Virtual Help Assistant

Seyed Ahmad Hosseini
s.arashhosseini@gmail.com

Animationsinstitut,
Filmakademie Baden-Wuerttemberg

1 Abstract

This paper describes an approach to create a virtual help assistant for artists working in the Visual Effects (VFX) industry. Having a virtual help assistant will have a beneficial impact on VFX production by increasing artists productivity and reducing workload on supporting parts of VFX pipeline (e.g. TDs, pipeline, IT support). This project leverages recent advances in Artificial Intelligence (AI) and Machine Learning (ML) technologies and aims to provide natural communication with artists, mainly by combining neural machine translation, speech recognition, speech synthesis, emotion recognition, gaze attention and a natural language toolkit. In order to have a natural conversation, the virtual assistant needs to react naturally to both VFX-related requests and general conversation requests, which are not related with VFX. The main challenge is to build a pipeline, connecting all the individual parts mentioned in the abstract to a low-latency workflow. An artist should interface the whole system through an avatar, which will feel naturally and easy to communicate with.

2 Neural Machine Translation (NMT), Speech Recognition and Synthesis

The core of the virtual assistant relies on TensorFlow's implementation of Neural Machine Translation (NMT) [6]. It is employed in the model training on acquired datasets and inference. Its usage in the system was inspired by work from Harvard NLP Group [4]. Because of good documentation and resources this library was preferred to other alternatives. Speech recognition is responsible for the system's input. It is crucial for recognizing artists question and for transcripting it into text. All further components of the virtual assistant depend on the quality of this step. 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. Some situations however may require direct text input, which is also supported in a form of a chat window. Further development will evaluate sign language gesture recognition as a possible source of artists input. 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. Most current end-to-end systems, including Tacotron, do not explicitly model prosody, even if models are trained on very meaningful records such as audiobooks, which often contain strong-voiced voices.

3 Dataset acquisition and Model training

For the avatar to respond naturally to general-topic questions and to answer correctly to DCC-related questions it needs to be trained with large-enough dataset. 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 and Help Documentation was used. A web-scraping approach was chosen for acquiring, extracting and categorising relevant information such as comments, examples and explanations. NLTK [3] was used to enhance information extracted from Autodesk Maya Documentation dataset. While general-knowledge datasets provide answers to general questions, for artist-DCC interactions an action from the assistant is required. For example the assistant should manage to extract relevant information from the documentation, but also trigger actions in DCC applications itself. Example cases include an artist asking for help with render settings or asking the assistant to create a constraint. The training procedure needs to extract parameters from artists questions and use them appropriately in the intended way. This procedure requires an intents dataset which provides mapping from user question to target function and its arguments. Quality and usefulness of the assistant heavily depend on the intents dataset. Creating many high-quality intents is labor intensive and requires a knowledge of the DCC application. In a VFX environment it would be a task for technical directors who can anticipate artists questions or extract them from studio issues database. This project tries to generate intents based on the documentation.

4 Avatar

The avatar is responsible for the system's visual output. It's the component which will receive most of the artists attention. Therefore it should feel natural and easy to talk to. 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. 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 model trained with deep neural network on FER2013 [2] dataset. The extracted emotion could influence avatars mood in face blendshape selection and could serve as automatic artists feedback indicator for the system.

5 Conclusion

The current implementation includes a working prototype. Further work will focus on improving individual components to provide more natural and smooth experience for artists and to make it more accessible. Mainly by employing Wavenet to avoid monotonous voice and improving intents to increase the usefulness of the assistant.

[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-expres 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. Accessed: 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] tesleslfast.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. URL http://dx.doi.org/10.21437/Interspeech.2017-1452.