

KLAIR: a Virtual Infant for Spoken Language Acquisition Research

Mark Huckvale¹, Ian S. Howard², Sascha Fagel³

¹ Department of Speech, Hearing & Phonetic Sciences, University College London, London, U.K.

² Computational & Biological Learning Laboratory, University of Cambridge, Cambridge, UK

³ Department for Language & Communication, Berlin Institute of Technology, Berlin, Germany

m.huckvale@ucl.ac.uk, ish22@cam.ac.uk, sascha.fagel@tu-berlin.de

Abstract

Recent research into the acquisition of spoken language has stressed the importance of learning through embodied linguistic interaction with caregivers rather than through passive observation. However the necessity of interaction makes experimental work into the simulation of infant speech acquisition difficult because of the technical complexity of building real-time embodied systems. In this paper we present KLAIR: a software toolkit for building simulations of spoken language acquisition through interactions with a virtual infant. The main part of KLAIR is a sensori-motor server that supplies a client machine learning application with a virtual infant on screen that can see, hear and speak. By encapsulating the real-time complexities of audio and video processing within a server that will run on a modern PC, we hope that KLAIR will encourage and facilitate more experimental research into spoken language acquisition through interaction.

Index Terms: speech acquisition, machine learning, autonomous agent, situated learning, toolkit

1. Introduction

Research into the machine acquisition of spoken language is still in its infancy. Most research has concentrated on sub-parts of the problem, such as the segmentation of the speech stream into words [1], the discovery of perceptual categories corresponding to phonological choices [2] or the development of articulatory gestures through imitation [3]. Furthermore this research has mostly been "off-line" in the sense that the caregivers' spoken interactions are recorded, stored in a corpus, and processed in batches at a later time.

Recently we have come to realise the importance of embodiment and real-time interactions as essential parts of spoken language acquisition [4] and to incorporate these elements within our research. In [5], for example, Howard and Messum showed that an infant can acquire the production of proto-words not by imitation but by being rewarded by a caregiver – the initial form of the words arising merely through the infant exploring the capabilities of its vocal tract. In ongoing work we are looking at the use of adult reformulations of infant sounds to address the problem the infant has in associating and matching adult sounds to its own vocal productions.

This focus on interaction also fits well with theoretical accounts that treat language acquisition as much a social process as a cognitive one [6]. In such accounts, language is acquired as a product of attempted meaningful communication with caregivers, rather than as the result of passive observation of language use among others. The implication is that better models of infant language acquisition and better performing machine learning systems will occur when interaction is given its proper value. For machine learning this will involve placing

our learning system within an agent that is situated in the world where it can both sense its environment and communicate about it with caregivers.

While many researchers may support the need for a situated autonomous agent for language acquisition, few researchers have the technical competence to implement such an agent. Existing tools in this area are extremely expensive and difficult to use; for example the iCub robot [7] costs thousands of pounds and requires an understanding of how to control an elaborate mechanical system. We believe the unavailability of suitable tools to support spoken language interaction has hindered research in the area. Thus to support and encourage further research we have developed KLAIR: a toolkit for Language Acquisition through Interactions in Real-time. 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. 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 and it can sense the position of parts of its virtual body. The server communicates with a background machine learning client that could run on the same or a separate computer. We aim to supply KLAIR free of charge to interested researchers.

In this paper we describe and justify the decisions that went into the design of the KLAIR server. We provide some technical details about its implementation and give suggestions for research directions that could be facilitated by the system.

2. KLAIR Design

In this section we describe the design goals and justify the design decisions of the KLAIR toolkit.

2.1. Why Interactive?

Fundamentally, KLAIR is designed to instigate and to capture interactions between a machine learning "agent" and a human caregiver. On-line interactions allow the agent to observe how its outputs affect the behaviour of the caregiver, and for the caregiver to adapt their responses to the properties of those outputs. In addition there is often an assumption in off-line learning that the system or the caregiver are not changing with time over a series of communication events. But infants can easily tell the difference between live and recorded interactions [8] precisely because the interactions are adaptive to the state of the parties.

There is a lot of work on language acquisition that concerns itself with discovering the underlying structure of language through passive observation of statistics of the surface form of language. For example, there are systems that "discover" words by looking for repeating spectro-temporal patterns in the speech signal [9]. The essential limitation of

this idea is that the learning agent does not know what objects are being referred to and cannot use information from the real world to establish whether two audio patterns are different versions of the name of some object or whether they refer to two different objects. For an agent to build a discriminative model of speech, which identifies lexical choices and establishes a phonology of language, it needs to discover through interaction which utterance components are "the same" and which are "different".

Interactions are also very useful to the agent in exploring how spoken communication works. From the earliest attempts at using facial expressions to indicate awareness to the use of vocalised sounds to elicit caregiver responses, interactions are essential to allow the agent to explore the range of useful motor outputs and to learn how to exploit them to gain reward in turn taking and dialogue.

2.2. Why Multi-modal?

If we situate an agent in an environment where both objects and spoken descriptions can be perceived, then the agent has the opportunity to learn both words and their meaning. It can then compute a probability distribution over words for some given object or event that can be used in both recognition and in expressive speech. Conversely, common characteristics of utterances used by caregivers in different situations may indicate conceptual links between objects: categories such as "toys" or "food" for example.

Another important aspect of multi-modality is perception of self – awareness that motor outputs have sensory consequences. Such exteroception and proprioception are essential in learning how to perform efficient motor control. Speech articulation is a skilled action [10] incorporating immediate compensation that can only be learned with feedback from the effects of the motor system. To learn control of a real vocal tract it is clearly essential to perceive its auditory effects, so that auditory goals can be maintained. But in addition it is important to know where the articulators are currently positioned, particularly in the presence of noise and perturbations, since in these cases feed-forward control will make inadequate predictions. Perrier [11] found evidence that mental representations of speech production are multimodal, associated with regions of the acoustic, orosensory and motor control spaces, with the acoustic modality having the highest level of priority.

Multimodality also provides additional channels for communication between agent and caregiver. Sensing the caregiver's response to an utterance will be useful to the agent for reinforcement learning and for turn-taking. Using the agent's facial expressions to indicate emotional states may help the caregiver match feedback to the agent's needs.

2.3. Why Embodiment?

It is important that we embody the agent for several reasons. Having a vocal tract, even a rough simulation of one, not only provides important constraints on the kinds of speech sounds that can be generated but also provides constraints on the process of learning how to speak. If we hope to draw parallels with infants, then a vocal tract driven by motor commands is a pre-requisite. In addition, a vocal tract model makes the link between speech sounds and the shapes of the jaw and lips producing a more convincing illusion that the virtual infant is actually speaking.

It would have been possible to build an interactive, multi-modal agent without a face, but there are a number of reasons to think that making the agent look somewhat like a human infant will be advantageous. Firstly, we want to encourage our

experimental subjects to talk to the agent in a (relatively) natural way, with an expectation that their responses will be similar to those they would have given to a real infant. An infant face will hopefully make their experience an engaging one, and may even provoke the use of motherese. Caregivers are also expert in decoding facial expressions and so these can be exploited by the agent to obtain required responses. For example, simply looking at the caregiver or looking at an object when speaking may indicate the difference between a command and an observation.

Additionally, embodiment demonstrates the importance we put on modelling the agent to at least some degree on the capabilities of a human infant. Of course at the current state of technology, such modelling is extremely crude. But human infants are the only systems known to us that fully solve the spoken language acquisition problem and it may be that comparisons between the agent and an infant can give us clues as to how to improve the agent's performance. If we fail to make the link to infant development and behaviour, then any discoveries we make can always be criticised as not being relevant to the human situation.

2.4. Why Real-time?

We have already pointed out that the advantage of interactive learning is that the communication between agent and caregiver can develop and adapt from one conversational turn to the next. To achieve natural interactions the system must operate in close to real-time, that is the response of the agent to a caregiver's utterance must be quick enough for a response to be associated with a stimulus. The system must be able to control its vocal tract at typical articulatory rates if it is to make speech-like sounds. To give the illusion of a virtual infant, movement of the jaw and lips must be synchronised with the spoken output. Slowed down speaking or understanding would disturb the naturalness of infant-caregiver dialogue.

2.5. Why Separate Learning Agent from Sensori-motor Server?

KLAIR is implemented as two separate components: a machine-learning (ML) client and a sensori-motor server. There are a number of advantages to this configuration. Firstly the sensori-motor server contains all the real-time audio and video processing, that is both complex and close to the PC hardware. To create the illusion of presence, the agent must look as though it is constantly aware of its surroundings even if it is performing a lot of background processing. Thus the server application needs to be autonomous to the extent that it constantly presents a body state that changes smoothly over time. The background client polls the server to receive past input and queues motor commands to be executed in the future. While it is better for the client to "keep up" with the conversation, it is not operating under the same time pressures as the foreground simulation.

Another advantage of the separation between client and server is that it allows other researchers to use the server even when they use different technologies and programming languages. We have chosen a very simple asynchronous function-call protocol to connect client and server which allows the client to be written in a number of different programming languages. The call protocol also operates over a computer network, allowing the client to reside on a different computer to the server, or even for the server to be controlled by multiple clients.

3. KLAIR Implementation

3.1. Configuration

The KLAIR toolkit is designed to run on a modern Windows PC with a microphone, speakers, webcam, mouse and screen. The central component is the sensory-motor server application that provides sensory input and motor output for a separate machine learning client application. The server runs multiple real-time processing threads, while the learning system runs asynchronously in the background and which polls the server to input audio, video and sensory signals, or to deliver facial expressions or vocal output. See Fig 1. The server can optionally log all I/O to disk for off-line processing.

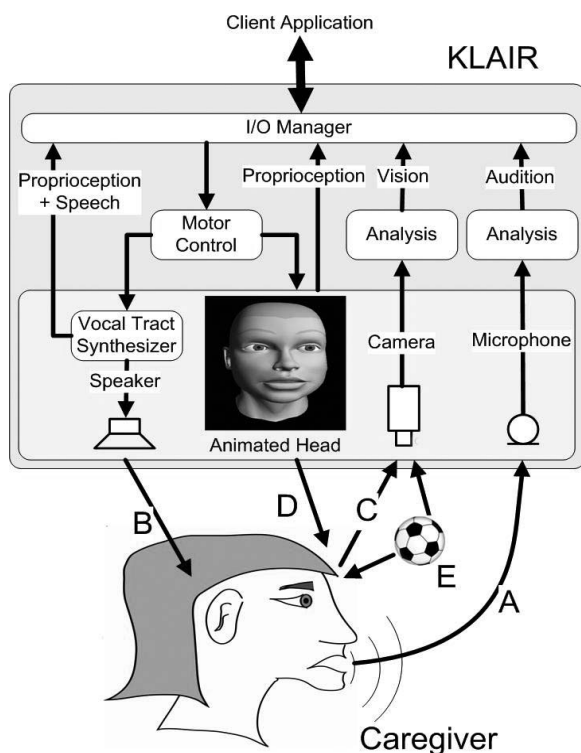


Fig 1. KLAIR and its interactions with a caregiver. KLAIR receives auditory input from a caregiver using a microphone A and corresponding visual input from a webcam C. These signals are processed and sent to the I/O manager. The I/O manager also receives incoming motor commands from a client application and passes them to a motor controller. Speech output B is generated using an articulatory synthesizer and its movements are synchronized to an animated head D. Objects in the environment E can be seen by both KLAIR and the caregiver.

3.2. Motor Output

The audio output stream is generated through an adaptation of the articulatory synthesizer of Shinji Maeda [12] to approximate an infant-sized vocal tract. This takes 10 articulatory parameters as input: JW: Jaw Position, TP: Tongue Position, TS: Tongue Shape, TA: Tongue Expansion, LA: Lip Aperture, LP: Lip Protrusion, LH: Larynx Height, NS: Velopharyngeal port opening, GA: Glottal Aperture, FX: Fundamental Frequency, VQ: Voice Quality, and PS: subglottal pressure. Dynamical smoothing of the parameters over time is applied using a critically-damped second-order spring-mass system.

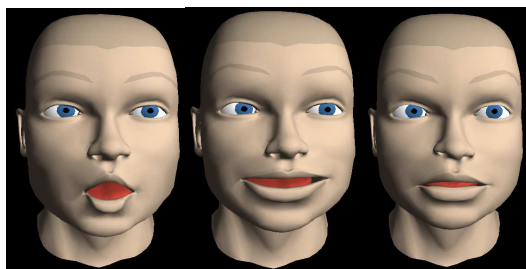


Fig 2. KLAIR Infant Talking Head (from left to right): open jaw plus reduced lip spreading with neutral expression, displaying pleasure, and displaying confusion.

The visual appearance of the server is as an infant talking head (see Fig 2). This is an OpenGL implementation of the talking head MASSY [13] adapted to appear as a virtual human infant. The head is controlled using three sets of animation parameters: one describing the articulatory movements of speech, one describing facial expressions, and one describing movements of the head and eyes. The 6 articulatory parameters are: vertical jaw opening, tongue advancement/retraction, vertical tongue dorsum position, vertical tongue tip position, vertical lip opening, and lip spreading/protrusion. These parameters have been designed to enable the display of visually distinguishable German and English phonemes. They are derived automatically from a linear combination of the control parameters of the articulatory synthesizer, where the coefficients are manually optimised for the best match of vocal tract shapes in both parameter models. There is the possibility for a unified parameter model in a future version of the toolkit.

The 23 facial expression parameters are: inner eyebrow riser, outer eyebrow riser (left/right), eyebrow depressor, upper and lower eyelid depressors (l/r each), cheek raiser (l/r), nose wrinkle (l/r), nose wings opener, upper and lower lip raisers and protruders, lip stretchers and depressors (l/r each), and jaw advancer and side shifter. A huge variety of facial expressions – e.g. for displaying emotional states – can be obtained by combining these facial expression parameters. The set of animation parameters is inspired by action units of the facial action coding system (FACS, [14]) but is in contrast to FACS designed for the generation of facial expressions instead of their description. Fig. 2 shows examples of the articulation and facial expressions of the infant talking head.

3.3. Sensory Input

The audio input stream delivers information about the loudness, pitch and timbre of sounds currently being acquired from the microphone or generated by the articulatory synthesizer. A psychoacoustic model of the auditory periphery is used to deliver estimates of loudness. An autocorrelation analysis provides estimates of pitch, while an auditory filterbank based on the channel vocoder provides estimates of spectral envelope across 24 frequency channels. Estimates are provided in real time at 100 frames/sec.

Video capture is performed using the Windows VFW interface. The lowest supported resolution is 320×240 pixels at 10 frames/sec. Captured frames are each converted to an RGB bitmap. Future enhancements of the toolkit may include some level of visual feature representation.

The client can also obtain proprioception and exteroception input from the server. Proprioception input returns the current vocal tract configuration, including information about articulator contact. Exteroception input converts mouse movements to a form of touch sensation,

depending on whether the mouse cursor is over the visualised face.

3.4. Machine Learning Client

The ML client contains all the machine learning components used in experiments with the toolkit. The client communicates with the server through asynchronous remote procedure calls (RPC). We anticipate a rather slow polling rate of about 10 calls/sec, transferring 10 audio frames and one video frame per call, but higher rates are possible. The server will maintain a short history of frames for when the client falls behind. We have endeavoured to make the interface as simple as possible and not to restrict the computer languages in which the executive may be programmed by toolkit users. In particular we provide a MATLAB interface to the RPC mechanism.

4. Research potential

Our current work being carried out with elements of KLAIR is in the modelling of a non-imitative account of the development of infant speech production that includes natural physiological constraints such as those imposed by speech breathing [15]. This requires interactivity since the model uses caregiver reformulations to first reinforce the discovery of simple speech sounds and then to learn a mapping between caregiver and infant speech.

However we believe KLAIR will be of interest to researchers in a much wider range of spoken language acquisition topics, for example:

Pre-linguistic development:

- Listening and responding to speech directed to the agent with head turns and facial expressions
- Learning to take turns in speaking

Perceptual development:

- Recognition of caregiver's voice
- Discrimination between utterance functions on the basis of prosody and voice quality
- Development of phonological categories in perception

Production development:

- Vocal control, breath control and prosody
- Babbling and imitation
- Development of an inventory of articulatory gestures
- Refinements of motor plans to match perceptual categories

Linguistic communication:

- Learning the names of objects & actions
- Use of speech to satisfy desires or needs of the agent

5. Conclusions

In this paper we have presented some arguments for the use of real-time interactions with a situated autonomous agent in spoken language acquisition research. We have presented the KLAIR toolkit which we hope will make research in this area much more accessible to new research workers. We have given examples of some applications where the toolkit may be used for experiments. The toolkit will be released for public download prior to September 2009.

In future versions of the toolkit we hope to add more functionality within the server to emulate lower-levels of processing. For example, we may add feature level representations on top of the raw acoustic and visual data, such as sparse coding. We would hope to add more realistic dynamical constraints to the articulatory synthesizer and to the talking head. We also hope to develop a library of common pattern recognition and machine learning algorithms to help

users build client applications. This and example clients will be shared through a dedicated web site for the KLAIR project.

6. Acknowledgements

With thanks to the many open-source developers that have made available example code for building Windows multimedia applications. Thanks to Shinji Maeda for making his articulatory synthesizer code available. Thanks to Piers Messum for discussions.

7. References

- [1] Roy, D., Pentland, A., "Learning words from sights and sounds: a computational model", *Cognitive Science* 26 (2002) 113-146.
- [2] Westermann, G., Miranda, E., "A new model of sensorimotor coupling in the development of speech", *Brain and Language*, 89 (2004) 393-400.
- [3] Guenther, F.H., "A neural network model of speech acquisition and motor equivalent speech production", *Biol Cybern*, 71 (1994) 43-53.
- [4] Messum, P., "The Role of Imitation in Learning to Pronounce", PhD Thesis, University College London, 2007.
- [5] Howard, I.S., Messum, P.R., "Modeling infant speech acquisition using action reinforcement and association", *Speech and Computer (SPECOM'2007)*, Moscow Linguistics University, 2007.
- [6] Vygotsky, L. S., Thought and Language, M.I.T. Press, 1985.
- [7] Metta, G., Sandini, G., Vernon, D., Natale, L., Nori, F., "The iCub humanoid robot: an open platform for research in embodied cognition", *PerMIS: Performance Metrics for Intelligent Systems Workshop*, Washington, 2008.
- [8] Stormark, K., Braarud, H., "Infants' sensitivity to social contingency: a 'double video' study of face-to-face communication between 2- and 4-month-olds and their mothers", *Infant Behavior and Development*, 27 (2004) 195-203.
- [9] Park, A., Glass, J., "Unsupervised word acquisition from speech using pattern discovery", *Proc. ICASSP-2006*.
- [10] Saltzman, E., Kelso, J. "Skilled actions: A task dynamic approach", *Psychological Review*, 94 (1987) 84-106.
- [11] Perrier, P., "Control and representations in speech production", *Papers in Linguistics, ZAS* 40 (2005) 190-132.
- [12] Maeda, S., "Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model", in Speech production and speech modelling, ed. W.J. Hardcastle and A. Marchal, Kluwer Academic Publishers, 1990, 131-149.
- [13] Fagel, S., Clemens, C., "An Articulation Model for Audiovisual Speech Synthesis: Determination, Adjustment, Evaluation", *Speech Communication* 44 (2004) 141-154.
- [14] Ekman, P., Friesen, W., Manual for the Facial Action Coding System, Consulting Psychologists Press, 1977.
- [15] Howard, I., Messum, P., "Modeling motor pattern generation in the development of infant speech production", *International Seminar on Speech Production*, Strasbourg, 2008.