Learning to imitate adult speech with the KLAIR virtual infant

Mark Huckvale, Amrita Sharma

Dept. of Speech, Hearing and Phonetic Sciences, University College London, London, UK.

m.huckvale@ucl.ac.uk, amrita.sharma.09@ucl.ac.uk

Abstract

Pre-linguistic infants need to learn how to produce spoken word forms that have the appropriate intentional effect on adult carers. One proposed imitation strategy is based on the idea that infants are innately able to match the sounds of their own babble to sounds of adults, while another proposed strategy requires only reinforcement signals from adults to improve random imitations. Here we demonstrate that knowledge gained from interactions between infants and adults can provide useful normalizing data that improves the recognisability of infant imitations. We use the KLAIR virtual infant toolkit to collect spoken interactions with adults, exploit the collected data to learn adult-to-infant mappings, and construct imitations of adult utterances using KLAIR's articulatory synthesizer. We show that speakers reinterpret and reformulate KLAIR's productions in terms of standard phonological forms, and that these reformulations can be used to train a system that generates infant imitations that are more recognisable to adults than a system based on babbling alone.

Index Terms: speech acquisition, infant speech

1. Introduction

1.1. Early infant word learning

We are interested in computational modelling of the processes of early word acquisition by infants. The challenge is to create an artificial system that learns to produce recognisable utterances from an infant-sized vocal tract without prior auditory, phonetic or phonological knowledge about speech communication. Infants' first words often relate to names of objects in their environment, which they must learn only through interactions with caregivers. To achieve this, the infant must learn how to control its own vocal apparatus, must monitor the utterances of its adult carers, and must learn how to articulate its own versions of these utterances adequate to achieve the appropriate effect in the adult listener.

Two general approaches to how the infant addresses these problems have been proposed. The first is based on the idea that the infant can (innately) judge how well its own productions match the adult forms. Learning to articulate the name of an object is then just a process of exploring different articulations and determining which motor sequence generates a sound that best matches the target form produced by the adult. This approach is best described in the work of Frank Guenther [1]. However, it has been suggested that such an approach fails to acknowledge the large differences between infant vocal tracts and adult vocal tracts, and assumes, without evidence, that the two are close enough for a mapping between adult sounds and infant articulations to be learned by some general associative learning process. In contrast, the approach by Howard and Messum [2] suggests that infant vocalisations are refined solely by reinforcement signals from the caregiver, and no matching or imitation is required. The infant explores a range of vocalisations using a non-linguistic auditory analysis based on salience, then determines through experience which seem to achieve the best effects with adults. This view does not require any mapping between adult sounds and infant sounds, so avoids the normalisation problem.

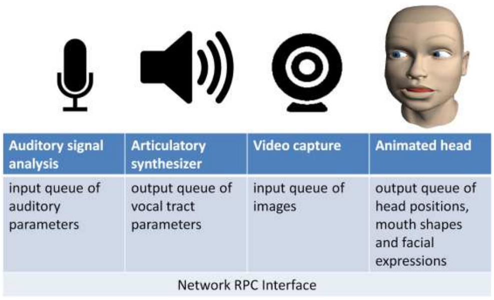
A third approach is also possible. Perhaps the infant, during vocal play with its carers, *notices* when adult forms are imitated versions of its own articulations. If these imitations are noted by the infant and compared to its own productions, then they might be an additional source of information for learning a mapping between adult forms and infant forms. A number of studies have shown that such adult imitations of infants do occur. For example Pawlby [3] found that 90% of the imitative exchanges between mothers and young infants were the adults copying the infant forms. The occurrence of adult imitations have been confirmed in a number of other studies [4,5,6]. Papoušek [5] showed that mothers imitated around 50% of young infant productions.

How might we evaluate experimentally the strengths and weaknesses of these three approaches? We believe that too much previous work in this area uses artificial data sets which have not been derived from realistic environments. We suggest that the only way to evaluate learning hypotheses is to actually simulate the experience of an infant in its interactions with adult carers. We propose that we should "embody" our learning system, let it articulate real sounds and let it listen to itself and to adults, then determine from that data what capacity it has for learning from its experiences.

For this kind of embodied evaluation to be practical we need access to a programmable infant - either an infant robot or a virtual simulation. In this work we use the KLAIR virtual infant.

1.2. KLAIR toolkit

The KLAIR toolkit was launched in 2009 [7] 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.



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 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 with synchronized mouth animation, look around its environment and change its facial expressions. Machine-learning and experiment-running clients control the server using remote procedure calls (RPC) over network links using a simple application programming interface (API). KLAIR is supplied free of charge to interested researchers from http://www.phon.ucl.ac.uk/project/klair/.

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; KLAIR supports clients written in C, MATLAB and .NET [8]. The software is also open source and freely available.

1.3. Aims of experiment

In this paper we describe an experiment that uses the KLAIR virtual infant to collect audio recordings of babbling and also recordings of adult reformulations of simple infant vocalisations. We then use the data collected to train a number of mappings between the infant articulatory space, the infant acoustic space and the adult acoustic space. We evaluate the utility of those mappings in a simple listening experiment in which virtual infant imitations of some adult utterances (generated by KLAIRs articulatory synthesizer) are rated for recognisability. In particular we compare learning accounts in which reformulations are noticed and exploited with an account in which babbling alone is used.

Section 2 of the paper describes the experimental methods used for collecting adult reformulations and presents some analysis of their acoustic properties. Section 3 of the paper describes the learning of the mappings, the generation of the imitated utterances and the listening experiment.

2. Quality of adult imitations

2.1. Objectives

The main goal of this experiment was to collect adult imitations of virtual infant nonsense productions to use as training materials for building acoustic and articulatory mappings. We also look at how accurately the imitated vowel qualities match the vowel qualities produced by the infant.

2.2. Data collection

To establish the range of vowel-like sounds that could be produced by KLAIR, the articulatory synthesizer parameters Jaw Position, Tongue Position, Lip Aperture and Lip Protrusion were systematically varied over a wide range. The subset of these articulations which gave rise to unconstricted vocal tracts were used to generate vowels, and the synthetic signals as self-monitored by KLAIR were recorded. The space of available vowels was then quantified by formant frequency measurements of the vowel productions, and that space was sampled to derive 25 vowel qualities, relatively uniformly spaced in units of Bark. See Fig 2.

The articulatory positions for these 25 vowel qualities (V_1-V_{25}) were then combined with articulatory gestures for /b/, /d/ and /m/ to create some simple one-syllable and two-syllable pseudo-words. For example, /bV_1/, /'mV_2bV_{14}/, /dV_{23}'mV_{12}/. The list was divided into 10 random sets of 75 words, with 25 one-syllable words, 25 two-syllable words with stress on the first syllable, and 25 two-syllable words with stress on the second syllable in each set.

Ten adult female subjects were asked to take part in the experiment. The task took place in a sound-conditioned booth. Each subject sat in front of a screen displaying KLAIR's animated head, and was asked simply to repeat back to KLAIR whatever speech productions KLAIR made. Sound was played through a loudspeaker behind the screen, and sound was recorded from a webcam microphone (Logitech 9000 Pro) sitting on top of the screen. Each subject heard 75 pseudo-words from KLAIR and in only 7 cases overall did a subject fail to produce some kind of imitation.

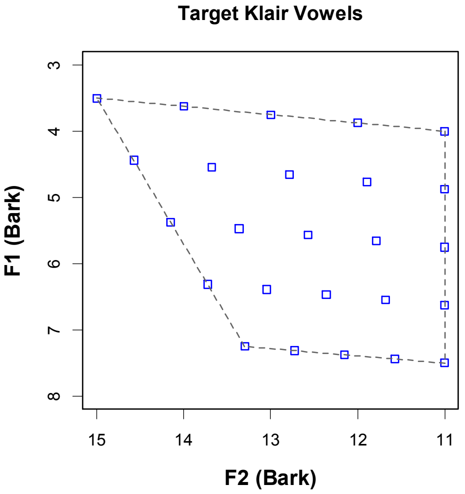


Fig 2. Locations for KLAIR's vowels in terms of the first two formant frequencies

2.3. Data analysis

The pseudo-word imitations were then analysed in terms of the formant frequencies of the vowels used in each syllable as compared to the formant frequencies used by KLAIR. This gave rise to 1243 vowel comparisons.

To study the overall reformulation behaviour across subjects, the vowel formant measurements for each subject were individually normalised by first converting the hertz values to Bark, then converting to z-scores. Fig 3 shows the position of KLAIR's original vowels together with the mean position across all subjects of the imitated vowels. The arrows link KLAIR's vowel locations to the mean for the subjects' imitations. Similar results were obtained for each individual speaker.

2.4. Discussion

It is clear from Fig. 3 that the vowel qualities of the adult copies do not match the quality of the infant productions particularly well. There are two main effects. Firstly the adult vowels seem more centralised, with less variability in F1 and particularly in F2 compared to KLAIR. This may simply be due to the averaging process and the variety of vowel qualities produced by the adults. The vowel space of each individual subject was similar to KLAIR's with standard deviations around 1.0-1.5 Bark. Secondly there is strong evidence that many different target vowels were collapsed into a few vowel categories. For example, almost all of KLAIR's open vowels were mapped to a single open central vowel, and KLAIR's close back vowels were all mapped to a single close central vowel. This is supporting evidence for the idea of *reformulation,* and may be due to the listeners remembering and reproducing the infant sound in terms of English phonological vowel categories. In Westermann and Miranda [9], this preference for certain phonetic forms of vowels has been proposed as a mechanism for infants to learn the phonological vowel categories in the carers' language. In this work, we do not make explicit use of this effect, although the preferences of the adult speakers may well influence the acoustic-articulatory maps learned subsequently.

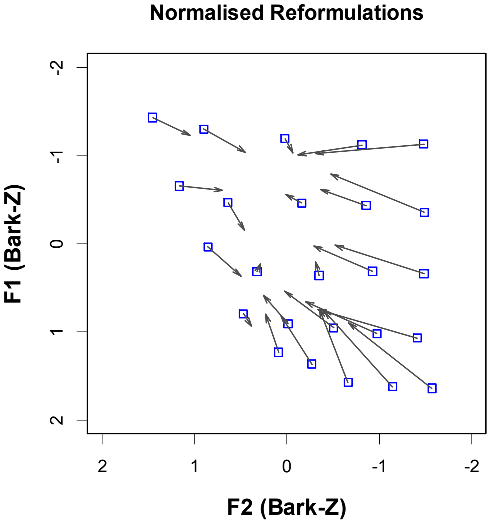


Fig 3. Shift in vowel locations measured from reformulations. Units: normalised Bark. Squares are infant vowels, arrows point to mean imitated vowels.

3. Quality of imitated utterances

3.1. Objectives

The goal of this experiment was to investigate whether knowledge gained from the adult reformulations of the virtual infant’s vocalisations is of use to improve the ability of the infant to imitate an adult utterance. We compare three essential strategies:

a) No normalisation. In this strategy, the virtual infant learns a map between its own speech sounds and its own articulations then uses that map to imitate the adult directly.

b) Auditory normalisation. In this strategy, the virtual infant uses the reformulations to learn an auditory map between adult sounds and infant sounds. This is then combined with the map learned in strategy a) to imitate the adult.

c) Articulatory normalisation. In this strategy, the virtual infant uses the reformulations to learn a map between the adult sounds and its own articulations. This map can be used directly to imitate the adult. This strategy was evaluated in a speaker-independent (SI) and speaker dependent (SD) form.

A schematic of the mapping relationships between the data sets is shown in Fig 4. The effectiveness of these strategies is evaluated using adult listeners to rate the recognisability of some imitated adult sentences.

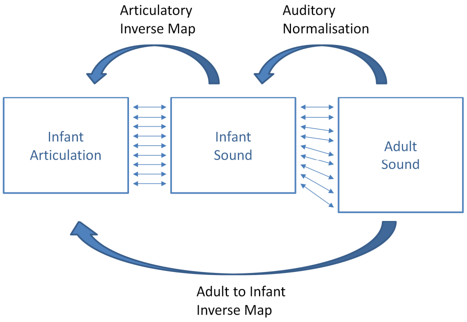


Fig 4. Schematic of data sets and learned mappings.

3.2. Data Modelling

All audio data is analysed into 12 mel-frequency cepstral coefficients plus energy at 100 frames/sec. A pitch contour is extracted, smoothed and interpolated (so that an F0 value is available at all times). The F0 value is stored in semitones. All audio data is normalised by subtracting the mean value calculated separately for each speaker.

Each adult imitation is time-aligned to the original infant synthetic version using the MFCC parameters together with a dynamic-programming search and a Euclidean distance metric. The alignment is used to generate the data for learning vector maps.

The mapping between data sets was performed on a frame-by-frame basis using a multi-layer perceptron with linear output units. For audio-to-articulatory inversion, the network had 3 frames of 14 audio parameters as input and 3 frames of 12 articulatory parameters as output. For audio-to-audio mapping, the network had 3 frames of 14 audio parameters as input, and 3 frames of 14 audio parameters as output. All networks had one hidden layer of 64 units. The use of multiple input frames allows the system to exploit time differences if needed. The use of multiple output frames provides a small degree of temporal smoothing.

Networks were trained by back-propagation; the learning rate was 0.1, and the momentum was 0.9. Networks were trained for 300 cycles through the training data. 100 cycles with an update every 1000 frames, 100 cycles with an update every 10,000 frames and 100 cycles with an update every 100,000 frames.

In the "No normalisation" condition, a network was trained between KLAIR's synthetic audio output and KLAIR's articulatory input for the pseudo-word utterances. This network was then applied to adult audio-recorded sentences to generate infant articulatory imitations.

In the "Auditory normalisation" condition, 10 networks were trained between the audio imitations of each adult speaker and KLAIR's synthetic audio output for the pseudo-word utterances. Each network was then applied to audio-recorded sentences of the selected speaker to generate equivalent KLAIR audio versions of the sentences, and these in turn were input to the previous network to create infant articulatory imitations.

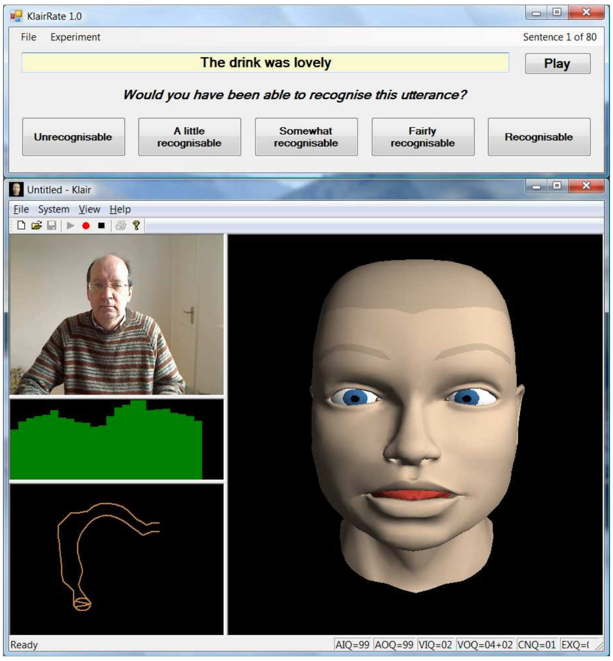
In the "Articulatory normalisation (SI)" condition, a network was trained between the natural audio of the adults' imitations and KLAIR's articulatory input for the pseudo-word utterances. A single network was trained for all 10 speakers. The network was then applied to adult audio-recorded sentences to generate infant articulatory imitations.

In the "Articulatory normalisation (SD)" condition, 10 networks were trained between the natural audio of each adult's imitations and KLAIR's articulatory input for the pseudo-word utterances. The appropriate speaker-specific network was then applied to adult audio-recorded sentences to generate the infant articulatory imitations.

3.3. Rating Experiment

In the rating experiment, two of the adult speakers from the first experiment were selected to provide 10 short sentences to act as the targets for imitation. Each of these were processed according to the four experimental conditions to generate a total of 80 test imitations.

Ten listeners (not involved in the first experiment) were asked to rate the recognisability of the imitated utterances. The utterances were produced by KLAIR from the articulatory parameters in real time, so that the listeners could also see KLAIR's articulation, see Fig 5. Listeners could also read the supposed target sentence. A five-point rating scale was used going from "Unrecognisable" to "Recognisable". Each listener rated 10 training utterances before rating the 80 test imitations in random order.



The drink was lovely

Would you have been able to recognise this utterance?

Fig 5. Screenshots of the rating experiment.

3.4. Data Analysis

Histograms of the listener ratings across the four learning conditions are shown in Fig 6. The mean rating for each condition is shown in Table 1. To examine the significance of the effect of condition, the rating histograms were divided into "low" and "high" counts using a threshold of 1.5. A chi-square test on low-high proportions aggregated across listeners shows a significant effect of condition (χ^2=27.2, df=3, p<0.001). Post hoc analyses of conditions taken in pairs show significant differences between all conditions except the two variants of Articulatory normalisation.

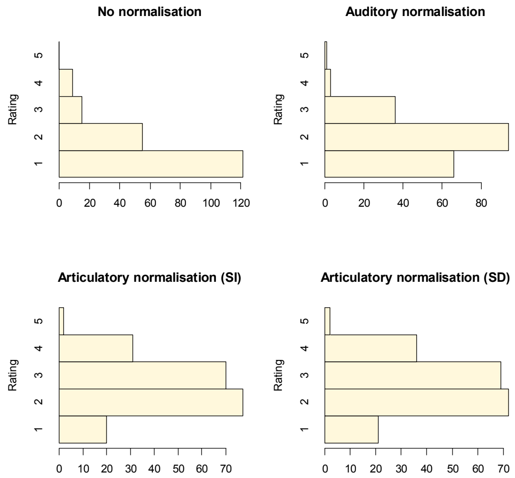


Fig 6. Histograms of ratings by condition.

Condition	Mean Rating
No normalisation	1.560
Auditory normalisation	1.895
Articulatory normalisation (SI)	2.590
Articulatory normalisation (SD)	2.630

Table 1. Mean recognisability ratings per training condition.

3.5. Discussion

All learning strategies that exploited the data from adult imitations were rated significantly higher than the strategy that did not. This is despite the fact that the imitations were not particularly accurate. The introduction of auditory normalisation did improve the recognisability of the infant imitations built using a mapping learned only from babble. This supports the idea that some normalisation process is required to address the differences between infant and adult vocal tracts. The articulatory normalisation strategies performed best, despite not making any use of the infant sound except as an index into the adult reformulations. A speaker-independent strategy seemed to work as well as a speaker-dependent strategy. This may have been because more training data was available in the speaker-independent case, and all our speakers were adult female.

4. Conclusions

In this paper we have shown how different hypotheses about the process by which infants acquire the ability to articulate first words may be evaluated through the use of a virtual infant interacting with adult carers. Our experiment generated real sounds through an infant-scaled articulatory synthesiser, and collected real audio responses from adult carers. Using only small amounts of data, we were able to build systems for imitating adult utterances using three different strategies, and showed that their effectiveness can be compared in a listening experiment. Although many aspects of the experiment remain highly artificial, we hope that we have shown how scientific investigations of infant speech acquisition may be explored using interactions with a virtual infant.

5. References

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

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

[3] Pawlby, S., "Imitative interaction". In H.R. Schaffer (Ed.), Studies in mother-infant interaction. London: Academic Press, 1977, 203-223.

[4] Veneziano, E., Sinclair, H., Berthoud, I., "From one word to two words: repetition patterns on the way to structured speech". Journal of Child Language, 17 (1990) 633-650.

[5] Papoušek, M., & Papoušek, H., "Forms and functions of vocal matching in interactions between mothers and their precanonical infants". First Language, 9 (1989) 137–158.

[6] Kokkinaki, T. and Vasdekis, V.G.S., "A cross-cultural study on early vocal imitative phenomena in different relationships". Journal of Reproductive and Infant Psychology, 21 (2003) 85-101.

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

[8] Huckvale, M., "Recording caregiver interactions for machine acquisition of spoken language with the KLAIR virtual infant", InterSpeech 2011, Florence, Italy.

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