INTERSPEECH 2021: Show & Tell Contribution
30 August – 3 September, 2021, Brno, Czechia

Expressive Robot Performance based on Facial Motion Capture

Jonas Beskow[1,2], Charlie Caper[1], Johan Ehrenfors[1], Nils Hagberg[1], Anne Jansen[1], Chris Wood[1]

[1]Furhat Robotics, Sweden
[2]KTH Royal Institute of Technology, Sweden
jonas@furhatrobotics.com, chris@furhatrobotics.com

Abstract

The Furhat robot is a social robot that uses facial projection technology to achieve a high degree of expressivity and flexibility. In this demonstration, we will present new features that takes this facial expressiveness further. A new face engine for the robot is presented which not only drastically improves the visual fidelity of the face and the eyes, it also adds increased flexibility when it comes to designing new robotic characters as well as modifying existing ones. Most importantly, we will present a new toolset and a workflow that allows users to record their own face motion and incorporate them into skills (i.e. custom robot applications) as gestures, prompts or entire canned performances.

Index Terms: social robotics, human-robot interaction, facial animation, motion capture

1. Introduction

Furhat is a platform built for social robotics research and development, constructed specifically with face-to-face interaction in mind. The robot is equipped with advanced built-in capabilities that includes speech, dialogue and computer vision components, allowing the robot to detect users and engage in spoken interaction (with multiple simultaneous users if desired), as well as an animated face, projected onto an (interchangeable) facial mask and a movable neck with three degrees of freedom that gives the robot a high degree of visual expressivity.

Furhat comes with an SDK for developing advanced human-robot interaction applications, referred to as *skills*. Skills offer a comprehensive way to author not only the spoken part of the interaction (speech in/out) but also the non-verbal, allowing the robot to react to actions (e.g., visual attention or a smile from the user) and also display facial *gestures*. The SDK includes several pre-defined gestures, and it allows developers to define custom gestures through code.

In this paper, we present new tools that makes it possible to record new gestures or audiovisual prompts for the robot via facial mocap on a regular iPhone.

2. The Furhat Face Engine

We introduce an all-new face engine that greatly increases the fidelity and expressive capabilities of the Furhat robot. The face engine is built on the industry standard Unity3D platform. The main task of the face engine is to draw the face of the robot, or more precisely, to load a pre-defined face rig and animate it according to parameters it receives in real-time from the main Furhat system. The face rig can be controlled using two independent sets of animation parameters. The first one is the standard Furhat animation parameters set, which consists of 46 parameters grouped by expressions (e.g., affecting the whole face, e.g. emotional expressions "Angry", "Sad" etc.), modifiers (affecting individual parts, e.g. "browUp", "blinkLeft" etc.) and phoneme shapes (for speech animation, e.g. "Ah", etc.). The second set of animation parameters is based on Apple ARKit and is described below.

In addition to animation parameters, the rig also has a set of character parameters that, in combination with different textures and overlays, can be used to create different characters based on the same rig, see below.

Furhat face rigs are constructed in the open source Blender3D and converted into Unity3D assets, which means it is possible to extend the system with custom models in order to create entirely different characters.

2.1. Graphical Fidelity and Cartoon Rendering

The face engine provides high graphical fidelity in the projected faces, while allowing for different character styles to be produced, ranging from cartoon/manga style to realistic.

The rendering engine also supports dynamic shadows and eye reflections that are synchronized with the head movements of the robot, in such a way that when the robot head rotates in physical space, the shadows and eye reflections on the face are updated accordingly.

2.2. Virtual Furhat

The new face engine is not only used for real-time animation on the robot face. It is also embedded in the robot simulation tool known as the *Virtual Furhat*. This tool produces a 3D rendering of the entire robot, including 3 DoF head movements (pan, tilt, roll), and accurately simulates the projection of face animation onto a switchable mask surface (see figure 1). The *Virtual Furhat* is part of the Furhat SDK, which is freely available for download[1] and runs on Windows, MacOS and Linux. Virtual Furhat can be used to preview animations, custom characters and test interactions (skills) built with the SDK.

2.3. Character Customization

The face engine allows for detailed customization of facial features in order to create new characters. Both the shape and texture can be controlled independently. Facial shape can be modified through a set of parameters that include size and positioning of facial features e.g., eye size, eye tilt, nose width,

[1] https://furhatrobotics.com/furhat-sdk/

mouth height etc. Facial texture can be modified in a flexible way through a set of overlays. These overlays make it possible to independently select between a variety of different eyebrows, eyes, lips etc as well as the overall skin texture. Individual color adjustment for each of the facial features is also possible.

The above character controls taken together allow for a wide range of flexibility in creating and customizing characters and matching them to different face masks. For certain kinds of characters, e.g., when there is a desire to produce a replica of an existing person, the texture controls and overlays might not provide enough expressive freedom. For such cases it is also possible to import custom textures.

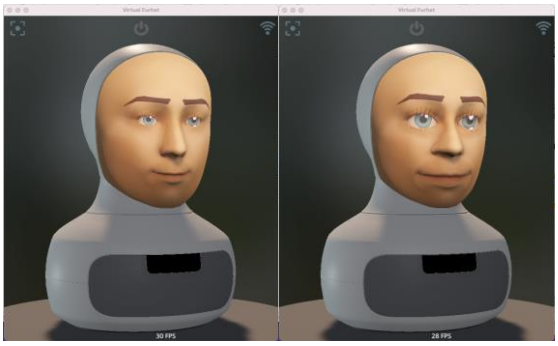

Figure 1: *Virtual Furhat - effect of different character parameter settings.*

2.4. Apple ARKit Blendshapes

The new face engine supports animation though Apple ARKit[2] blendshapes, which is a set of 52 low-level blendshapes for face animation controlling different parts of the face (eyes, jaw, mouth, cheeks, nose and tongue). Using the front-facing camera on an iPhone X (or later model), ARKit is able to estimate these blendshapes from the face of the user in real time, so any iPhone app built on top of ARKit has direct access to these values and can either stream them (e.g., for live animation) or store them in memory.

In the Furhat face engine, ARKit parameters can be used interchangeably with the standard set (described above) or in combination, in which case they have an additive effect. The ability to animate the robots face and head via ARKit parameters opens up a host of possibilities to developers, researchers and users of the Furhat robot. Recording of gestures and prompts and using these as building blocks in a robot skill makes it possible to build highly expressive interactive content. It is also possible to record entire performances, e.g., of speech or singing. Other possibilities that are not part of this demonstration include live streaming of face data, e.g., for telepresence interaction and behavior generation models based on machine learning, facilitated by the fact that face animation data can be captured from users in high quality and in large quantities.

3. Robot Show and Tell

In this demonstration we showcase the new facial capabilities of the Furhat robot. We focus on the ability to use an iPhone for facial motion capture, and taking captured performances consisting of face parameters, audio and video (for reference), and editing/converting these recordings into gestures (with or without sound) that can be used either as part of an interactive skill, or as a stand-alone performance. See figure 2 for an illustration. Below we describe the workflow, step by step

1. A facial performance is captured using an iPhone X (or later model). Capture is done using a third-party app (*Live link face* from Epic Games[3]) which saves face motion (52 blend shapes + 3 DoF head rotations @ 60 fps) to a CSV file. It also saves a video including sound.
2. The captured data is transferred from the phone to a computer and imported to a custom Gesture Editing tool. This tool makes it possible to select what portion of the recording to use, both in time (by selecting start/end position with a slider) and in terms of parameters (by selecting groups of blendshapes to include: eyes, mouth, head motion etc.). One can also select whether or not to include sound. The gesture can be previewed using the Virtual Furhat. Finally, it is exported as a JSON file. If audio is included, a separate WAV file is exported.
3. The last step is to invoke gestures from within a skill on the robot. When a skill is built in the SDK, the exported JSON/WAV files are placed in the resource folder of the skill, which makes them accessible for invocation from the application logic.

Aside from the gesture recording/editing functionality, it will be possible during the demonstration to experiment with character shape- and texture settings and explore how they interact with different masks on the robot.

Figure 2: *Facial motion capture using ARKit and mapping to the robot*

[2] https://developer.apple.com/augmented-reality/arkit/

[3] https://apps.apple.com/se/app/live-link-face/id1495370836