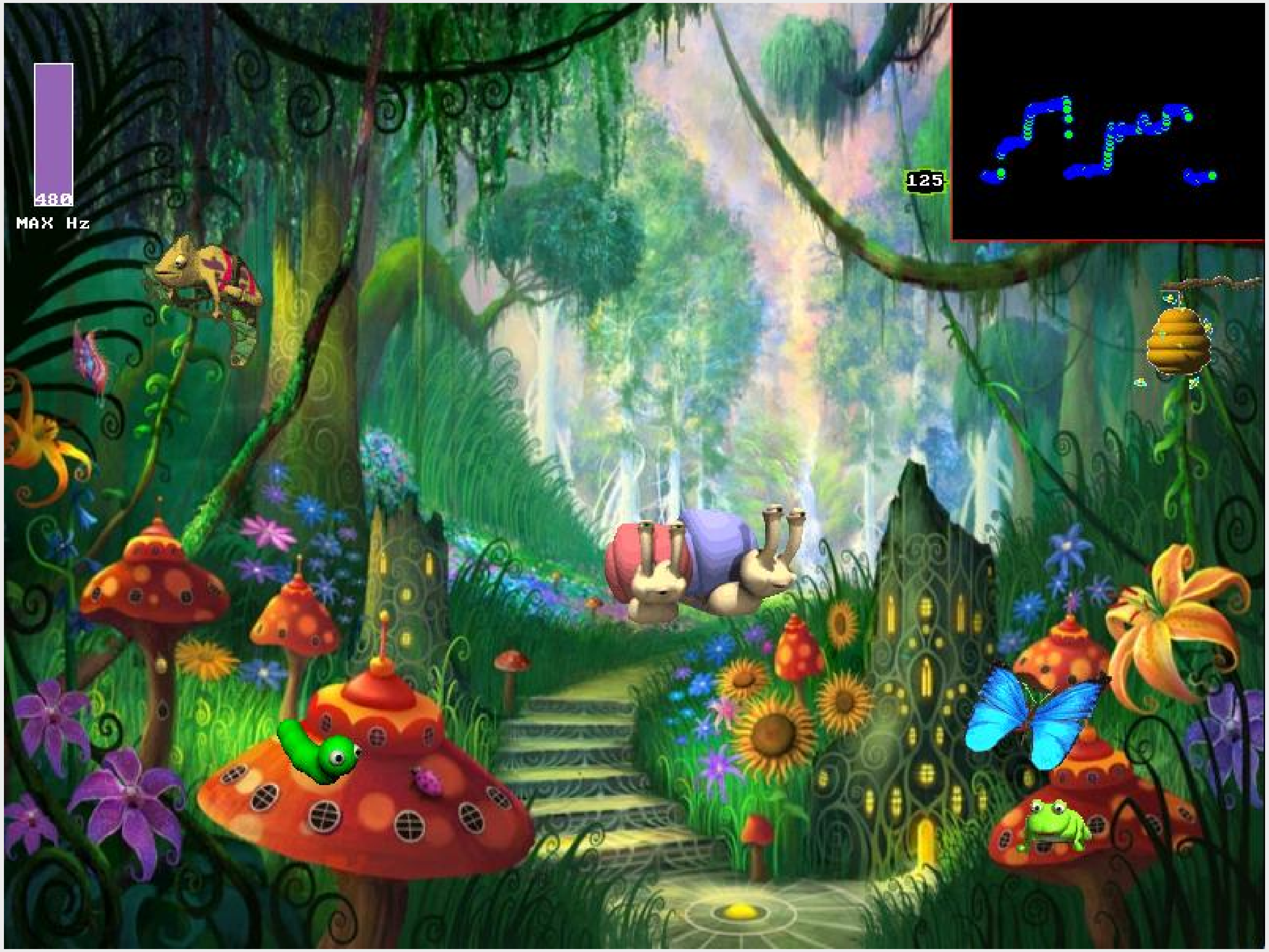
Pre-linguistic skills

- Breathe



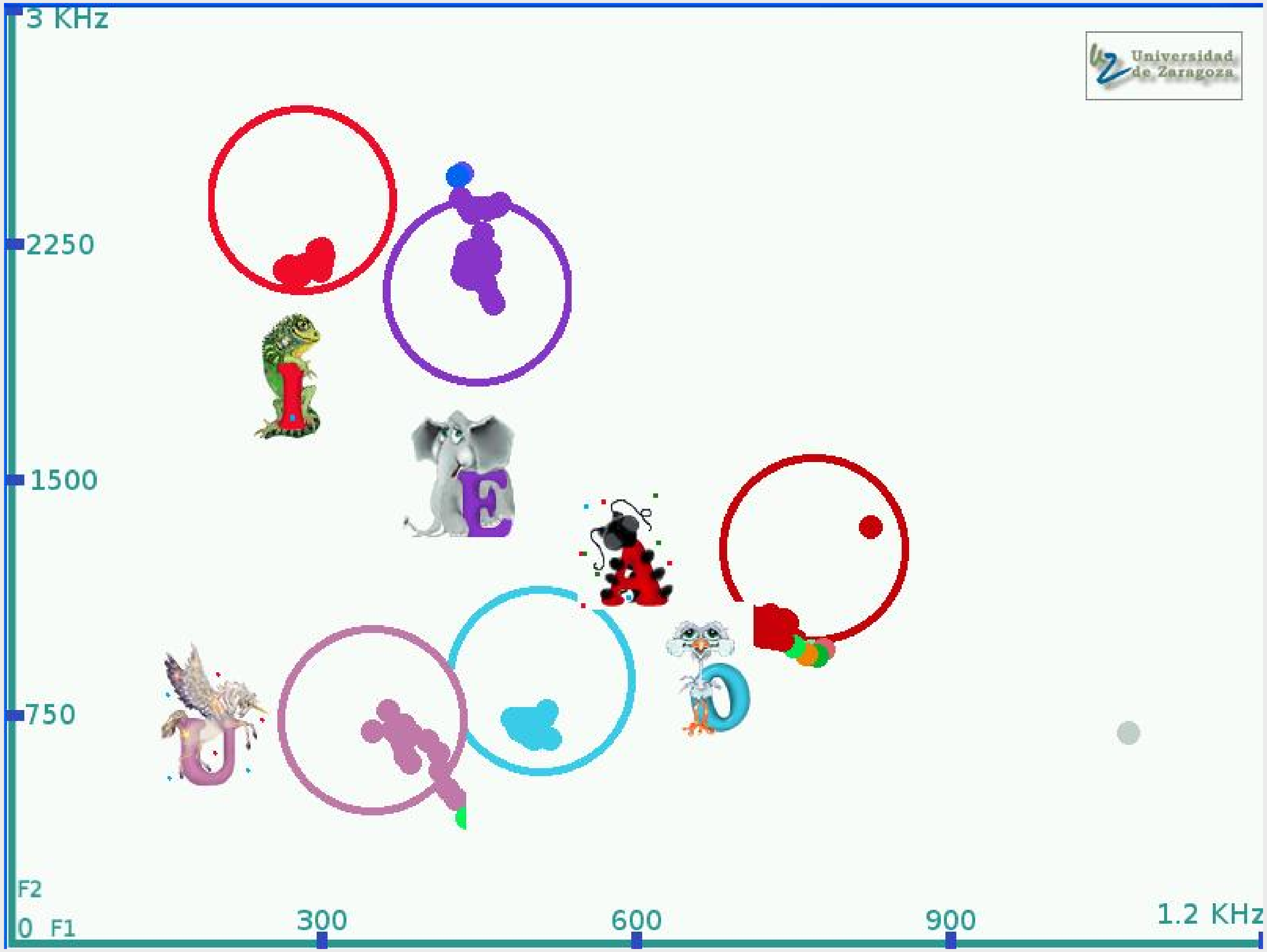
Pre-linguistic skills

- Tone



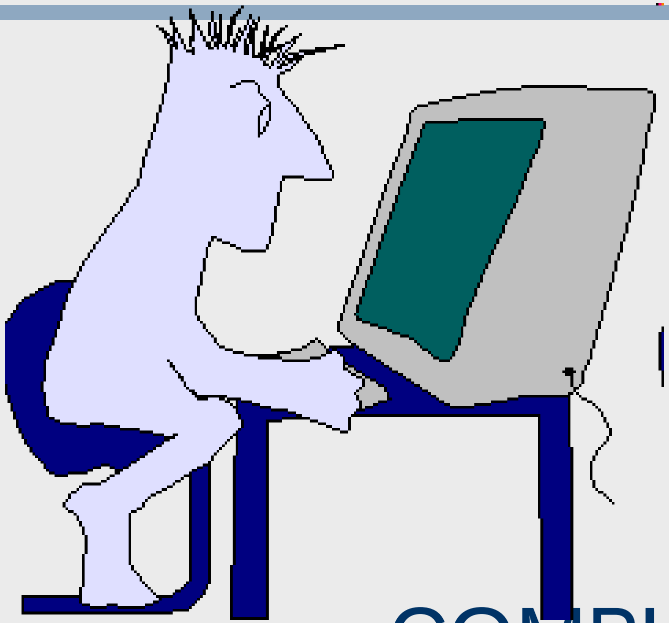
Pre-linguistic skills

- Vocalization



Examples

- Now, practice



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

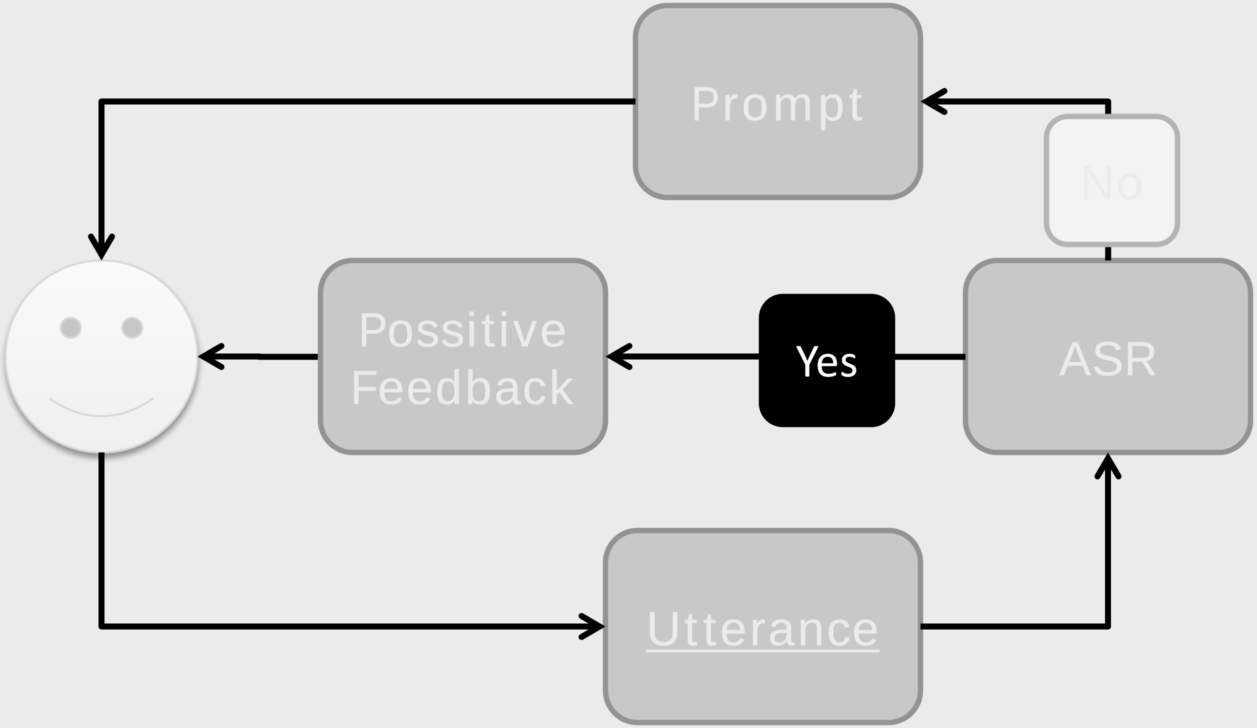
Articulatory skills

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

- Word or phoneme based feedback

Evaluation - Alternatives

- Whole word evaluation - ASR

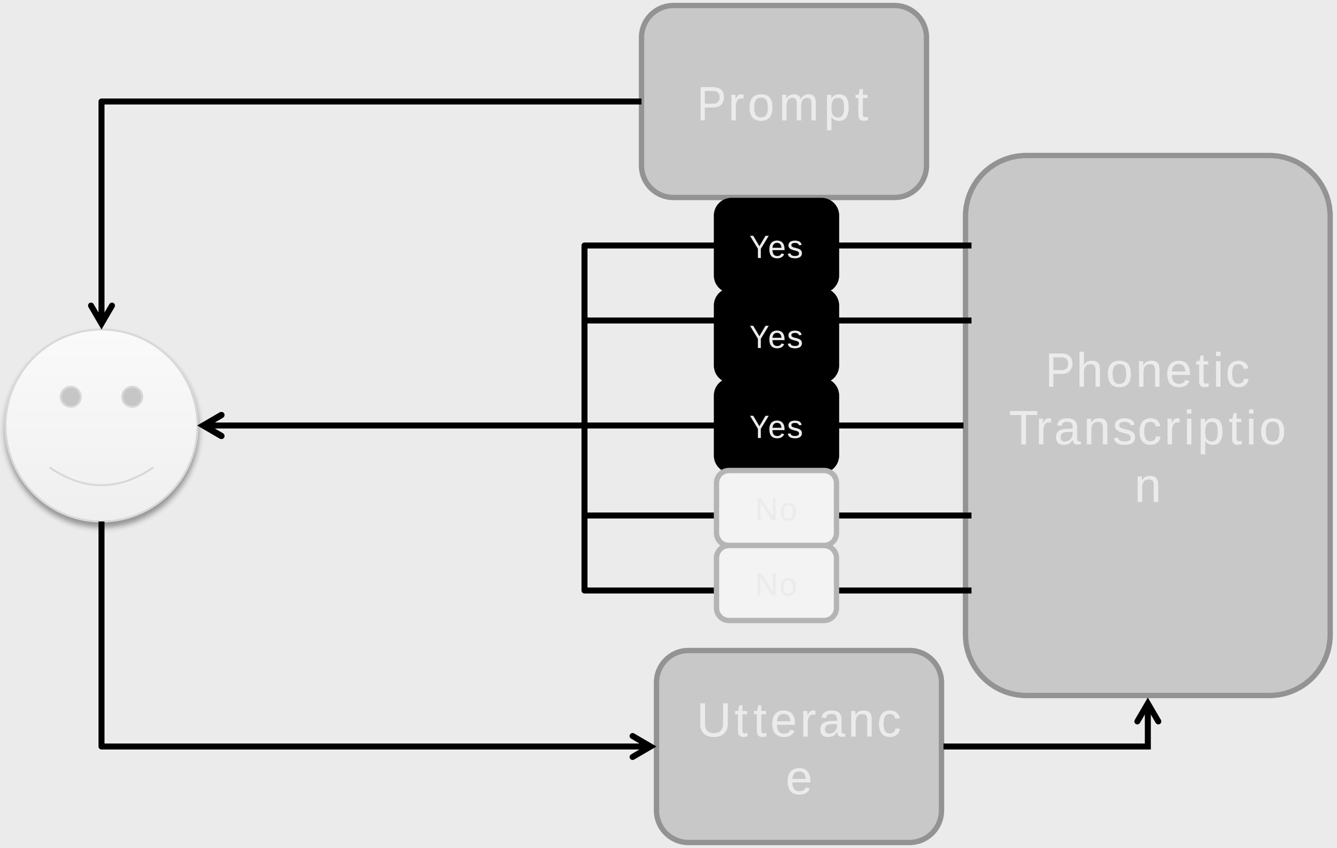


Evaluation - Alternatives

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

Evaluation - Alternatives

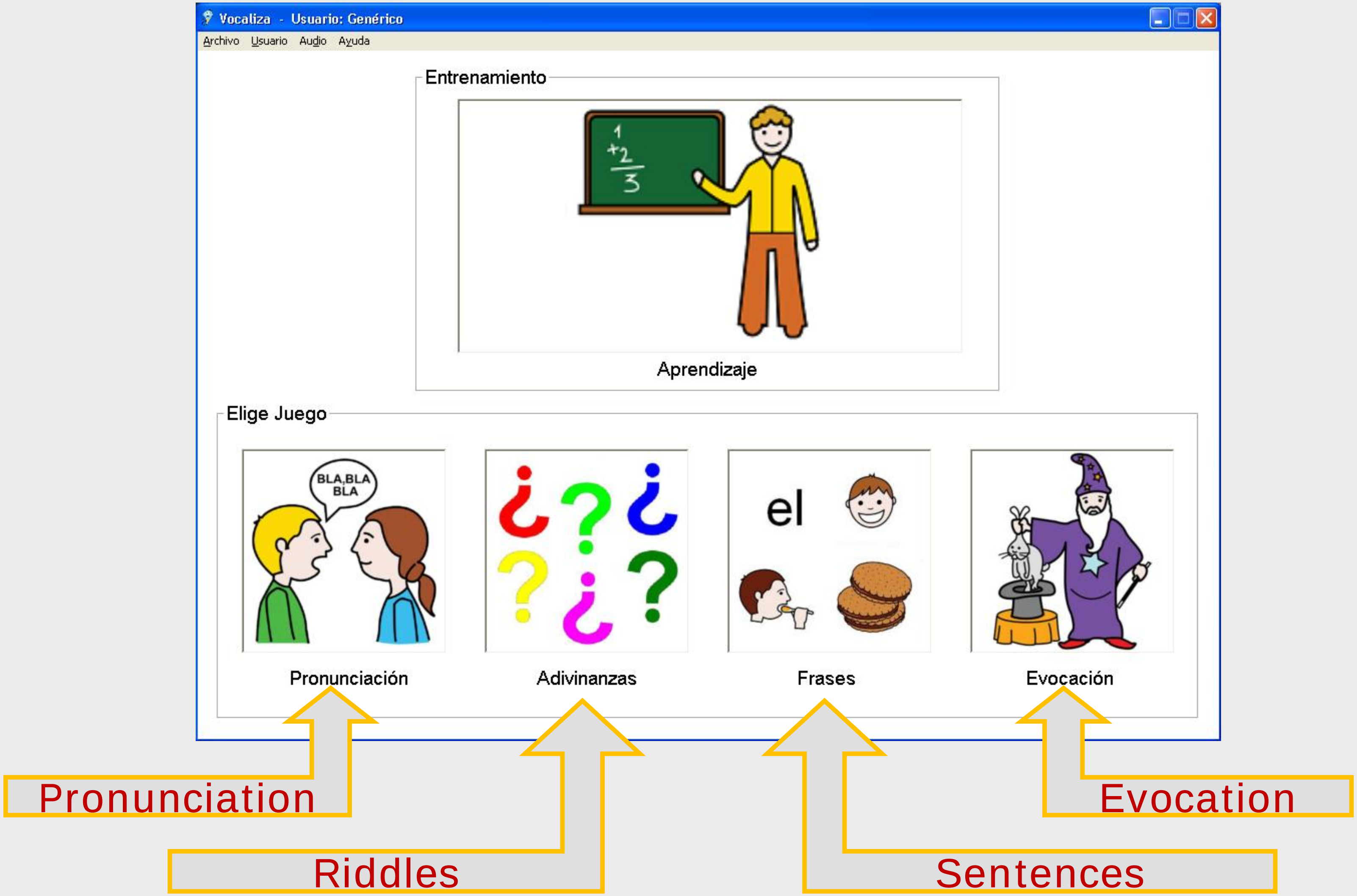
- Phoneme evaluation



Evaluation - Alternatives

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

Articulatory skills



Pronunciation

Riddles

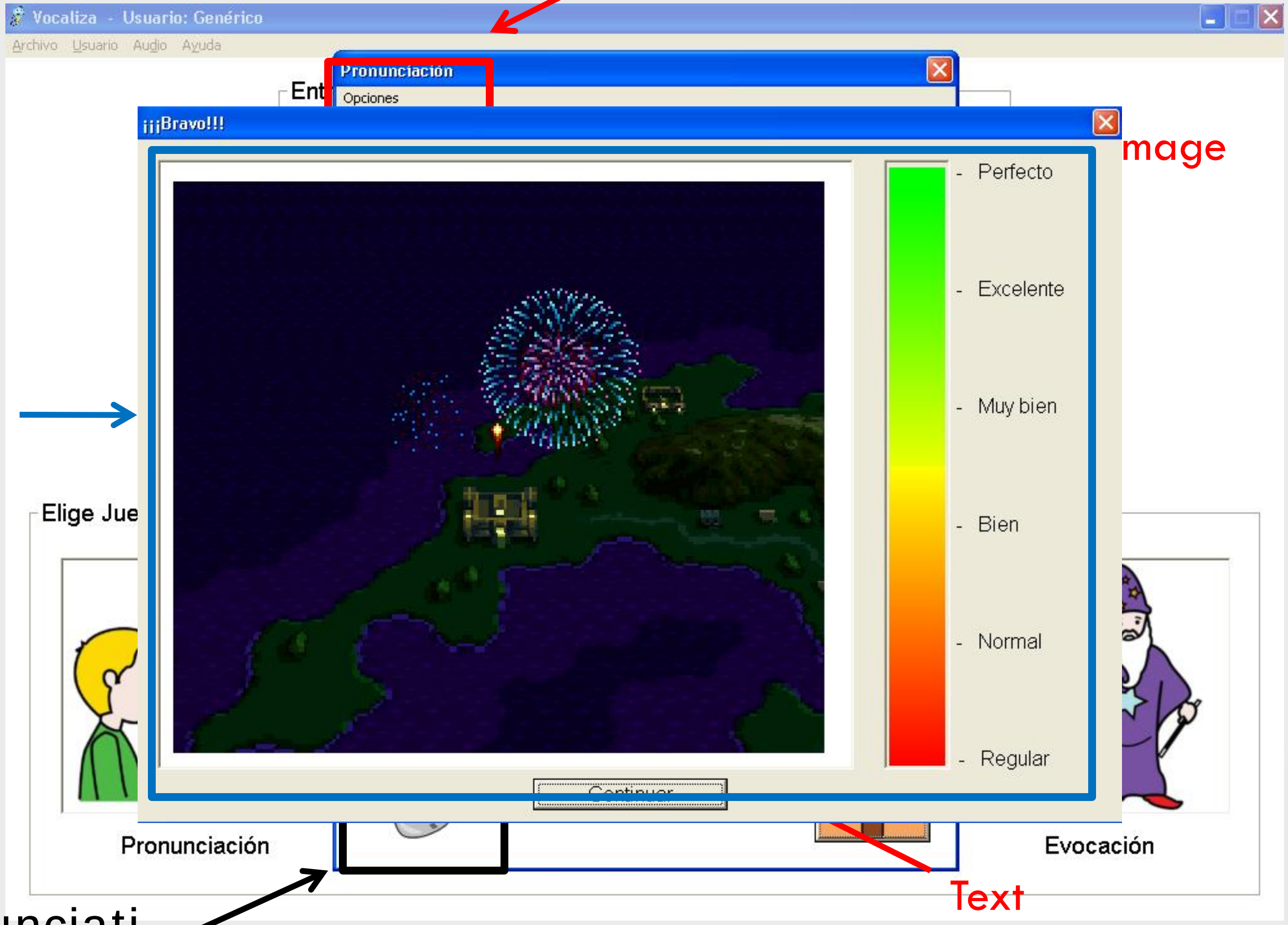
Sentences

Evocation

Articulatory skills

- Pronunciation

Audio



mage

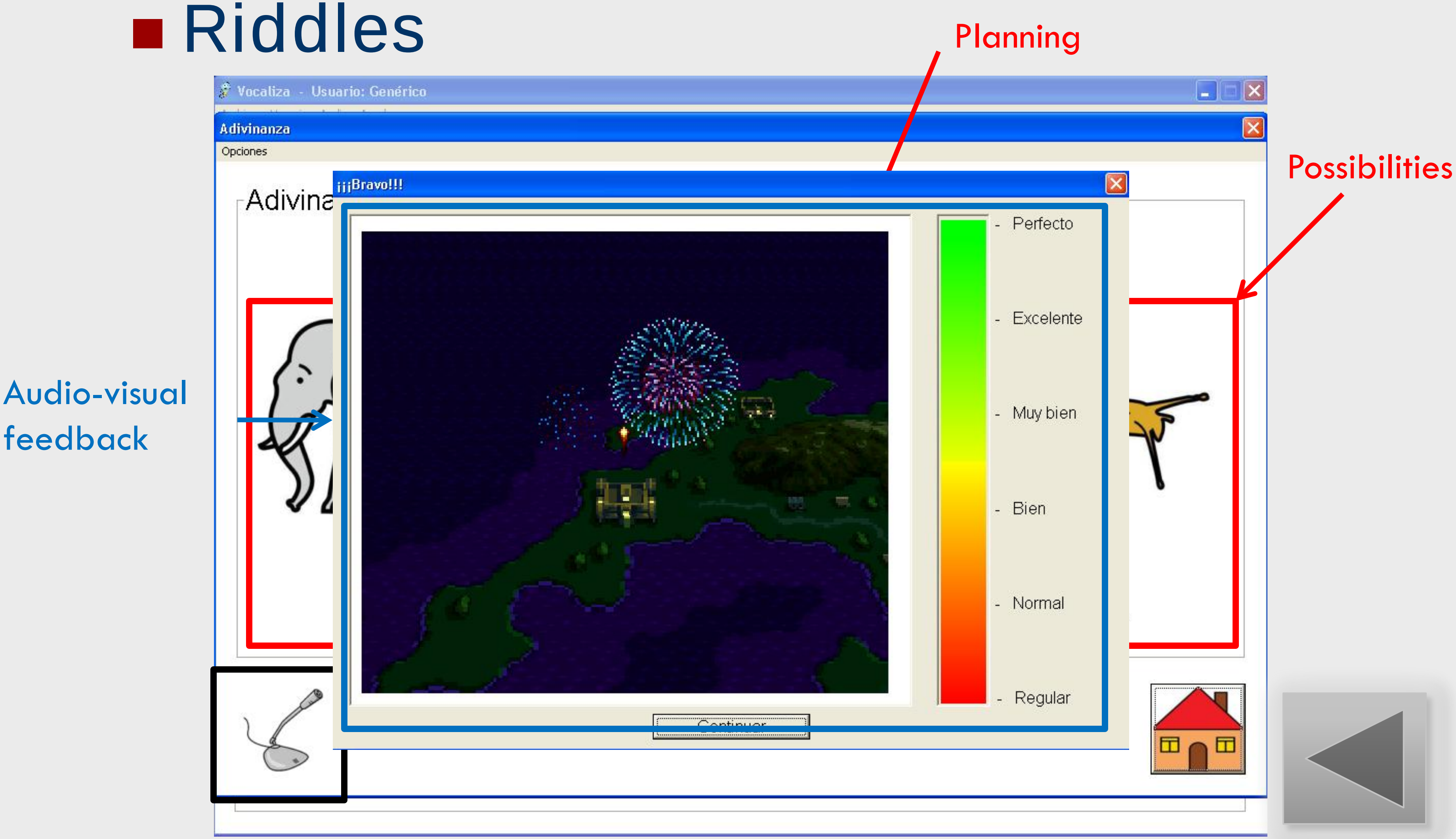
Audio-visual feedback

Text

Pronunciati
on

Articulatory skills

- Riddles



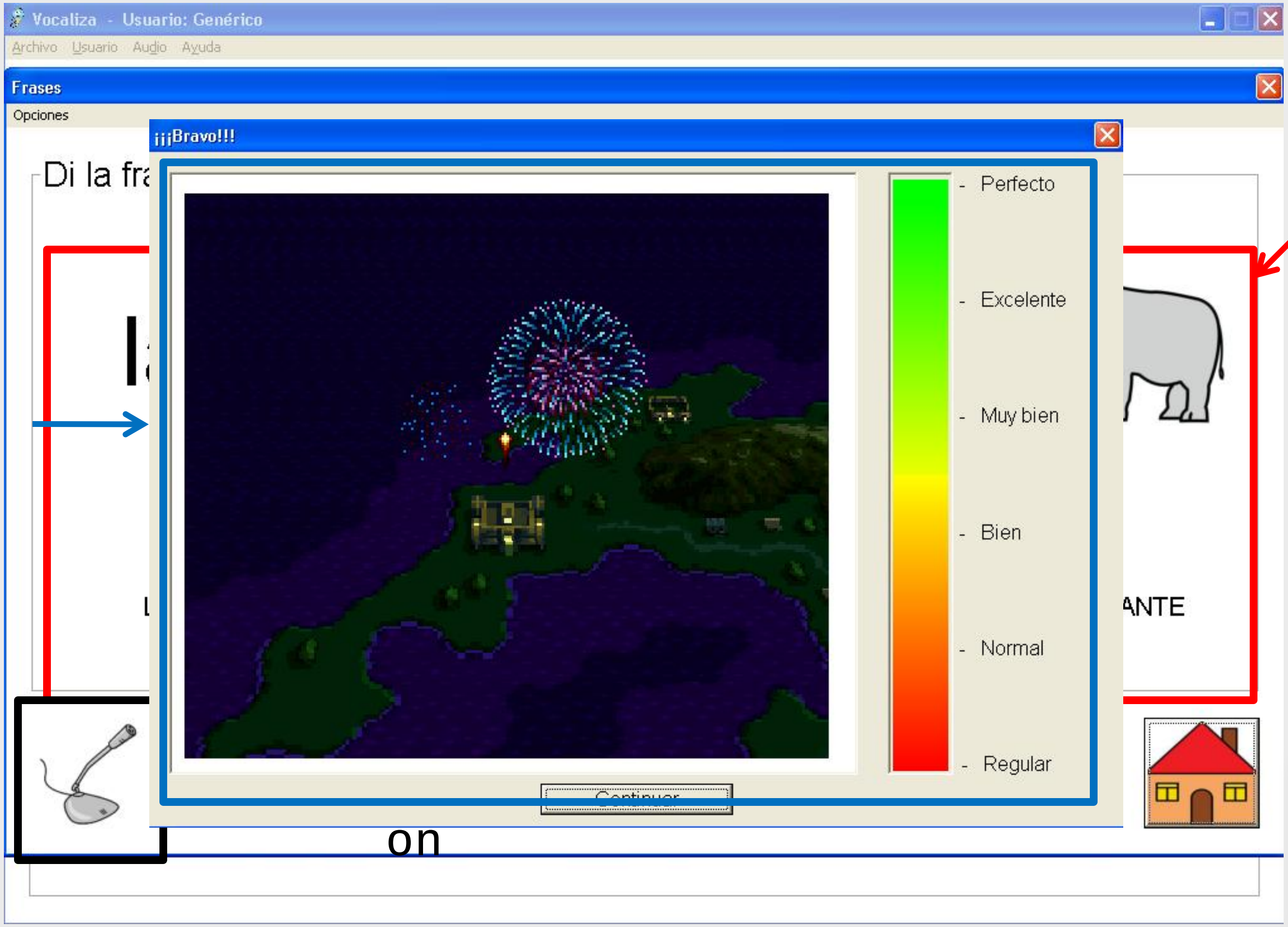
Planning

Possibilities

Audio-visual feedback

Articulatory skills

- Sentences

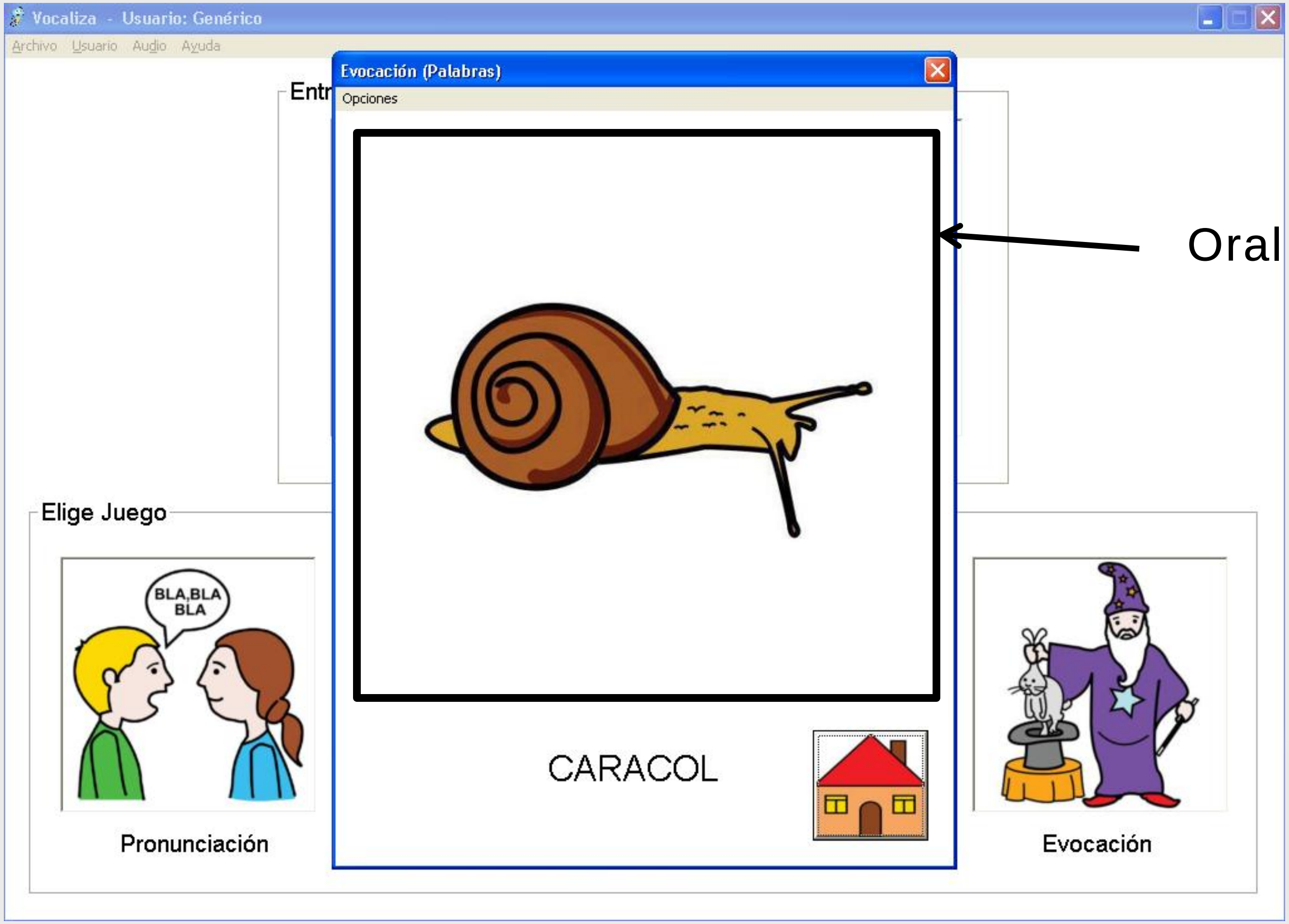


Sentences

Audio-visual feedback

Articulatory skills

- Evocation



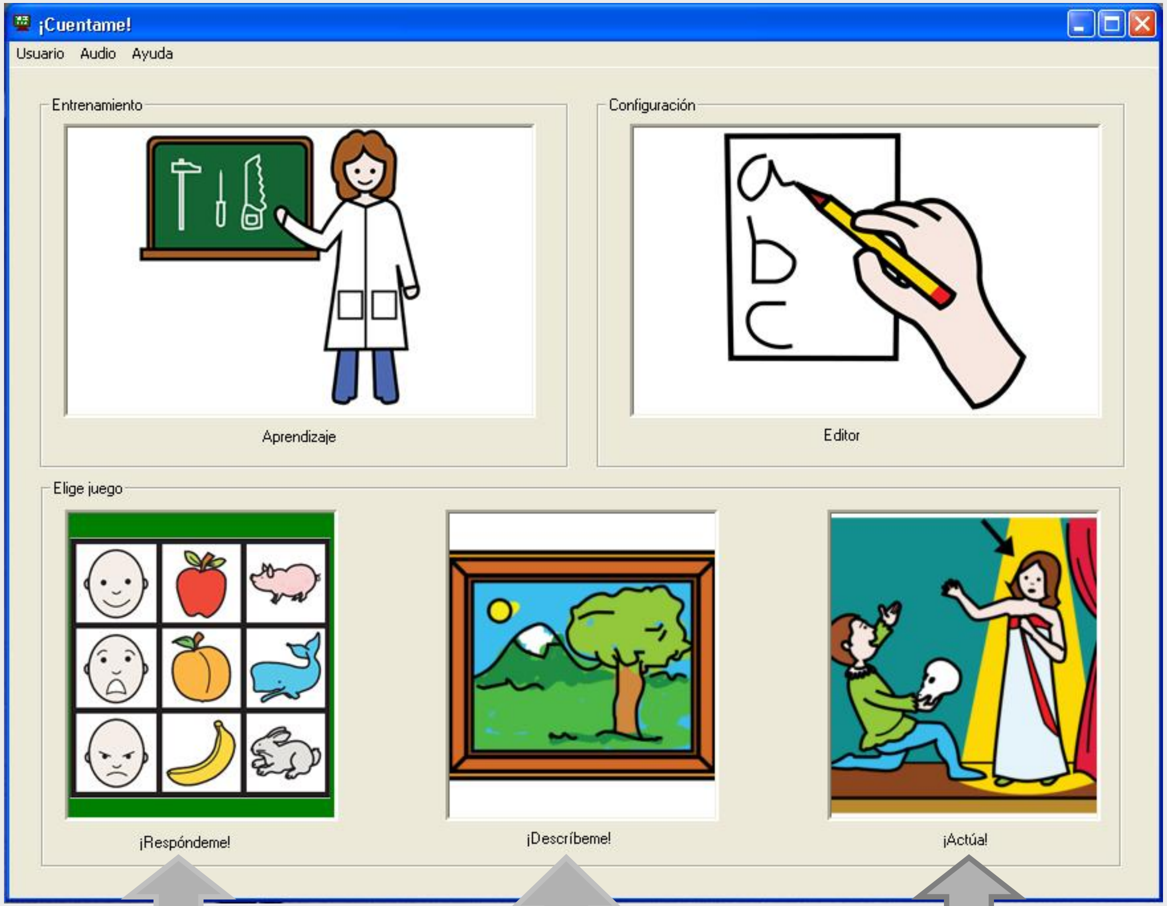
Oral input

Language

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

- Creation of sceneries to be solved by speech

Language



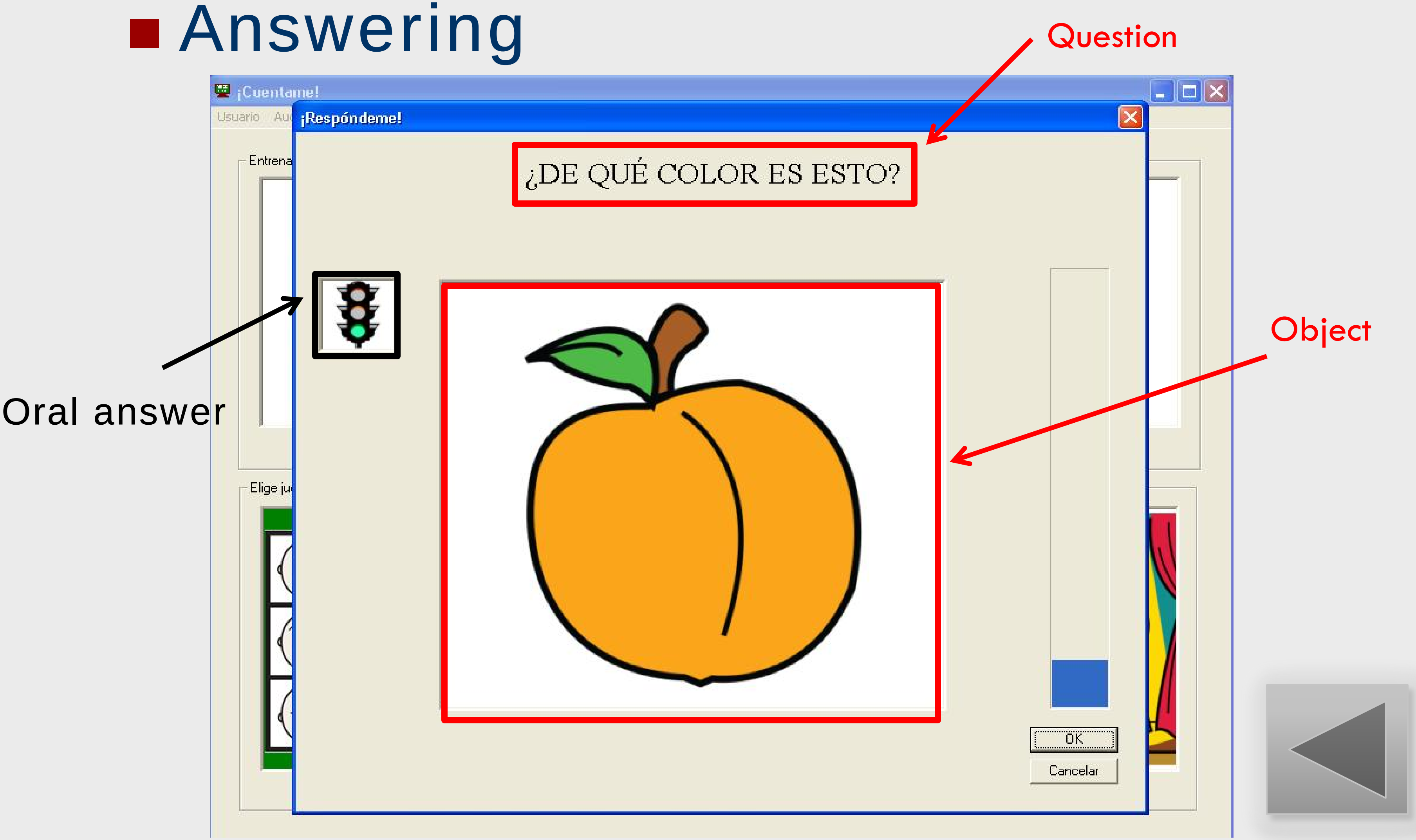
Answering

Description

Acting

Language

- Answering



Question

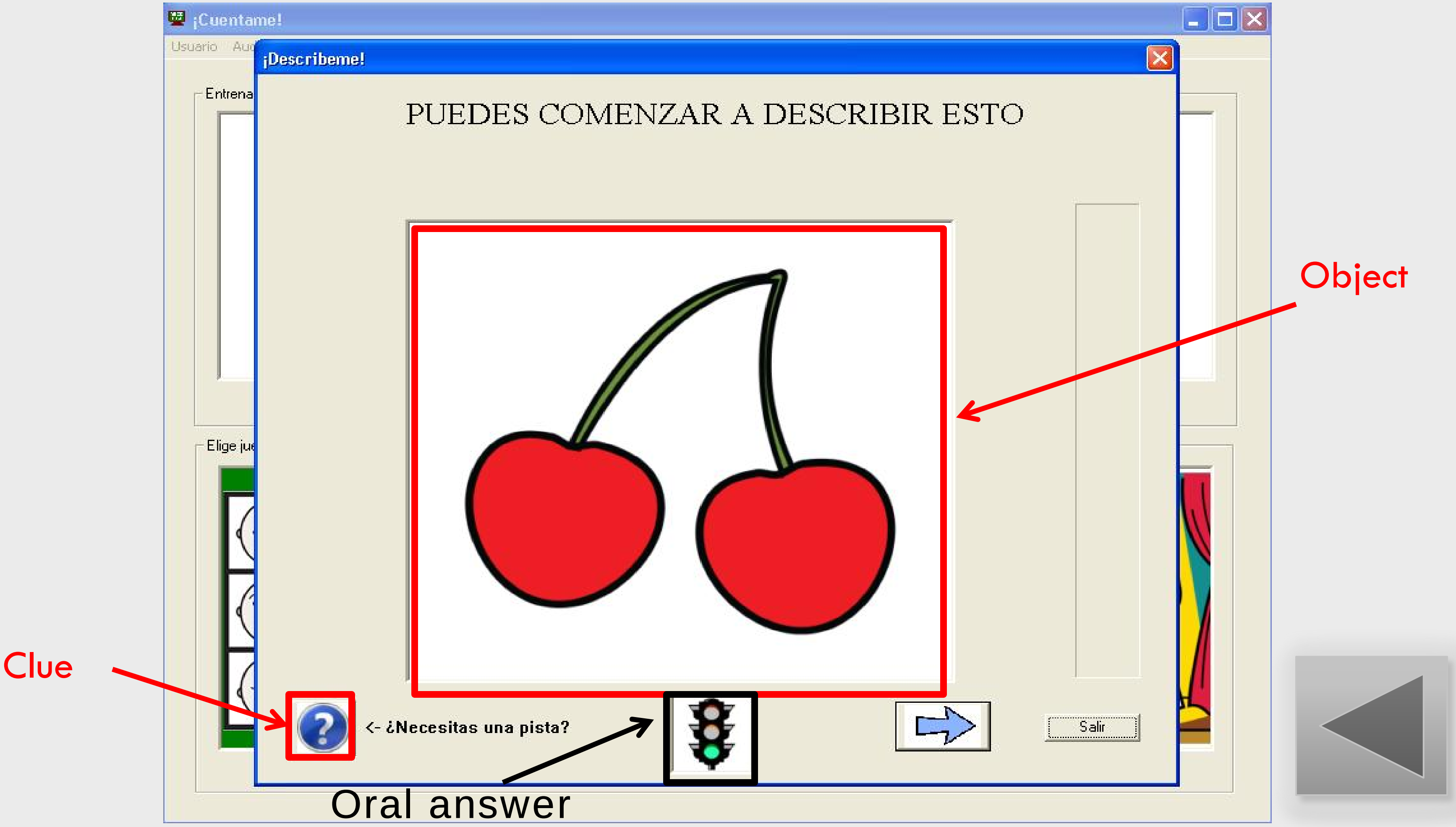
Object

Oral answer

¿DE QUÉ COLOR ES ESTO?

Language

- Description

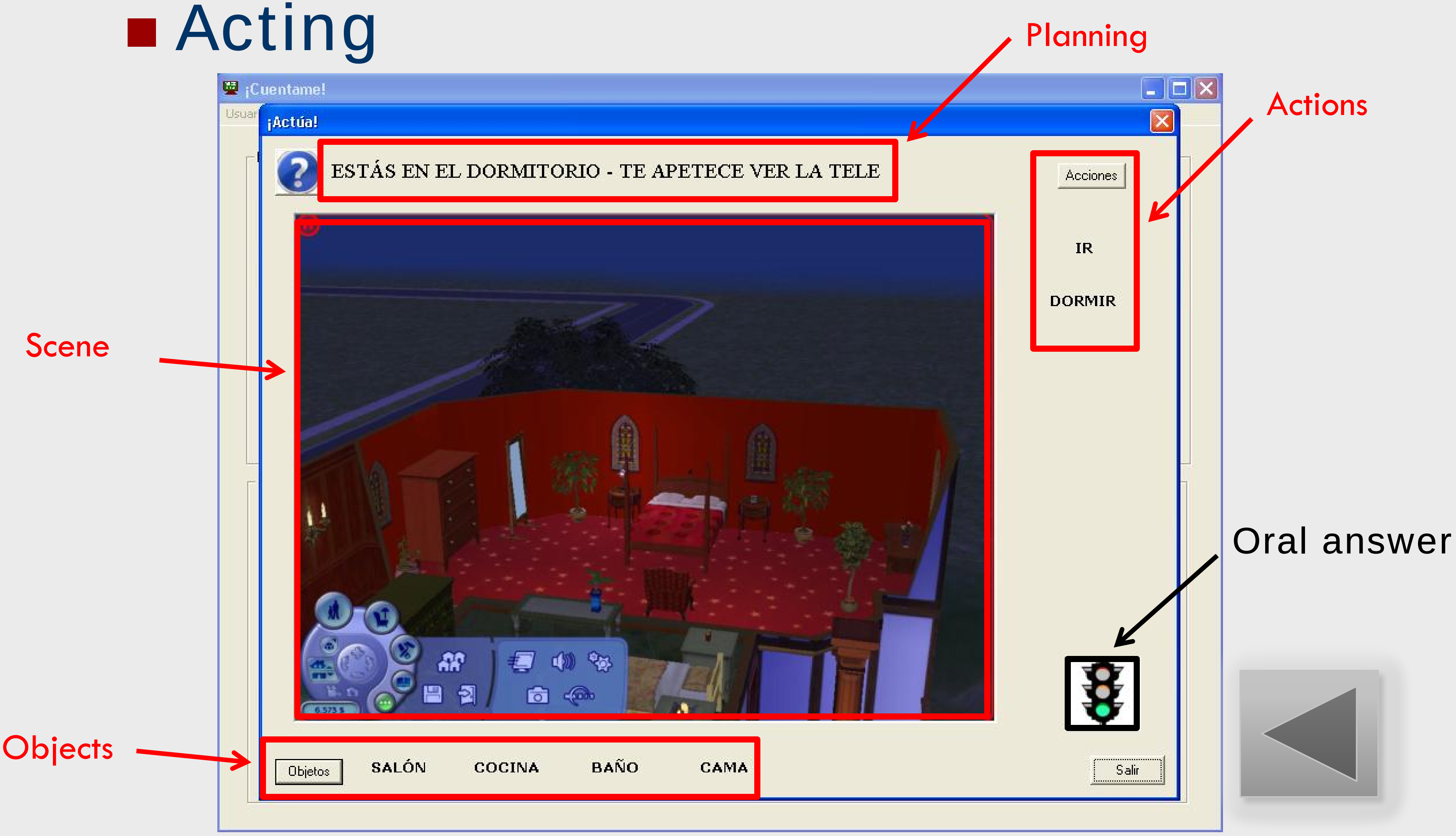


PUEDES COMENZAR A DESCRIBIR ESTO

<- ¿Necesitas una pista?

Language

- Acting



ESTÁS EN EL DORMITORIO - TE APETECE VER LA TELE

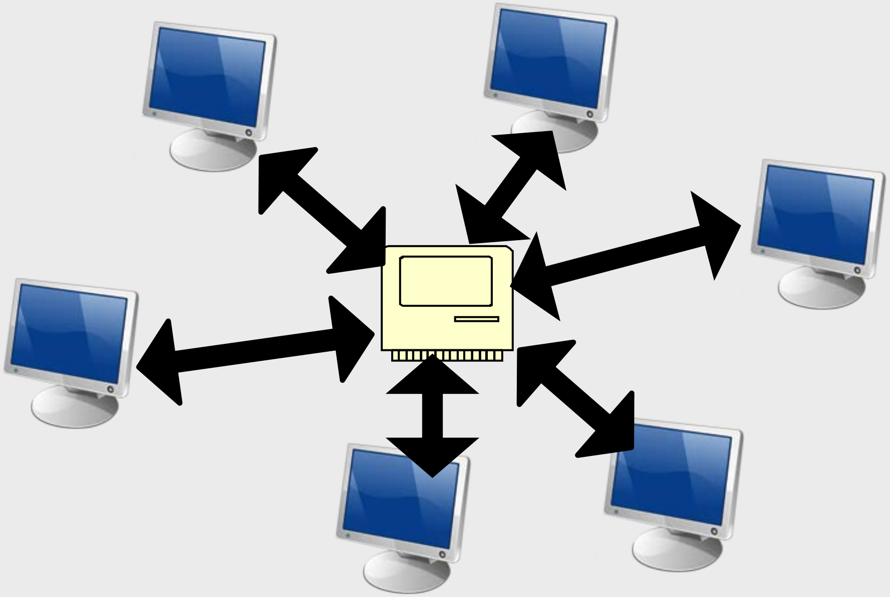
Planning

Actions

Scene

Objects

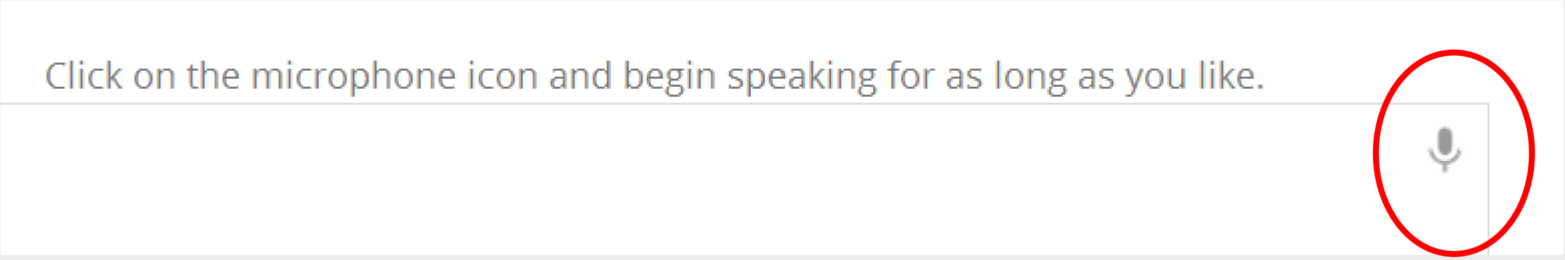
Oral answer



Using voice to drive the web

Overview

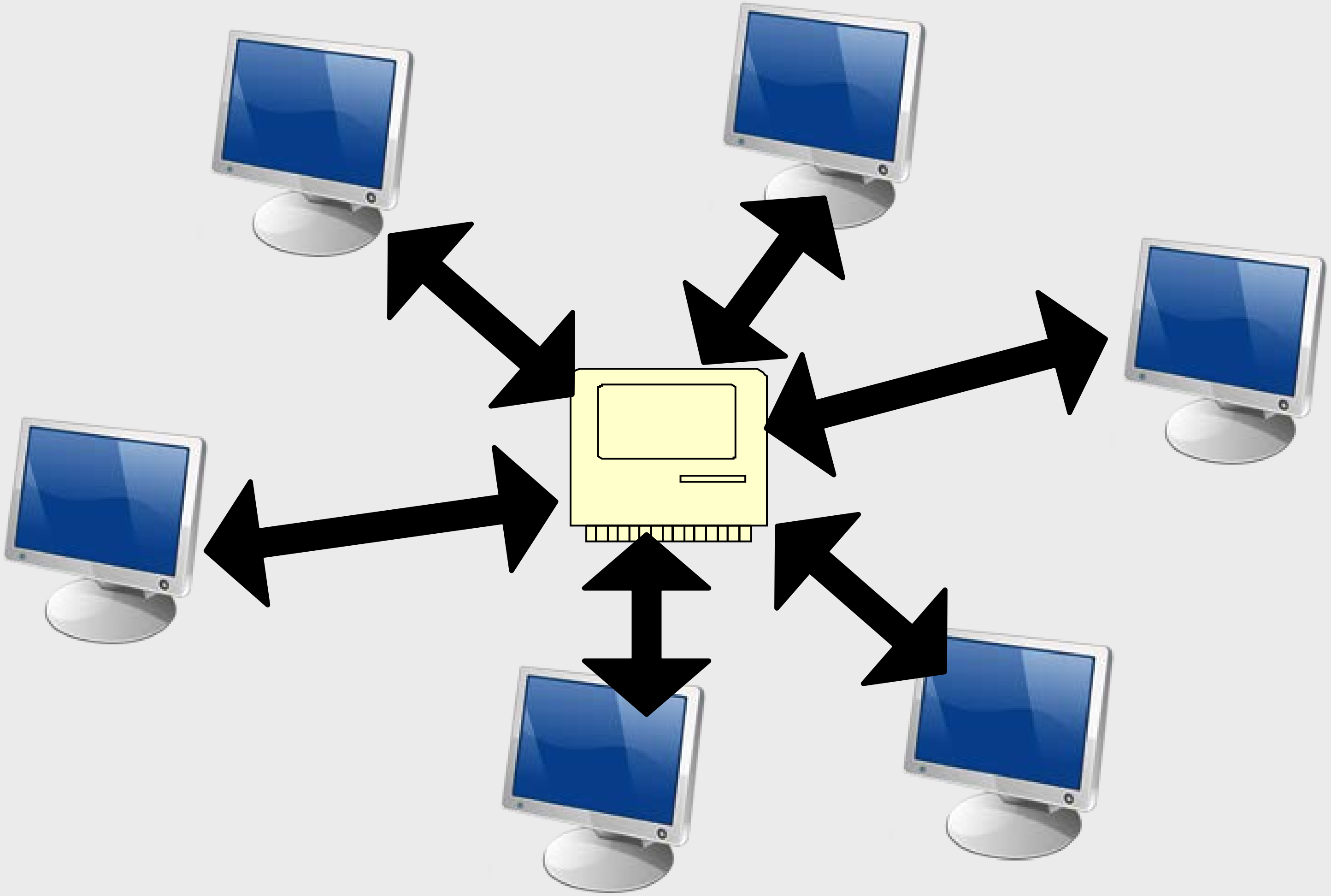
- Distributed frameworks
- Web Speech API
- Google implementation



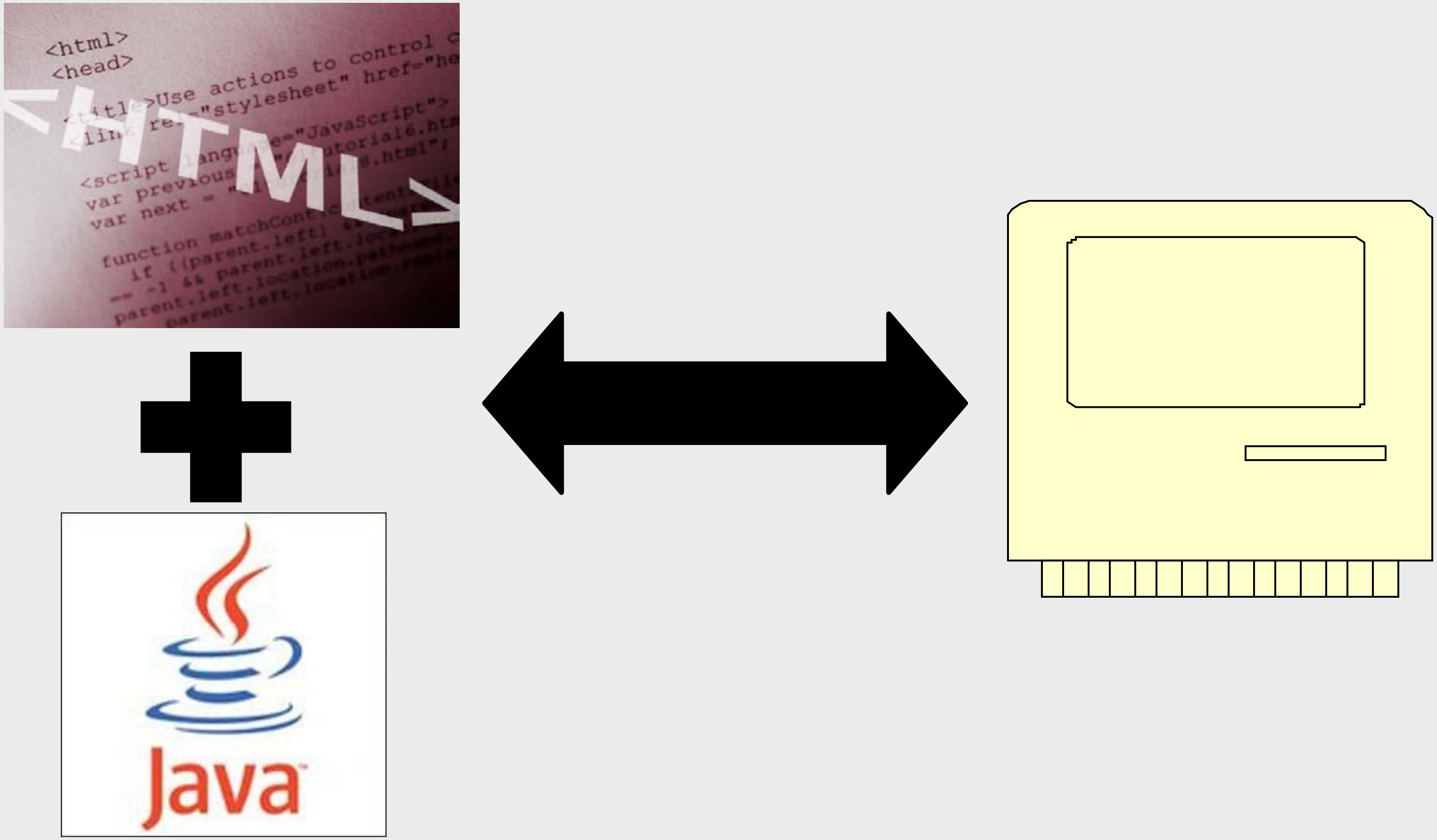
Click on the microphone icon and begin speaking for as long as you like.

Distributed frameworks

Client-server framework



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

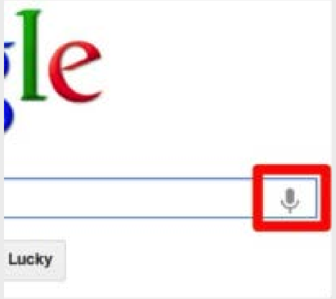
WEB SPEECH API

In-browser speech recognition

The Speech Input API

In late 2010, shortly after the W3C HTML Speech Incubator Group was formed, Google submitted the Speech Input API Specification for consideration. This spec centered on the addition of a speech attribute to the HTML input element.

In early 2011, Google added support for the Speech Input API to Chrome, with the x-webkit-speech vendor-prefixed attribute. Adding this attribute to any text input field causes Chrome to add a microphone icon to that field.



Some examples in www.trabhci.eu

```
<input type="text" id="address"  x-webkit-speech
onwebkitspeechchange="codeAddress();"/>
```

WEB SPEECH API

The Web Speech API

Rather than propose HTML elements and attributes for consideration, as Speech Input had, the Web Speech API spec was focused solely on new JavaScript APIs for speech recognition.

In the spring of 2012, the Speech API W3C Community Group was formed to produce a JavaScript Speech API that addressed many of the use cases identified by the W3C Speech Incubator group's final report.

The Web Speech API consists of three main feature areas:

Speech recognition via the SpeechRecognition object

Text-to-speech synthesis via the SpeechSynthesis object

The creation of custom grammars via the SpeechGrammar object

The Web Speech API Specification was finalized in October of 2012.

WEB SPEECH API

- The Web Speech API aims to enable web developers to provide, in a web browser, speech-input and text-to-speech output features that are typically not available when using standard speech-recognition or screen-reader software.
- The API itself is agnostic of the underlying speech recognition and synthesis implementation and can support both server-based and client-based/embedded recognition and synthesis.
- The API is designed to enable both brief (one-shot) speech input and continuous speech input.
- Speech recognition results are provided to the web page as a list of hypotheses, along with other relevant information for each hypothesis.

WEB SPEECH API

This specification supports, among others, the following use cases:

- Voice Web Search
- Speech Command Interface
- Continuous Recognition of Open Dialog
- Speech Translation
- Speech Enabled Email Client
- Dialog Systems
- Multimodal Interaction
- Multimodal Search

WEB SPEECH API

The SpeechRecognition Interface

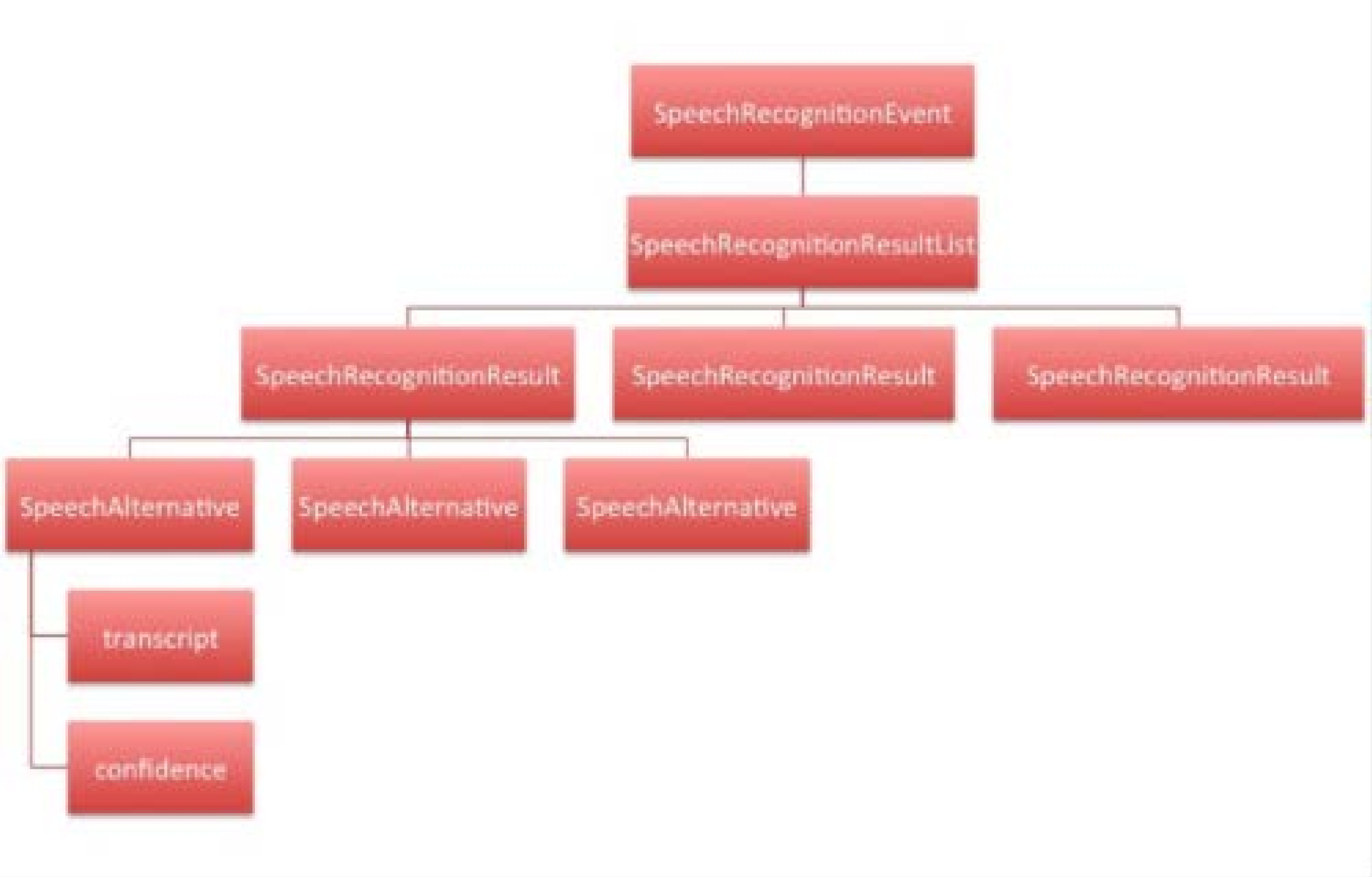
```
[Constructor]
interface SpeechRecognition : EventTarget {
// recognition parameters
attribute SpeechGrammarList grammars;
attribute DOMString lang;
attribute boolean continuous;
attribute boolean interimResults;
attribute unsigned long maxAlternatives;
attribute DOMString serviceURI;
// methods to drive the speech interaction
void start();
void stop();
void abort();
// event methods
attribute EventHandler onaudiostart;
attribute EventHandler onsoundstart;
attribute EventHandler onspeechstart;
attribute EventHandler onspeechend;
attribute EventHandler onsoundend;
attribute EventHandler onaudioend;
attribute EventHandler onresult;
attribute EventHandler onnomatch;
attribute EventHandler onerror;
attribute EventHandler onstart;
attribute EventHandler onend;
};
```

WEB SPEECH API

```
// Item in N-best list
interface SpeechRecognitionAlternative {   };

// A complete one-shot simple response
interface SpeechRecognitionResult { };

// A collection of responses (used in continuous mode)
interface SpeechRecognitionResultList {  };

// A full response, which could be interim or final,
part of a continuous response or not
interface SpeechRecognitionEvent : Event {  };

// The object representing a speech grammar
[Constructor]
interface SpeechGrammar { };

// The object representing a speech grammar collection
[Constructor] interface SpeechGrammarList {  };
```

WEB SPEECH API

The SpeechRecognitionEvent and its child objects



WEB SPEECH API

The SpeechSynthesis Interface

```
interface SpeechSynthesis {
readonly attribute boolean pending;
readonly attribute boolean speaking;
readonly attribute boolean paused;

void speak(SpeechSynthesisUtterance utterance);
void cancel();
void pause();
void resume();
SpeechSynthesisVoiceList getVoices(); };

interface SpeechSynthesisUtterance : EventTarget {
attribute DOMString text;
attribute DOMString lang;
attribute DOMString voiceURI;
attribute float volume;
attribute float rate;
attribute float pitch;
attribute EventHandler onstart;
attribute EventHandler onend;
attribute EventHandler onerror;
attribute EventHandler onpause;
attribute EventHandler onresume;
attribute EventHandler onmark;
attribute EventHandler onboundary; };
```

WEB SPEECH API

How to use Google TTS

```
function speak(output, lang) {
// (Use a TTS API to speak output in lang)
var sintesis="http://translate.google.com/translate_tts?";
if(output.length>0){
            outputs=output.replace(/\s/g,"+");
            sintesis=sintesis+"q="+outputs+"&tl="+lang;
// create  HTML
            var salida = "<iframe rel='noreferrer' src='" + sintesis+ "'></iframe>";
// show
            document.getElementById("TTS").innerHTML = salida;
            }
}

<div id="TTS" style="position:absolute;left:-1000px"></div>
```

WEB SPEECH API

Vendor prefix

```
<button id="button" onclick="toggleStartStop()"></button>
  <div style="border:dotted;padding:10px">
    <span id="final_span"></span>
    <span id="interim_span" style="color:grey"></span>
  </div>
<script type="text/javascript">
  var recognizing=false;
  var recognition = new webkitSpeechRecognition();
  recognition.continuous = true;
  reset();
  recognition.onend = reset;


  recognition.onresult = function (event) {
      var final = "";
      for (var i = 0; i < event.results.length; ++i) {
          final += event.results[i][0].transcript;
      }
      final_span.innerHTML = final;
    }

  function reset() {
                  recognizing = false;
                  button.innerHTML = "Click to Speak";
  }
```

```
function toggleStartStop() {
                 if (recognizing) {
                   recognition.stop();
                   reset();
                 }
                 else
                 {
                   recognition.start();
                   recognizing = true;
                   button.innerHTML = "Click to Stop";
                 }
  }
</script>
```

WEB SPEECH API

Play with demos in the trabhci web page

More info

http://www.adobe.com/devnet/html5/articles/voice-to-drive-the-web-introduction-to-speech-api.html

http://updates.html5rocks.com/2013/01/Voice-Driven-Web-Apps-Introduction-to-the-Web-Speech-API