Paolo Baggia
Loquendo

Speech Technologies and Standards

DIT Seminars, Povo, Trento

June 7th, 2006

Loquendo Today

- Global company formed in 2001 as a spin-off from the Telecom Italia R&D center with over 30 years experience in Speech Technologies
- A Telecom Italia Group company
- HQ in Turin (Italy) and Sales offices in Miami (US), Madrid (Spain), Munich (Germany) and Paris (France), and a Worldwide Network of Agents.
- Market-leader in Italy, Spain, Greece with a strong presence in Europe, growing presence in North and Latin America

Strategy:

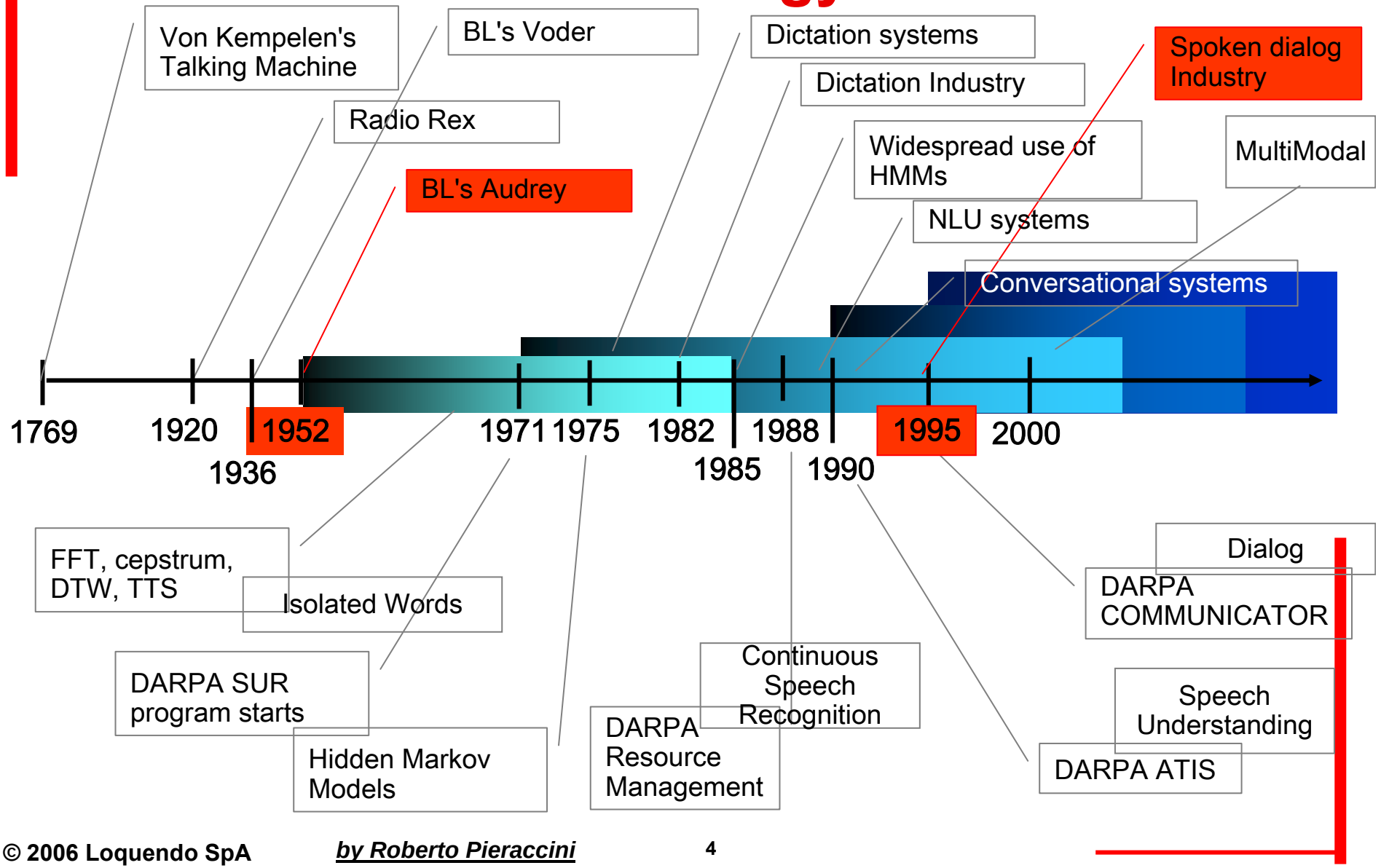
- Complete set of speech technologies on a wide spectrum of devices
- Ready for challenging future scenarios: Multimodality, Security
- Partnership as a key factor

"Best Innovation in Multi-Lingual Speech Synthesis" Prize AVIOS-SpeechTEK West 2006

Plan of my talk

- **The Past**
  - Evolution of Speech Technologies (mainly ASR, TTS)
  - First Speech Applications (Touch Tones, IVRs)
- **The Present – The Impact of Standards**
  - The (r)evolution of VoiceXML
  - A Constellation of W3C Standards
  - Work in progress
- **The Future**
  - Voice and Video Applications
  - Speech and SemanticWeb
  - Towards Multimodality
- **Q&A**

A Brief History of Spoken Language Technology

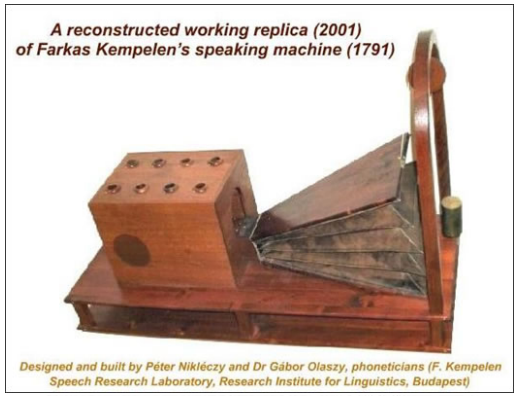


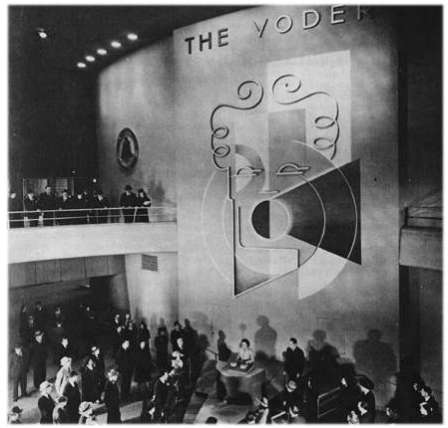
Evolution of Speech & Language

- Foundations lay in many fields:
  - Computer science, electronic engineering, linguistics and psychology
- '50 – '70 : Two paradigms
  - Symbolic: Formal syntax, AI techniques
  - Stochastic: Bayesian method, corpora (es. Brown Corpus)
- '70 – mid '80 : Four separate fields
  - Stochastic:
    - Dynamic Programming, Viterbi algorithm
    - hidden Markov models
    - Back-forward algorithm for training HMMs
    - Language Modeling (bigrams, trigrams, smoothing techniques)
  - Logic-based: DFG, functional grammars, LFG
  - Natural Language Understanding: Winograd, Woods, many others
  - Discourse Modeling: Grosz, Sidner, Allen
- mid '80 - '90 : Return of empiricism
  - Return to finite state modeling: AT&T Bell Labs
  - Rise of probabilistic models: IBM J. Watson research center
- mid '90 – now : The fields come together
  - The separation is partially blurred
  - Probabilistic and corpus driven in natural language processing

Evolution of Speech Synthesis

- **Early days:**
  - ***From***: **von Kempelen's Talking Machine (1791)**
  - **Sir Charles Wheatstone (1835), Faber's Euphonia (1835), etc.**
  - ***To***: **Dudley's Voder at NY World Fair 1939**
- **'50 – '60:**
  - **Audio Spectrograms, first articulatory and formant approaches**
- **mid '70 – '90:**
  - **KlattTalk, MIT Talk, DECTalk**
  - **Kurzweil Machine for the visually impaired**
  - **IBM and AT&T progresses**
  - **➔ Very intelligible TTS, but still robotic**
- **mid '90:**
  - **Concatenation by Unit Selection ➔ Intelligible and very natural voices**
- **Today**
  - **Lots of application fields for TTS: SF-embedded, mobile devices, navigation systems, IVRs (!) ➔ but also search engines, domotic, multimodal, etc.**





R&D in Italy

- **First attempts: Luigi Stringa, ELSAG**
- **R&D:**
  - ***Olivetti*** **(Speech lab in Turin, closed in 1990)**
    - **Speech dictation, Speech synthesis (VoxPC)**
  - ***IRST*** **(lab in Povo, Trento, by Istituto Trentino di Cultura)**
    - **Speech recognition, dictation, speech synthesis, spoken dialog**
  - ***CSELT*****, then *TILAB***
    - **Speech recognition, Speech synthesis, Speaker Verification**
    - **Speech recognition on chip**
    - **First large speech applications deployed**
      Servizio 1412 e 12 (Telecomitalia), FS_INFORMA (Trenitalia)
- **Speech Technologies and Platforms:**
  - **LOQUENDO** **(spin-off in 2001)**
    - **Engines: speech recognition, speech synthesis, speaker verification**
    - **Embedded speech technologies**
    - **VoiceXML platform (VoxNauta) and MRCP-based server (Loquendo Speech Suite)**

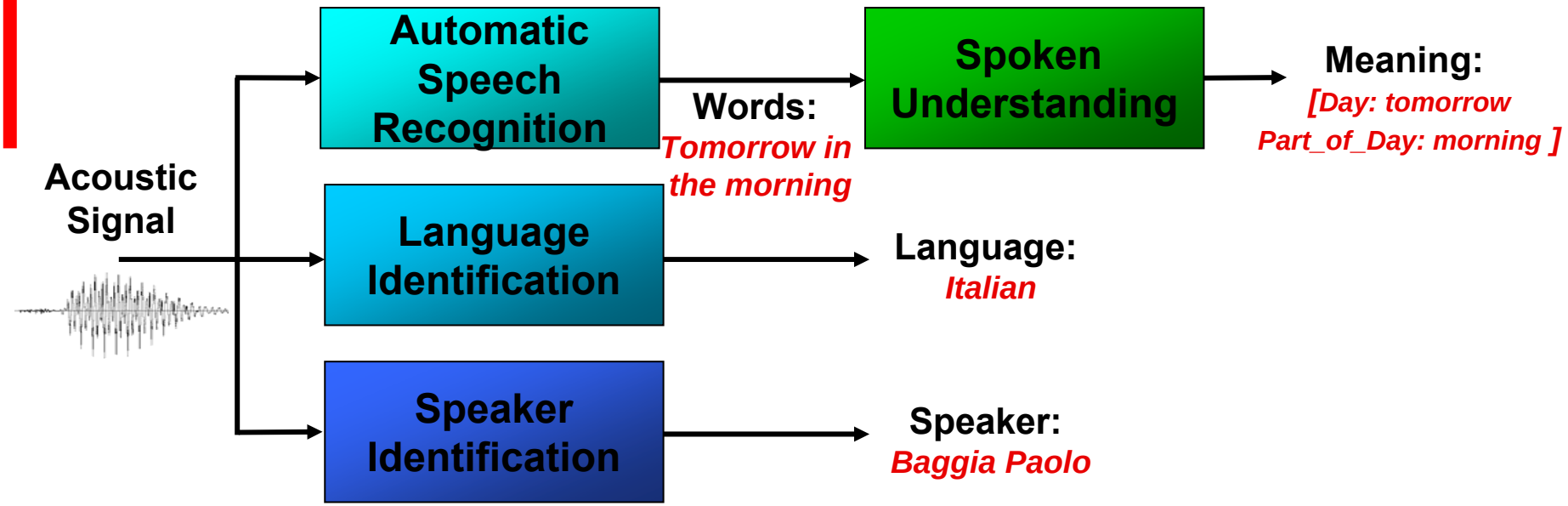
Loquendo
VOCAL TECHNOLOGY AND SERVICES

Introduction to Speech Technologies

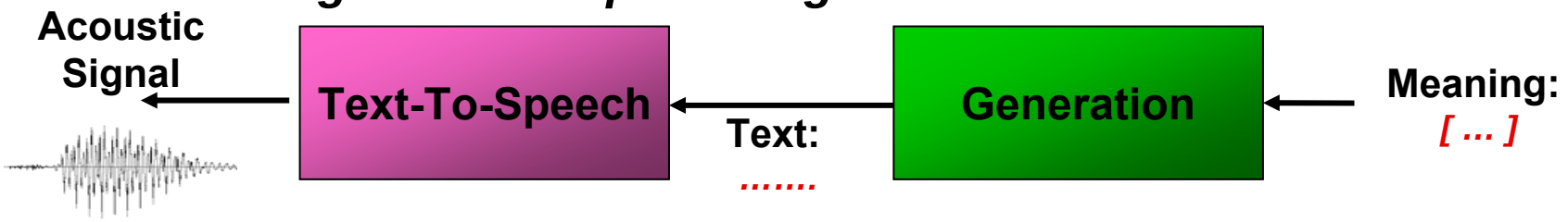
- ASR: Automatic Speech Recognition
- TTS: Text-To-Speech
- SIV: Speaker Identification Verification
- Embedded Speech Technologies

Introduction to Speech Technologies

- **Goal:** ***To extract information from a speech signal***

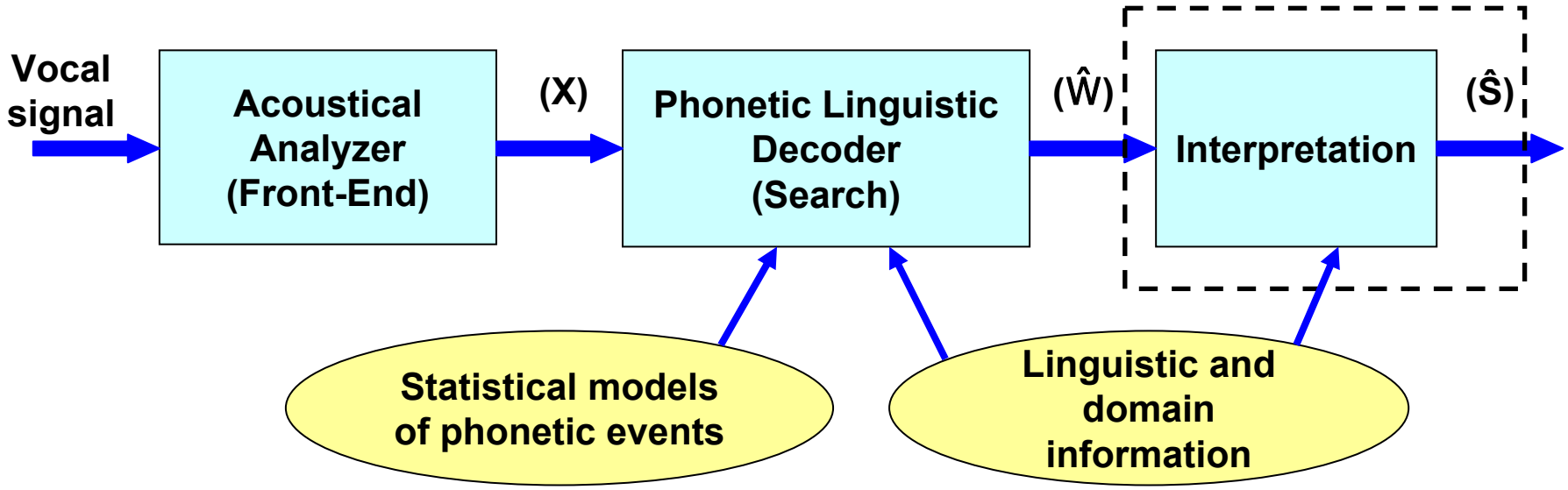


- **Goal:** ***To generate a speech signal from content***



ASR: Automatic Speech Recognition

ASR block diagram



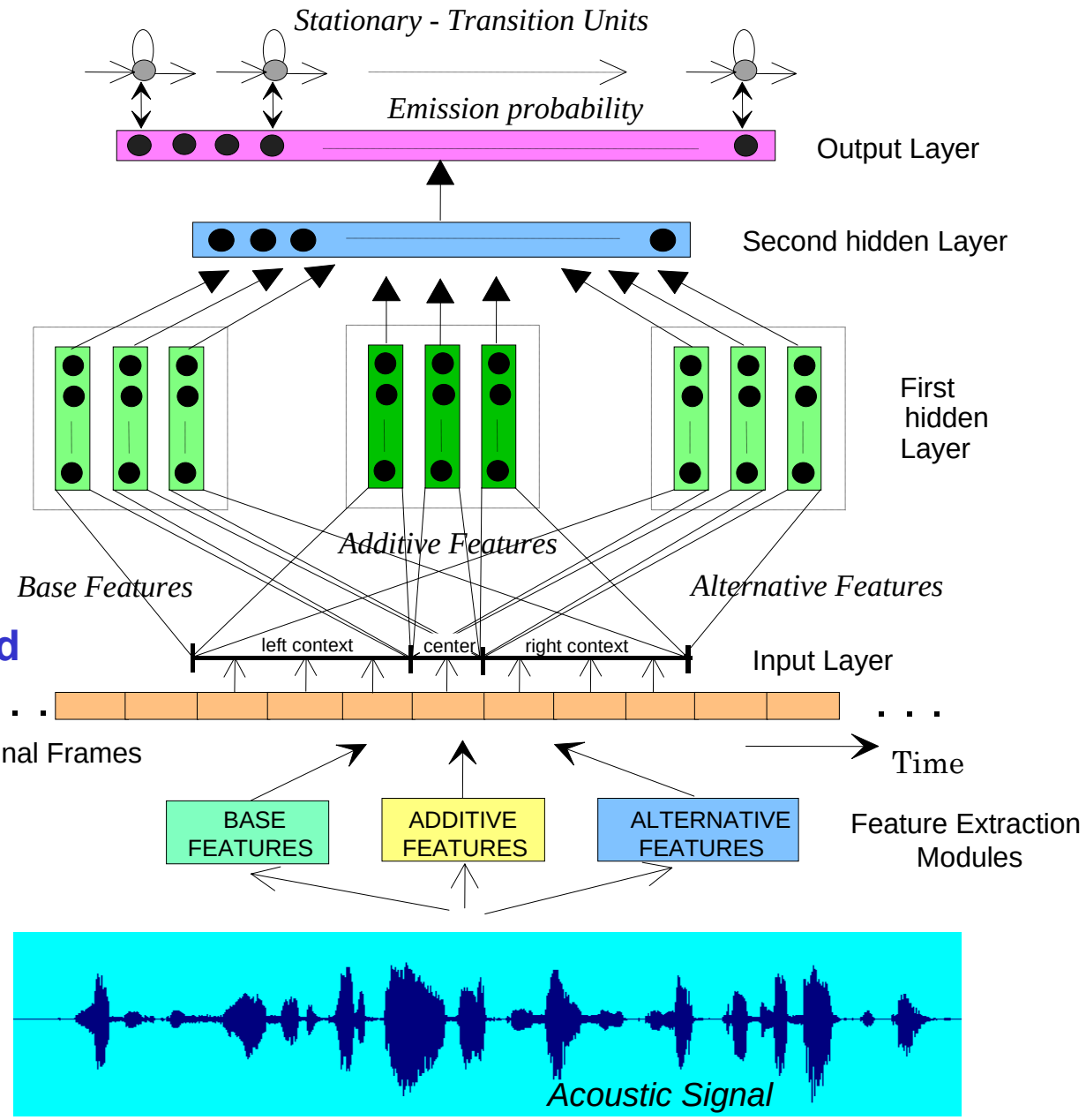
(X) Parametric description of the voice signal
(Ŵ) Sequence of recognized words
(Ŝ) Semantic representation

Automatic speech recognizers: main features

Features			
Training mode	Focused on a unique speaker	Speaker independent	
Pronunciation mode	Isolated	Continuous	
Environment condition	Controlled	Difficult	
Recognition Domain	Constrained	Customizable	Dictation

- **Current telephonic recognizers are speaker independent, supporting continuous speech input; the recognition domain must be defined by the application developer**
- **The dictation on a general domain is today available with speaker adapted desktop recognizers:**
  - **The recognizer models are focused on the target speaker voice**
  - **The environment must be good (high quality microphone, low noise)**
  - **However, the domain must not be unconstrained … to obtain high performance specific additional information should be supplied**

HMM-NN Hybrid Models



Innovations

- **Multi-source NN**
- **MLP topology**
- **Acceleration method**

Base Feature: **MFCC**

Alternative Features :
Rasta-PLP, Ear-Model

Additive Features :
Gravity Centers,
Frequency Derivatives

ASR Typical Performance

- The accuracy and efficiency performance on a vocabulary depends on the vocabulary size and confusability

Telephonic environment

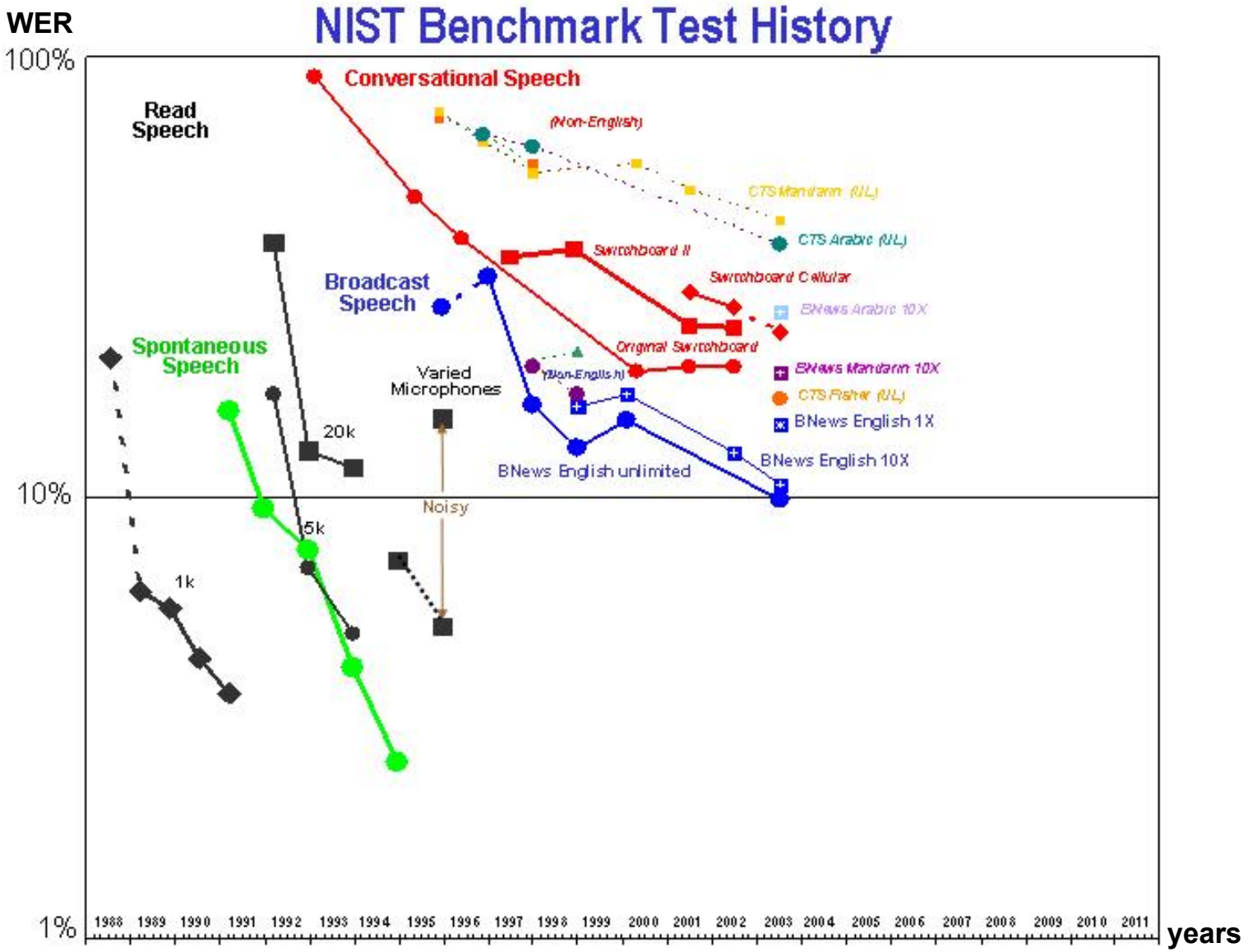
98.5 ÷ 99% digits
(PSTN DB)

95 ÷ 97% 500 words vocabulary
(Railways Stations: FS-Informa)

85 ÷ 90% 10,000 words vocabulary
(Italian cities: Telecom Italia 12)

- Kind of applications:
  - Multiple choice menus
  - Large vocabulary recognition (e.g Telecom Italia 12)

ASR performance (by D. Pallet, NIST)

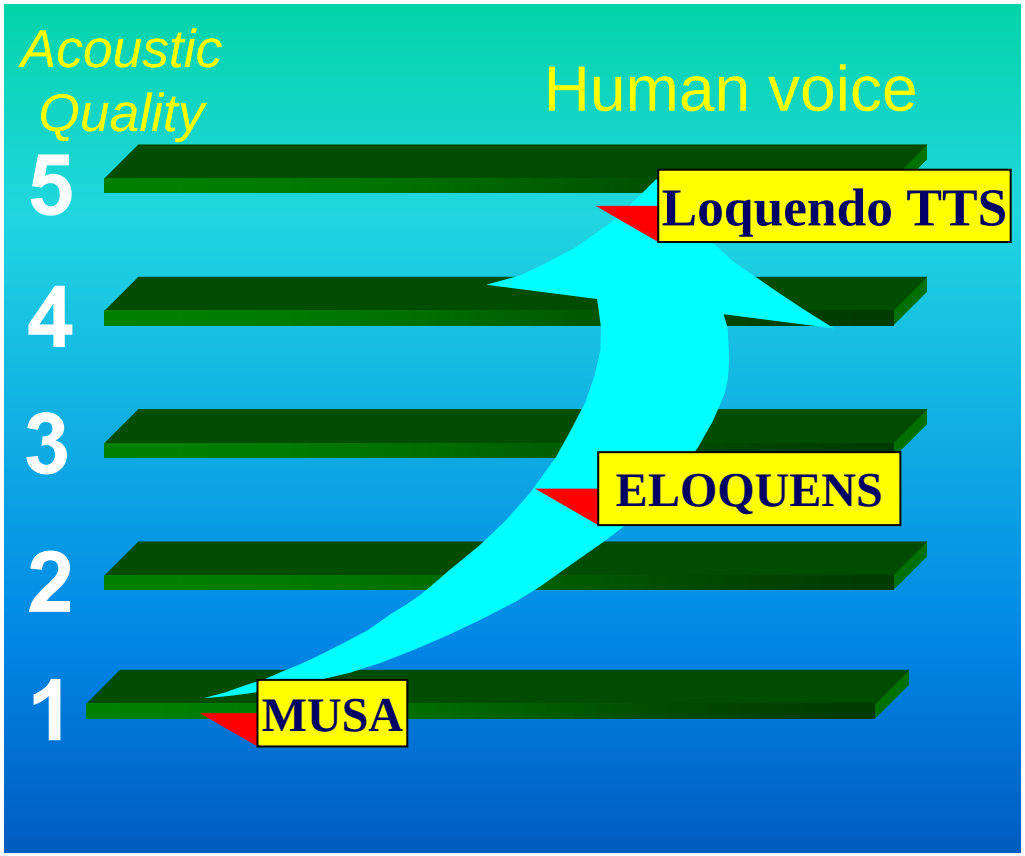


Hot Issues in ASR

- **Robustness**
  - to the environment:
    noise (in-car, in house), side speech (mobile)
  - to the channel:
    VoIP channel, mobile channel
  - to the speaker:
    individual peculiarities (adaptation techniques)
- **Spontaneous Speech**
  - Interruptions, restarts, extralinguistic phenomena
- **Speech search, audio indexing**
- **Speech for very large applications**
  - Integration of many knowledge sources

TTS: Text-To-Speech

Evolution of TTS quality



Loquendo TTS

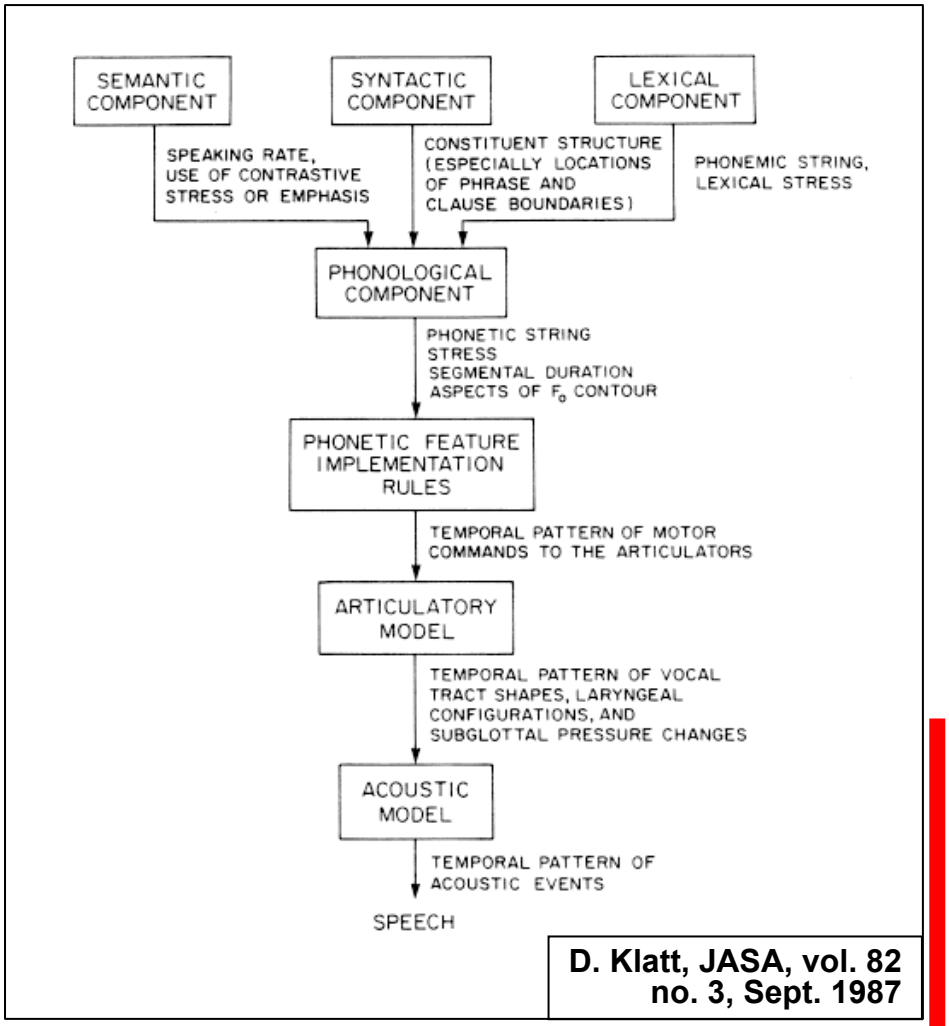
	Timber	Intelligibility
Loquendo TTS	natural	excellent
ELOQUENS	acceptable	good
MUSA	robotic	sufficent

1980 1990 2000

1st 2nd 3rd *generation*

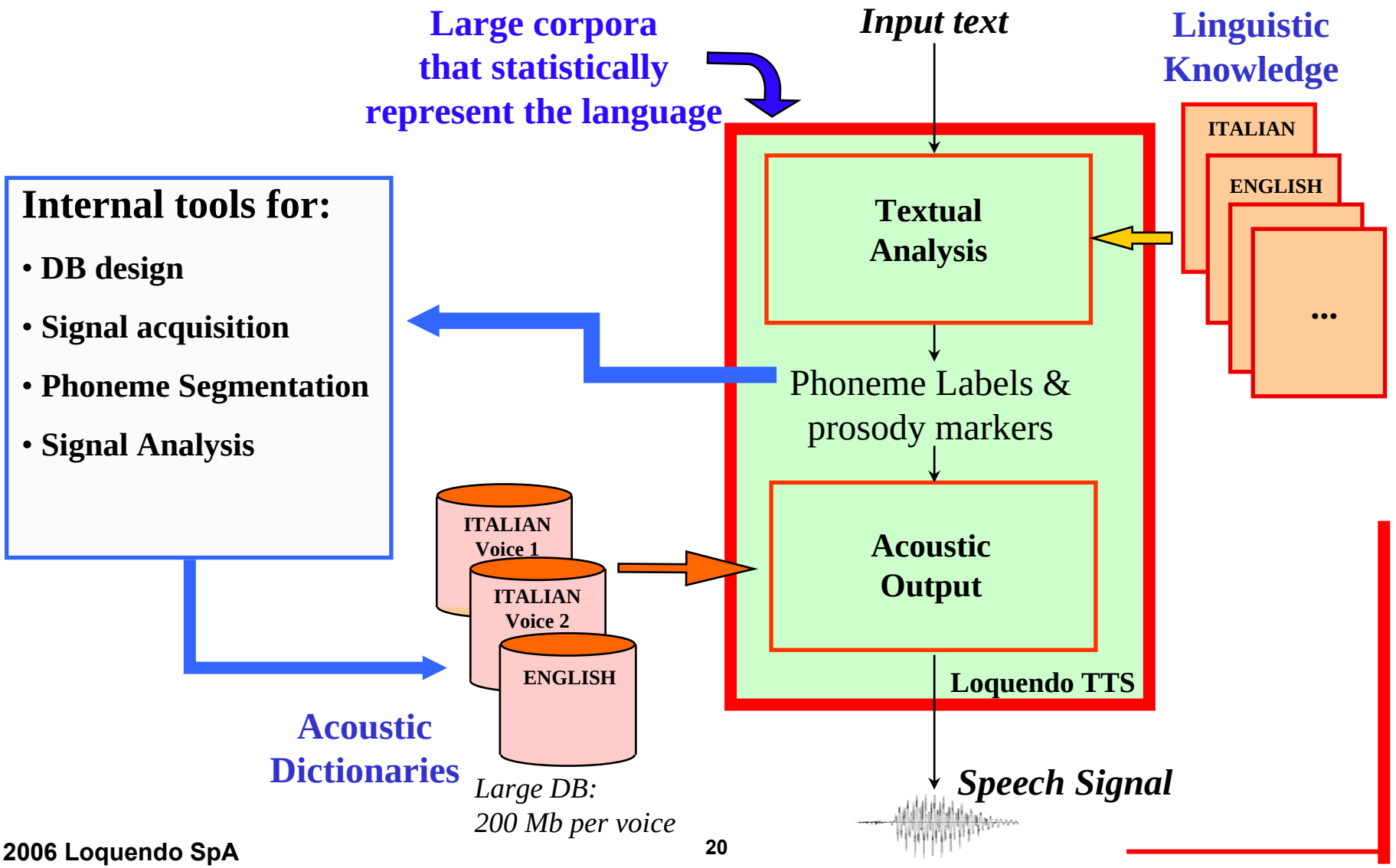
Three Families of Speech Synthesizers

- Articulatory synthesizers
- Formant synthesizers
- Concatenative synthesizers
  - Diphone based (concatenation of small units)
  - Unit Selection (concatenation of larger units)



D. Klatt, JASA, vol. 82 no. 3, Sept. 1987

Corpus based, Unit Selection TTS



Loquendo TTS Features

- **Interactive demo of the 39 voices in 17 languages**
  http://www.loquendo.com/en/demos/demo_tts.htm

- **Innovative features released in the TTS engine:**
  - **Expressiveness**
    Conventional figures, different styles, expressive cues
  - **Customizable voices**
    Change in timbre, save and reuse a customized voice
  - **Mixing TTS and music**
    Music and TTS are synchronized

SMS reading

r u don nethng 2mor? would b gr8 2cu.

i'll b @ pub b4 8.

ringl8r. b4n xoxox susan

Mixed Language Capability

Loquendo TTS automatically detects each change of language and gives the user the option to:

- Change voice, choosing from the voices available, based on the language detected.
- Keep the same voice as the original language, and apply phonetic mapping between the newly detected language and the original one.

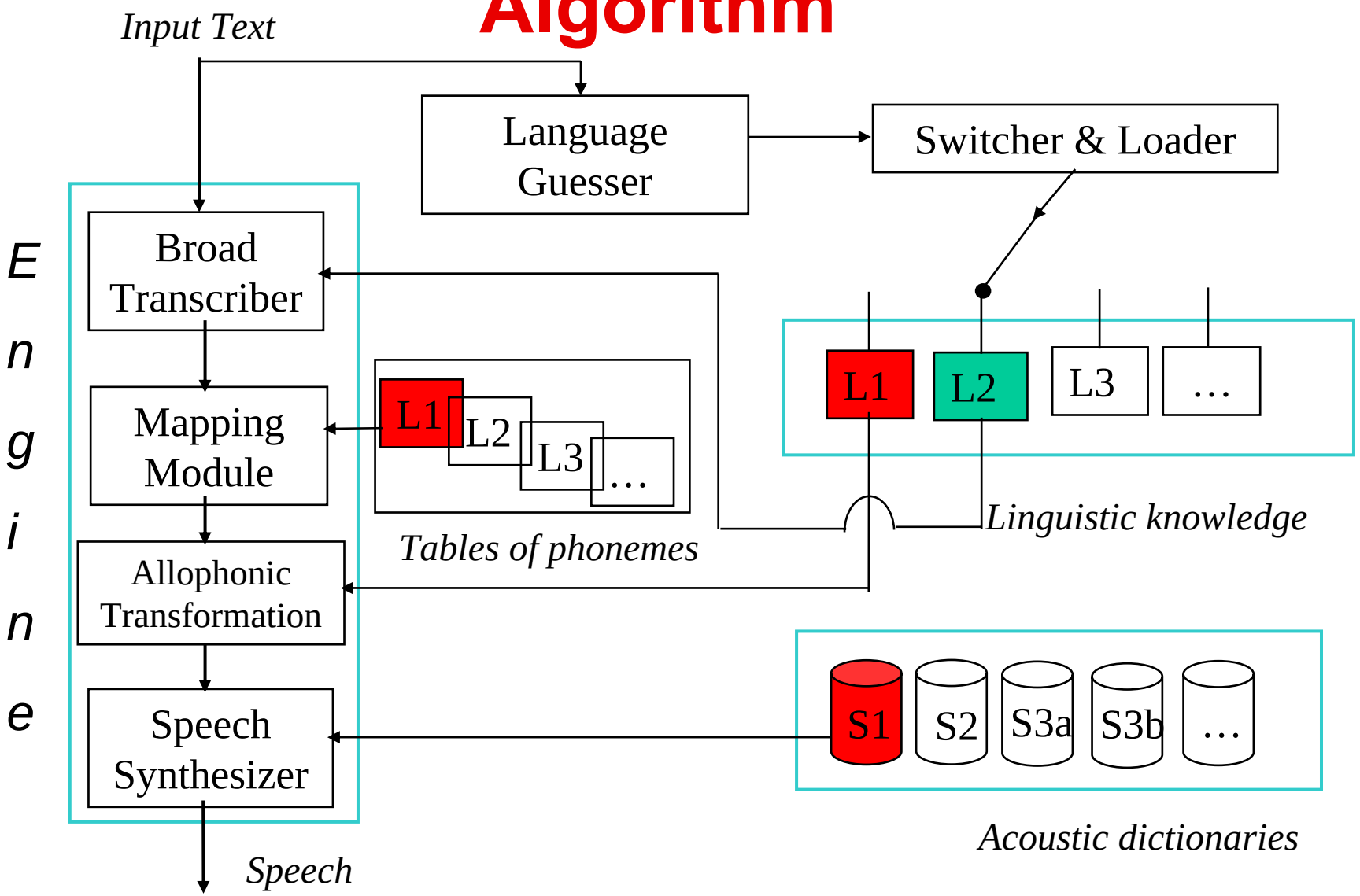
Many movies have been produced and filmed in Mexico, even though in many instances you would never know it. We can quote, for example: "Y tu mamá también", or "La última noche", that are examples of films shot in Mexico.

 *Change of voice*

 *Mixed language*

To deal with mixed-language text: web content, movie titles, addresses, e-mails, …

Flow of Phonetic Mapping Algorithm



An Example of Phonetic Mapping Algorithm

Hello *lang=French* **Mme Françoise Dupois,** *lang=* may I help you?

həlˈəʊ | madˈam fχɑ̃swˈɑz dypwˈɑ | mˈeɪ aɪ hˈelpʰju

French-to-English Phoneme Mapping

həlˈəʊ | maːdˈaːm fɹaːnswˈɑːz dupʰwˈɑː | mˈeɪ aɪ hˈelpʰju

English Allophone Processing

həlˈəʊ | maːdˈaːm fɹaːnswˈɑːz dupʰwˈɑː | mˈeɪ aɪ hˈelpɾju

English Phonetic Transcription

These examples are in IPA codes, if troubles (install Gentium fonts)
http://scripts.sil.org/gentium

Research Topics on TTS

- **How to modify the prosodic contour in a high quality TTS?**
  - Voice Morphing, Express emotions
    Ellen Eide et alii, in Narayanan, Alwan, 2005
  - More generally, how to modify or label the speech signal to simplify modifications
- **Cross-fertilization between ASR and TTS areas**

  Ostendorf, Bulyko, in Narayanan, Alwan, 2005
- **Speaking style and domain dependencies**

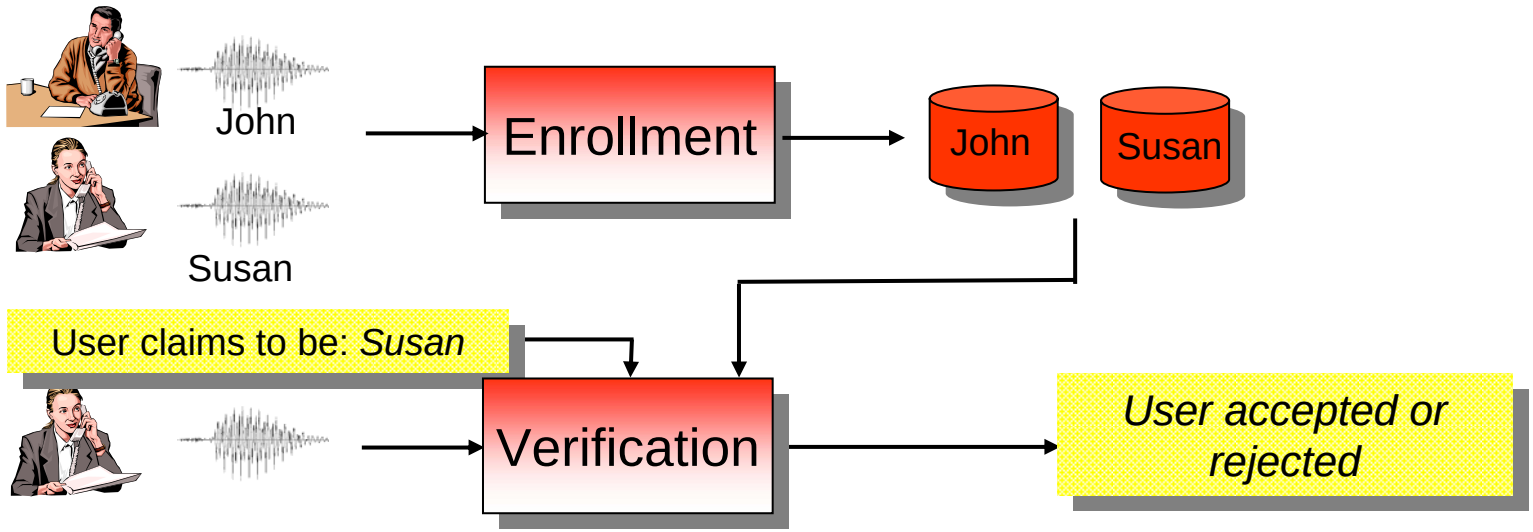
Suggested reading:

S. Narayanan, A. Alwan (eds.), "Text to speech synthesis: new paradigms and advances", Prentice Hall, 2005.

SIV: Speaker Identification and Verification

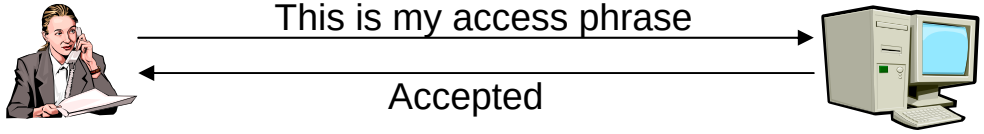
Speaker Verification

- **A non intrusive biometric measure**
  - **Every person has a unique voice print**
  - **High user acceptance**
    - **No password or pin codes to remember**
  - **Easy to integrate**
    - **It doesn't require any specific equipment**

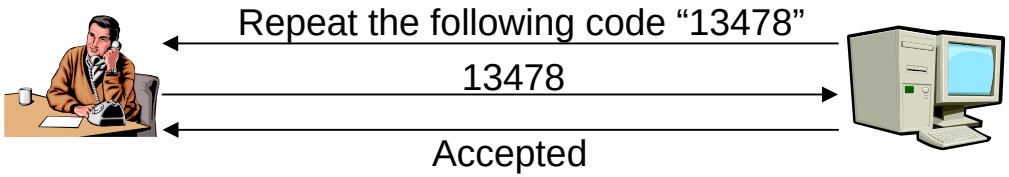


Speaker Verification Modes

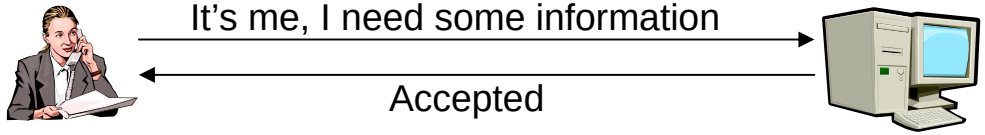
- **Text dependent**
  - Carried out on the basis of a specific text
  - User can choose text or the system can suggest it



- **Text prompted**
  - The system provides a “random” text for the user to repeat
  - No possibility of fraudulent access by means of recordings



- **Free Speech**
  - User is free to say anything



Multi-Verification and Identification

User is aware of verification

- **phone banking**
- **access to personal information**
- **…**

- **Verification: “Is Susan speaking?”**
  - User states identity
  - System verifies user’s assertion
- **Identification: “Who’s speaking?”**
  - The system knows a set of possible speakers (max. 50)
  - Users don’t state anything about their identity
  - System identifies the user that is currently speaking

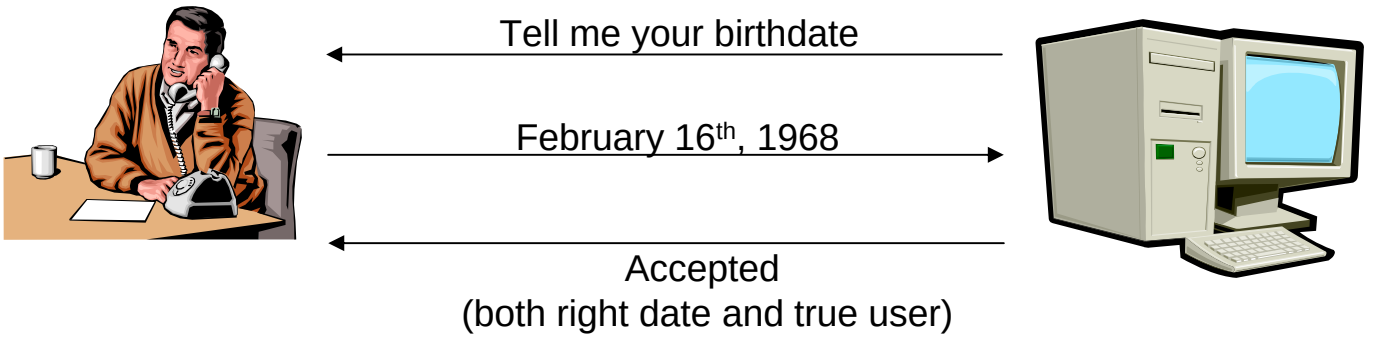
Products are specifically dedicated to identification

User is not aware of identification

- **hidden monitoring**
- **wire tapping**
- **verifying people on probation**
- **…**

Getting the Best from SV and ASR

- **Loquendo SV leverages ASR's acoustic models and NLP**
- **Verifies both the user's voice and the content of the vocal password**
- **System's overall security level is considerably increased**
- **Field of vocal password specified along with recognition grammars**
- **Highly precise application customization**



Voice Biometrics Security

Authenticated Access

Presence Monitoring

Forensic Analysis

LEA Investigations

Counter-terrorism Intelligence

CRM Applications

Loquendo
VOCAL TECHNOLOGY AND SERVICES

Free Speech Identification Technology

Accurate and easy-to-use tool based on Speaker Verification & Identification

- Voiceprint comparison within phone calls
- Language-independent mode
- Stand-alone and Networked solutions

Embedded TTS & ASR

Loquendo TTS and ASR: a Complete Product Line from Server to Smartphone

>1GHz	CPU	100MHZ
>256MB	RAM	4MB
>250MB	ROM	3MB

PC/Server
- Windows XP
- Windows 2000
- Windows 2003
- Linux
- Unix
- Solaris

Industrial PC
- Windows XP embedded Ver.
- Linux emb.

Tablet PC
- Windows XP tablet ver.
- Linux emb.

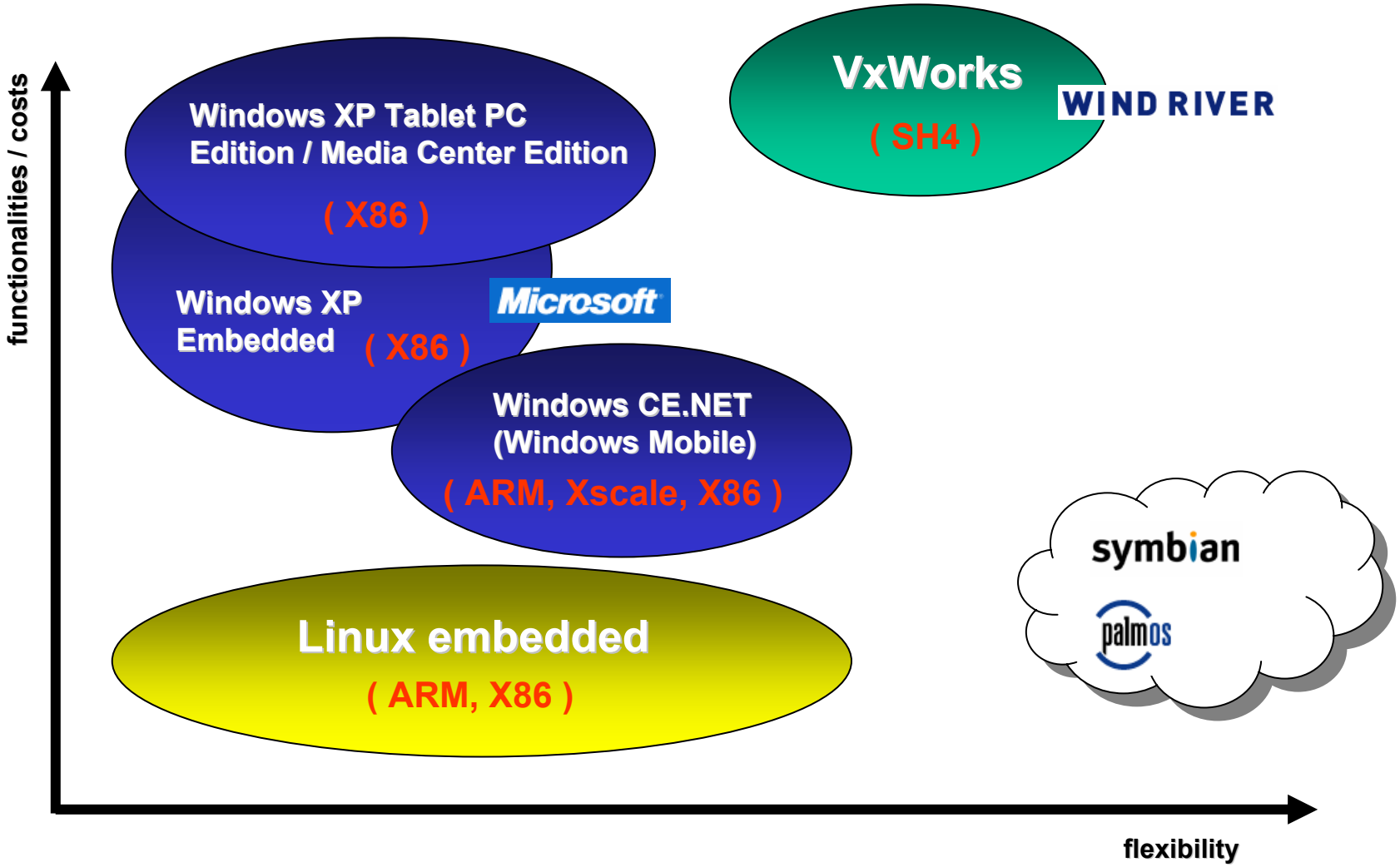
PDA
- Ce.net
- Linux emb.
- PPC03

Automotive
- VxWorks
- Ce.net
- Linux emb.

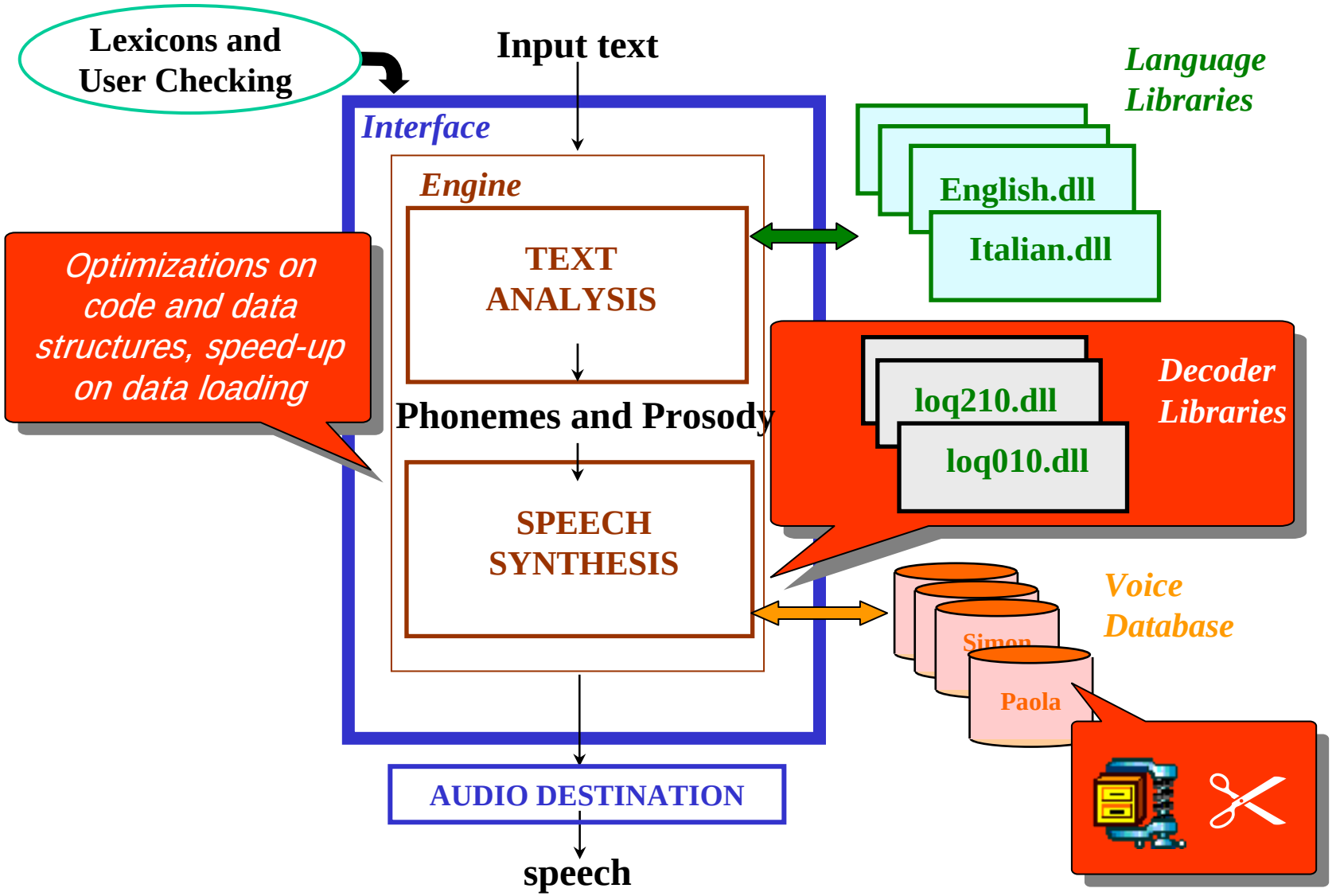
Set Top Box
- Linux emb.

Smartphone (TTS)
- Win Mobile 2003
- Symbian 7

Operating Systems and CPUs

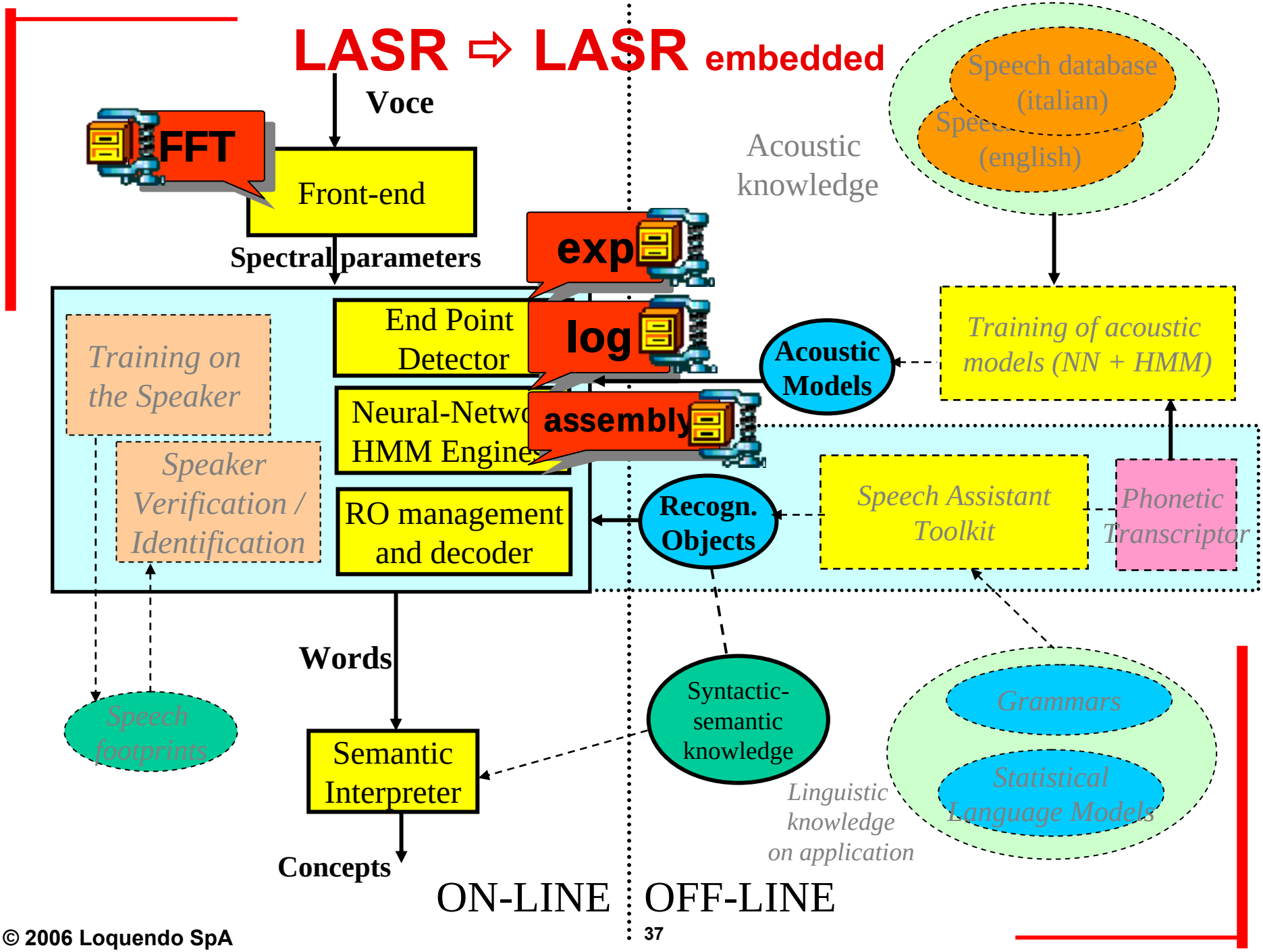


LTTS ⇨ LTTS embedded



Loquendo Embedded TTS Requirements

- **CPU:** Intel Pentium III, ARM, StrongARM, x86, Xscale, SH4
- **RAM:** 5 MB
- **Operating systems:** Windows NT/2000/XP, XP Embedded, XP Tablet PC, PocketPC 2002/2003, SmartPhone 2003, Windows CE.NET 4.2 (ARM, x86, XScale), VxWorks, Linux, Symbian
- **Interfaces:** Loquendo API, SAPI 5
- **Supported languages:** Italian, Castilian Spanish, French, German, Brazilian, Portuguese, Dutch, U.K. and U.S. English, Greek, Chilean, Argentinean, Swedish, Catalan, Mexican, Mandarin Chinese



Loquendo Embedded ASR Requirements

- **CPU:** Intel Pentium III, ARM, x86, XScale 624 Mhz or equivalent
- **RAM:** 20 MB
- **Operating systems:** Windows NT/2000/XP, XP Embedded, PocketPC 2003, Windows CE.NET 4.2 (ARM, x86, XScale), Linux
- **Interfaces:** Loquendo API
- **Grammar formalisms:**
  - JSGF (Java Speech Grammar Format)
  - W3C SRGS 1.0 (XML and ABNF Form)
  - RAU (Rule-based Automatic Understanding), next SISR
- **Supported languages:** Italian, Castilian Spanish, French, German, Brazilian, Portuguese, Dutch, U.K. and U.S. English, Greek, Chilean, Argentinean, Swedish, Catalan, Mexican

Minimal bibliography – ASR/NLP

- **ASR:**
  - X. Huang, A. Acero, H.-W. Hon, "**Spoken Language Processing: A Guide to Theory, Algorithm and System Development**", 980 pag.; Prentice Hall; ISBN: 0130226165; 1st edition (2001).
  - X. Huang, A. Acero, H.-W. Hon, R. Reddy, "**Spoken Dialogues with Computers: Signal Processing and Its Applications**", 702 pag.; Academic Press; ISBN: 0122090551; 1st edition (1998).
  - H.A. Bourlard, N. Morgan, "**Connectionist Speech Recognition – A Hybrid Approach**", 348 pag.; Kluwer Academic Publishers; ISBN: 0792393961; 1st edition (1994).
  - L. Rabiner, B.-H. Juang, "**Fundamentals of Speech Recognition**", 496 pag.; Prentice Hall; ISBN: 0130151572; 1st edition (1993).

- **Natural Language and Dialog:**
  - D. Jurafsky, J.H. Martin, "**Speech and Language Processing (2nd Edition)**"; Prentice Hall; ISBN: 0131873210; 2nd edition (2006).
  - D. Dahl, "**Practical Spoken Dialog Systems (Text, Speech and Language Technology)**"; Springer ed.; ISBN: 1402026757; 1st edition (2005).
  - J. Allen, "**Natural Language Understanding (2nd Edition)**"; Addison-Wesley Pub Co; ISBN: 0805303340; 2nd edition (1995).

Minimal bibliography – TTS/VUI

- **Text-to-Speech:**
  - S. Narayanan, A. Alwan , "**Text to Speech Synthesis : New Paradigms and Advances**", Prentice Hall PTR; ISBN: 013145661X; (2004).
  - T. Dutoit, "**An Introduction to Text-To-Speech Synthesis (Text, Speech and Language Technology, V. 3)**", Kluwer Academic Publishers; ISBN: 0792344987; (1997).
  - J. Allen, "Overview of Text-to-speech systems" , in Furui, Sondhi "**Advances in Speech Signal Processing**", Marcel Dekker; ISBN: 0824785401; (1991).
  - D.H. Klatt, "Review of text-to-speech conversion for English", **J. Acoustic. Soc. Am.**, vol. 82, n. 3, Sept. 1987.
  - J.L. Flanagan, "**Speech Analysis Synthesis and Perception**", Springer-Verlag; New York; (1972).

- **Vocal User Interfaces:**
  - B. Balentine, D. P. Morgan, W. S. Meisel, " **How to Build a Speech Recognition Application: Second Edition: A Style Guide for Telephony Dialogues**", 414 pag.; Enterprise Integration Group; ISBN: 0967127823; 2nd edition (2001).
  - S. Weinschenk, D. T. Barker, " **Designing Effective Speech Interfaces**", 406 pag.; John Wiley & Sons; ISBN: 0471375454; 1st edition (2000).

Loquendo's Selected Papers – ASR/NL/LM

- **ASR:**
  - R. Gemello, F. Mana, S. Scanzio, P. Laface, R. De Mori, "**Adaptation of Hybrid ANN/HMM models using hidden linear transformations and conservative training**", in ICASSP-2006.
  - R. Gemello, F. Mana, D. Albesano, R. De Mori, "**Multiple resolution analysis for robust automatic speech recognition**", Computer Speech & Language, Vol. 20, No. 1, 2006.
  - R. Gemello, F. Mana, R. De Mori, "**Automatic Speech Recognition With a Modified Ephraim-Malah Rule**", IEEE Signal Processing Letters, Vol. 13, No. 1, 2006.
  - D. Colibro, L. Fissore, C. Vair, E. Dalmasso, P. Laface, "**A Confidence Measure Invariant to Language and Grammar**", in INTERSPEECH-2005.
  - D. Colibro, L. Fissore, C. Popovici, C. Vair, P. Laface, "**Learning Pronunciation and Formulation Variants in Continuous Speech Applications**", in ICASSP-2005.
  - C. Popovici, M. Andorno, P. Laface, L. Fissore, M. Nigra, C. Vair, "**Learning New User Formulations in Automatic Directory Assistance**", in ICASSP-2002.
  - R. Gemello, D. Albesano, F. Mana, L. Moisa, "**Multi-source Neural Networks for Speech Recognition: a Review of Recent Results**", in IJCNN-2000.
  - L. Fissore, F. Ravera, P. Laface, "**Acoustic-Phonetic Modeling for Flexible Vocabulary Speech Recognition**", in Eurospeech-1995.

- **Natural Language/Language Modelling:**
  - C. Popovici, P. Baggia, P. Laface, L. Moisa, "**Automatic Classification of Dialogue Contexts For Dialogue Predictions**", in ICSLP-98.
  - C. Popovici, P. Baggia, "**Specialized Language Models Using Dialogue Predictions**"; in ICASSP-97.
  - P. Baggia, C. Rullent, "**Partial Parsing as a Robust Parsing Strategy**", in ICASSP-93.
  - P. Baggia et al., "**Improving Speech Understanding Performance through Fedback Verification**", Speech Communications, Vol. 11, 1992.

Loquendo's Selected Papers – TTS/Dial

- **TTS:**
  - P. Massimino, A. Pacchiotti, "**An automaton-based machine learning technique for automatic phonetic transcription**", in INTERSPEECH-2005.
  - P. Massimino, "**From Marked Text to Mixed Speech and Sound**", in ICAD-2005.
  - L. Badino, "**Chinese Text Word-Segmentation Considering Semantic Links among Sentences**", in ICSLP-2004.
  - L. Badino, C. Barolo, S. Quazza, "**A General Approach to TTS Reading of Mixed-Language Tests**", in ICSLP-2004.
  - Zovato, A. Pacchiotti, S. Quazza, S. Sandri, "**Towards emotional speech synthesis: a rule based approach**", in Speech Synthesis Workshop, Pittsburgh, 2004.
  - S. Quazza, L. Donetti, L. Moisa, P.L. Salza, "**ACTOR®: a multilingual unit-selection speech synthesis system**", in ISCA Tutorial and Research Workshop on Speech Synthesis, 2001.
  - M. Balestri, A. Pacchiotti, S. Quazza, P.L. Salza, S. Sandri, "**Choose the best to modify the least: a new generation concatenative synthesis system**", in Eurospeech-1999.
  - S. Quazza, P.L. Salza, S. Sandri, A. Spini, "**Prosodic Control in a Text-to-Specch system for Italian**", in ESCA Workshop on Prosody, in Working Papers, Lund University Department of Linguistics, Vol. 41, 1993.

- **Spoken Dialog and Evaluation:**
  - P. Baggia, G. Castagneri, M. Danieli, "**Field trials of the Italian ARISE Train Timetable System**", Speech Communication, Vol. 31, 2000.
  - P. Baggia, M. Danieli, "**CSELT Approaches to Spoken Dialogue**", in SPECOM-1998.
  - M. Danieli, "**On the use of Expectation and Repairing Human-Machine Miscommunications**", in AAAI *Workshop on Detecting, Preventing, and Repairing Miscommunications*, 1996.
  - M. Danieli, E. Gerbino, "**Metrics for evaluating dialogue strategies in a Spoken language system**", in AAAI Spring Symposium *Empirical Methods in Dialogue and Discourse*, 1995.

The Present –
The Impact of Standards

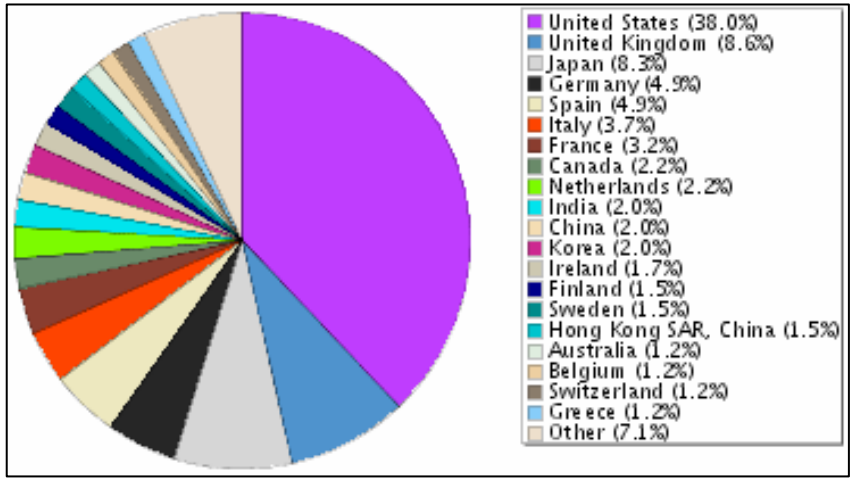
- The (r)evolution of VoiceXML
- A Constellation of W3C Standards
- Next steps

Why do Standards Matter?

- Are speech standards well known in research institutions? In Europe? In US? In other countries?
- Research may be advantaged to build upon standards.
- Results of research must be available soon and fit well in a broader picture.
- Increase the contributions from the research institutions to the standards bodies.

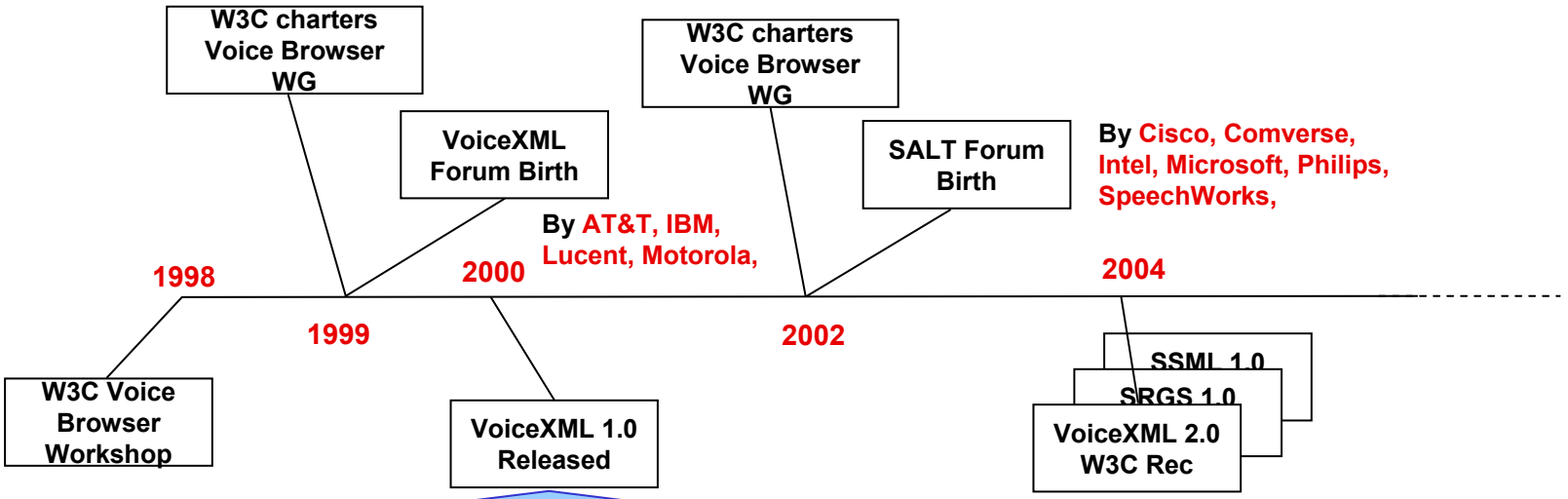
(World Wide Web consortium) in Pills

- Founded in 1994, by Tim Berners-Lee with a mission to lead the Web to its full potential
- 400 members all over the world, 50 Working, Interest and Coordination Groups
- Staff based in MIT (USA), ERCIM (France), Keio Univ (Japan)
- W3C is where the framework of today's Web is developed (HTML, CSS, XML, DOM, SOAP, RDF, OWL, VoiceXML, SVG, XSLT, P3P, XML)
- Special attention to Web Accessib. and Internationalization issues



To learn more, please visit:
http://www.w3.org and click on "Join W3C"

The (r)evolution of VoiceXML 1998 - 2004



Preparing to announce VoiceXML 1.0
Friday Feb. 25th, 2000
Lucent, Naperville, Illinois

Left to right: Gerald Karam (AT&T), Linda Boyer (IBM), Ken Rehor (Lucent), Bruce Lucas (IBM), Pete Danielsen (Lucent), Jim Ferrans (Motorola), Dave Ladd (Motorola).

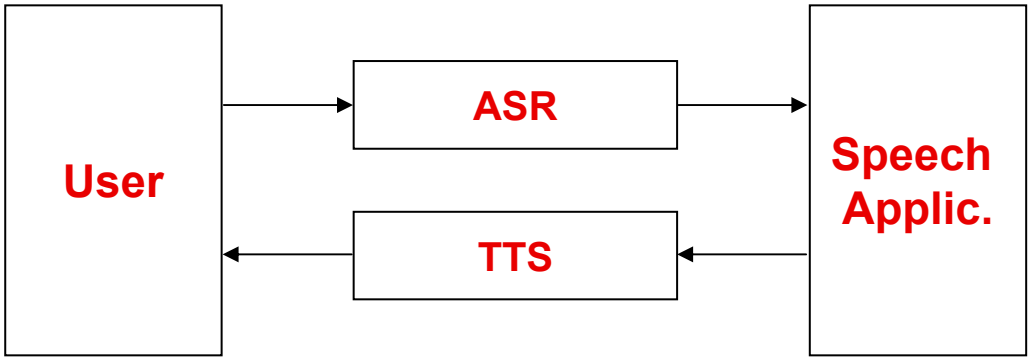
The VoiceXML Impact

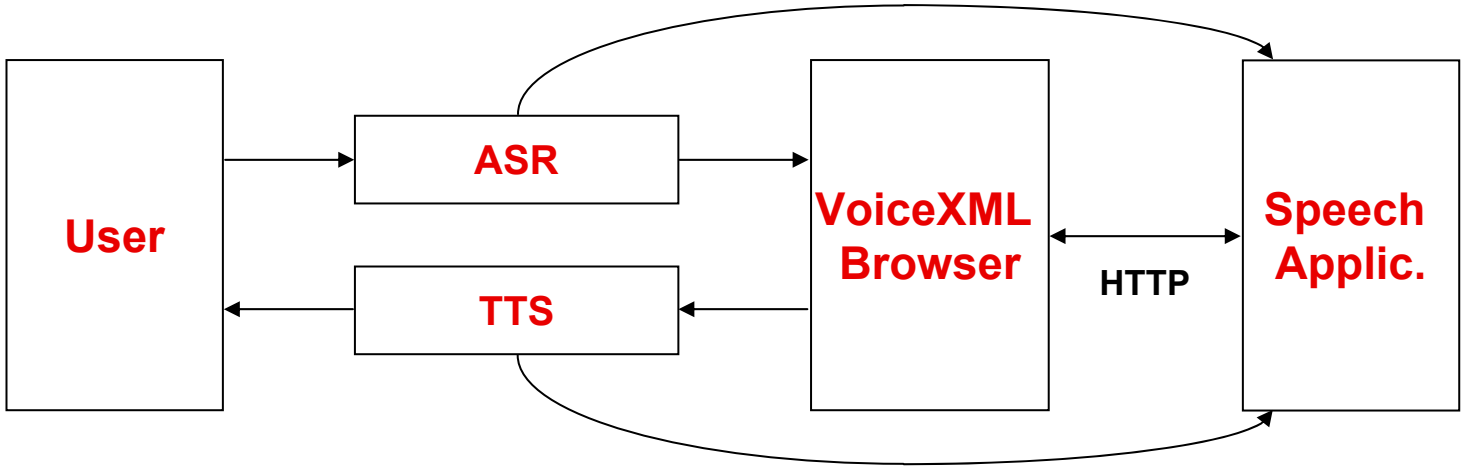
- **Changed the landscape of IVRs and, more generally, of speech applications**
  - From proprietary to standard-based speech applications

Why?

- **Takes the web paradigm to the core of speech applications development**
- **Powerful abstraction – Easy to author**
- **Delegates to other standards for ASR, TTS, call control:**
  - **Speech Recognition Grammar Specification (SRGS)**
  - **Semantic Interpretation for Speech Recognition (SISR)**
  - **Speech Synthesis Markup Language (SSML)**
  - **Pronunciation Lexicon Specification (PLS)**
  - **Call Control XML (CCXML)**

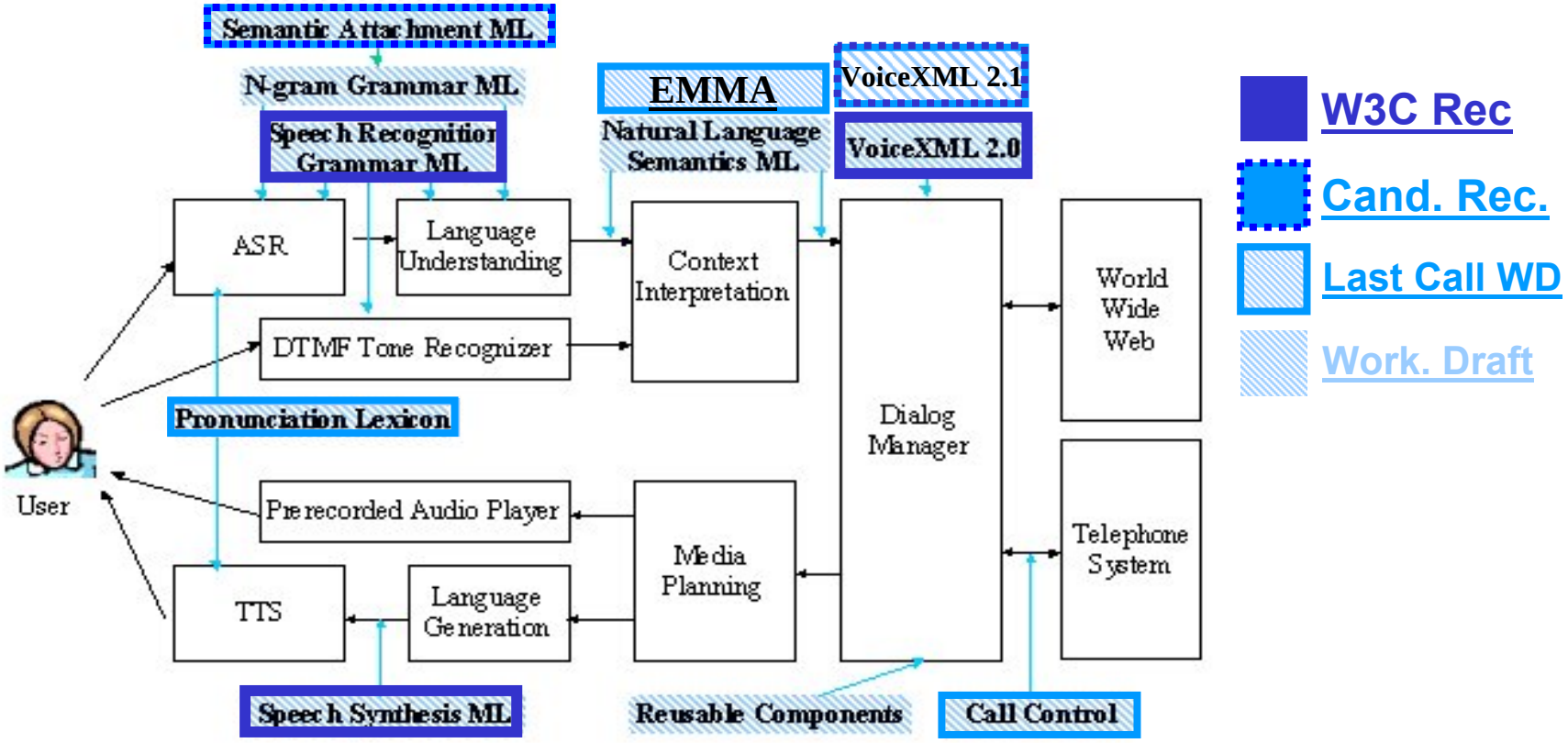
The VoiceXML Architecture



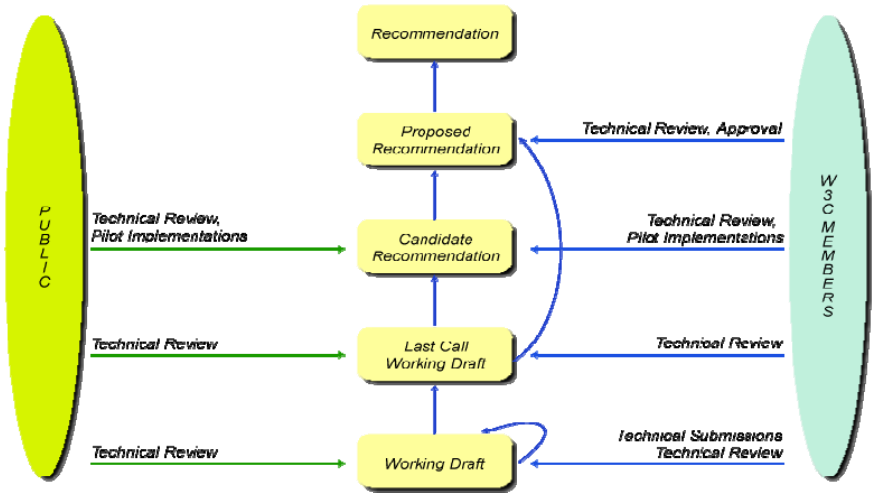


Speech Interface Framework (by Jim Larson, 2000)

- **A Constellation of W3C Standards for:**
  - **ASR and DTMF:** **SRGS, SISR and N-grams, PLS**
  - **TTS and prerecorded audio:** **SSML, PLS**
  - **Dialog management:** **VoiceXML 2.0, 2.1**
  - **Call Control:** **CCXML**



Status of W3C Speech Interface Languages



W3C Recommendation

Proposed Recommendation

Candidate Recommendation

Last Call Working Draft

Working Draft

Requirements

VoiceXML 2.0

SRGS 1.0 (speech grammars)

SSML 1.0 (speech synthesis)

SISR 1.0 (semantic interpret-ration)

PLS 1.0 (pronunc. lexicon)

CCXML 1.0 (telephony, call control)

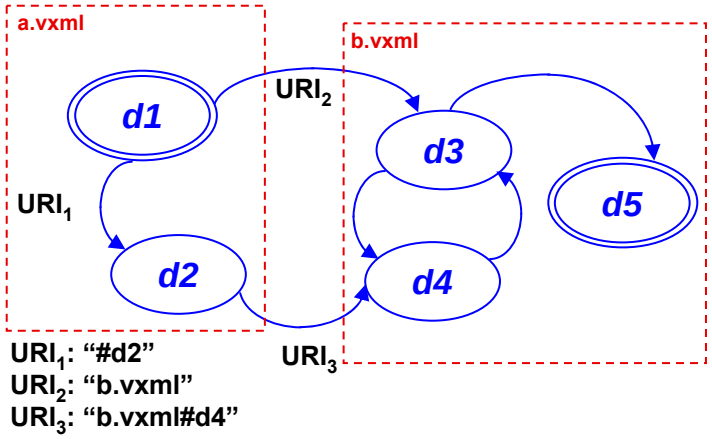
VoiceXML 2.1

VoiceXML 2.0 Features

- **Menus, forms, sub-dialogs**
  - <menu>, <form>, <subdialog>
- **Input**
  - Speech recognition <grammar>
  - Recording <record>
  - Keypad <grammar mode="dtmf">
- **Output**
  - Audio files <audio>
  - Text-To-Speech <prompt>
- **Variables (ECMA-262)**
  - <var>, <assign>, <script>
  - scoping rules
- **Events**
  - <nomatch>, <noinput>, <help>, <catch>, <throw>
- **Transition and submission**
  - <goto>, <submit>
- **Telephony**
  - Connection control <transfer>, <disconnect>
  - Telephony information
- **Platform**
  - Objects
- **Performance**
  - Fetch

http://www.w3.org/TR/voicexml20/

VoiceXML 2.0 – Dialog Trans. Network



VoiceXML Dialog:

- either a <menu> or a <form>
- <form> may contain many <field>s
- <field> selection is determined by FIA algorithm

- **System guided**
  - A single dialog is active at a time
  - In a form, field by field acquisition
  - Optional presence of general commands: help, exit, cancel, application defined

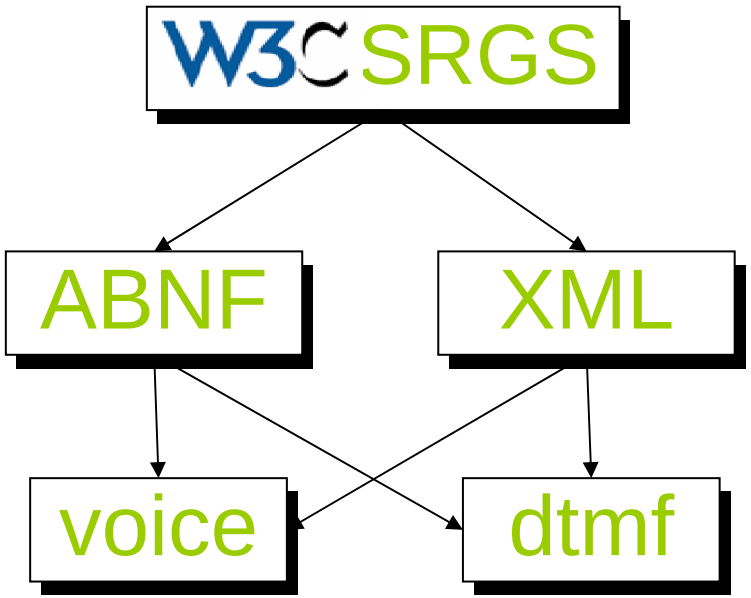
- **Mixed initiative**
  - Dialogs with form level grammars
    ➔ all the fields may by recognized in a single sentence
  - More than one dialog active at the same time

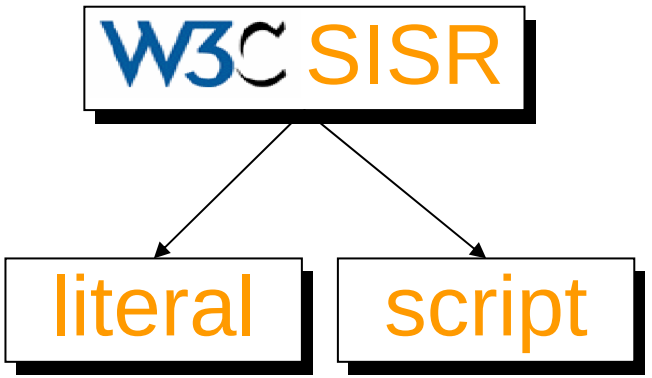
Standards Push on Speech Grammars

SYNTAX
Defines constraints on admissible sentences for a specific recognition turn

Speech grammar

SEMANTICS
Describes how to produce results after an utterance is recognized

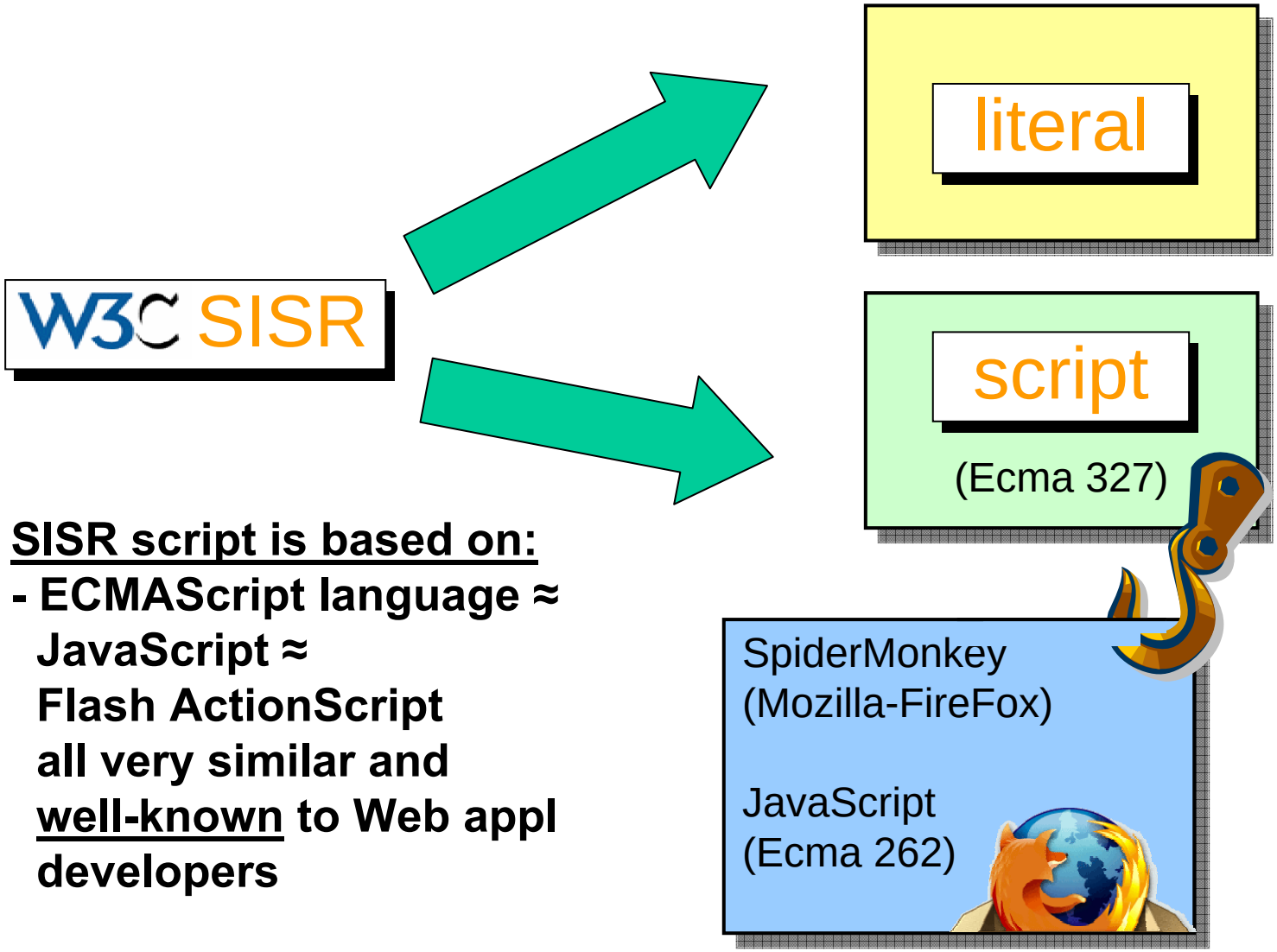




http://www.w3.org/TR/speech-grammar/

http://www.w3.org/TR/semantic-interpretation/

Evolution of the Semantic Interpretation in SISR



SISR script is based on:

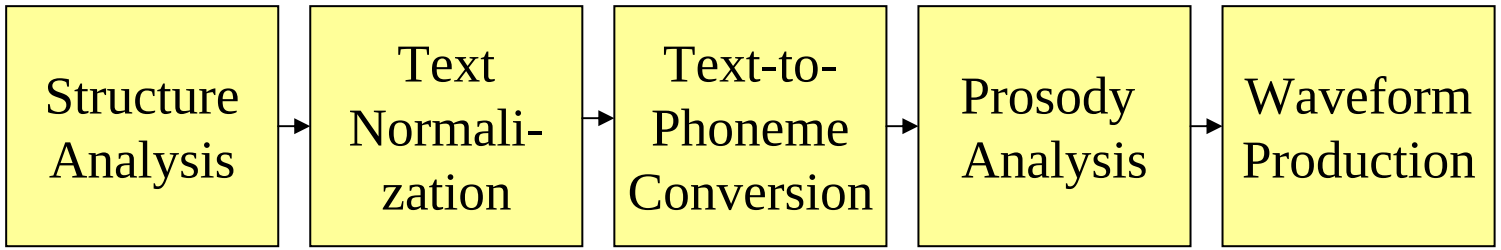
- **ECMAScript language ≈ JavaScript ≈ Flash ActionScript all very similar and well-known to Web appl developers**

Examples SRGS Grammars with SISR

	SRGS XML	SRGS ABNF
SISR literal	<?xml version="1.0" encoding="UTF-8"?> <grammar xml:lang="en-US" version="1.0" xmlns="http://www.w3.org/2001/06/grammar" tag-format="semantics/1.0-literals"> <rule id=”main” scope=”public”> <token>Torino</token> <tag>10100</tag> </rule> </grammar>	#ABNF 1.0 iso-8859-1; mode voice; tag-format <semantics/1.0-literals>; public $main = Torino {10100} ;
SISR script	<?xml version="1.0" encoding="UTF-8"?> <grammar xml:lang="en-US" version="1.0" xmlns="http://www.w3.org/2001/06/grammar" tag-format="semantics/1.0"> <tag>var a=7;</tag> <rule id=”main” scope=”public”> <token>Torino</token> <tag>out=”10100”;</tag> </rule> </grammar>	#ABNF 1.0 iso-8859-1; mode voice; tag-format <semantics/1.0>; {var a=7;}; public $main = Torino {out=“10100”;} ;

Speech Synthesis with SSML

- **SSML is a markup language (XML-based) for encoding content to be synthesized**
- **SSML allows the enhancement of TTS rendering:**
  - Explicit structure of content (<s>, <p>)
  - Text normalization: substitutions (<sub>, <lexicon>)
  - Emphasize and pause (<emphasis>, <break>)
  - Phonetic pronunciation (<phoneme>, <lexicon)
  - Computation of prosodic parameters (<prosody>)
  - Change voice on gender, age, language (<voice>)
  - Include audio files and marker events (<audio>)
  - Generation of marker events (<mark>)



http://www.w3.org/TR/speech-synthesis/

Pronunciation Lexicon Specification (PLS)

- **A standard Pronunciation Lexicon format (XML-based)**
- **PLS designed to be used in both TTS and ASR**
  - SRGS grammars already allow access to external lexicons
  - SSML documents allow reference to more than one external lexicon
- **PLS allows:**
  - Multiple pronunciations for ASR
  - Pronunciation expressed in IPA or other phonetic alphabet
  - Textual substitutions (acronyms, ambiguities, etc.)

http://www.w3.org/TR/pronunciation-lexicon/

```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0"
  xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
  alphabet="ipa" xml:lang="it-IT">
  <lexeme>
    <grapheme>file</grapheme>
    <phoneme>fa&#x026A;l</phoneme>
    <!--  The pronunciation of the English word: "file"
          in an Italian text. In IPA it is: "faɪl". -->
  </lexeme>
  <lexeme>
    <grapheme>EU</grapheme>
    <alias>Unione Europea</alias>
  </lexeme>
</lexicon>
```

Other Related Standards …

- **CCXML – Call Control for Voice Browser**
  http://www.w3.org/TR/ccxml/
- **VoiceXML 2.1 - 8 features to extend VoiceXML 2.0**
  http://www.w3.org/TR/voicexml21/

- **Work in progress on a new generation of standards:**
  - Easy to integrate in other W3C languages, i.e. voice enabling SMIL, SVG, XForms
  - Extended features, i.e. speaker verification, etc.
- **Extensive interest on dialog authoring languages or tools, which generate VoiceXML or the future extended VB languages.**

Final Remarks - VoiceXML

The changed landscape for speech application development:

- Virtually all the IVRs today support VoiceXML applications
- New options related to VoiceXML:
  - SIP-based VoiceXML platforms (Loquendo, Voxpilot, Voxeo, VoiceGenie)
  - Large hosting of speech applications (TellMe, Voxeo)
  - Development tools (VoiceObjects, Audium, SpeechVillage, Unisys, etc.)

➔ Further changes may come from the CCXML adoption

… but:

- Mainly system driven applications are actually deployed
- New challenges to incorporate more powerful dialog strategies, mixed-initiative are under discussion.

V3 or VoiceXML 3.0:

- First result a standard markup for encoding state charts (SCXML, http://www.w3.org/TR/scxml/)
- Many pending extensions to VoiceXML: Speech biometrics, Media control, etc.

Final Remarks – Speech Standards

- **ASR standards:**
  - Adoption of the PLS lexicon and possible extensions
  - N-grams in a standard format?
    - There is a first working draft, then no more progress. Should we restart it?
    - Standard mixtures of N-grams and SRGS grammars
    - Uniform semantic interpretation? Based on SISR?

- **TTS standard - SSML:**
  - Adoption of the PLS lexicon and possible extensions
  - Work on Internationalization of SSML:
    - For Eastern languages (Workshop Beijing – Nov. '05)
    - For Middle-East, Eastern EU languages, India (Work. Crete – May '06)
    - Other workshops? India?
  - More fine-grained control of TTS

VoiceXML Resources

- **Voice Browser Working Group (spec, FAQ, implementations): http://www.w3.org/Voice/**
- **VoiceXML Forum site (resources, education, interest groups): http://www.voicexml.org/**
- **VoiceXML Forum Review: http://www.voicexmlreview.org/**
  - **Interesting articles related to VoiceXML and more**
  - **Example code in the sections "First Words" and "Speak & Listen"**
- **Ken Rehor's World of VoiceXML http://www.kenrehor.com/voicexml**
- **Online documentation related to VoiceXML Platforms**
  - **Loquendo Café, Voxeo (http://www.vxml.org/), TellMe, VoiceGenie**
- **Many books on VoiceXML:**
  - Jim Larson, "VoiceXML Introduction to Developing Speech Applications", Prentice-Hall, 2002.
  - A. Hocek, D. Cuddihy, "Definitive VoiceXML", Prentice-Hall, 2002

VoiceXML Forum

- **Goal:**
  Support and promote the diffusion of VoiceXML and related standards
- **Home-page: http://www.voicexml.org**
- **Members:**
  - Board of Directors (4 founders: AT&T, IBM, Lucent, Motorola and 10 sponsor members, including Loquendo)
  - Promoters: 28 members
  - Supporters: 117 members
- **Current activities:**
  - VoiceXML Platform Certification:
    - 14 platforms have been certified
      http://www.voicexml.org/platform_certification/certified_platforms.html
  - VoiceXML Developer Certification:
    - Near 100 certified people around the world
  - Committees (with mailing lists):
    - **Education**:
      Publication of the VoiceXML Review online, courses and Webinars
    - **Marketing**:
      Presence in conferences, journals, etc.
    - **Tools**:
      Standardize logs for ASR and VoiceXML Platforms to allow the development of tools
    - **Biometrics**:
      Work on having speech to meet the other biometric standards
    - **Technical**:
      Core of the Forum's technical activities and planning + MRCP group

Loquendo
VOCAL TECHNOLOGY AND SERVICES

The Future

- Voice and Video Applications
- Speech and SemanticWeb
- Towards Multimodality

Voice and Video Applications

Interactive Video Services

Renewed interest in video, examples:

- 3G handsets (10% today, 60% in 2010)
- 3G network ➔ IMS architecture
- Fixed-mobile convergence

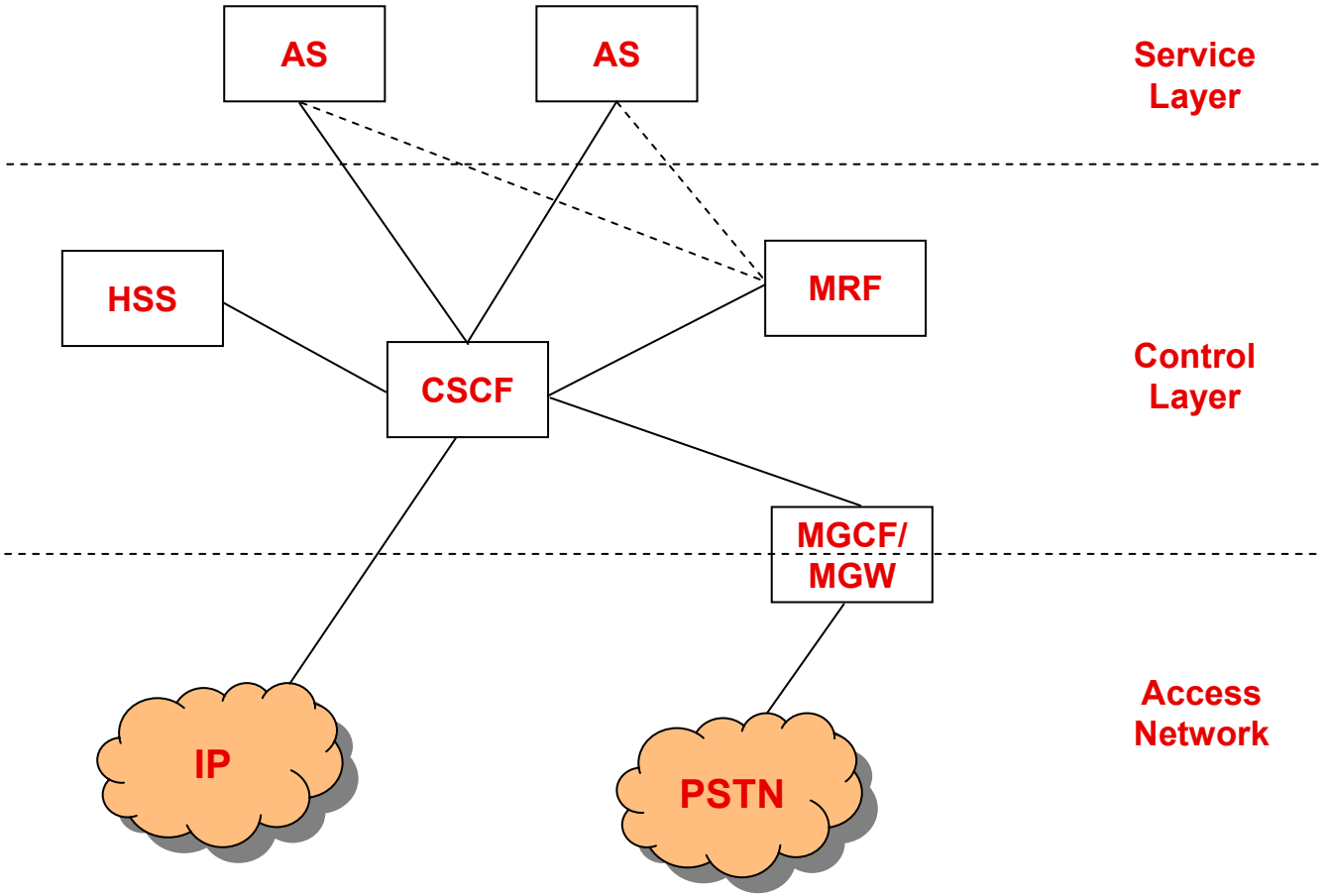
Renewed interest in video, examples:

- Video content streaming
- Video contact center
- Video blogging
- Video mail
- DVB-H back channel

3GPP IMS Architecture

Legenda:

- AS – Application Server
- CSCF – Call Session Control Function
- HSS – Home Subscriber Service
- IP – Internet Protocol
- MGCF – Media Gateway Control Function
- MRF – Media Resource Function
- PSTN – Public Switch Telephone Network



IMS - Media Resource Function

- **Centralized media processing capabilities**
- **Outside IMS is generically called Media Server**
- **Two sub-nodes:**
  - MRFC – Media Resource Function Control
    exposes a SIP Interface to CSFC
    invoked by AS via CSFC
  - MRFP – Media Resource Function Processing
    includes media processing capabilities
    exposes SIP interface for control and RTP for media transport

Examples of Media processing services:

- **Media streaming origination: announcements, TTS messages**
- **Media streaming processing: DTMF/speech recognition, audio/video recording, transcoding**
- **Media mixing (audio/video conferencing)**

VoiceXML for Interactive Video

- **VoiceXML powerful language for programming voice and video**
- **VoiceXML is largely independent of media**
- **Media processing delegated to external server**

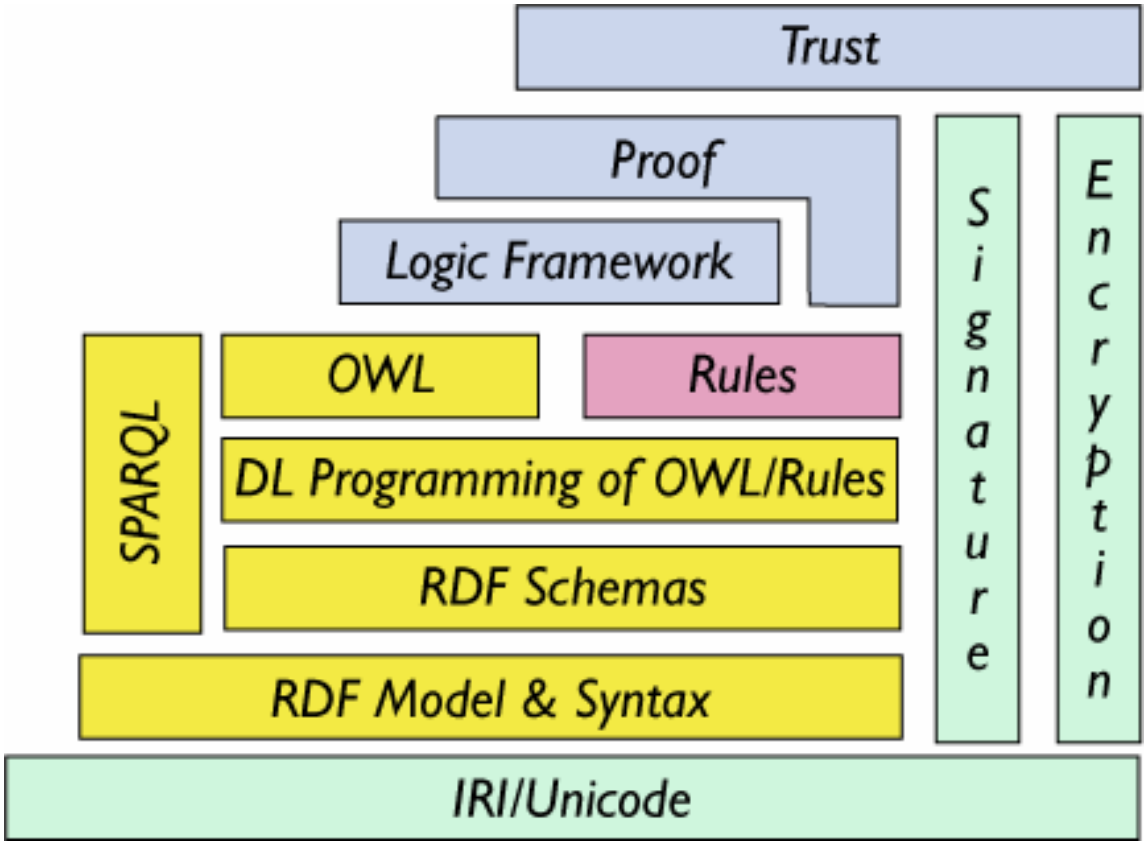
- **VoiceXML 3.0 pending extensions:**
  - Generalize <audio> to <media> element
  - Generalize <record> from audio to video

- **Big issue:**
  - How to add more powerful synchronization?
    ➔ the adoption of a subset of SMIL 2.1 is under evaluation

Speech and SemanticWeb

W3C SemanticWeb

A stack of W3C activities:



Speech Applications and SemanticWeb

- **Large applications depend on extensive knowledge**
- **Quality of speech recognition strictly depends on the access to real world data**
- **Issues:**
  - SemanticWeb as a source of lexical domain knowledge
  - How to encode results of speech and natural language in a SemanticWeb language

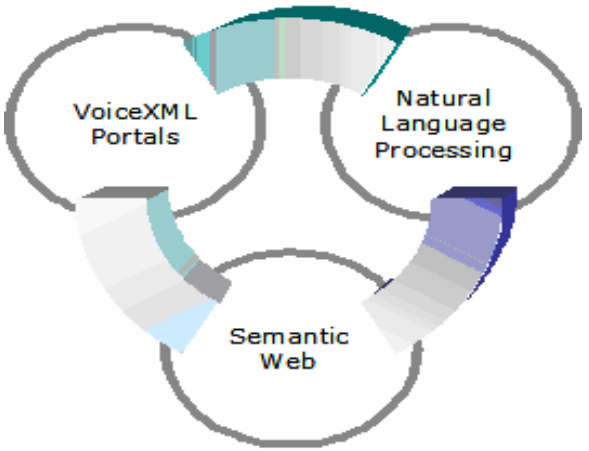
HOPS ICT Project

HOPS = Enabling an Intelligent Natural Language Based Hub for the Deployment of Advanced Semantically Enriched Multi-channel Mass-scale Online Public Services

HOPS is a three-year project focused on the deployment of advanced ICT enabled "voice-enabled front-end public platforms" in Europe permitting access for European citizens to their nearest Public Administration. Started in 2003.

Technologies involved:

- **VoiceXML Platform**
- **Natural Language Processing**
- **Semantic Web Technologies**



Integration between technologies

Critical Expectations

- It is still difficult to find a common basis among different research areas, especially if the difficult issues might be tackled on more than one area
- Some HOPS results have been obtained by enriching the semantic results of text/speech parsing to access an ontology
- Open Issues:
  - ➔ How to enrich grammars with real data?
  - ➔ Is the SemanticWeb useful for speech?

Towards Multimodality

Multiple Interaction Modes

- Visual
  - ✓ Persistent
  - ✓ Visual effects (graphics, animations, ...)
  - ✓ Limited on small devices
  - ✓ Difficult to read, hands-free
- Vocal
  - ✓ Transient
  - ✓ Audio effects (prosody, emphasis, music, ...)
  - ✓ Effective on small devices
  - ✓ Sensitive to environmental noise
- Keyboard
  - ✓ Standard Keyboard / Phone Keypad (DTMF)
  - ✓ Difficult on small devices
- Stylus/Pen
  - ✓ Handwriting
  - ✓ Need two hands on small devices

Use of speech can complement other modalities, especially on small devices

Class of Devices

- Thick Clients (Desktop, ...)
  - ✓High processing and available memory, big display and standard keyboard
  - ✓Typically local applications
  - ✓Limited need for alternative modalities, but useful for special needs or accessibility reasons

- Medium Client (PDA, ...)
  - ✓ Medium processing and memory, medium size display, difficult input by stylus or keypad
  - ✓ Embedded capabilities of speech and handwriting recognition
  - ✓ Both local and remote applications
  - ✓ *Medium/high need for speech modality*

- Thin Clients (Smart Phones, ...)
  - ✓ Limited processing and memory, small display, keypad, audio
  - ✓ Limited resources for complex speech recognition, OK for TTS
  - ✓ Server based applications
  - ✓ Greater need for alternative modalities

W3C MMI Framework

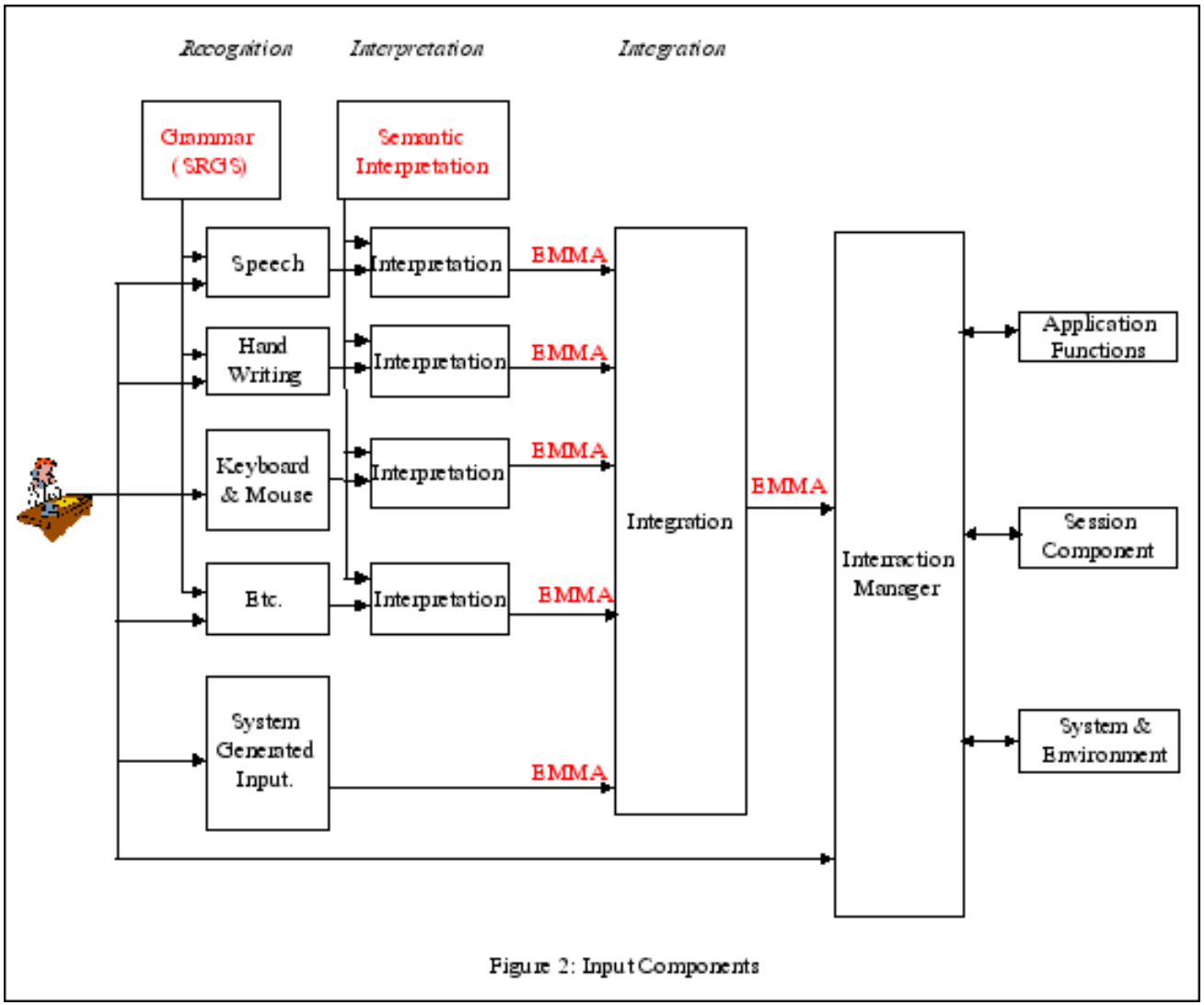


Figure 2: Input Components

W3C Multimodal Interaction WG

- **Home-page: http://www.w3.org/2002/mmi/**
  52 participants of 29 organizations, plus 1 Invited Speaker

- **First contributions:**
  - MMI Use Cases: http://www.w3.org/TR/mmi-use-cases/
  - MMI Interaction Framework: http://www.w3.org/TR/mmi-framework/

- **Ongoing activities:**
  - EMMA specification: http://www.w3.org/TR/emma/
    XML data exchange format between input processors and Interaction Management. Very rich annotations for input modalities, such as speech, handwriting or gesture.
  - InkML specification: http://www.w3.org/TR/InkML
    XML data exchange format for ink entered with an electronic pen or stylus as part of a multimodal system.
  - MMI Architecture & Interf.: http://www.w3.org/TR/mmi-arch/
    A loosely coupled architecture that focuses on providing a general means for modality components to communicate with each other, and with an Interaction Manager.
  - MMI Appl. Devel. Feedbacks: http://www.w3.org/TR/mmi-dev-feedback/
    After several years of multimodal application development, application developers provide detailed feedbacks about what they like, dislike, and want to see improve and continue

AURORA Distributed Speech Recognition

ETSI

3GPP

Client Devices

[Wireless] Packet Data Network

IP Network

Content Servers

Voice Gateway / Server:
- VoiceXML / SALT / X+V Browser
- Speech Resources (ASR, TTS, etc.)

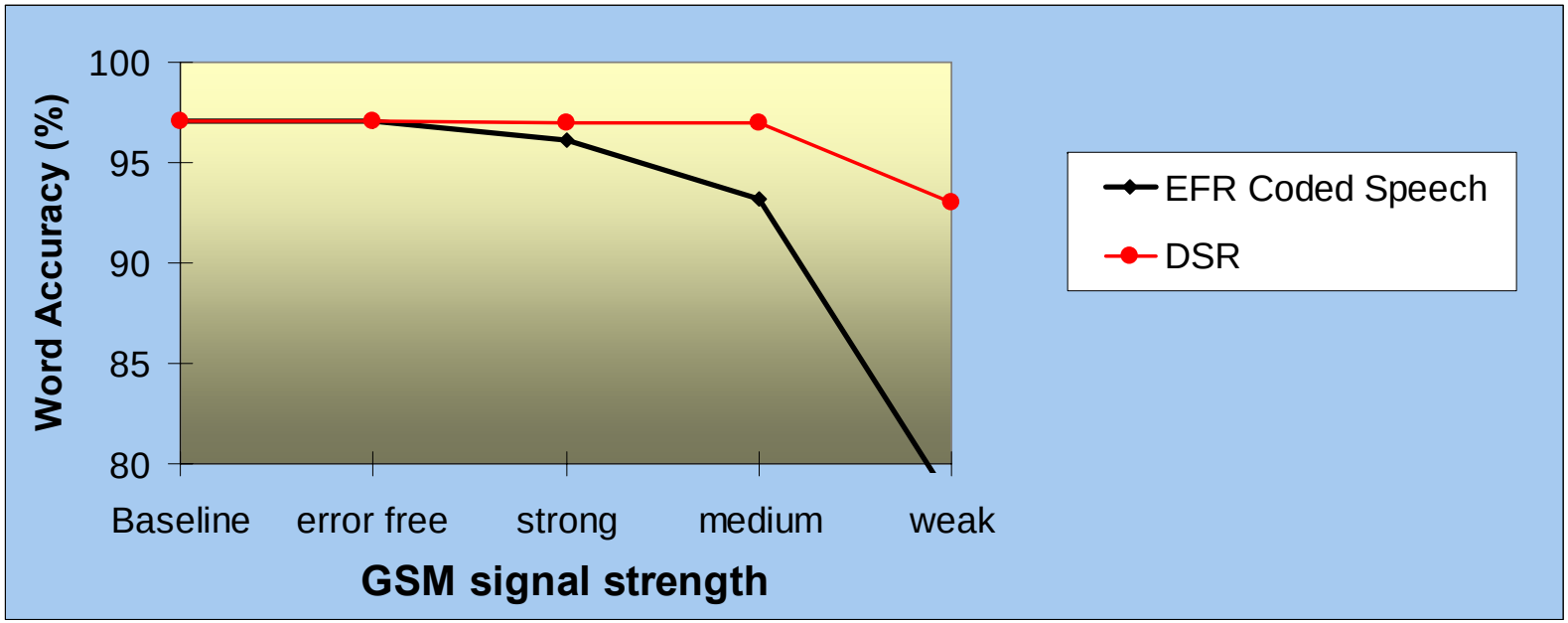
Conventional

Speech Coder → Circuit Switched / Mobile Voice Channel → Speech Decoder → ISDN → ASR Front-end → ASR Decoder

DSR

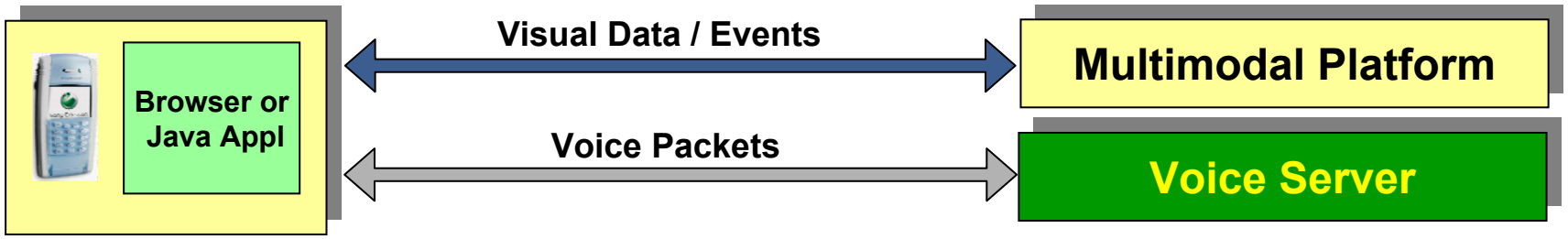
ASR Front-end → Packet Data Channel / e.g. GPRS or 3G → ASR Decoder

Benefits of DSR Aurora



- **Improves performance over wireless channels**
  - **Minimises impact of codec & channel errors**
  - **Consistent performance over coverage area**
- **Improved performance in background noise**
  - **53% reduction in error rate**
- **Ease of integration of combined speech and data applications**
  - **Use packet data channel for both DSR and other data**

Server-based MMI Architecture



Main features of a Server-based MMI architecture:

- **Device client is very thin (not a real browser)**
- **Voice packets are sent through a data connection**
- **Press to talk for activating ASR (input)**
- **Intermix of TTS prompts and visual data/charts (output)**
- **Heavy processing is done on the server side (ASR/TTS)**
- **Rich speech interaction is possible with VoiceXML platform**
- **Events and data are exchanged between client and server to synchronize the visual and vocal interactions**
- **Synchronization is done on the server side (by MMI Platform)**

Q & A

- **[ASR] Why is the ASR Front End module based on 'old' NN (like MLP)? Why not use a SVM (Support Vector Machine) approach?**
  - ➔ *Since '90, MLP+HMM was tested as an alternative to CD+HMM, where the MLP calculates the posteriori probabilities. The SVMs are computationally heavier and present some drawbacks, nevertheless first study of the use of SVMs in speech recognition are under way. See: Collobert, Bengio, ICML'04 for a study on the links between these classification algorithms.*
- **[ASR] Is there a processing of the emotions in the ASR?**
  - ➔ *This is a very interesting research topic, but currently we are working on emotion only for the TTS.*
- **[TTS] Request for details about on-going research into emotional TTS?**
  - ➔ *Suggested readings: Eide et al., SSW5-2004; Zovato et al., ICSLP-2004*
- **[TTS] Relative degradation on embedded TTS is due to DB reduction or coding?**
  - ➔ *Mostly to DB reduction, but it depends on voice/lang too.*
- **[SemanticWeb] It might be important for large speech applications.**
  - ➔ *Yes, but the powerful advantages should be demonstrated. Previous attempts to encode common sense knowledge failed.*

Thank you!

Credits

- Dave Burke, CTO Voxpilot, author of SISR
- Deborah Dahl, chair of Multimodal Interaction Working Group
- Jim Larson, co-chair Voice Browser Working Group
- David Pearce, Motorola, author of DSR Aurora spec
- Roberto Pieraccini, CTO Tell-Eureka, ex-colleague & friend
- Ken Rehor, chairman VoiceXML Forum

\+ Many researchers and colleagues in Loquendo!

Thank you!