Virtual Help Assistant

Seyed Ahmad Hosseini - s.arashhosseini@gmail.com

FILMAKADEMIE BADEN-WÜRTTEMBERG

animationsinstitut

Introduction

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

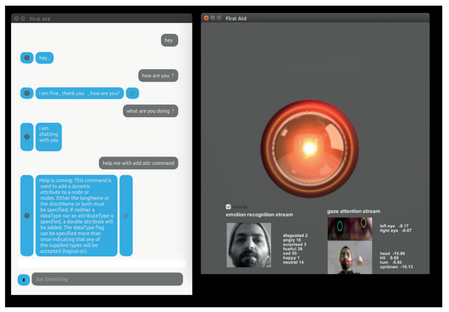
Dataset acquisition and model training

- For general conversation and knowledge, multiple datasets were 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.

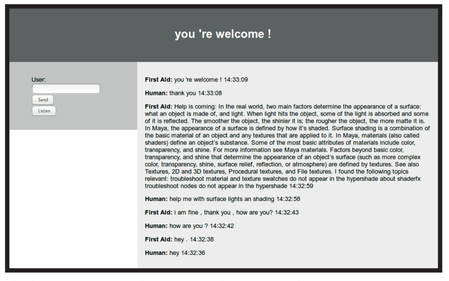
dataset	*weight*
maya_guide_set.txt	0.340
cornell.txt	0.188
reddit.txt	0.158
maya_python_set.txt	0.140
first_aid_person.txt	0.071
scenario.txt	0.060
knowledge.txt	0.012
maya.txt	0.009
computing.txt	0.007
datetime.txt	0.003
narrative.txt	0.003
elementary.txt	0.002
unk.txt	0.001

Figure 2: The system includes three different frontends:



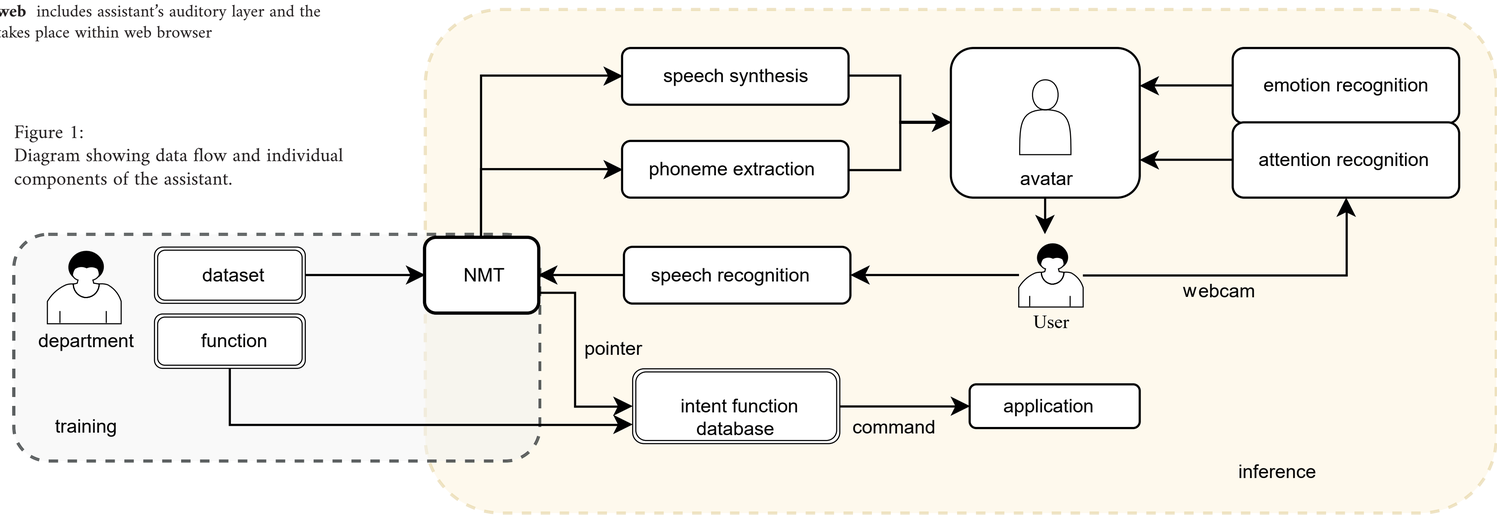
a) **Assistant-full** includes all perception layers, Qt based UI and Avatar

b) **Assistant-light** includes assistant's auditory layer and Avatar in Unity game engine.



c) **Assistant-web** includes assistant's auditory layer and the conversation takes place within web browser

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 synchronised 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 avatars mood in face blendshape selection and could serve as autoacaptured for each phoneme in seven basic emotions.
- The capture process utilizes Apple **ARKit** which outputs the needed coefficeints 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.

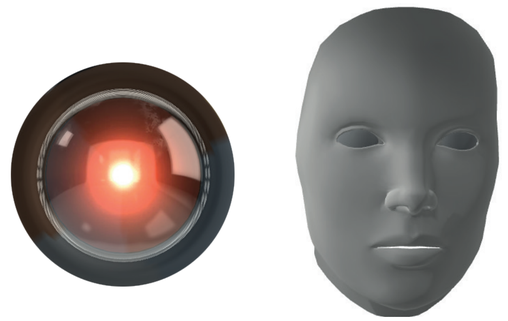


Figure 3: Two different avatars which are currently implemented.

References

[1] The carnegie mellon university pronouncing dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict, Accessed: 2018-10-10
[2]Fer2013.https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data. Accessed: 2018-10-10.
[3] Natural language toolkit.https://www.nltk.org/. Accessed: 2018-10-10.
[4]Opennmt.http://forum.opennmt.net/t/english-chatbot-model-with-opennmt/184 .Acessed: 2018-10-10.
[5]Tacotron. https://github.com/keithito/tacotron.git. Accessed: 2018-10-10.
[6] Neural machine translation. https://github.com/tensorflow/nmt/tree/tf-1.4 . Accessed:2018-10-10.
[7] Cristian Danescu-Niculescu-Mizil and Lillian Lee. http://www.cs.cornell.edu/%7Ecristian/Cornell_Movie-Dialogs_Corpus.html 2011. Accessed: 2018-10-10.
[8] jason@pushshift.io. Reddit's entire publicly available comment dataset. https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/ . Accessed: 2018-10-10.
[9] teslesIfast.com. English conversations. https://www.eslfast.com/robot/ . Accessed: 2018-10-10.
[10] Yuxuan Wang et al.Tacotron: Towards end-to-end speech synthesis, 2017. http://dx.doi.org/10.21437/Interspeech.2017-1452
[11] https://github.com/ArashHosseini/autodesk_maya_docs_dump/tree/master/py_cmds .Acessed: 2018-12-5.
[12] https://github.com/ArashHosseini/autodesk_maya_docs_dump/tree/master/guide .Acessed: 2018-12-5.
[13] https://developer.apple.com/documentation/arkit/arfaceanchor/blendshapelocation .Acessed: 2018-12-5.
[14] http://www.cstr.ed.ac.uk/projects/blizzard/2011/lessac_blizzard2011/ .Acessed: 2018-12-5.
[15] Sequence to Sequence Learning with Neural Networks, Sutskever et al., 2014
[16] WaveNet: A Generative Model for Raw Audio , van den Oord et al.,2016a
[17] A Neural Conversational Model , Vinyals and Le, 2015