The KLAIR toolkit for recording interactive dialogues with a virtual infant

Mark Huckvale

Department of Speech, Hearing and Phonetic Sciences,
University College London, U.K.

m.huckvale@ucl.ac.uk

Abstract

The goals of the KLAIR project are to facilitate research into the computational modelling of spoken language acquisition. Previously we have described the KLAIR toolkit that implements a virtual infant that can see, hear and talk. In this demonstration we show how the toolkit can be used to record interactive dialogues with caregivers. The outcomes are both an audio-video recording and a log of the "beliefs" and "goals" of the infant control program. These recordings can then be analysed by machine learning systems to model spoken language acquisition. In our demonstration, visitors will be able to interact with KLAIR and try to teach it the names of some toys.

Index Terms: speech acquisition, computer models of language acquisition

1. Introduction

The KLAIR toolkit was launched in 2009 [1] with the aim of facilitating research into the machine acquisition of spoken language through interaction. The main part of KLAIR is a sensori-motor server that implements a virtual infant on a Windows PC equipped with microphone, speakers, webcam, screen and mouse, see Fig 1. The system displays a 3D talking head modelled on a human infant, and can acquire and process audio and video in real-time. It can also speak using an articulatory synthesizer, look around its environment and change its facial expression. Machine-learning and experiment-running clients control the server over network links using a simple API. KLAIR is supplied free of charge to interested researchers from http://www.phon.ucl.ac.uk/project/klair/.

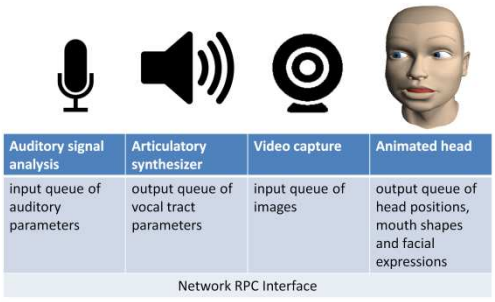

Fig 1. KLAIR server architecture

Our hope is that the KLAIR toolkit will encourage researchers, students and hobbyists to do research and build new applications related to spoken language acquisition. The KLAIR server contains all the real-time audio and video processing including auditory analysis, articulatory synthesis, video capture and 3D head display. Data acquisition and control of the server can be performed over an exposed API by client applications. The server maintains processing and analysis queues which mean that clients do not have "keep-up" with flows of data. In addition client applications can be written in any language that supports remote procedure calls; the toolkit currently supplies interfaces for MATLAB and Microsoft's .NET languages.

In this demonstration we show how KLAIR can be driven by a client application designed to encourage caregivers to engage in infant-directed dialogues. The outcomes are tagged audio-video recordings of caregivers' speech suitable for machine learning studies.

2. Demonstration

2.1. Visual Appearance

Subjects ("caregivers") interact with KLAIR through a screen displaying the 3D infant head surrounded by 3D models of toys, see Fig 2. Caregivers can interact with the infant by speaking, by touching or clicking on the toys, or by touching or clicking on the head. KLAIR can respond by looking at the caregiver, by looking at a toy, or by looking around generally.

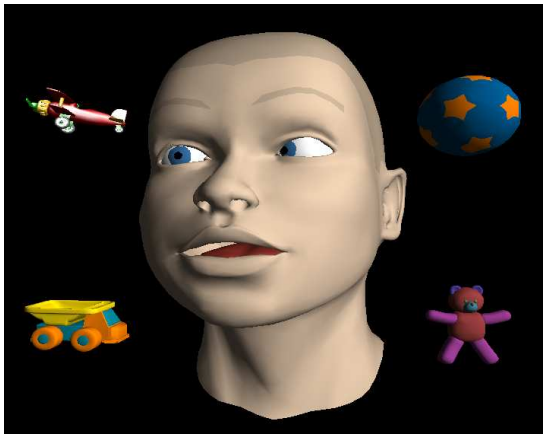

Fig 2. 3D Toys give the caregiver and the virtual infant something to talk about.

When the dialogue starts, KLAIR is asleep and still. Some event wakes up the infant who looks around and takes an interest in movement in the web cam, movement in the toys, or some speaking event.

Initially the response of the infant to speech from the caregiver is rather random: the infant might look out of the screen, look at a toy or ignore the person. However by indicating a toy and producing an utterance that appears to be the name of the toy, over time the infant appears to associate toys with names. It might respond, for example, by looking at the toy that has been named. The infant will also come up with spoken utterances of its own, which may be random, may be directed at the person or may be the name of a toy. Caregivers can also give reward and correction utterances which affect the

mental state of the infant. More success and more speaking by the caregiver will make the infant happier, while lack of success and little speaking will make it less happy. Eventually the infant becomes tired and goes back to sleep.

2.2. System Configuration

The system has two parts: the KLAIR server which performs all the real-time processing and interaction with the caregiver, and a background client application which handles events, maintains the status of the infant and decides on actions. The two parts communicate over a network link using remote procedure calls (RPC). The server provides access to its input audio queues and to its output queues for synthesis, head position and facial expressions using a simple API. In addition the server is responsible for making a real-time video and audio recording of the interaction. It also keeps a log of events, which include mouse clicks and touches on the 3D models and on the head, together with actions that change head position or expression. The log file and video file are saved together for subsequent alignment and analysis.

Once connected to the server, the client application requests the loading of the 3D model toys into the virtual space occupied by the head. Once the experiment is started the client application pipes microphone audio from the server into the local SAPI speech recogniser. The recogniser is configured with a grammar which identifies commonly used utterances and assigns semantic tags which are then used to drive the dialogue management and learning system.

2.3. Dialogue Management & Learning

The actions of the client program are centred around events, states and actions, as listed in Tables 1, 2 & 3.

Event	Description
Speech detected	Caregiver started speaking
Toy name message	ASR hears name of toy
Reward message	ASR hears reward
Correction message	ASR hears correction
Attention message	ASR hears request for attention
Toy model touched	Caregiver has touched a toy
Klair touched	Caregiver has touched infant
Timeout	No interaction for time period

Table 1. Events that drive the client program

State	Description
Percentage time through test	Record of how much is learned
Names identified for toys	What names are being used for the toys by the caregiver?
Names used for toys	What phones are being used to speak toy names
Is speaking	Infant is currently speaking
Is listening	Infant is currently listening
Satisfaction index	How well is interaction going

Table 2. State variables in the client. Responses to events are conditioned on current state.

Action	Description
Look at caregiver	Position head and eyes
Look at a toy	Position head and eyes
Look around	Position head and eyes
Speak the name of a toy	Articulate a known word
Babble	Articulate a random word

Table 3. Possible actions taken by the client in response to events and the current state

In the current experiments no actual learning takes place. Instead the client is programmed such that it is able to satisfy the goals of the interaction with the caregiver, and then that ability is revealed slowly as the dialogue proceeds. So the client starts out knowing the names of each toy and how to articulate their names. However the client's responses through the infant are deliberately randomised and distorted at first to make it appear to the caregiver that the infant does not know anything. As the interaction progresses, the amount of added randomisation is reduced, so that the infant appears to be adapting his behaviour to the caregiver.

3. Future Work

We are currently using KLAIR to collect virtual infant-caregiver dialogues. Our initial goals are quite modest: to determine whether caregivers are willing to "suspend disbelief" and interact with the virtual infant using strategies found for interactions with human infants. As a secondary goal, we will see how well the caregivers notice the adaptive behaviour of the infant and whether as a consequence they adapt their own language behaviour. We hope to make the audio-video recordings of these interactions available to others for further study.

Once we have shown that our experimental setup is a reliable source of linguistic interactions, we can start to look at the machine learning challenges of on-line speech acquisition. In future work we would like to implement computational models of the acquisition of speech perception and production such as [2, 3] but in real-time using the interactional framework provided by KLAIR.

We hope that this demonstration will encourage more researchers to use the KLAIR toolkit to experiment in this area.

4. Acknowledgements

Thanks to Sascha Fagel for help in programming the talking head. Thanks to Priya Sharma for her enthusiasm for the project which has encouraged me to pursue yet more software development on KLAIR.

5. References

[1] Huckvale, M., Howard, I., Fagel, S., "KLAIR: a Virtual Infant for Spoken Language Acquisition Research", Interspeech 2009, Brighton, U.K.

[2] Guenther, F.H., "A neural network model of speech acquisition and motor equivalent speech production", Biological Cybernetics, 71 (1994) 43-53.

[3] Westermann, G., Miranda, E., "A new model of sensorimotor coupling in the development of speech", Brain and Language, 89 (2004) 393-400.