IL PARLATO
E
I SISTEMI
TEXT-TO-SPEECH

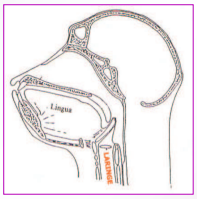
Rachele Sprugnoli

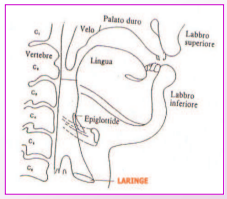
L'espressione fonico-acustica

Risultato di una complessa storia evolutiva.

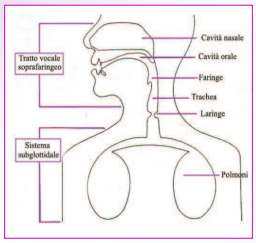
Circa 250.000 anni fa:

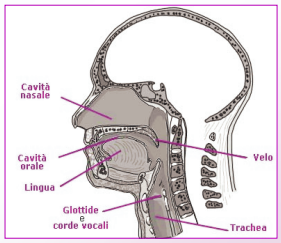
- discesa della laringe verso il basso;
- sistema "a due canne".





Struttura dell'apparato fonatorio





NB.

Fonazione come FUNZIONE SECONDARIA installata su un insieme di organi inizialmente disegnato per la respirazione, l'alimentazione e la percezione degli odori.

Fonazione come SCELTA tra le possibili modalità espressive (ad es. gestualità, mimica, produzione di manufatti) per ragioni di EFFICIENZA SEMIOTICA

Vantaggi della fonazione

- **Simultanea** ad altri comportamenti;
- Può essere eseguita e ricevuta in **condizioni ambientali difficili**;
- Ha una larga **modulabilità**;
- Può essere **ricevuta** da più riceventi **contemporaneamente**;
- Può essere prodotta in **modo continuo**;
- È "**portatile**";
- È **rapida**.

Come si forma un suono

I. I polmoni si espandono e si contraggono immettendo ed espellendo aria;

II. Se le corde vocali sono chiuse, l'aria esercita una pressione su di esse → l'aria esce sotto forma di sbuffo;

III. Si forma il cosiddetto TONO LARINGEO;

IV. Il tono laringeo viene modificato dai diaframmi che incontra → si originano i suoni linguistici come li percepiamo uditivamente.

NB. Vocali vs. Consonanti

Grafemi, foni, fonemi

Vari modi in cui si possono descrivere i suoni di una lingua:

- PRONUNCIA → ***grafemi***, digrammi, trigrammi
- FONETICA → ***foni***
- FONEMATICA (o fonologia) → ***fonemi***

grafema = segno che costituisce l'unità grafica minima;

fono = suono di una lingua considerato in base alle sue caratteristiche fisiche (articolatorie, acustiche, percettive);

fonema = suono di una lingua considerato in base alla funzione distintiva che ha in un determinato sistema linguistico. Unità minima di una parola, non dotata di significato.

Fenomeni che riguardano l'enunciato (1)

Fenomeni SOPRASEGMENTALI (funzione distintiva)

Accento soprasegmentale di parola

ES.

prìncipi VS. princìpi

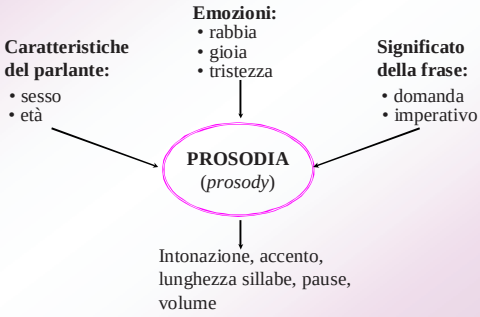
Intonazione soprasegmentale di enunciato

ES.

Chiudi la porta! VS. Chiudi la porta?

Fenomeni che riguardano l'enunciato (2)

Il concetto di prosodia



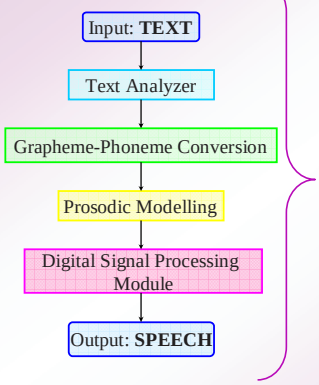
Cos'è un sistema Text-To-Speech

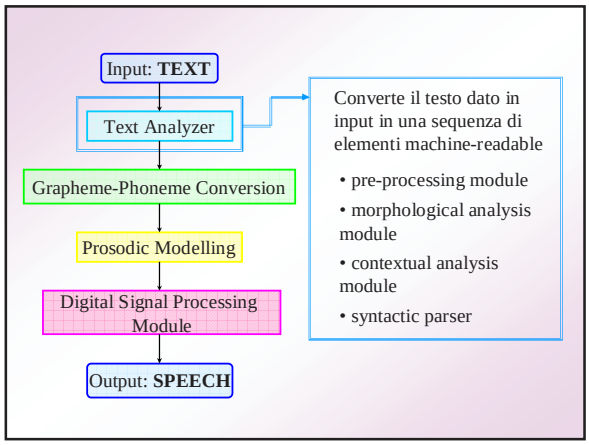
"The automatic production of speech, through a grapheme-to-phoneme transcription of the sentences to utter." (Thierry Dutoit)

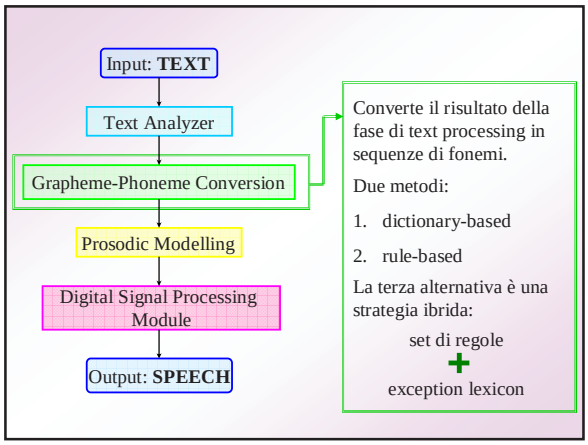
NB.

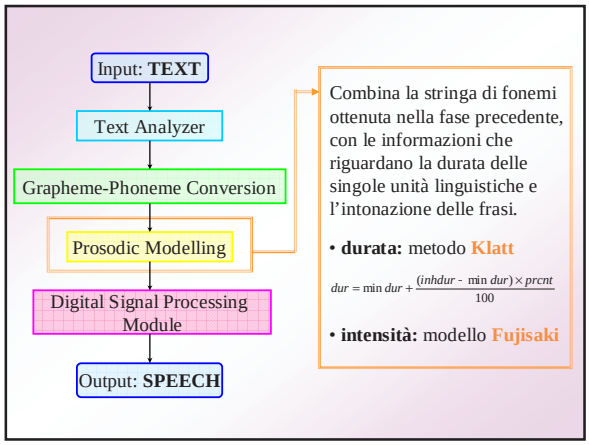
TTS **vs.** mangianastri o lettori CD
TTS **vs.** Voice Response System

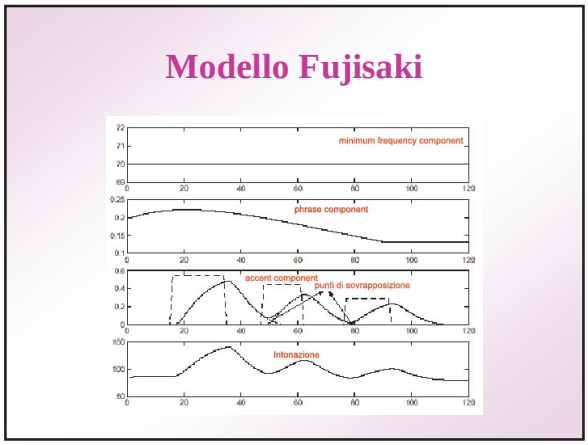
Schema di un sistema TTS

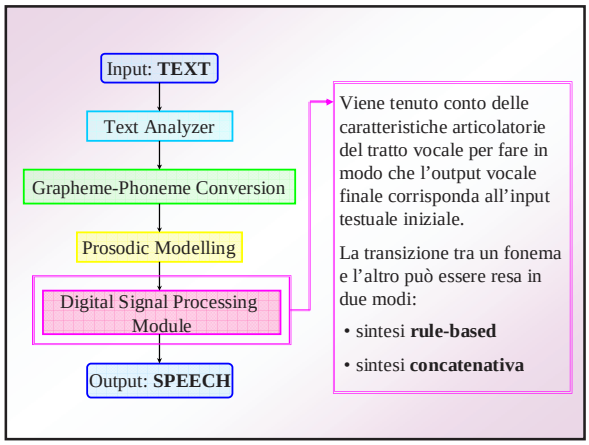



Input: TEXT
Text Analyzer
Grapheme-Phoneme Conversion
Prosodic Modelling
Digital Signal Processing Module
Output: SPEECH
Converte il testo dato in input in una sequenza di elementi machine-readable
• pre-processing module
• morphological analysis module
• contextual analysis module
• syntactic parser


Input: TEXT
Text Analyzer
Grapheme-Phoneme Conversion
Prosodic Modelling
Digital Signal Processing Module
Output: SPEECH
Converte il risultato della fase di text processing in sequenze di fonemi.
Due metodi:
1. dictionary-based
2. rule-based
La terza alternativa è una strategia ibrida:
set di regole
+
exception lexicon


Input: TEXT
Text Analyzer
Grapheme-Phoneme Conversion
Prosodic Modelling
Digital Signal Processing Module
Output: SPEECH
Combina la stringa di fonemi ottenuta nella fase precedente, con le informazioni che riguardano la durata delle singole unità linguistiche e l'intonazione delle frasi.
• durata: metodo Klatt
$dur = \min dur + \frac{(inhdur - \min dur) \times prcnt}{100}$
• intensità: modello Fujisaki


Modello Fujisaki
minimum frequency component
phrase component
accent component
punti di sovrapposizione
Intonazione



Sintesi rule-based

Formant Synthesis

FORMANTS = frequenze caratteristiche del parlato grazie alle quali i vari suoni vengono identificati dall'ascoltatore.

• Vengono simulate le caratteristiche spettrografiche dei diversi suoni tramite composizioni di vari risonatori, ciascuno accordato su frequenze specifiche di formanti diversi.

• Le posizioni dei formanti nei diversi fonemi e le loro variazioni nel passaggio da un fonema al successivo sono descritte da specifiche regole.

Sintesi concatenativa

I sistemi più recenti utilizzano la concatenazione di difoni.

DIFONI = segnale che connette la seconda metà di un fonema con la prima metà del fonema successivo.

unità fonetiche che realizzano la transizione tra fonemi

utilizzandoli si risolve il problema della distorsione causata dalla dispersione nei punti di concatenazione.

Sintesi articolatoria

• In teoria è il metodo migliore per generare il parlato in maniera artificiale in quanto punta a riprodurre il sistema fonatorio umano usando modelli computazionali biomeccanici.

• Nella pratica 2 difficoltà: – ottenere modelli tridimensionali del tratto vocale;

– modellare il sistema con un insieme limitato di parametri

richiede una notevole capacità di elaborazione

VISUAL TTS

Visual TTS is the synchronization of a facial image, or **talking head**, with synthesized speech.

- incremento dell'intelligibilità del sistema
- aggiunta di informazioni nel processo comunicativo tra utente e sistema

- 1972, Parke crea il primo modello facciale 3D
- 1974, Parke sviluppa il primo modello parametrico in 3D

Esempi

August
varie espressioni

Lucia:
tristezza

Lucia:
gioia

Valutazione dei sistemi TTS

Qualità del parlato ottenuto in output da un sistema TTS rispetto ai parametri di **naturalezza** e di **intelligibilità**:

più livelli di analisi

più metodi per testare proprietà diverse:

- Segmental evaluation methods;
- Comprehension tests;
- Prosody evaluation;
- Field tests

Applicazioni

- servizi di telecomunicazione;
- insegnamento delle lingue;
- aiuto ad ipo-vedenti e non-vedenti;
- strumenti per la logopedia;
- libri e giocattoli parlanti;
- monitoraggio vocale;
- interfacce "hands-free";
- ricerca

Linguaggi di markup (1): SSML

Speech Synthesis Markup Language

Sviluppato dal Centre for Speech Technology Research (CSTR) dell'Università di Edinburgo nel 1995.

ES.

```
<ssml>
<define word= "edinburgh" pro="EH1 D AH0 N B ER2 OW0"
  format= "cmudict.1.0">
<phrase> I saw the man in the park <phrase> with the telescope
</phrase>
<phrase> I saw the man <phrase> in the park with the telescope
</phrase>
<phrase> The train is now standing on platform <emph> A </emph>
<language="italian"> <phrase> continua in italiano </phrase>
</ssml>
```

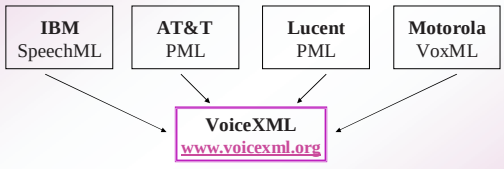
Linguaggi di markup (2): VoiceXML

Voice eXtensible Markup Language

È un "***dialog markup language***" (definizione del W3C), basato sulla sintassi XML, usato per sviluppare applicazioni in cui la comunicazione uomo-computer non si realizza solo per mezzo di un interfaccia grafica (ad es. link e bottoni accessibili attraverso tastiera e mouse) ma anche e soprattutto mediante un'interfaccia vocale.

Ancora VoiceXML…

Un po' di storia…



W3C:
Marzo 2000: Recommendation versione 1.0;
16 Marzo 2004: Recommendation versione 2.0;
23 Marzo 2004: primo Working Draft della versione 2.1

…e ancora…

Dal punto di vista tecnico…

Per l'accesso vocale sono necessari due moduli lato-server:

1. uno per il riconoscimento vocale (Speech Recognition);
2. uno per la traduzione del testo in voce (TTS).

Dal punto di vista pratico…

Un documento VoiceXML è costituito da una serie di "dialoghi".

Ne vengono definiti 2 tipi:

- form, per l'output e l'input vocale;
- menu, presenta una selezione di voci tra cui scegliere.

VoiceXML, esempio

ES.

```
<?xml version="1.0"?>
<vxml version="2.0">
<menu>
<prompt>Scegli una sezione<enumerate/> </prompt>
<choice next="http://www.sports.example/start.vxml">
  Sport </choice>
<choice next="http://www.weather.example/intro.vxml">
 Tempo </choice>
<choice next="http://www.news.example/news.vxml">
 News </choice>
<noinput>Per favore scegli una sezione <enumerate/></noinput>
</menu>
</vxml>
```

VoiceXML, portali vocali

Offrono accesso vocale a vari servizi quali titoli di borsa, liste di spettacoli teatrali e cinema, notizie del giorno, lettura e scrittura di e-mail usando la voce:

- TellMe **www.tellme.com**
- Quack **www.quack.com**
- BeVocal **www.bevocal.com**
- Hey Anita **www.heyanita.com**

Linguaggi di markup (3): SALT

Speech Application Language Tags

"SALT extends existing web markup languages to enable multimodal and telephony access to the web" (SALTforum)

Consente agli sviluppatori di inserire comandi vocali all'interno di pagine HTML, XML o XHTML già esistenti.

le applicazioni vocali possono convivere parallelamente ai mezzi tradizionali di input/output

Linguaggi di markup (4): (X+V)

XHTML + VoiceXML

- Ideato da IBM, OPERA e Motorola
- **W3C:** *Novembre 2001*: versione 1.0
  Gennaio 2003: versione 1.1

Integra XHTML con VoiceXML 2.0 per realizzare l'interazione vocale

Bibliografia ragionata (1)

Studi di linguistica:

❖ Simone R., *Fondamenti di linguistica*, Roma-Bari, Laterza, 1990

❖ Muliačić Ž., *Fonologia generale e fonologia della lingua italiana*, Bologna, il Mulino, 1971

❖ Sobrero A.A., *Introduzione all'italiano contemporaneo: le strutture*, Roma-Bari, Laterza, 1999

Bibliografia ragionata (2)

Informazioni generali sui sistemi TTS e sulla sintesi vocale:

❖ Tutorial per un sistema TTS: http://www.ias.et.tu-dresden.de/sprache/lehre/multimedia/tutorial/index.html

❖ Speech synthesis: http://www.fact-index.com/s/sp/speech_synthesis.html

❖ A Text-to-Speech Primer: http://www.portset.co.uk/text_to_speech_primer.htm

❖ A Short Introduction to Text-to-Speech Synthesis: http://tcts.fpms.ac.be/synthesis/introtts.html

❖ Review of Speech Synthesis Technology: http://www.acoustics.hut.fi/~slemmett/dippa/

Bibliografia ragionata (3)

❖ **Bilotta E.**, *Interfacce multimodali ed aspetti psicologici dell'interazione uomo-computer*: http://galileo.cincom.unical.it/Pubblicazioni/editoria/libri/HCI-ele/coper.html

❖ **Klatt D.H.,** *Review of text-to-speech conversion for English*: http://www.mindspring.com/~ssshp/ssshp_cd/dk_737a.htm

❖**Ibarra I.O. – Curatelli F.**, *A Brief Introduction to Speech Analysis and Recognition, An Internet Tutorial*: http://www.mor.itesm.mx/~omayora/Tutorial/tutorial.html

Bibliografia ragionata (4)

VoiceXML:

❖ Guida del W3C: http://www.w3.org/Voice/Guide/

❖Kleis Communication Technologies: http://www.kleis.it/portali-vocali/1#1

❖ Community italiana di VoiceXML: http://www.vxmlitalia.com/

❖ ITportal: http://www.itportal.it/special/internet/voicexml/default.asp

Bibliografia ragionata (5)

SALT:

❖ SALT Forum:
http://www.saltforum.org/

SSML:

❖ Speech Synthesis Markup Language Version 1.0
http://www.w3.org/TR/speech-synthesis/

XHTML + VoiceXML:

❖ XHTML+Voice Profile 1.0
http://www.w3.org/TR/xhtml+voice/

Bibliografia ragionata (6)

Linguaggi di markup per le applicazioni vocali:

❖ Standard per applicazioni multimodali
http://www.vxmlitalia.com/Baggia.pdf

Rivista on-line:

❖ Speech Technology Magazine:
http://www.speechtechmag.com/

Pagina di riferimenti on-line:

❖ Home pages related to phonetics and speech sciences:
http://fonsg3.let.uva.nl/Other_pages.html