Recording caregiver interactions for machine acquisition of spoken language using the KLAIR 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 paper we describe how the toolkit has been enhanced and extended to make it easier to build interactive applications that promote dialogues with human subjects, and also to record and document them. Primary developments are the introduction of 3D models, integration of speech recognition, real-time video recording, support for .NET languages, and additional tools for supporting interactive experiments. An example experimental configuration is described in which KLAIR appears to learn how to say the names of toys in order to encourage dialogue with caregivers.

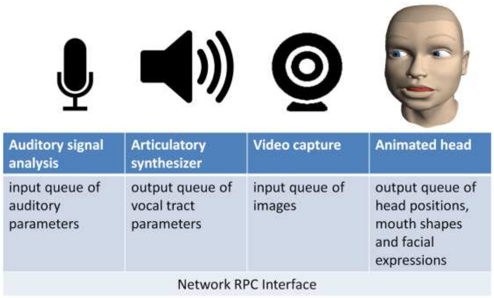
Index Terms: speech acquisition, computer models of language acquisition

1. Introduction

Research into the machine acquisition of spoken language from audio is still in its infancy. Most studies treat the problem as one of pattern discovery from passively acquired audio, where machine learning algorithms are applied off-line to data collected by "listening in" to human infant-caregiver dialogues [e.g. 5, 6]. With my colleagues Ian Howard and Piers Messum we have proposed that such an approach misses out the essential aspect of language that is key to its acquisition. Human infants learn how language is used to describe and control the world by interacting meaningfully with caregivers, not by listening in on what other people say to each other. If we are to model that acquisition in machines, we need to embody the learning process within a system that interacts with human caregivers [3].

There are however, many practical difficulties faced by researchers wanting to engage in research in this area. Building an interactive machine learning system for speech acquisition is far from easy even without considering the linguistic issues. To learn language through interaction, the machine needs to engage with caregivers in real-time. It needs to react when the caregiver starts speaking, and to respond quickly through speaking itself. The machine needs to be able to sense objects and events in the world, so that dialogues are *about* something. The machine needs to have a form that encourages caregivers to want to interact with it, and it must demonstrate that it notices and reacts to actions by the caregiver. Not only does the system need to learn on-line, it needs to do so fast enough that caregivers notice and adapt their behaviour to encourage further learning. Lastly, the very people most interested in language acquisition: the linguists, psycho-linguists and speech scientists, rarely have the technical skills to implement a system that is more akin to social robotics than experimental phonetics.

The KLAIR toolkit was launched in 2009 [4] 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 modern Windows PC equipped with microphone, speakers, webcam, screen and mouse, see Figs 1 & 5. The system displays a talking head modelled on a human infant, and can acquire audio and video in real-time. It can speak using an articulatory synthesizer, look around its environment and change its facial expressions. 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/.



Auditory signal analysis	Articulatory synthesizer	Video capture	Animated head
input queue of auditory parameters	output queue of vocal tract parameters	input queue of images	output queue of head positions, mouth shapes and facial expressions
Network RPC Interface			

Fig 1. KLAIR server architecture

The KLAIR toolkit makes it much easier to create applications designed to collect infant-caregiver interactions for the study and modelling of 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. Client applications can be written in any language that supports remote procedure calls. For example, the original KLAIR toolkit provided a MATLAB interface. The software is also open source and freely available.

Since the original release of KLAIR we have been planning how to use it to research into elementary, pre-linguistic, human-infant interactions. We realised that we first needed to explore the extent to which human subjects would be willing to act as caregivers for a virtual infant. Would subjects be willing to engage in dialogues at all? What infant behaviours would most encourage them to do so? Would there be similarities between virtual-infant-directed speech and infant-directed speech? Would our subjects notice adaptive behaviour in the infant, and would they adapt their own behaviour accordingly? This paper describes some of the ways we have extended the KLAIR toolkit to support these research questions.

2. KLAIR Development

To achieve the goals of our research required additional functionality to be built into the KLAIR Toolkit. Some extra functionality has been added to the server application, some changes have been made to the API, and some new utilities have been added. These are described below.

2.1. 3D models

The first challenge was to consider what would be the goals of the interaction that we were going to ask caregivers to perform. Previous work in social robotics has provided human subjects with physical objects that they can show to and describe to the robot [1]. Since the visual processing capabilities of KLAIR are rather limited, and because to implement visual object recognition would divert us from our main goals, we chose instead to implement virtual objects that are "in sight of" both the infant the caregiver. These objects not only create foci of attention for the caregiver and for the infant, but also form the basis for language games based around the names of the objects. The caregiver can then be asked to describe the objects or to teach the infant their names, and we can assess how the linguistic behaviours of the caregiver are adapted to the responses of the infant.

To implement the virtual objects, support was added to KLAIR for the display of 3D models within the virtual space occupied by the talking head, see Fig 2. The models are 3D graphical objects stored in WaveFront OBJ file format, and displayed through the OpenGL library used for the head. The objects can be loaded remotely from client applications, and they can be positioned in the 3D space, and given translational and rotational velocities. For our current work we have chosen models representing children's toys that we have obtained from free sources of models found on the internet.

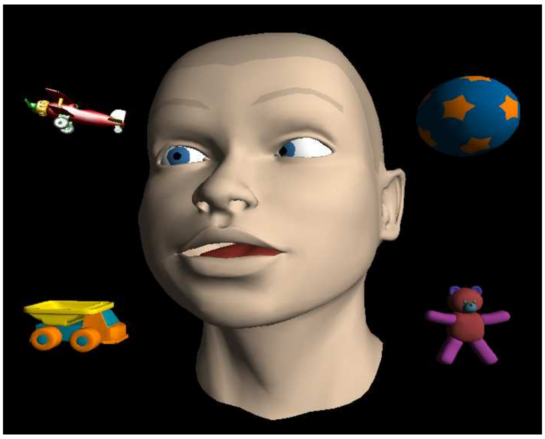


Fig 2. 3D Toys give the caregiver and the virtual infant something to talk about.

To allow caregivers to interact with the models, we also detect mouse clicks and touch-screen presses on the objects, and these signals can be used to raise events in the client tools to trigger infant responses. For example, touching a model might make it rotate and the infant head might turn to look at it.

2.2. Dialogue recording and documentation

Since our research goals involve analysing how our caregivers interacted with the infant, we needed to have a means to document the interaction. First we implemented real-time video capture from the webcam to disk to capture the caregiver's behaviour. This is supplemented by a stereo audio signal comprising the caregiver's microphone signal and the infant's articulatory synthesis. To document the behaviour of the virtual infant in response to caregiver action, the server also maintains a log of events, which automatically include mouse clicks and touches on the 3D models and on the head, together with actions that change head position or expression. The client application can also add event messages to the log, so as to record the "state of mind" of the agent at different stages in the interaction. The log file and video file are saved together for subsequent alignment and analysis.

2.3. Support for .NET

While KLAIR is written in Visual C++ we realised from the outset that we could not expect researchers in the field of spoken language acquisition to have appropriate skills to program KLAIR at this level. Our first additional API was a MATLAB interface programmed through MEX functions. This allowed client applications to be programmed and run on remote computers talking to the KLAIR server over the network without getting involved in the complexities of remote procedure calls.

While MATLAB might be a suitable API for the creation of machine learning applications, it seems less suited to the creation of simple GUI applications for running experiments. Also since it is a commercial product, it is not so popular among students and hobbyist experimenters. Because of this, we have added an interface to the .NET languages created and supported by Microsoft. These include Visual Basic and C#. The .NET languages are modern, compiled, object-oriented programming languages that use a common runtime library. Microsoft provides free "Express" editions of the languages which include a sophisticated integrated development environment. Much tutorial material is also available for learning how to program in these languages.

The KLAIR API for .NET is built around a native library KlairLink.dll, which can be called from within .NET managed code using the P/Invoke mechanism. Example code shows how easy this is in practice.

The provision of the .NET API now allows students and hobbyists to build client applications to control KLAIR using only free software.

2.4. Speech recognition

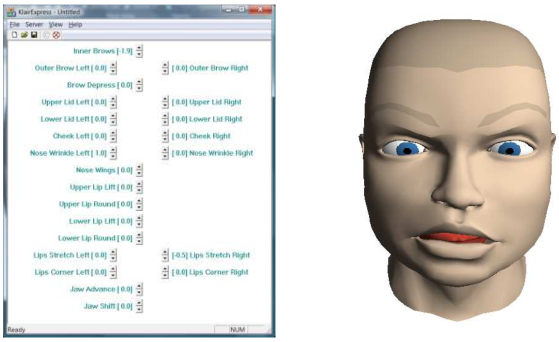
In the longer term, KLAIR will need to learn how to recognise speech as part of the language acquisition process. However, even for our current application we need a mechanism for the infant to detect and respond to different utterances spoken by the caregiver. Fortunately the speech recognition system integrated into the Windows operating system exposes an API through the Common Object Model (COM) which can be accessed from within .NET applications. Thus the .NET API we have provided for KLAIR also gives us access to a real-time speech recogniser.

One problem however, is that the recognition needs to be integrated into client applications, but the audio signal is collected from the caregiver interacting with the server. Thus the KLAIR server now also supports export of the unprocessed microphone signal across the network so that the audio stream can be presented to the recogniser on the client.

The Windows Speech-API recognizer can be constrained using a supplied finite-state grammar or with a dictation grammar. The use of semantic tags in a grammar allows the recogniser to combine a large number of utterance variants into a small number of events meaningful to the client application. The recogniser can also raise events in the client when the caregiver starts and stops speaking, which might be used to signal the head to look out of the screen, for example.

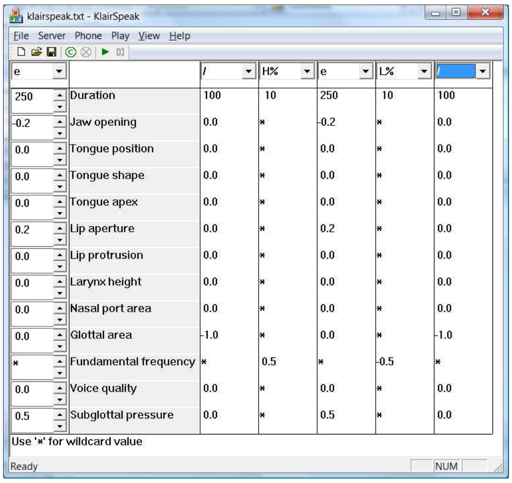
2.5. Additional tools

Two additional tools have been added to the KLAIR toolkit to aid in the running of experiments. The first, KlairExpress, aids in the development of suitable facial expressions, see Fig 3. The second, KlairSpeak, aids in the development of a phone inventory for articulatory synthesis, see Fig 4.



Inner Brows [-1.9]
Outer Brow Left [0.0]
[0.0] Outer Brow Right
Brow Depress [0.0]
Upper Lid Left [0.0]
[0.0] Upper Lid Right
Lower Lid Left [0.0]
[0.0] Lower Lid Right
Cheek Left [0.0]
[0.0] Cheek Right
Nose Wrinkle Left [1.0]
[0.0] Nose Wrinkle Right
Nose Wings [0.0]
Upper Lip Lift [0.0]
Upper Lip Round [0.0]
Lower Lip Lift [0.0]
Lower Lip Round [0.0]
Lips Stretch Left [0.0]
[-0.5] Lips Stretch Right
Lips Corner Left [0.0]
[0.0] Lips Corner Right
Jaw Advance [0.0]
Jaw Shift [0.0]

Fig 3. KlairExpress can be used to design custom expressions.



e		/	H%	e	L%	/
250	Duration	100	10	250	10	100
-0.2	Jaw opening	0.0	*	-0.2	*	0.0
0.0	Tongue position	0.0	*	0.0	*	0.0
0.0	Tongue shape	0.0	*	0.0	*	0.0
0.0	Tongue apex	0.0	*	0.0	*	0.0
0.2	Lip aperture	0.0	*	0.2	*	0.0
0.0	Lip protrusion	0.0	*	0.0	*	0.0
0.0	Larynx height	0.0	*	0.0	*	0.0
0.0	Nasal port area	0.0	*	0.0	*	0.0
0.0	Glottal area	-1.0	*	0.0	*	-1.0
*	Fundamental frequency	*	0.5	*	-0.5	*
0.0	Voice quality	0.0	*	0.0	*	0.0
0.5	Subglottal pressure	0.0	*	0.5	*	0.0

Use '*' for wildcard value

Fig 4. KlairSpeak can be used to build and test an articulatory phone inventory

3. Example Application

In this section we present an overview of the hardware and software configuration used in our current experiments to collect caregiver-infant interactions.

3.1. Hardware configuration

The server runs on a 3GHz Windows PC connected to a touchscreen and webcam sitting in a sound-treated room, see Fig 5. A client application runs on a 2GHz Windows laptop outside the room. The two PCs are connected by wired network. Audio is recorded using the web cam microphone (Logitech Pro 9000) to make the act of being recorded less obvious. Throughout the recording session a real-time audio-video capture to disk runs on the server.

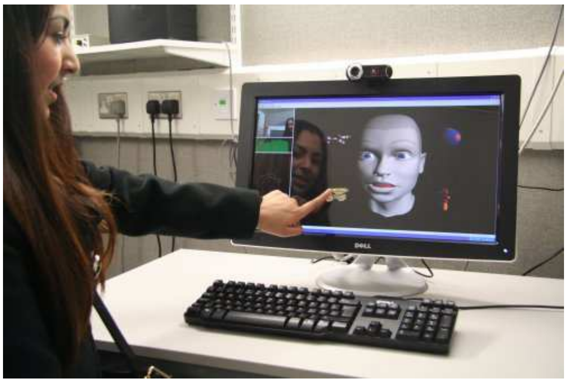


Fig 5. Recording configuration with touchscreen

3.2. Client Configuration

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 client program manages interactions with caregivers using a set of linked 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.

3.3. How KLAIR "learns"

In the current experiments no 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.

4. Future Work

Experimental dialogues are currently being collected using the enhanced KLAIR toolkit. 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 on-line machine learning challenges of speech acquisition. In future work we would like to implement computational models of the acquisition of speech perception and production such as [2, 7] but in real-time using the interactional framework provided by KLAIR.

We hope that the additions we have made to KLAIR will also encourage more researchers to experiment in this area.

5. 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.

6. References

[1] Fischer K., Kilian F., Rohlfing K. & Wrede B., " Mindful Tutors: Linguistic Choice and Action Demonstration in Speech to Infants and Robots", Interaction Studies, 12 (2011) 134-161.

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

[3] Howard, I., Messum, P., “Modeling the development of pronunciation in infant speech acquisition”, Motor Control 15(1) (2011) 85-117.

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

[5] Roy, D., Pentland, A., "Learning words from sights and sounds: a computational model", Cognitive Science 26 (2002) 113-146.

[6] ten Bosch, L., van Hamme, H., Boves, L., Moore, R., "A computational model of language acquisition: the emergence of words", Fundamenta Informaticae, 90 (2009) 229-249.

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