EmotionML:
A language for embodied conversational agents

Davide Bonardo
davide.bonardo@loquendo.com

Loquendo
We Speak. We Listen. We Understand.

Outline

Loquendo
VOCAL TECHNOLOGY AND SERVICES

- Introduction
- The COMPANIONS project
- Architecture
- Emotions: Input and Output
- Multimodal Emotional Input
- Output Specification
- Multimodal Integration
- Multimodal Emotional Output
- Interoperability with other standards
- Conclusions

Introduction

Loquendo
VOCAL TECHNOLOGY AND SERVICES

- Affective dialog interfaces provide functionalities aimed at:
  - Recognizing the emotional state of the user through:
    - Text
    - Speech
    - Physiological sensors
    - Video
  - Providing the most appropriate feedback to the users by generating coherent emotional responses through different modalities
    - Text
    - Speech
    - Facial expression
    - Gestures
- Handling different modalities both in input and output requires a common and efficient formalism to represent, and eventually merge, data.

The COMPANIONS Project

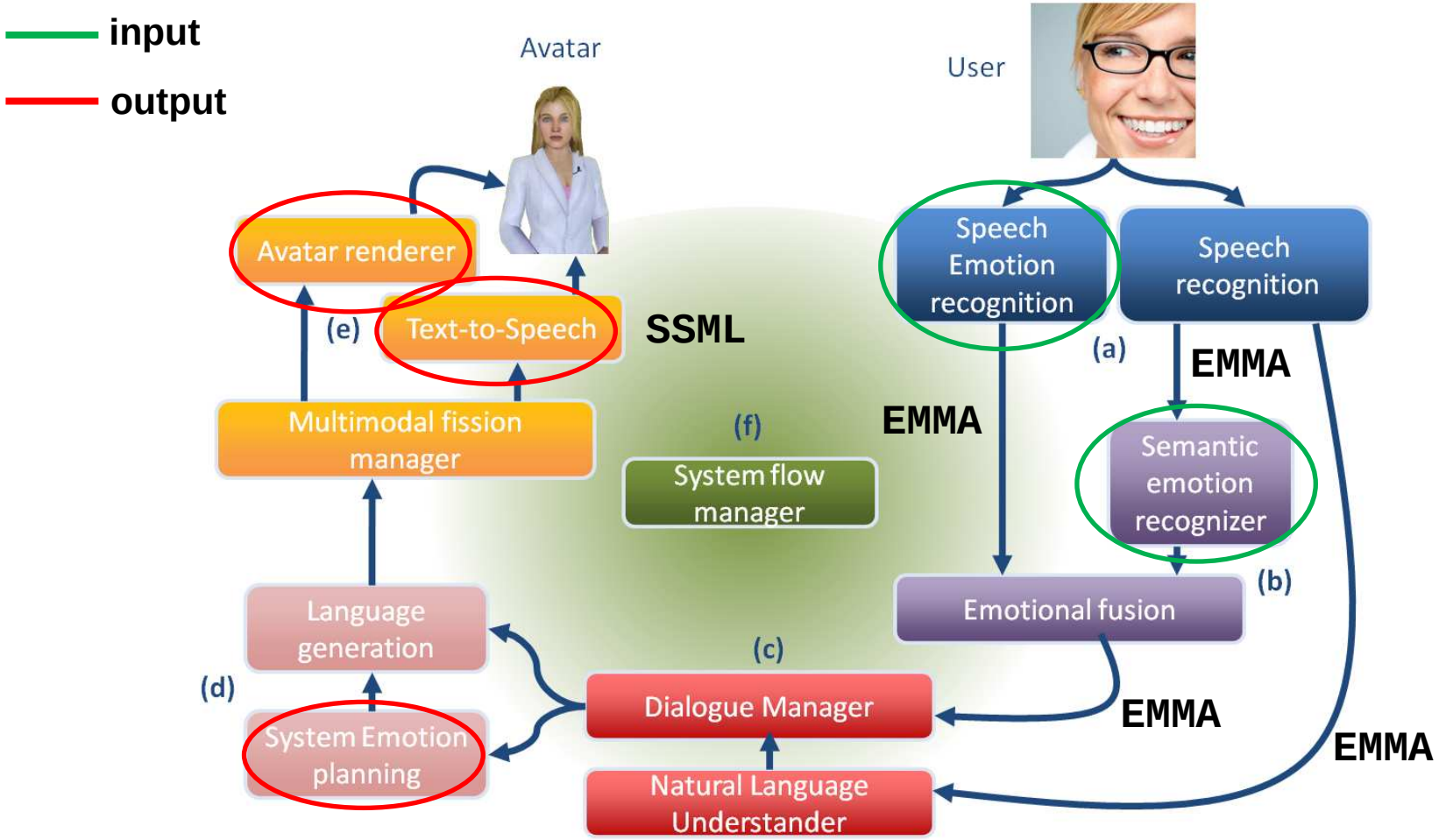
Loquendo
VOCAL TECHNOLOGY AND SERVICES

- We will consider a real use case: the system developed within the project Companions (sponsored by EU – FP6)
- The goal of the project is the development of a dialogue system with the ability to perceive and express emotions.
- The system is based on an embodied conversational agent (ECA) with some expressive features.
- Scenario: ‘How was your day?’ (HWYD)
  - based on the idea of a user who freely discusses his/her working day in a typical office environment, with the avatar providing advice and comfort in a natural ‘social’ dialogue situation.

companions

www.companions-project.org

Architecture



Emotions: Input and Output

Loquendo
VOCAL TECHNOLOGY AND SERVICES

- A simplified Emotional Model is used to adapt the avatar's emotional behavior to the user's emotional state

- INPUT: recognition of emotional state, two emotion detectors
  - acoustic level (EmoVoice – University of Augsburg)
    - *neutral*
    - *positive-passive, positive-active*
    - *negative-passive, negative-active*
  - semantic level (Sentiment Analyzer – University of Oxford)
    - *neutral,*
    - *positive,*
    - *negative*

T. Vogt, E. André and N. Bee, "EmoVoice - A framework for online recognition of emotions from voice"

Karo Moilanen and Stephen Pulman, "Sentiment Composition"

- OUTPUT: generation of emotional behavior
  - Avatar's body gestures and facial expressions
  - **Expressive speech synthesis**
    - *neutral,*
    - *positive,*
    - *negative*

"Application of expressive TTS synthesis in an advanced ECA system" by Jan Romportl, Enrico Zovato, Raul Santos, Pavel Ircing, Jose Relano Gil, and Morena Danieli

Multimodal Emotional Input

Loquendo
VOCAL TECHNOLOGY AND SERVICES

- EmotionML could be used to represent in an unique framework the two input representations:

```
<emotionml xmlns="http://www.w3.org/2009/10/emotionml"
dimension-set="http://www.example.com/emotion/dimension/FSRE.xml">

  <emotion start=1268647200 modality="voice">
  <!—- Positive-Active -->
        <dimension name="valence" value="1.0" confidence="0.6"/>
        <dimension name="arousal" value="1.0" confidence="0.6"/>
        <reference uri="asr_output.wav#t=26,98"/>
  </emotion>

  <emotion start=1268647200 modality="text">
        <!—- Positive -->
        <dimension name="valence" value="1.0" confidence="0.9"/>
        <reference uri="asr_output.txt#t=26,98"/>
  </emotion>
</emotionml>
```

Speech

Text

Output Specification

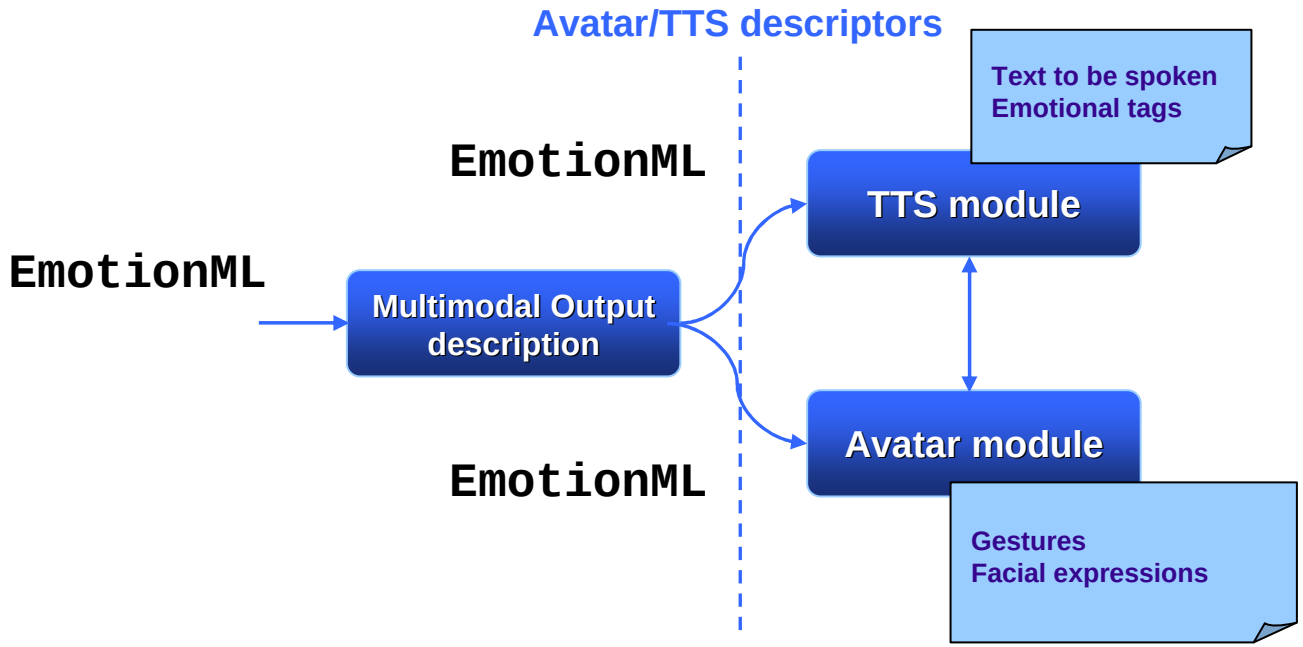
- The system output is defined by 3 labels generated by the Dialog Manager and the Emotion Planning module:
  1. Performative:
     - *suggest, propose, warn, agree, criticize, advice, confirm, incite, inform, greet, wait*
  2. Affect:
     - A subset of OCC Categories:
     - *neutral, embarrassment, happy-for, relief, skeptical, mellow*
  3. Emphasis:
     - *weak, medium, strong*

- This information is stored in an XML message:

```
<message>
  <header> (data) </header>
    <payload>
      <ECAOutput
         perform  = "greet"
         affect   = "happy_for"
         emphasis = "medium"
         text     = "I had a very good day!" />
    </payload>
</message>
```

Multimodal Integration

- Distinction between planning and generation:
  - The results of emotion planning is included in the “affective” and “emphasis” labels
  - Emotion descriptors drive the generation of emotional behaviors through speech and avatar rendering



Multimodal Emotional Output

- EmotionML could be effectively used to represent the target emotion, i.e. the results of the modules involved in emotion planning.
- In the example below, a “default” vocabulary for emotion categories is chosen (defined by the “category-set” attribute)
- Intensity of affective (emphasis) is also specified

```
<emotionml xmlns=http://www.w3.org/2009/10/emotionml
category-set="http://www.example.com/emotion/category/OCC.xml">
      <emotion>
               <category name="happy-for"/>
               <intensity value="0.5"/>
      </emotion>
</emotionml>
```

Interoperability with other standards

- One of the key-points of EmotionML will be its interoperability with other standards (e.g. EMMA, SSML, etc.)
- The EmotionML elements could be embedded into other markup, improving the efficiency of the integration process.
- For example: the emotion specification seen in the previous example could be inserted in a SSML document. The TTS parser will then get the information on how to render output speech.

```
<?xml version="1.0"?>
<speak version="1.1" xmlns="http://www.w3.org/2001/10/synthesis"
    xmlns:emo="http://www.w3.org/2009/10/emotionml" xml:lang="en-US">
     <s>
           <emo:emotion category-set="http://www.example.com/emotion/category/OCC.xml">
                     <emo:category name="happy-for"/>
                     <emo:intensity value="0.5"/>
           </emo:emotion> I had a very good day!
    </s>
</speak>
```

Conclusions

Loquendo
VOCAL TECHNOLOGY AND SERVICES

- We have described how the EmotionML specification could be exploited in a real case, an affective dialogue system, to appropriately handle emotion input and output representations

- As from the last working draft, the EmotionML is flexible enough in describing emotions according to suggested vocabularies derived from existing models of emotions

  How to control the coherence between the values expressed in the vocabularies and the values used in the <emotion> elements?

- There's still however a significant gap between what technologies are capable of reaching in the fields of emotion recognition and generation and the potentiality of EmotionML in terms of emotion description

  How to make the language "attractive" for different technological applications, despite their different levels of complexity?