Multilingual Aspects in Speech and Multimodal Interfaces

Paolo Baggia
Director of International Standards
Loquendo Today

Do we need multilingual applications?

Voice is different from text?

Current Solutions – a Tour:

Speech Interface Framework Today

Voice Applications

Speech Recognition Grammars

Speech Prompts

Pronunciation Lexicons

Discussion Points
Company Profile

- Privately held company (fully owned by Telecom Italia), founded in 2001 as spin-off from Telecom Italia Labs, capitalizing on 30yrs experience and expertise in voice processing.

- Global Company, leader in Europe and South America for award-winning, high quality voice technologies (synthesis, recognition, authentication and identification) available in **30 languages** and **71 voices**.

- Multilingual, proprietary technologies protected over 100 patents worldwide

- Financially robust, break-even reached in 2004, revenues and earnings growing year on year


- Flexible: About 100 employees, plus a vibrant ecosystem of local freelancers.
International Awards

**Market leader-Best Speech Engine Speech Industry Award**

**2010 Speech Technology Excellence Award CIS Magazine**

**2008 Frost & Sullivan European Telematics and Infotainment Emerging Company of the Year Award**

**Loquendo MRCP Server: Winner of 2008 IP Contact Center Technology Pioneer Award**

**Best Innovation in Automotive Speech Synthesis Prize AVIOS-SpeechTEK West 2007**

**Best Innovation in Expressive Speech Synthesis Prize AVIOS-SpeechTEK West 2006**

**Best Innovation in Multi-Lingual Speech Synthesis Prize AVIOS-SpeechTEK West 2005**
Do We Need Multilingual Applications?

Yes, because …

- **We live in a Multicultural World**
  - Movement of students/professionals, migration, tourism

- **Monolingual Contexts**
  - Air Traffic, International Projects, International Agencies often require a common language, such as English, French, Arabic or Mandarin Chinese

- **Multilingual Speakers**
  - Where the region has more than one national language, extreme case India with 20 official languages
Voice vs. Text

Voice is different from text, because …

- Takes into account the reader:
  - S/he might be native speaker, bilingual, second language, or novice for a given language

- A speaker can have an accent:
  - Each speaker has an accent, soft or strong. The accent can cross borders and regions.

- Recognition vs. Synthesis:
  - Different perspectives on the same area

The role of audio material in the Web arena is increasing constantly.
Speech Interface Framework - End of 2010
(by Jim Larson)

Semantic Interpretation for Speech Recognition (SISR)
N-gram Grammar ML
Speech Recognition Grammar Spec. (SRGS)
Pre-recorded Audio Player
Speech Synthesis Markup Language (SSML)
DTMF Tone Recognizer
Pronunciation Lexicon Specification (PLS)
Language Understanding
Language Generation
ASR
Dialog Manager
Context Interpretation
Natural Language Semantics ML
VoiceXML 2.1
VoiceXML 2.0
EMMA 1.0
N-gram Grammar ML
User
VoiceXML 2.0
VoiceXML 2.1
World Wide Web
Telephone System
Reusable Components
Call Control XML (CCXML)
Semantic Interpretation for Speech Recognition (SISR)
N-gram Grammar ML
Speech Recognition Grammar Spec. (SRGS)
Pre-recorded Audio Player
Speech Synthesis Markup Language (SSML)
DTMF Tone Recognizer
Pronunciation Lexicon Specification (PLS)
Language Understanding
Language Generation
ASR
Dialog Manager
Context Interpretation
Natural Language Semantics ML
VoiceXML 2.1
VoiceXML 2.0
EMMA 1.0
N-gram Grammar ML
User
VoiceXML 2.0
VoiceXML 2.1
World Wide Web
Telephone System
Reusable Components
Call Control XML (CCXML)
A Tour of W3C Speech Standards

W3C Voice Browser standards are the basis for all the voice development in the Web:

- **Grammars for Speech (and DTMF)** – SRGS 1.0 (2004), SISR 1.0 (2007)
- **Prompts** – SSML 1.0 (2004), SSML 1.1 (2010)
- **Pronunciation Lexicon** – PLS 1.0 (2008)
- **Input Results** – EMMA 1.0 (2009)

More to come: VoiceXML 3.0, SCXML 1.0, EmotionML 1.0, etc.
Naming a Language is not a trivial task!

- **IANA Language Subtag Registry** –
  http://www.iana.org/assignments/language-subtag-registry
  Searching Tool: http://rishida.net/utils/subtags/

- **IETF BCP-47** –
  About Language Subtags:
  http://www.w3.org/International/articles/language-tags/Overview.en.php

- **Examples**:  
  - zh-yue – Cantonese Chinese (macrolanguages)  
  - ar-afb – Gulf Arabic  
  - es-005 – South American Spanish  
  - ca-es-valencia – Valencian spoken language
VoiceXML 2.0 & 2.1

<?xml version="1.0" encoding="UTF-8"?>
<vxml xmlns="http://www.w3.org/2001/vxml"
     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
     xsi:schemaLocation="http://www.w3.org/2001/vxml http://www.w3.org/TR/voicexml20/vxml.xsd"
     version="2.0" xml:lang="en-US">
  <form>
    <field name="drink">
      <prompt>Would you like coffee, tea, milk, or nothing?</prompt>
      <grammar src="drink.grxml" type="application/srgs+xml"/>
    </field>
    <block>
      <submit next="http://www.drink.example.com/drink2.asp"/>
    </block>
  </form>
</vxml>

Notes:
- xml:lang inheritance
- VoiceXML 2.0 mandates RFC 3066 (before RFC 1766)
- Now, by Errata extensions to IRI and BCP 47
Speech Recognition Grammars – SRGS 1.0

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

<grammar version="1.0" xml:lang="en-US"
    mode="voice" root="main">
    <rule id="main">
        <one-of>
            <item> yes please </item>
            <item> no thanks </item>
        </one-of>
    </rule>
</grammar>

Notes:
- xml:lang inheritance
- SRGS 1.0 mandates RFC 3066 (before RFC 1766)
- Now, by Errata extensions to IRI and BCP 47
ABNF 1.0 ISO-8859-1;

// Default grammar language is US English
language en-US;

// Single language attachment to tokens
// Note that "fr-CA" (Canadian French) is applied to only
// the word "oui" because of precedence rules
$yes = yes | oui!fr-CA;

// Single language attachment to an expansion
$people1 = (Michel Tremblay | André Roy)!fr-CA;

// Handling language-specific pronunciations of the same word
// A capable speech recognizer will listen for Mexican Spanish and
// US English pronunciations.
$people2 = Jose!en-US | Jose!es-MX;

/**
 * Multi-lingual input possible
 * @example may I speak to André Roy
 * @example may I speak to Jose
 */

Notes:
Language tags attached to rules and words. Instruction to transcribe the word in a different language to extend coverage.
SSML 1.1 – lang element

http://www.w3.org/TR/speech-synthesis11/

- **lang element** -

- Indicates the natural language of the content
- *May be used when there is a change in the natural language*

- **Attributes:**
  - `xml:lang` is a required attribute specifying the language
  - `onlangfailure` the desired behavior upon language speaking failure

- **When the language change is associated with the structure of the text, it is recommended to use the `xml:lang` attribute on the respective** `p`, `s`, `token`, and `w` **elements**

```xml
<?xml version="1.0"?>
<speak version="1.1" xmlns="http://www.w3.org/2001/10/synthesis"
     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
     xsi:schemaLocation="http://www.w3.org/2001/10/synthesis
                        http://www.w3.org/TR/speech-synthesis11/synthesis.xsd"
     xml:lang="en-US">
  The French word for cat is <w xml:lang="fr">chat</w>.  
  He prefers to eat pasta that is <lang xml:lang="it">al dente</lang>. 
</speak>
```
Phonetic Mapping – TTS Sample

Phonetic Mapping
Applies the foreign language grapheme-to-phoneme transcription-rules to the foreign text, and then maps the transcribed phonemes onto those of the voice's native language in order to access its acoustic units

- Approximate Pronunciation (speaker maintains her/his native-tongue phonological system when pronouncing foreign words)

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Italian</th>
<th>French</th>
<th>German</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>German Voice</td>
<td><img src="image" alt="Audio" /></td>
<td><img src="image" alt="Audio" /></td>
<td></td>
<td><img src="image" alt="Audio" /></td>
<td><img src="image" alt="Audio" /></td>
</tr>
<tr>
<td>Italian Voice</td>
<td><img src="image" alt="Audio" /></td>
<td></td>
<td><img src="image" alt="Audio" /></td>
<td><img src="image" alt="Audio" /></td>
<td><img src="image" alt="Audio" /></td>
</tr>
<tr>
<td>French Voice</td>
<td><img src="image" alt="Audio" /></td>
<td><img src="image" alt="Audio" /></td>
<td></td>
<td><img src="image" alt="Audio" /></td>
<td><img src="image" alt="Audio" /></td>
</tr>
<tr>
<td>Spanish Voice</td>
<td><img src="image" alt="Audio" /></td>
<td><img src="image" alt="Audio" /></td>
<td><img src="image" alt="Audio" /></td>
<td></td>
<td><img src="image" alt="Audio" /></td>
</tr>
</tbody>
</table>
On the Wikipedia Web site I found that SSML is an acronym, which can stand for more than one thing, for example:

<lookup ref="markup"> SSML, an XML-based markup language for speech synthesis applications.
<lookup ref="league"> SSML, a football league in England.
<lookup ref="ship"> SSML, National Research Laboratory, funded by the Korea Science and Engineering Foundation.
</lookup>

But today we are going to speak about SSML.
SSML 1.1 – voice element

- **The `xml:lang` attribute** (present in SSML 1.0) has been removed.
- **languages** OPTIONAL attribute indicating the list of languages the voice is desired to speak. The value MUST be:
  - the empty string ""
  - or a space-separated list of languages, with OPTIONAL accent indication per language.

- **Each language/accent pair is** of the form "language" or "language:accent", where both language and accent MUST be an Extended Language Range [BCP47], except that the values "und" and "zxx" are disallowed.

- **For example:**
  - languages="en:pt fr:ja" can legally be matched by any voice that can both read English (speaking it with a Portuguese accent) and read French (speaking it with a Japanese accent). Thus, a voice that only supports "en-US" with a "pt-BR" accent and "fr-CA" with a "ja" accent would match.
  - languages="fr:pt" there is no voice that supports French with a Portuguese accent, then a voice selection failure will occur.
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0"
 xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://www.w3.org/2005/01/pronunciation-lexicon
 http://www.w3.org/TR/2007/CR-pronunciation-lexicon20071212/pls.xsd"
 alphabet="ipa" xml:lang="en-US">
  <lexeme>
    <grapheme>Sepulveda</grapheme>
    <phoneme>səˈpʌlvədə</phoneme>
  </lexeme>
  <lexeme>
    <grapheme>W3C</grapheme>
    <alias>World Wide Web Consortium</alias>
  </lexeme>
</lexicon>

Notes:
- PLS documents are monolingual: a single xml:lang declaration

Proposal to create ➔ IANA Registry for Phonetics Alphabets
• Speech technologies enable multilinguality to be addressed in a wide variety of sectors and applications

• The use of standards facilitates the development of speech multilingual applications

• Use of BCP-47 and IANA Language Subtag Registry

• Need of Registry for Phonetic Alphabets
THANK YOU

for clarifications or questions:

paolo.baggia@loquendo.com

My GoogleTalks available on YouTube:
• Introduction to Speech Technologies (March 2008)
• Voice Browser and Multimodal Interaction In 2009 (March 2009)