Speech Communication 86 (2017) 1–23

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

# An online model for vowel imitation learning

Heikki Rasilo[a,b,*], Okko Räsänen[a]

[a] Department of Signal Processing and Acoustics, Aalto University P.O. Box 13000 00076 Aalto Finland
[b] The Artificial Intelligence Laboratory, Vrije Universiteit Brussel Pleinlaan 2 1050 Brussels Belgium

ARTICLE INFO

*Article history:*
Received 23 November 2015
Revised 27 October 2016
Accepted 31 October 2016
Available online 3 November 2016

*Keywords:*
Imitation
Correspondence problem
Normalization problem
Vowel learning
Speech acquisition
Associative algorithm
Weakly supervised learning

ABSTRACT

When infants learn to pronounce speech sounds of their native language, they face the so-called correspondence problem – how to know which articulatory gestures lead to acoustic sounds that are recognized as native speech sounds by other speakers? Previous research suggests that infants might not learn to imitate their parents via autonomous babbling because direct evaluation of the acoustic similarity between the speech sounds of the two is not possible due to different spectral characteristics of the voices caused by differing vocal tract morphologies. We present a novel robust model of infant vowel imitation learning, following a hypothesis that an infant learns to match their productions to their caregiver's speech sounds when the caregiver imitates the infant's babbles. Adapting a cross-situational associative learning technique, evidently present in infant word learning, our simulated language learner can cope with ambiguity in caregiver's responses to babbling as well as with the imprecision of the articulatory gestures of the infant itself. Our fully online learning model also combines vocal exploration and imitative interaction into a single process. Learning performance is evaluated in experiments using Finnish adults as caregivers for a virtual infant, responding to the infant's babbles with lexical words and, after a learning stage, evaluating the quality of the vowels produced by the learner. After 1000 babble-response pairs, our virtual infant is seen to reach a satisfying vowel imitation accuracy of 70–80%.



## 1. Introduction

Infants begin to vocalize early in their lives, creating sounds such as crying and gurgling that do not sound like recognizable native language sounds. Later, at around 6 months of age, infants start to babble sounds that are reminiscent of native language sounds, and during the second year of their lives they normally produce their first recognizable words. Initially, infants are capable of learning phonemic systems of every language, but they end up converging to the language used in their environment.

Infants must overcome a large array of problems in order to converge to a functioning vocal system. Language learning occurs in an extremely complicated environment, consisting of several speakers with different voices and possibly different languages or accents. The infant must learn to physically control its vocal tract, and somehow find a way to produce vocalizations that the parents consider as matching to their vocalizations. This problem is complicated by the fact that the sizes of the vocal tracts of the child and the parents differ radically in size and morphology, producing completely different acoustic sounds between which a direct comparison cannot be performed. This problem is often called the *normalization problem*. Although there is some empirical evidence that infants tune their auditory perceptual systems to discriminate contrasts in their native language (Polka and Werker, 1994), and that linguistic (Goldstein and Schwade, 2008) and social non-imitative (Goldstein et al., 2003) feedback by parents guide infants towards more mature speech sound production, the mechanisms underlying learning the solution to the normalization problem are currently not well understood.

In this paper, we investigate the normalization problem in early language acquisition using a computational model of a child. This virtual infant does not initially know the acoustic outcomes of its articulatory gestures nor does it have perceptual categories corresponding to native language speech sounds. Instead, it learns a set of vocalic productions and their correspondence to adult speech sounds in an interactive learning scenario with an adult caregiver. We use a group of human test subjects to act as caregivers for the learner, responding to the infant's vocal exploration with spoken words that partially match phonetic features of the infant's verbal output, thereby extending the earlier work (Section 1.2) by allow-

*Abbreviations:* LeVI, learning virtual infant; CG, caregiver; CM, concept matrix; DCM, dynamic concept matrix; MFCC, Mel-frequency cepstral coefficients; VQ, vector quantization; C, consonant; V, vowel; F0, fundamental frequency; F1, first formant frequency; F2, second formant frequency.

* Corresponding author at: Department of Signal Processing and Acoustics, Aalto University P.O. Box 13000 00076 Aalto Finland.
*E-mail addresses:* heikki.rasilo@aalto.fi (H. Rasilo), okko.rasanen@aalto.fi (O. Räsänen).

http://dx.doi.org/10.1016/j.specom.2016.10.010
0167-6393/

ing uncertainty in the caregiver responses as well as articulatory inaccuracy in the infant's babbling. We describe a novel learning mechanism that is able to cope with the above-mentioned learning challenges, and show how the virtual infant can learn to use its own vocal tract to reproduce vowel sounds produced by a caregiver so that the caregiver recognizes the reproduction as the original vowel – i.e., how the infant learns to *imitate* vowels based on interaction with the caregiver.

### 1.1. Learning to imitate

Although imitation of human actions may appear trivial, every imitator is faced with a so-called "correspondence problem". When we perceive an action performed by another human, for example by seeing a hand movement, we only perceive the result of the underlying motor commands, but not the motor commands themselves. The imitator has to somehow deduce how it's own motor commands result in a similar action.

According to so-called specialist theories, there are innate mechanisms allowing for imitation. The specialist theories are mainly supported by experimental findings that neonates are able to imitate facial gestures such as tongue protrusion, and, due to their young age, the imitation ability is not likely to be a product of learning (e.g. Meltzoff and Moore, 1977). However, the validity of this result has been questioned (see, e.g., Anisfeld, 1996; Jones, 2006): for example, tongue protrusion has been found to be a general response to different kinds of stimuli, such as listening to music (Jones, 2006). Recently, Oostenbroek et al. (2016) have shown in a comprehensive study of 106 infants at ages of 1, 3, 6 and 9 week that the infants did not imitate any of the 11 gestures studied. A study by Kuhl and Meltzoff (1996) shows that when infants between 12 and 20 weeks saw video recordings of a female talker producing /a/, /i/ or /u/ vowel sounds, their vocalic responses were reminiscent more of the particular vowel they heard, rather than the other two vowels, arguing for early vocal imitation. Infant vocalizations were classified by an experienced scorer into eight vowel categories /æ, a, ʌ, i, ɪ, ɛ, ʊ, u/, that were further grouped in three /a/, /i/ or /u/ like categories (/æ, a, ʌ/), (/i, ɪ, ɛ/) and (/ʊ, u/) correspondingly. However, despite their young age, these infants had already experienced significant amount of experience from face-to-face interaction with their caregivers, making the study incapable of disentangling any factors of learning during the first three to five months of age. However, the study still shows that adults can interpret early infant vocalizations as different vowels, providing the basis for contingent parental responses (see below). In general, it is currently unclear what proportion of the early imitation capabilities are truly driven by innate mechanisms.

In contrast to the specialist theories, the so-called generalist theories claim that imitation abilities are learned based on general learning and motor control mechanisms (see, e.g., Brass and Heyes, 2005). For instance, the generalist Associative Sequence Learning model (ASL, Heyes and Ray, 2000) states that imitation is enabled when a link between the perception (e.g. visual information) of someone else's action and the imitator's own motor commands leading to the same action are formed due to associative learning. These links can be formed during observation of one's own actions in relation to actions performed by the model to be imitated. When actions are *transparent,* the sensory feedback of the executed and observed actions are similar (e.g. visually observed hand movements). However, some actions to be imitated can be *opaque*, meaning that the sensory feedback of the observed action is not similar to the sensory feedback when the action is performed (e.g. when learning to imitate facial expressions, the learner cannot see his own face and thus cannot visually evaluate if the motor commands performed lead to the action to be imitated). In these cases, associative learning can occur for example with the help of mirrors or imitative social partners. Several behavioral and neuroimaging studies support the view that prior learning strengthens the activation of corresponding motor representations when actions are observed, thus supporting the generalist view (see Brass and Heyes, 2005, for a review).

There is also evidence that caregiver imitations may play a role in infant vocal imitation learning. This was probably first observed by Pawlby (1977) in a study where parents' spontaneous interactions with their infants were recorded from 17 until 43 weeks of age. Out of all imitative sequences, 79% were initiated by the infants and imitated by the mothers. Out of the speech sounds in all three reported age groups, mothers' imitation frequency was about 10 times that of the infants', and infants' speech sound imitation frequency increased slowly when approaching the age of 43 weeks. Pawlby concludes that "*Paradoxically our study suggests that the whole process by which the infant comes to imitate his mother in a clearly intentional way is rooted in the initial readiness of the mother to imitate her infant.*" (p. 220).

Other research suggests that infants learn to imitate their parents gradually, i.e., the ability to imitate vocalic sounds does not seem to be innate or is at least significantly reinforced over the development. Kokkinaki and Kugiumutzakis (2000) have studied imitative patterns between parents and their two-to-six month-old infants and showed that vocal imitative sequences occurred on average 3.7 times during a 10 min period, out of which on average 66.6% were performed by the parent; parents imitated their children significantly more than vice versa. Kokkinaki and Vitalaki (2013) have studied imitative interaction (vocal imitation, non-speech sound imitation, facial expression imitation and hand movement imitation) between two-to-ten-month-old infants and their mothers and grandmothers. Vocal imitation was the most frequent imitation type for all studied subject groups, corresponding to 80% of all imitations for mothers M1 in Group 1 (Group1 infants had no frequent contact with grandmothers), and 75% for mothers M2 and 73% for grandmothers G2 in Group 2 (infants who had had frequent contact with grandmothers). Non-speech sound imitation frequencies were 8%, 4% and 5% for M1, M2 and G2, respectively. Of all imitative actions 79% (M1), 66% (M2) and 70% (G2) were produced by the mothers or grandmothers. The respective imitation frequencies were, on average, 3.7, 3.0 and 3.1 imitations during a 10 min interaction session.

Masur and Rodemaker (1999) studied vocal, verbal and action imitation of infants when they were 10, 13, 17 and 21 months old. At 10 months, parents imitated their children's vocalizations rarely, but still about six times more than vice versa on average, and continued to exceed the infants' imitation rate in all age groups. Between 13 and 17 months, there was a radical increase for both infants and parents in their frequency of verbal imitation (parents still exceeding the infants' imitation rate), indicating that around this time period infants learn to match at least some of their vocalizations with parental speech, and learning to produce their first words seems to encourage imitative responses from their parents.

Recently, Yurovsky et al. (2016) analyzed the amount of lexical alignment in caregiver responses to infant vocalizations. More specifically, they analyzed 417 native English infant-caregiver dyads from CHILDES (MacWhinney, 2000) with children between 12–60 months of age, measuring how much children and adults re-use the expressions they just heard from each other. Their main finding was that parents align to their children significantly more than children align to their parents, and that the amount of parental alignment decreases monotonically with child's age, reaching adult-to-adult levels around the age of 54 months. Although their data did not cover infants younger than 12 months, the data supports the previous studies by showing that parents