Interspeech 2018
2-6 September 2018, Hyderabad

HoloCompanion: An MR Friend for EveryOne

Mallikarjuna Rao Bellamkonda [1], *Annam Naresh*[1], *Rushabh Gandhi*[2], *Mithun DasGupta*[1],

[1]Microsoft, Hyderabad, India
[2]BITS, Hyderabad, India
{annaresh,mallib, migupta}@microsoft.com, gandhirushabh2311@gmail.com

Abstract

Chat bots are becoming ubiquitous in our day to day life. The advent of the summer of AI has brought us all in close contact with intelligent agents such as Cortana, Siri and Alexa. We envisage a world, where these bots have there physical existence within the realm of Mixed Reality (MR). We present the first 3D chit-chat bot called the HoloCompanion. This bot has a personality, can chat with anyone and about any topic and has articulated lip, eye and head movements.
Index Terms: ASR, Chit-chat bot, TTS, human-computer interaction, lipsync.

1. Introduction

Chat bots are highly prevalent means of interaction with devices these days. Voice has been claimed to be the keyboard of the future, and as such these bots have an increasing role to play in our lives. Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home are all examples of chat interfaces which listen to humans and try to infer the action intent expressed by the humans. There has been larger quanta of development in transactional bots, which try to fulfill a specific intent such as ticket booking or hotel reservations. Chit-chat or non-transactional conversation on the other hand is more subjective and hence more difficult to comprehend owing to the open-endedness of such conversations.

There has been some interesting development in the space of chit-chat bots and Microsoft has led this front across different geographies. Xiaoice[1] in China, Xo[2] in USA, Rinna[3] in Japan and Ruuh[4] in India are existing chit-chat agents across different countries.

2. Chit-chat within Microsoft HoloLens

Microsoft HoloLens is a self-contained, holographic computer, enabling users to engage with digital content and interact with holograms in their world. Mixed reality brings people, places, and objects from the physical and digital worlds together. This blended environment becomes the canvas, where people can create and enjoy a wide range of experiences. Interacting with holograms in mixed reality enables users to visualize and work with their digital content as part of their real world. Holograms are responsive to the user and the world around the user. Microsoft HoloLens enables them to interact with content and information in the most natural ways possible.

1. Gaze: Built-in sensors let you use your gaze to move the cursor so you can select holograms. Turn your head and the cursor will follow.
2. Gesture: Users can use simple gestures to select, drag, drop and interact with the holograms.
3. Voice: Use voice commands to navigate, select, open, command, and control the applications.

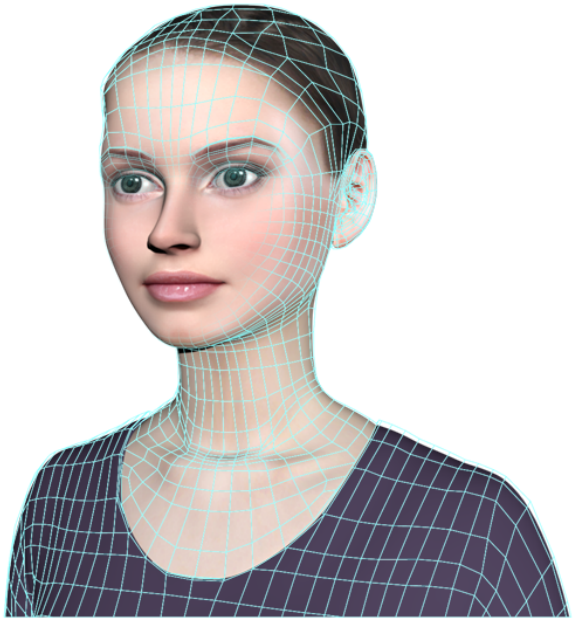

Figure 1: *Avatar model used in HoloCompanion.*

We propose to bring a chat bot into the augmented reality realm. The articulation of the chat through lip sync, blink and head movement is close to real time with no post-processing as is common in the animation industry. This is the first such attempt to the best of our knowledge. This app hosts a human like avatar which interacts with the user in real time. Users speech inputs are processed and the avatar provides an intelligent response to the user's statements, in a turn by turn conversation.

The primary building blocks for the HoloCompanion are mentioned in the following subsections.

2.1. Speech Recognition

HoloLens provides an inbuilt state-of-the-art voice recognition system (ASR) and dictation analyzer system (DAS) which is used to recognize spoken sentences. The dictation analyzer works in near real time to understand the spoken word stream, and formulate a proper sentence which resembles user utterance.

[1]http://www.msxiaoice.com/
[2]http://www.zo.ai/
[3]http://www.rinna.jp/
[4]https://www.facebook.com/Ruuh/

2.2. Chit-chat

We implement a system similar to the neural dialogue system work presented by Sordoni et al. [1]. Their work improved on the machine translation (MT) based approach proposed by Ritter at al. [2] to incorporate multi-turn conversation as evident in Twitter conversations. The challenge of context sensitive response generation was addressed by using continuous representations or embeddings of words and phrases to compactly encode semantic and syntactic similarity. Embedding-based models afford flexibility to model the transitions between consecutive utterances and to capture long-span dependencies in a domain where traditional word and phrase alignment typically fail to perform. The chit-chat server is hosted as an Azure[5] end point and invoked through a web call with the users speech transcript.

2.3. TTS

We use default Bing text to speech service to convert the response of the chat service to voice output. The speech service is obtained from Bing Speech API [6]. Users can select multiple voice patterns (2 male and 2 female) from the current set.

2.4. Speech to Gestures

We learn a sequence labeling model to predict the probable visemes based on the speech input. We used LSTM based sequence model with the following simple architecture: input layer, 1-LSTM layer and output layer. The input layer takes a sequence of d-dimensional feature vectors, $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_1, \ldots, \mathbf{x}_T)$, where $\mathbf{x}_t \in R^d$ and LSTM layer produces a higher-order feature representation, denoted by $\mathbf{h} = (\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_T)$. In output layer, each frame is assigned with a viseme and denoted by $\mathbf{y} = (y_1, y_2, \ldots, y_T)$, where $y_t \in \{1, \ldots, 22\}$ (sample visemes are shown in Figure 2).

2.4.1. Training Details

Our training corpus consists of anonymized voice search queries. The total queries used for experiment are 1,35,000. We have used Microsoft Speech Synthesis tool [7] to generate speech files and corresponding visemes. We have divided total dataset into train, validation and test sets with 70%, 10%, and 20% respectively. We compute 39-dimensional Mel Frequency Cepstral Coefficients (MFCC) features (including Delta and double Delta) with a 25 ms window and shifted every 10 ms at the current frame, t. Ihe average response time of our chit-chit API is 2.5 sec, hence we chose T as 250 (250 $\times$10ms= 2.5 sec). We padded sequences whose time steps are less than 250 and truncated if it is more than 250. Our model is compiled with *categorical crossentropy* as loss function and *rmsprop* is used for model regularization. The batch size is set to 1000.

2.4.2. Accuracy Results

Classification accuracy is used for model evaluation. The experimental results obtained on validation data for different hidden units 32, 64, 128, 258 are 89%, 90%, 92%, 88% respectively. The best model chosen is based on accuracy on validation data i.e., hidden units equal to 128. Finally, the model is evaluated on test data and observed accuracy 90%. We have deployed our model to azure and created an endpoint to consume it in our MR application.

We have also integrated eye and neck movements into our avatar.

2.5. 3D model

We Designed and developed the 3D model and UV unwrapped[8] in Autodesk Maya[9]. Texture maps were created by using Adobe Photoshop[10]. From the base model we modeled multiple target blend shapes for phonetics, eye lids and eyebrows as shown in Fig. 2. The final Asset was translated from Autodesk Maya to Unity 3D. Shading and lighting attributes were developed within Unity 3D. Finally the unified asset is published as a package to be used for animation generation.

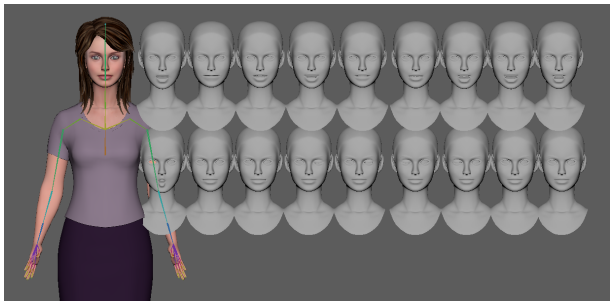

Figure 2: *Avatar details created in Autodesk Maya.*

3. Conclusions

We present an end to end chat companion in a mixed reality environment. The basic parts of such a system is described herein. The individual pieces are still deployed as web APIs. The way forward to reduce the web call latency would be to bring most of these capabilities onto the device. We are working towards a next generation companion which would be completely self contained within the device, thereby reducing the dependence on data bandwidth.

4. Acknowledgements

We would like to thank the Bing team to help with procuring the device as well as helpful guidance and suggestions to make this work better. Niranjan Nayak and Pradepp Rabindranath to help setup the project, Mayank Singh for help with the deployment to other devices and the Garage[11] staff for previous stage demos.

5. References

[1] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J. Nie, J. Gao, and B. Dolan, "A neural network approach to context-sensitive generation of conversational responses," *CoRR*, vol. abs/1506.06714, 2015. [Online]. Available: http://arxiv.org/abs/1506.06714

[2] A. Ritter, C. Cherry, and W. B. Dolan, "Data-driven response generation in social media," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 583–593.

[5] https://azure.microsoft.com

[6] https://azure.microsoft.com/en-us/services/cognitive-services/speech/

[7] https://msdn.microsoft.com/en-us/library/gg145021(v=vs.110).aspx

[8] https://en.wikipedia.org/wiki/UV_mapping

[9] https://www.autodesk.in/products/maya/overview

[10] https://www.adobe.com/in/products/photoshop.html

[11] https://www.microsoft.com/en-us/garage/