Virtual Help Assistant  
Seyed Ahmad Hosseini - s.arashhosseini@gmail.com

Introduction

• This project describes an approach to create a virtual helpful assistant for artists working in the Visual Effects (VFX) industry.
• The developed virtual helpful assistant is meant to increase artist productivity and reduce workloads on supporting parts of the VFX pipeline (e.g. technical directories, IT support).

Core technologies

• The assistant heavily relies on techniques described in “Sequence to Sequence Learning with Neural Networks” [15] and “A Neural Conversational Model” [17].
• The assistant’s core utilizes TensorFlow’s implementation of Neural machine translation (NMT) [6], its usage in the system was inspired by work from Harvard NLP Group [4].

Speech recognition and synthesis

• Speech recognition is responsible for the system’s input. After evaluating performance and accuracy of other alternatives it was clear that Google’s Cloud Speech-to-Text is the most feasible solution. Mainly because of low-latency responses, input languages and automatic punctuation.
• Speech Synthesis [5] is responsible for the system’s auditory output. In particular, end-to-end architectures, such as the Tacotron [10] systems can both simplify voice building pipelines and produce natural-sounding speech. The Nancy Corpus was used for training. [14]

Dataset acquisition and model training

• For general conversation and knowledge, multiple datasets was used: ELS Fast’s English Conversations [9], Cornell movie-quotes corpus [7] and Reddit’s [8] comments dataset.
  • In our case the hosting DCC application is Autodesk Maya. For help dataset acquisition an Autodesk Maya Python [11] and Help Documentation [12] was used. NLTK [3] was used to enhance information extracted from the Autodesk Maya documentation dataset.
  • Quality and usefulness of the assistant heavily depends on the intents dataset. Creating many high-quality intents is labor-intensive and requires a knowledge of the DCC application.
• The result of NMT inference are pointers to target functions, which also need to be implemented by a technical director (TD) or support department.
• Intents define industry application of the assistant. The assistant is not tied to VFX industry and could be easily modified to fit into other sectors by producing related intents.
• In a VFX environment it would be a TD who can anticipate artists questions or extract them from studio issues database.

Table 1: Datasets used for the model training and their weights.

<table>
<thead>
<tr>
<th>dataset</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>maya_guide_set.txt</td>
<td>0.340</td>
</tr>
<tr>
<td>cornell.txt</td>
<td>0.188</td>
</tr>
<tr>
<td>reddit.txt</td>
<td>0.158</td>
</tr>
<tr>
<td>maya_python_set.txt</td>
<td>0.140</td>
</tr>
<tr>
<td>first_aid_person.txt</td>
<td>0.071</td>
</tr>
<tr>
<td>scenario.txt</td>
<td>0.060</td>
</tr>
<tr>
<td>knowledge.txt</td>
<td>0.012</td>
</tr>
<tr>
<td>maya.txt</td>
<td>0.009</td>
</tr>
<tr>
<td>computing.txt</td>
<td>0.007</td>
</tr>
<tr>
<td>datetime.txt</td>
<td>0.003</td>
</tr>
<tr>
<td>narrative.txt</td>
<td>0.006</td>
</tr>
<tr>
<td>elementary.txt</td>
<td>0.002</td>
</tr>
<tr>
<td>unk.txt</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Figure 1: Diagram showing data flow and individual components of the assistant.

Avatar

• The avatar is featured as 3d animated mesh in Unity game engine.
• The project decided to use a stylized representation of the avatar to avoid likeness issues related to uncanny valley phenomenon. The Avatar’s face animation is blendshape-driven.
• Face blendshape coefficients are derived from phonemes extracted from answers using the CMU Pronouncing Dictionary [1] accessed through NLTK and their blending is synthesised with synthesized speech.
• The system contains face blendshapes for each phoneme in seven different emotions, according to Paul Ekman “basic emotions”.
• The virtual assistant system also tracks artists head orientation and gaze attention, which is determined using Dlib and OpenCV libraries. The system uses it to determine artists attention and the avatar responds accordingly. For example it enables the avatar to switch from idle state (looking around, without focus) to active state (focusing on the artist).
• The system is estimating artists emotion too. It is using a deep convolutional neural network on FER2013 [2] dataset. The extracted emotion could influence artists mood in face blendshape selection and could be used to autodecide for each phoneme in seven basic emotions.
• The capture process utilizes Apple ARKit which outputs the needed coefficients for a captured face expression for a given phoneme in a specific mood. This workflow captures subtle skin movements which produces more realistic results when compared to hand-designed blendshapes. This capture process needs to happen only once and can be re-used on different meshes with the same set of blendshapes [13].

Further work

The current implementation includes a working prototype. Further work will focus on improving individual components to provide more a natural and smooth experience for artists and make it more accessible, mainly by employing Wavenet [16], a technique to avoid the avatar’s monotonous voice and clarify it’s intents, to increase the usefulness. In addition, the use of sign language recognition would be possible extension.

References
