An online model for vowel imitation learning[1]

Heikki Rasilo[1,2], Okko Räsänen[1]

[1]Department of Signal Processing and Acoustics, Aalto University
P.O. Box 13000
00076 Aalto
Finland

[2]The Artificial Intelligence Laboratory, Vrije Universiteit Brussel
Pleinlaan 2
1050 Brussels
Belgium

Heikki Rasilo, corresponding author: heikki.rasilo@aalto.fi, tel. +32483047487

Co-author:
Okko Räsänen, okko.rasanen@aalto.fi, tel. +358504419511

[1] Used 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.

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%.

Keywords: Imitation, correspondence problem, normalization problem, vowel learning, speech acquisition, associative algorithm, weakly supervised learning

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 & Werker, 1994), and that linguistic (Goldstein & Schwade, 2008) and social non-imitative (Goldstein, King & West, 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 allowing 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 & 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 & Heyes, 2005). For instance, the generalist Associative Sequence Learning model (ASL, Heyes & 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 & 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 ten minute 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-minute 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, Doyle and Frank (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 are consistently providing contingent responses to their children's communicative attempts.

In total, the findings from the above studies suggest that the total amount of imitative feedback can be notable over the time-course of verbal development, and that caregiver imitation may play a central role in early language acquisition by providing equivalence cues between infant's own vocalizations and the corresponding vocal gestures in adult language.

In the present work, we adopt the generalist view of imitation learning and investigate its potential in infant vowel acquisition by utilizing a computational simulation. We argue that, without some kind of innate perceptual normalization, the actions behind vowels spoken by a caregiver are clearly opaque. In addition, due to different articulatory morphologies, the vowel to be imitated is perceptually different from the vowel that the infant is capable of producing. No matter how much the infant babbles and experiments with its articulatory system, it is not able to reproduce the acoustic characteristics of its caregiver's speech. Therefore, the goal is to study whether vocal imitation learning is possible under these circumstances when the infant uses generic associative learning mechanisms to link its own articulatory actions to the auditory perceptions of the caregiver's imitative responses to the infant's vocalization. Also, our goal is not to settle to debate between specialist and generalist theories, but to investigate whether the more parsimonious generalist theory can account for early vocal learning.

1.2 Computational models of vocal imitation learning

In this section we describe existing computational studies of speech imitation learning. Since real speech learning situations are very complicated and time consuming, most of the studies use an array of simplifications in order to keep experiments and simulations within realistic limits. However, simplified learning simulations have a risk of reducing the learning models' cognitive plausibility when compared to cognitive processes in real language learning situations.

Among the existing studies, Markey's HABLAR model (1994), and neurocomputational models of Kröger, Kannampuzha and Neuschaefer-Rube (2009) and Guenther (1995) offer solutions to the problem of learning the coupling between speech perception and production, but do not provide solutions to the normalization problem between different sounding speakers. Also, an imitation learning model by Murakami, Kröger, Birkholz and Triesch (2015) uses the same vocal tract model for adult and infant sounds. In a study of Westermann and Miranda (2004), a computational model's vocal production is seen to adapt to an external speaker's

vowel sounds during a babbling phase, but the spectral ranges of the external speakers and the model are similar, thus enabling direct spectral matching.

A number of computational studies use imitation by a caregiver as a method to teach the infant the mapping between the two differing voices. However in many of these studies (e.g. Miura, Yoshikawa and Asada, 2007; Ishihara, Yoshikawa, Miura and Asada, 2008; Vaz, 2009) the learner is provided with a pre-defined set of vocal primitives as a basis for babbling, and the caregiver responds the infant's vocal productions with the same phonetic content as was produced by the infant. In one of the first studies in the field, Yoshikawa, Koga, Asada and Hosoda (2003) used a physical vocal tract model randomly producing vowel sounds, and a human experimenter repeating the robots' productions.

Ananthakrishnan and Salvi (2011) propose a method for learning the mapping between an infant's and an adult's acoustic domains by learning a topological mapping between the two. The best parameters for the mapping are found using non-imitative feedback by the caregiver. Both the caregiver's and the infant's speech sounds are synthesized, and the acoustic signals created for both are produced by a similar babbling procedure. In normal learning scenarios the phonetic characteristics of speech produced by an adult and a speech learning infant are acoustically very different, and using synthesized caregiver speech sounds might result in less variation in speech when compared to real speakers.

Plummer (2012) approaches vowel normalization as a manifold alignment problem, where the infant maps the caregiver's and its own vowel sounds in a speaker independent mediating space. The manifolds are aligned using synthesized caregiver-infant vowel pairs. Imitation is discussed as a pairing method but in the experiment imitation data is selected manually from the two acoustic spaces.

Hörnstein and Santos-Victor (2007) have taught a humanoid robot to recognize and reproduce Portuguese and Swedish vowels. First, the robot learned a neural network to map vocalic sounds to articulatory motor representations in an initial autonomous babbling phase. Second, the robot tried to imitate vowels spoken by a human caregiver with the learned mappings and the caregiver reinforced successful imitations with positive feedback, ultimately teaching the robot nine Portuguese vowel categories. Next, the robot was trained with vowel samples extracted from Portuguese words spoken by several speakers. During training, the robot produced one of its vowels at a time, and a corresponding vowel sample by a caregiver was played back to the infant to model imitation (see also Hörnstein, 2013). When a test set of vowel samples from different speakers was mapped back to the motor representations and classified in the nine possible vowel categories, close to 60% vowel recognition accuracy was achieved without using visual information. This is one of the few studies that uses real recorded human speech when evaluating vowel learning performance, but the learning situation is simplified by manually reinforcing the robot to learn exactly nine vowel categories.

In the study of Hörnstein, Soares, Santos-Victor and Bernardino (2007) the mapping from caregiver's speech to the robot's motor commands is learned when a human caregiver imitates the babbled articulatory trajectories of the robot. However, the caregiver's imitations have to be manually time-aligned to the robot's babbles in order to avoid incorrect pairings. The study also offers a solution to automatically find nine vowel categories from the robot's speech. In their experiments, native speakers first listened to a set of 900 vowel sounds generated by the robot and either approved or rejected the sounds as native Portuguese vowels. When agglomerative clustering was then applied to the approved sounds (281 out of 900), nine vowel categories were

found. However, the original dataset was created from the nine prototype vowels by adding 10% of white noise to the articulatory parameters, potentially simplifying the clustering task.

Miura, Yoshikawa and Asada (2008) approach variation in human interaction by using a weakly supervised learning technique to teach a virtual infant vowels spoken by a human caregiver. In the learning scenarios, both speak several speech sounds and only part of the caregiver's sounds are imitations. The learning mechanism is auto-regulated so that the infant actively detects which sounds in the caregiver's utterances are actual imitations. Depending on the training and the testing phase difficulty, the infant learns to classify five Japanese vowels in a test data set with approximately 60–100% accuracy. The simulation does not contain an articulatory exploration phase, but the infant is given 15 vowel primitives, three per vowel category, which all have different probabilities of being imitated correctly by the caregiver.

The importance of pairing caregiver's imitative signals to the learner's initial productions is also discussed in the study of Hörnstein, Gustavsson, Santos-Victor and Lacerda (2008), where an automatic imitation classifier, classifying utterances into imitations or non-imitation, is proposed. In vowel learning experiments, the use of the classifier is simulated by adding incorrect caregiver's imitations to the set of correct ones, reducing the recognition rate of the learner.

Huckvale and Sharma (2013) have given a virtual infant a predefined set of phonemes to create babbles that are imitated by human caregivers. When new listeners evaluate the infant's imitation of new sentences, the results indicate the importance of imitation by the caregivers to guide the infant's articulation. In this work, caregivers' imitations are exact reformulations of the infant's initial utterances and phonetically aligned to the production data.

Howard and Messum (2014, see also 2011) have studied vocal learning by having their virtual infant first discover new motor patterns in an unsupervised exploration phase. After the exploration, English, French and German speaking participants were asked to respond to the infant's productions naturally. The infant associated its motor patterns to the resulting auditory responses, thus enabling imitation of novel utterances. On average, 78% of the infant's utterances were responded to by the caregivers, and out of the responses, 94% were reformulations of the original sound. In the infant's imitation phase, the infant tried to imitate 219 English words, 219 French words and 237 German words. Participants evaluated which of the infant's productions were successful imitations, resulting in 55 accepted imitations on average. For analysis purposes, the accepted infant imitations and the corresponding adult utterances were then classified into so-called archiphoneme categories: 5 for vowels and 18 for consonants. The similarity between the infants' and the caregiver's archiphoneme vowel categories from the imitated utterances was approximately 58% on average, and for consonant categories approximately 39% (estimated from the figures in Appendix S4 of Howard & Messum, 2014). Their work suggests that self-driven articulatory exploration and imitation by human caregivers can be a successful method to learn speech sound production and imitation.

In our previous study (Rasilo, Räsänen & Laine, 2013), we trained a virtual infant with a simulated caregiver in two phases. In the first phase, the infant received only positive/negative type feedback on its babbles and converged to the same phonetic system as its caregiver without being given the exact number of vowel or consonant categories in advance. In the second phase, the infant learned to imitate both vowels and consonants of the caregiver by associating its babbles with imitative

words spoken by the caregiver. However, we assumed that the caregiver is able to invert the infant's speech into exact articulations and thus guide the infants babbling towards the correct articulations of the Finnish phonemes. In natural situations, the underlying articulatory configuration stays mostly invisible for the parent, and caregivers have to make judgements on the infant's babbles based on their acoustic characteristics. Also, synthesized caregiver speech reduced the amount of natural variation in speech sounds.

In summary, the previous computational studies suggest that imitation of infants' babbles by caregivers may work as a mechanism to overcome the correspondence problem and help infants to later imitate their caregivers. However, most studies simplify real learning situations considerably. The infant may be provided with an initial set of phones or vocal primitives, consequently avoiding the vocal exploration phase, whereas human infants have to explore their range of speech sounds and converge to the sound system of their native language. In some studies the exact number of vowels spoken by the caregiver is known, which seems like an unrealistic assumption since unsupervised discovery of the number and characteristics of vowel categories from continuous speech has proven to be a very challenging task, and may be influenced by constraints from other modalities such as articulation or the lexicon (see Räsänen, 2012, for a review). In addition, the use of synthesized caregiver speech reduces the acoustic variability from real speech, making the mapping problems easier to cope with. In some studies (Hörnstein, Soares, Santos-Victor and Bernardino, 2007; Howard and Messum, 2014; Huckvale and Sharma 2013), where continuous babbles are created by using articulatory trajectories, the babbled utterances are time-aligned to the caregiver's imitations using dynamic programming, so that the learner ends up with one-to-one correspondences with babbled speech sounds and speech sounds in the responses (not to be confused with *linguistic alignment* where the expressions used by the conversational partner are re-used in following sentences). Time-alignment of imitation-response pairs reduces the ambiguity that might otherwise be present in the caregiver's responses when compared to the babble (regarding ordering or durations of phonemes for instance). Also, none of the studies discussed use integrated vocal exploration and interaction, but these two processes occur in two distinct phases where the infant first explores and stores speech sounds (or alternatively is given a set of vocal primitives), and later interacts with the caregivers using these representations. However, at least in some studies this was done to shorten the interaction phase, not necessarily affecting the model performance (as in Howard & Messum, 2014).

In our current work we relax the requirement for the infant to know any vowel categories – of the caregiver or the infant itself – prior to learning and the requirement for an accurate time-alignment of caregiver's imitative signals to the initial babbles. Although parental responses to infants' babbles can be exact reproductions, they are also often expansions such as "ball" after a child babbled "ba" (e.g. Tamis-LeMonda, Bornstein & Baumwell, 2001) or "Ma-ma. Yes, and da-da is working." after a child babbled "Ba-ba" (see Gros-Louis, West, Goldstein & King, 2006). In a study by Vigil, Hodges and Glee (2005), an average of 1.21 imitations and 8.37 expansions by parents were recorded in a 10 minute session with two-year old children with normal language development. An imitation was defined as "*a repetition of partial or exact imitation of preceding utterance*" and an expansion was defined as "*repetition of the child's preceding word approximation or verbalization and completes the utterance by adding one or more morphemes or words*" (p. 114). By extending the principles of so-called cross-situational learning, a well-documented learning principle in infant

word learning (Smith & Yu, 2008), to phone category acquisition, we can allow the existence of additional phones in the caregiver's imitative utterances while still coping with the noise in the imitations.

We also propose a mechanism that helps to cope with variability in the infants' own productions. Chung et al. (2012) found that when asked to repeat real words including vowels /a/, /i/ and /u/, 2-year-old infants' vowel productions had more variability than 5-year-olds' and adults' productions, and concluded that "*The variability observed for the 2-year-olds' productions suggests that they are still in the process of learning to produce adult-like vowels, even for productions that were transcribed as correct*" (p. 452). It is also clearly established that children's motor control for speech production is less developed than adults', and their articulation shows more variability (see, e.g., Smith & Zelaznik 2004; Walsh, Smith & Weber-Fox, 2006). In this study, we allow articulatory variation in the infants' babbles during the learning phase (the infant is not able to exactly reproduce its previous productions), and show that auditory clustering of the infants' own babbles maintains the possibility of associative vocal learning.

We use human participants acting as caregivers and, according to our knowledge, describe the first fully online vowel learning experiment where the articulatory exploration for vowels is intertwined with the interaction phase. The infant's vowel imitation performance is also accurately measured based on the judgements made by the participants acting as caregivers, providing direct information about the rate of mutual agreement between the final vowel systems of the learner and the caregiver.

2 Overview of the learning method

The goal of our experiment is to teach a Learning Virtual Infant (LeVI) to imitate eight Finnish vowels occurring in words spoken by human participants who act as LeVI's caregivers. This means that LeVI has to learn to map between three initially distinct representations: auditory representations of adult speech, auditory representations of LeVI's own speech, and the articulatory gestures responsible for its own speech.

From here on, a human participant is abbreviated as CG ("caregiver"). LeVI does not initially know anything about the auditory perceptual characteristics of its own possible vocalic productions or the number or characteristics of the "correct" Finnish vowel categories. Every participant interacts with a new initialization of LeVI, i.e., the interaction sessions are not dependent on each other, in order to measure the robustness of the learning strategy for different participants and of the randomized vocal exploration process.

The basic framework of the learning process is as follows, also illustrated in the flowchart in Figure 1:

Training phase:

- LeVI babbles a random vocalic sound, or tries to reproduce a sound it has already once produced. These sounds are clustered into categories based on their auditory similarity and using a distance threshold for same/different auditory distinction (either a new category is created, or the babble is assigned to an existing category). From now on, these categories are referred to as *LeVI auditory categories* (LAC), since they are characteristic to LeVI's vocal

exploration and differ from the caregiver's native vowel categories. The category to which the babble is assigned is called the *activated LAC* in Figure 1. From here on, in the present work we use the word "babble" to mean a single vocalic sound produced by LeVI.

After assigning the babbled sound into a LAC, its articulatory characteristics are compared to the articulations responsible for the previous babbles stored in the same LAC. LAC-specific articulatory parameters can have one or more subcategories (clusters) since multiple distinct articulatory configurations can lead to similar acoustic outcomes. Based on the distance and a threshold in the articulatory space (not to be confused with the auditory threshold above; see section 2.3), every new babble is either assigned to an existing articulatory category corresponding to the activated LAC or a new articulatory category is created.

- CG listens to the babble, classifies the vowel into a Finnish vowel category, and responds by pronouncing a Finnish CVCV-word (C=consonant, V=vowel), where only one of the vowels is the CG's interpretation of LeVI's babbled vocalization. Every response thus contains additional phonemes and temporal ambiguity regarding the matching sound.
- LeVI associates CG's response to its own babble using a weakly supervised, associative, learning method. The correct associations between CG vowels and LACs arise across several interaction trials as the ambiguity present in individual trials decreases with increasing statistical evidence (= cross-situational learning, XSL).

Testing phase:

- Finally, after a number of training trials, LeVI tries to imitate vowels from a new set of CG's CVCV-words based on the learned associations. CG is asked to classify LeVI's imitative vocalizations, unaware of the original word and the vowels that LeVI imitated. LeVI's imitation accuracy is measured based on successful imitations as judged by CG - if CG annotated the imitated vowel as the same vowel as in the original imitated word, the imitation was successful.

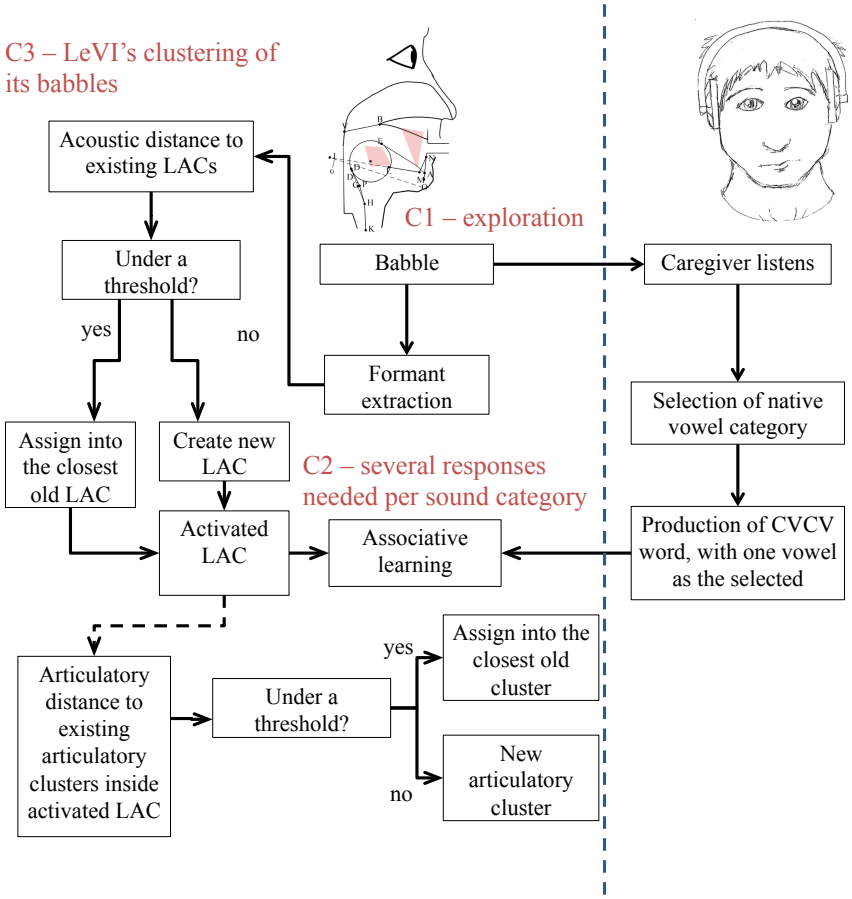

Figure 1. Flow chart illustrating LeVI's learning process

The approach described has several challenges that need to be solved (marked with red color in corresponding locations in Figure 1):

1. How should LeVI explore its high-dimensional articulatory space efficiently in order to find vowel sounds that are interpreted as native vowels by CG (C1 at Figure 1)? As will be described later, uniformly sampled random productions from the possible articulatory parameter space produces mostly centralized vowels. Consequently, vowel categories that can be produced only in relatively small articulatory regions (such as Finnish vowels /i/ or /a/), may be difficult to find based on random exploration.
2. Because CG's responses are noisy due to ambiguity in phonetic content as well as variation typical of normal speech, it is not possible to learn any of the vowel sounds babbled by LeVI based on one babble-imitation pair. LeVI needs multiple imitative responses per babbled sound in order to learn robust associations to the acoustic cues of any vowel (C2 at Figure 1). In addition to vocal exploration, LeVI thus needs to also repeat earlier babbles. Repetition of the old babbles would benefit from concentrating on acoustic regions corresponding to prototypical CG's vowel categories, rather than regions that may not have a clear interpretation by the caregiver.
3. As explained above, several of CG's responses need to be associated with each of LeVI's articulatory or acoustic babbling-targets to overcome ambiguity. If LeVI has any trouble remembering or reproducing its old babbles exactly as they were during previous interactions, LeVI needs a way to judge which of the old babbles the new imitative response should be

associated with. LeVI thus needs to somehow cluster its own productions, and associate CG's imitative responses in the clusters created (C3 at Figure 1).

The main aspects of the learning components that were found to improve LeVI's learning performance during model development are described below. The technical details can be found in the Appendices. The model is designed considering the sources of information that are realistically available for real infants during speech learning. The learning components used in the present work are rough preliminary implementations that were found to improve learning performance – the algorithms and their parameter values are not systematically optimized for the learning task. Instead, we first started from initial guesses for model parameters and used them to collect interaction data from nine Finnish subjects. These data were then used to tune the model parameters towards better performance. The final performance of the resulting model was then evaluated using an *evaluation set* of four new participants. In addition, seven participants from the initial set of nine speakers were asked to retake the training and the testing phases after the tuning of the model. We will refer to this set of seven talkers as the *development set* in the remainder of this paper. Results for the two groups tested are reported separately in section 4.

LeVI's learning and imitation ability is always dependent on the participant. We did not measure how well LeVI, after learning from one participant, would be able to imitate vowels spoken by another participant. In addition, for automatic evaluation of different learning components, the words that were originally recorded by the participants for LeVI's training were used as CG's imitative responses. CG's perceptual behavior was simulated by classifying LeVI's babbles with an automatic classifier. Results from these simulations are discussed in Section 5.4.

2.1 The articulatory model

The speech production mechanism of the virtual infant used in the current experiments consists of a vocal tract model that is implemented as an adapted version of the Mermelstein's (1973) vocal tract model. A detailed description of the model can be found in Rasilo (2012). In short, only stable vocal tract configurations are used as articulatory targets and each configuration is defined by nine vocal tract parameters: tongue body x and y -coordinates, tongue tip x and y –coordinates, jaw angle, hyoid x-coordinate, lip protrusion, lip opening and velum opening. The coordinates of the articulatory parameters are transformed into cross-sectional areas of the acoustic tube model, and synthesized using the Kelly-Lochbaum transmission-line (Kelly & Lochbaum, 1962). In order to give vocalizations a more natural feel, each vocalization is synthesized with the F0 starting at a uniformly sampled random value between 300 and 320 Hz, rising to a uniformly sampled random value between 0 and 10 Hz higher than the starting F0, and finally decreasing to a uniformly random value between 10 and 20 Hz below the starting F0. The vocal tract length is linearly scaled to a length of 10 cm from the original adult values (15–18 cm) in order to represent an infant vocal tract and to induce non-correspondence to adult vocal tracts (see, e.g., Vorperian et al., 2005). We acknowledge that vocal-tract growth is non-uniform in reality (e.g. Vorperian et al., 2009) and linear scaling of the vocal tract to infant size is a rough approximation. However, the main purpose is to induce the most essential

acoustic differences between the adult participants' and LeVI's voices, and the linearity is not utilized in the mapping process in any manner[1].

2.2 Clustering of LeVI's own productions

As mentioned before, LeVI learns acoustic models for CG's vowels by associating its own productions to the imitative responses by CG. Because of the ambiguity in CG's responses, robustness of LeVI's acoustic model for a particular vowel sound only arises due to repeated accumulation of acoustic information from several CG's responses. The acoustic information from multiple responses has to be gathered in one distinct model representing the sound to be learned. If the articulatory parameters of individual babbles were simply associated to the responses by CG, LeVI would end up with pairs of individual points between the articulatory and auditory space, and an additional mechanism would be needed to extract any categorical structure shared by these pairs of exemplars (e.g., high-dimensional probability density estimation), assuming that the words could be converted into fixed-dimension vector representations in the first place. On the other hand, if LeVI can represent its own speech sounds in terms of categories containing multiple individual babbles, the corresponding CG responses can be analyzed in the context of each category. As long as the auditory content in CG responses to a certain LAC is statistically biased towards a specific speech sound (i.e., there is an above-chance level of imitation or lexical alignment in parental responses), the category can gradually become representative of a single phoneme most consistently occurring in CG responses when analyzed in distributional manner. Note that this consistently occurring phoneme can (and should) have normal acoustic variation so that the category becomes robust to different realizations of the same phoneme. However, due to variation in LeVI's babbles, it is not trivial how these categories should be defined.

To clarify via an example, if CG's responses were associated with their preceding babbles (b1 – b4): b1→"kisa", b2→"kate", b3→"sato", b4→"loma", without any clustering the learner would end up with four unreliable word recognizers, one per babble, and no recognizer would react strongly to individual vowel sounds. If the four babbles were assigned to one category c = {b1, b2, b3, b4} and all four CG's response were associated with c, statistically the proportion of vowel /a/ would end up dominating the category, and we could deduce that the category c represents the vowel /a/ in CG's speech.

If LeVI were to have infinite articulatory accuracy and thus an ability to repeat babbles in exactly the same manner, it could associate several CG responses directly to the motor commands performed. When motor commands cannot be perfectly reproduced, as in the present simulations, it is possible that small variations in the articulatory domain causes relatively large changes in the acoustic output (in sensitive regions of the vocal tract, for example close to constrictions). If this small amount of noise in LeVI's articulation causes a big shift in the babbled vowel's spectral characteristics, CG may end up interpreting the vowel as a different vowel than the original motor command that LeVI aimed at would have produced. As a result of this effect, it is beneficial for LeVI to cluster its own productions in the auditory domain.

[1] In principle, the present learning framework allows mapping between any two arbitrary vocal systems without requiring any acoustic similarities at all. The only requirement is that the functionally equivalent sounds in these two systems consistently co-occur in similar interaction contexts.

We assume that caregivers interpret infants' vocalizations based on their acoustic characteristics, and it is thus beneficial for infants to learn vocalic categories that are consistently perceived as a single phoneme by their caregivers. The clustering principle is illustrated in Figure 2.

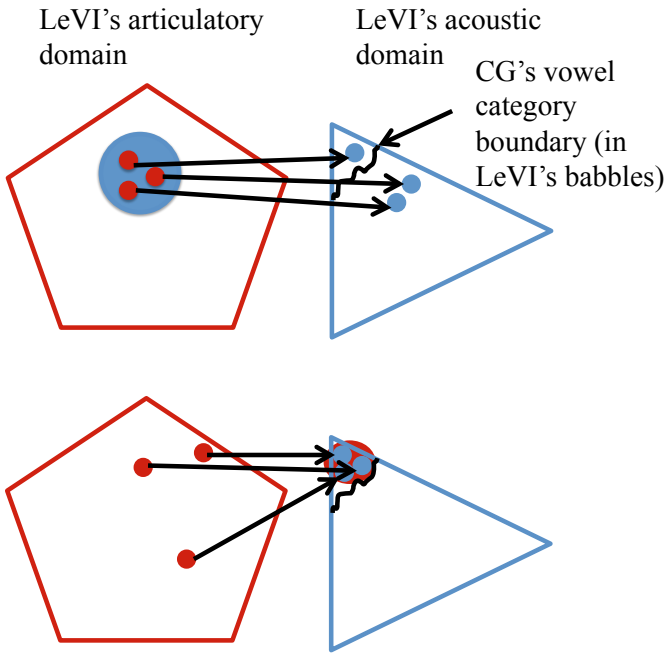

Figure 2. Illustration of LeVI clustering its own productions. If clustering were performed in the articulatory domain (above), the acoustic outputs from the articulations inside an articulatory cluster may cross phonemic category boundaries, as interpreted by CG. If the clustering is performed in the auditory domain as in this work (below), LeVI can learn several ways to articulate each sound while CG's interpretations of babbles from the same auditory category (LAC) remain more consistent**.**

When LeVI babbles a vocalic sound, its articulatory and acoustic characteristics are stored either in a new LAC or in an already existing one. During the simulations LeVI ends up learning a large number of LACs, out of which some will react more strongly to vowels spoken by CG, and thus could be considered closer to the real native language vowel categories. The LACs created could be considered as "proto-vowels" that might evolve to the correct number of native vowel categories during further stages of speech learning that also includes feedback from the concurrent lexical learning (a topic not discussed in the present paper). We use a simple threshold on the acoustic distance between the babbled sound and the existing LACs, based on extracted two first formant frequencies, in order to determine whether a new LAC should be created for the latest babble. Details of the category creation process are discussed in Appendix A.1.

2.3 Articulatory properties of LeVI's vocalizations

As described above, LeVI creates clusters (LACs) for its babbled sounds based on their acoustic representations. LACs store the articulatory and acoustic features of the included babbles. Since it is possible that a LAC includes several differing articulatory configurations, additional articulatory clusters are created for every LAC.

In this work, after the babble has been assigned into a LAC based on their acoustic distances, an articulatory distance threshold (= 40) is used to either assign the babbled articulatory configuration into an existing articulatory cluster or into a new articulatory cluster, specific to the activated LAC. The distances are calculated using non-normalized articulatory parameter vectors between the articulatory vector of the current babble and the centroids of each articulatory cluster (calculated as the mean of all articulatory vectors stored inside an articulatory cluster). This is a preliminary approximation and aims to avoid a situation where LeVI keeps trying to reproduce an articulatory configuration in a non-linear region, where acoustic outcomes of nearby articulations are unlikely to fall close to the original acoustic output.

Whenever LeVI tries to reproduce one of its previous babbles, it aims to produce an articulatory parameter vector that is the mean of all the articulatory configurations stored in the *largest* articulatory cluster belonging to the activated LAC. This is done because it is likely that the articulatory category with the most members lies in a rather stable articulatory region – small changes in articulation lead to small changes in the acoustic domain – and is thus the easiest to reproduce. The articulatory clustering is illustrated in Figure 3.

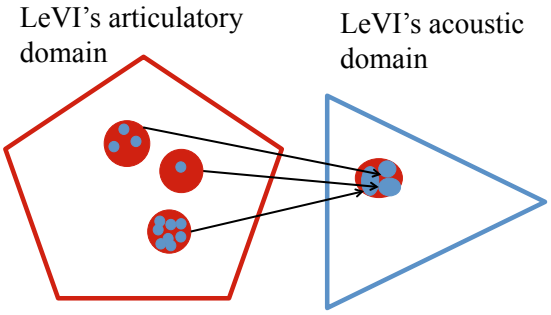

Figure 3. Illustration of a LAC and related articulatory categories. LAC on the acoustic domain (right) and corresponding articulatory clusters (left). Each articulatory cluster can consist of several articulatory parameter vectors. When LeVI tries to reproduce the babble, it tries to produce the centroid of the articulatory cluster with the most members.

2.4 LeVI's vocalization behavior

LeVI's vocal behavior has to serve two functions. First, LeVI has to explore its articulatory domain in order to find new LACs. Second, LeVI should reproduce already existing categories in order to learn robust acoustic models for vowels occurring in CG's imitative responses. In our current experiments, LeVI creates an exploratory babble with a probability 0.20 and a reproductive babble with a probability 0.80.

When LeVI creates an exploratory babble, it is beneficial for LeVI to try to expand the reach of its productions in the acoustic domain. Due to a non-linear mapping between articulatory and acoustic parameters, uniformly sampling articulatory configurations from the ranges of articulators' parameter values leads to concentration of babbled sounds in the middle of the acoustic region. In order to produce speech sounds that also fall in the border regions of the acoustic domain (for example, the borders of the vowel triangle spanned by the two first formant frequencies), LeVI must find articulations consisting of extreme values for several articulators. For example, the production of /i/ requires the tongue body to be a in maximally frontal position, tongue tip close to the hard palate, minimally protruded lips and a maximally widened hyoid region of the vocal tract. Therefore LeVI has a

mechanism that attempts to produce new babbles as far as possible from the already known articulatory configurations while still respecting the physical constraints of the vocal tract. In order to calculate potential new articulatory configurations far from existing ones, LeVI makes use of pre-defined ranges for each articulatory parameter (in practice: configurations that lead to audible sounds). Note that even though LeVI knows its possible articulatory ranges, it does not know their acoustic counterparts – LeVI thus cannot expand its acoustic domain directly in desired directions in the acoustic domain since it does not know which articulations lead to which acoustic outputs. Details of the implementation are given in Appendix A.2.

Since each LAC should converge to robustly recognize a particular vowel of CG's speech, LACs should have several imitative responses by CG associated to them. This is possible due to LeVI's reproductive babbles. However, since some LACs may be better exemplars of CG's vowels than others, LeVI should also guide its own babbling towards LACs that are consistently recognized as a single vowel by CG across several babbles. For example, if LeVI were to create a LAC in a region that lies in between two CG's vowel categories, CG's interpretations of sounds babbled from the LAC may not be consistent (see Figure 4). For example, if one LAC lies between the CG's perceptual categories for Finnish /a/ and /o/, it is possible that CG sometimes recognizes a babble from this category as /a/ and sometimes as /o/. In such cases, LeVI should have some way of biasing its babbling towards "pure" LACs that get more consistent feedback. In Appendix A.3. we describe how the weights of LACs are obtained in order to take into account both of the above aspects, the number of babbles per category and biasing babbling towards categories in less ambiguous regions, when reproductive babbles are created.

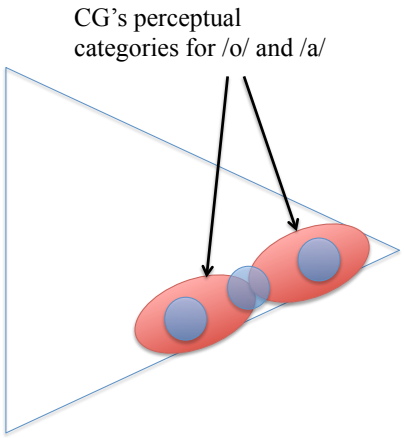

Figure 4. LACs (blue) and CG's perceptual categories for two vowels (red), as interpreted from LeVI's babbles. If LAC is created on the border of the two CG's categories, feedback from CG can be inconsistent, and the resulting acoustic model noisy. If LACs fall inside CG's perceptual categories, they will be consistently interpreted as the same vowel, and the models will converge better to recognize the corresponding CG's vowels.

In order to account for inaccuracies and variability in articulatory gestures targeted at the same categories (c.f., Chung et al., 2012; Smith & Zelaznik, 2004; Lee et al., 1999), a small amount of noise is added to the articulatory parameters. We also include a simple learning mechanism that increases LeVI's babbling accuracy based on babbling experience within regions of the articulatory domain. The implementation of the inaccuracy component is described in appendix A.4.

2.5 LeVI's speech recognition and imitation

Whenever LeVI babbles a vocalic sound, CG responds to the babble with a Finnish CVCV-word, where one of the vowels is the same vowel sound that the participant identified in LeVI's babble[2]. An associative acoustic model is initialized for every LAC created during the interaction. Acoustic features of CG's responses are stored in these models, and model identity is defined by the LAC assigned to the produced babble. Thus, in principle, the acoustic model cannot distinguish between the two presented vowel sounds or consonants after observing only one babble-response-pair since the model would be equally sensitive for all of the sounds. The convergence of the model towards the correct speech sounds occurs only through repetitive babble-response pairs that partially correlate with the vocalizations. Therefore training of the recognizer can be considered as weakly supervised.

In this study, we use a weakly supervised *Concept Matrix* algorithm (Räsänen & Laine, 2012) to associate the acoustic features of CG's responses to LACs. The algorithm approximates high-order Markov structure in speech signals as a mixture of bi-gram statistics at different temporal lags. In this work we use a dynamic version of the algorithm, DCM (Rasilo & Räsänen, 2015) since preliminary simulations with the algorithm indicated significant increase in the vowel learning accuracy compared to the original CM algorithm. In practical terms, in DCM, LeVI tries to detect where the babbled vowel sound lies in CG's response based on the training obtained until the current training utterance, and updates the acoustic models more strongly at the hypothesized locations. Detailed description of the algorithm used is given in Appendix A.5.

Whenever LeVI hears speech from the participant, it recognizes the speech using the acoustic models of each LAC learned thus far. The recognition procedure gives an activation value for each LAC for each time instant in the input signal. LeVI can try to imitate the participant's speech sound at a given time instant by selecting the most strongly activated category (see equation A-7) and reproducing the corresponding articulatory parameter vector in the same way as for babbling, described in section 2.2.3.

3 Experiments

The learning experiments were conducted with human participants in two separate sessions. As explained above, the basic idea behind the experiments is that LeVI babbles a sound with an open vocal tract configuration and voiced excitation. The participant interprets the sound as the most likely Finnish vowel sound and responds with a CVCV word containing the same vowel. Since the algorithm is designed to work on-line (babbling and interaction phases are combined), we had to design the experiment so that time requirements for all participants would stay within reasonable limits. We wanted each participant to complete a *training phase* for LeVI, consisting of 1000 babble-response pairs in order to show substantial learning. After the training, participants were asked to evaluate vowel identities in LeVI's imitations in a *testing phase* in order to get an accurate subjective measure of LeVI's imitation abilities.

The learning mechanisms of the model, including all fixed hyperparameters (e.g., clustering threshold values for acoustic and articulatory category creation and

[2] With the exception of two words in the 80-word training set. "söpö" and "höpö" have the same syllable twice but we accepted the words due to the rarity of real Finnish CVCV words with /ø/ and another vowel.

parameters chosen for the speech recognizer; see Appendix A), were fine-tuned on a development set collected from nine participants completing both the training and testing phases of the interaction with an early version of LeVI. The parameters were tuned in order to find parameter values leading to optimal LeVI's imitation performance, while using the speech utterances recorded by the nine participants as CG's responses to LeVI's babbles in automated simulations. During development, CG's behavior was simulated using the data collected, as is also done during the automated analysis in section 4.4. Since exhaustive grid search over all parameter combinations was not computationally feasible, the parameter values reported represent a setup that was found to provide consistent performance across CGs in the development data.

In order to test the generality of the model and its parameters, the final performance was evaluated using an evaluation set of four new participants that completed both the training and test phases while all hyperparameters of LeVI were kept fixed. Additionally, seven of the participants of the development set retook the training and testing phases after the model was established, but without pre-recording a new set of words for the CVCV-responses. The model development and evaluation phases are summarized below:

Model development:
9 participants (word recordings, training + testing phases)
Model evaluation:
4 new participants (word recordings, training + testing phases)
7 participants from development set (old word recordings, new training + testing phases)

3.1 Word recording

In order to save time in the training phase, all participants pre-recorded all the CVCV words used in the training phase. The recorded wordlist consisted of 160 unique Finnish words (see Appendix C for full word lists). Eighty of the words were repeated 14 times (non-consecutively) in order to have natural variability in the acoustic word forms (word set #1). Thirteen of the repetitions from the word set #1 formed the *training set* for the training phase of LeVI and one production per word was left out for testing. An additional set of 80 words, each word spoken once, was recorded for testing purposes only (word set #2) in order to see if the vowel imitation performance is affected by the lexical word forms. Out of these 80 words, 76 were novel words not present in word set #1. We accepted four overlapping words ("jänö, köhä, möly, tykö") between sets #1 and #2 because real Finnish CVCV words with vowel /ø/ were too rare to have a completely new set of words. Similarly, word set #2 includes a non-word ("käte") for vowel balancing purposes. However, "käte" occurs in Finnish as a part of longer words such as "kätevä". Thus, every participant recorded a total of 1200 words. The training set consisted of 1040 words, out of which words were drawn depending on the vowel the participant annotated, while the testing set consisted of 80 tokens from the word set #1 and 80 tokens from the word set #2.

3.2 Training phase

During the training phase, at every interaction, LeVI babbled a vocalic sound according to the description in section 2.4, and updated LACs accordingly (see section 2.2). The participants were only asked to select which vowel occurs in LeVI's vocalizations. After every selection, the system selected a previously recorded word from the training set including the annotated vowel, and played it to LeVI. In word selection, the system tried primarily to select a word sample that had not been used for training in previous interactions (from the 1040 training word samples in total). However, some words may have ended up being used more than once if LeVI ended up babbling a certain vowel sound more times than it existed in the recorded words of the training set. LeVI associated CG's response to the activated LAC using the weakly supervised learning algorithm. After 50, 250 and 1000 babble-response pairs LACs and associated acoustic models were saved in system memory. These were then used to evaluate LeVI's imitation accuracy in the corresponding moments during LeVI's training. The training phase procedure is illustrated in Figure 5.

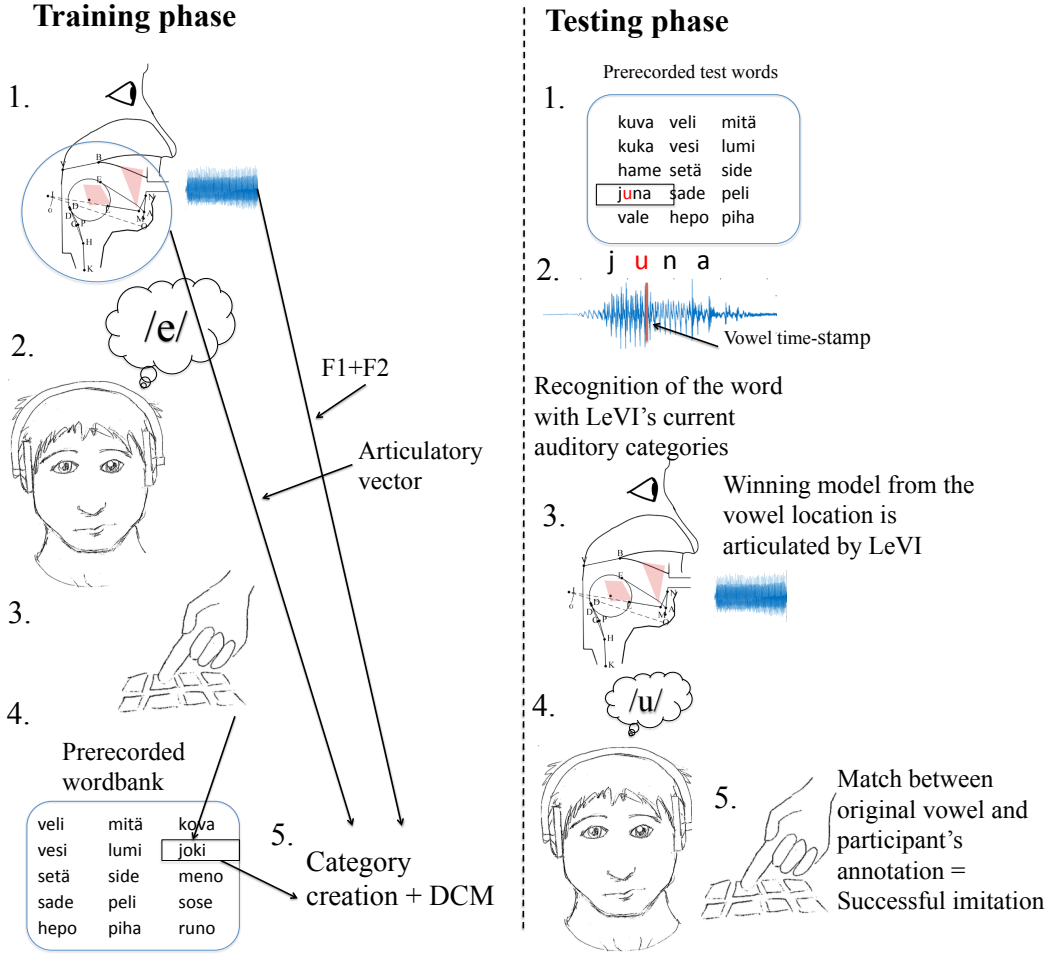

Figure 5. The process in the training phase and testing phase. **Training phase**: 1) LeVI babbles a sound, 2) The participant evaluates the sound and 3) assigns the sound to a Finnish vowel category, 4) The system selects a word that has the assigned vowel, pre-recorded by the same participant, 5) the acoustic word is played to LeVI and LeVI updates its recognizers and LACs using the F1 and F2 values of the babble, the articulatory vector and the response by the participant. **Testing phase**: 1) A pre-recorded test word and one of its vowels are randomly selected for evaluation by the algorithm, 2) The word is recognized by LeVI and the most activated LAC on the vowel time instant is selected, 3) LeVI vocalizes with the

articulatory configuration related to the winning LAC, 4) The vocalization is played back to the participant for evaluation, 5) If the assigned vowel matches with the original vowel in the selected test word, LeVI's imitation was successful.

3.3 Testing phase

The first author manually searched for the approximate locations of the vowel sounds from both recorded word sets for each participant. After the training phase of 1000 annotations finished, LeVI was set to imitate the complete test set. First, LeVI recognized the words in the test set according to equations (A7-A10) of the Appendix A by using its LAC-based acoustic models for CG speech. The most activated LAC, *winner*(*v*), was chosen for each vowel (see equation A-11), where *v* is the hand-selected time moment for the vowel sound. After recognition, LeVI imitated the vowels using the mean of the articulations stored in the largest articulatory cluster of the winning LAC (articulatory noise was not added in the testing phase, only in training), and the acoustic outputs of the imitations were stored. The vowels for word set #1 were imitated corresponding to LeVI's status after 50, 250 and 1000 interactions, and the vowels for word set #2 only after 1000 interactions. This led to a total of 640 (four times 80 words, two vowels per word) imitations.

After LeVI had imitated all the vowel sounds (during which the participant was having a five-minute break), the testing phase began. The word imitations were played one-by-one to the participant. The participant was unaware that the last 640 vocalizations were imitations – the annotation procedure was equal to the training phase, as seen by the participant. If the participant annotated an imitated vowel as the original vowel in the word that had been initially recorded by the participant herself, the imitation was considered successful. This last phase was used to measure LeVI's imitation accuracy, and thus the validity of the learning algorithm of this study. The testing phase procedure is illustrated in Figure 5.

3.4 Participants

The development set consisted of nine Finnish speaking participants (six male, three female, average age 29.7 years), and the evaluation set consisted of four Finnish speaking participants (three male, one female, average age 29.5 years). Participants were paid 20 euros in cash after finishing both parts of the experiment.

3.5 Procedure

Each participant was asked to record the word set of 1200 words using a graphical user interface developed in Matlab. When participants pressed "Record", 10 words were presented on the screen one by one, and the participants were asked to read the words aloud. If they made a mistake during recording, they were asked to rerecord the word list. When the recording was successful, the audio was saved and a new list of 10 words was shown. The recordings of the first nine participants were performed in an anechoic chamber. The second recording session for the four new participants took place in a sound-isolated listening room.

Both recordings were performed with Rode NT1-A microphones with a Motu Ultralite MK3 preamplifier. The duration of the recording session was about 60 minutes per participant. Participants were asked to speak with normal speaking voices, recommended to have pauses and offered drinks in order to avoid fatigue.

The training and testing phases were performed on a different day than the word recording for each participant. Before the experiment, the participants were shown a 17-second video of random babbling by LeVI to get an idea of how the

infant's voice sounded like. In the annotation phase, a babble sound lasting for 0.25 seconds was played to the participant via headphones, and the participant was asked to select the vowel heard on the keyboard.[3] The annotation of 1640 babbled sounds took approximately 40–60 minutes per participant.

4 Results

We report results separately for the seven participants from the development set (D1-D7) who retook the annotation phase with the tuned learning model, and the four evaluation set participants (E1-E4) whose speech was newly recorded and who interacted with the final version of the model. Figure 6 shows formant frequencies of all the vowels babbled by LeVI in the training phase for all 11 participants. The color indicates which vowel was selected by the participant for the corresponding babble. The regions for the participants' perceptual categories can be observed from the figure. The overlap of the perceptual categories on their border regions indicates that vowel discrimination at these regions is more difficult, at least when measured in terms of the two first formants. It is also visible in the figure that some babbles in the perceptual region for /o/ or /ø/ are annotated as /e/. Analysis of the spectra of the babbles indicates that a weak second formant is detected at lower frequencies in these cases, but the presence of a clear third formant at around 3000 Hz (typical for the second formant for /e/) caused participants to annotate the babble as /e/. Labeling mistakes (pressing the wrong key by accident) were minimal (total of 14 errors, on avg. 1.27 per participant, recorded with self-evaluation during the annotation phase).

The figure also shows that there are two blank areas in the region for vowel /æ/. This is likely caused by the sampling algorithm that seems to avoid articulatory regions that would have produced these specific formant values, possibly because a small change in articulatory parameters would have caused large acoustic distances in these regions, therefore leading to a low priority in the sampling process (see Appendix A.2 for details).

[3] In Finnish, all eight possible vowels have a corresponding letter on the alphabet (and a key on the Finnish keyboard) to make this arrangement possible. With other vowel-wise phonetic languages like Italian and Spanish, the same arrangement should work. With non-phonetic languages the interaction phase should have a different arrangement.

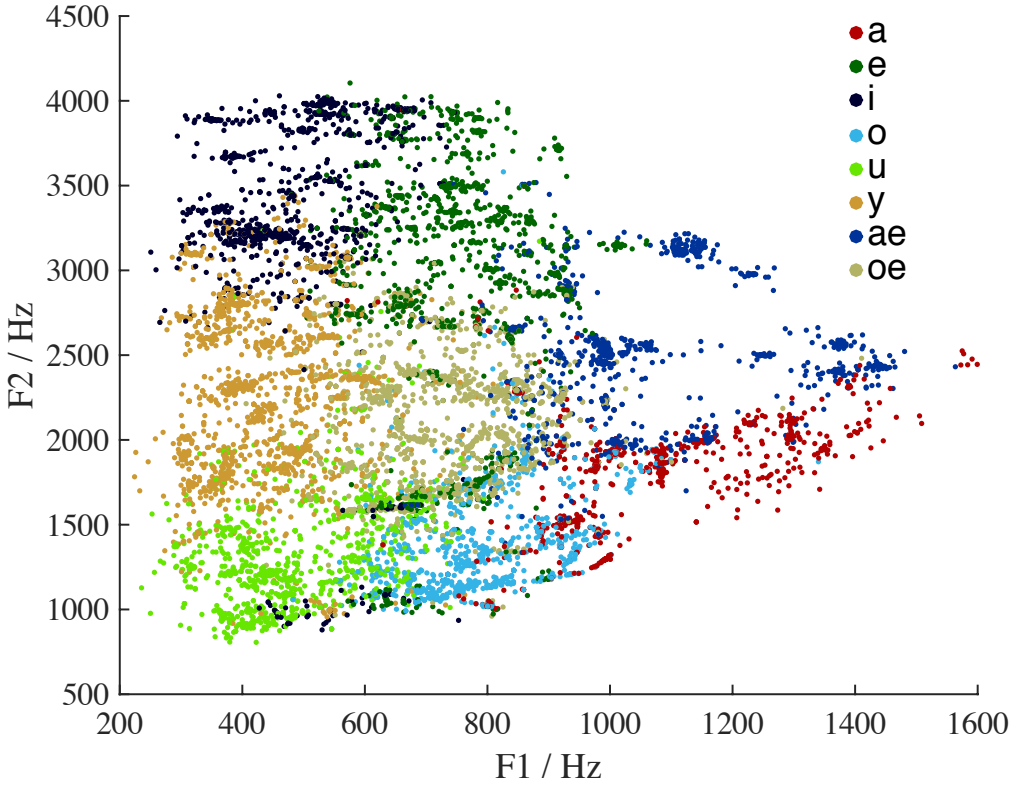

Figure 6. All vowels babbled by LeVI in the training phases for all participants. The colors illustrate which vowel was selected for the corresponding babble by the participant.

4.1 LeVI's vowel imitation accuracy

LeVI's average imitation accuracy, calculated as the number of vowels successfully imitated divided by the total number of vowels at the three time instances during the training, is shown in Figure 7. It can be seen that LeVI learns to imitate the vowels of the evaluation set better than the seven participants of the development set. This improvement in learning is probably due to smaller amount of noise in the recorded signals – the word recordings of the development set have a considerable noise component due to a microphone picking up the fan of the recording laptop. The functioning of the model with the four new speakers also suggests that the model implementation choices and free parameter values have not been selected to suit only the speakers in the development set, but the model should function similarly with novel voices. LeVI's imitations of the vowels in the words of the word set #2 for the participants in the evaluation set are also available in the supplementary material, where each recorded word is followed by LeVI's imitation of the two corresponding vowels.

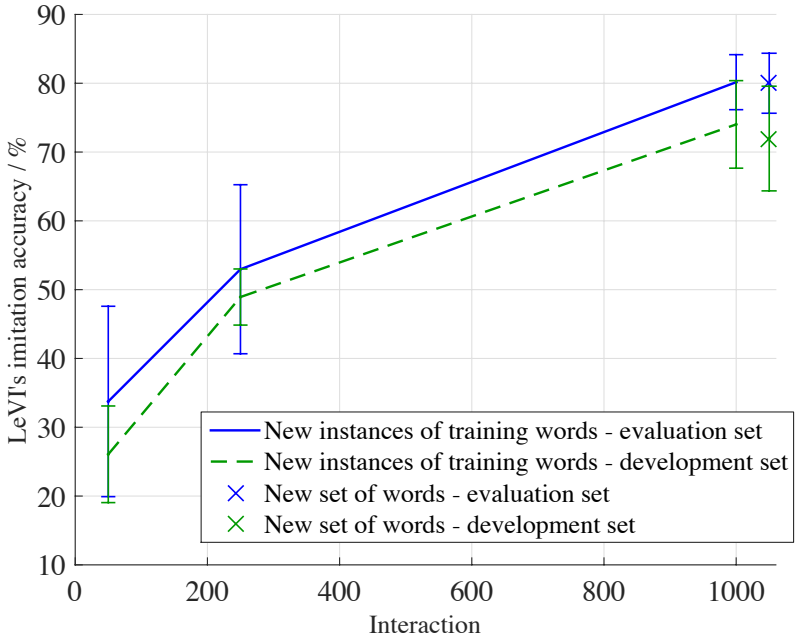

Figure 7. LeVI's imitation accuracy in three phases during learning, averaged over participants. Held-out part of the word set #1 is shown with solid and dashed lines for the evaluation set and the development set respectively, and word set #2 with crosses (shifted 50 steps to the right for clearness). Vertical bars denote ±1 standard deviations across the participants.

Table 1 shows the results in more detail, including vowel-specific imitation scores as well as scores specific to the position of the vowel in the tested CVCV-words. For the evaluation set, the average unweighted accuracy was 79.4% for the first vowel in the CVCV words and 71.6% for the second vowel. For the development set, the accuracies were 73.3% and 63.4% for the two positions, respectively. For both, the development and the evaluation set, and for both vowels, learning of /æ/ was the most difficult, having an average imitation success rate of only 55.6% and 66.7% in the evaluation set. This is probably due to the difficulty of exploring articulations that lead to vowel sounds produced in the perceptual area for /æ/ in the F1-F2 domain as explained above. It is likely that further improvements in the exploration algorithm could increase imitation accuracy.

The higher accuracy in imitating the first vowel of the words is likely due to the fact that in Finnish word stress is always placed on the first syllable of a word and the first vowel thus undergoes less reduction in its pronunciation. Observation of the words for the development set showed that, in some cases, the energy of the speech signal during the last vowel was very small and the proportion of the background noise was considerably high at the ends of the words. For some participants, the word ends were suppressed and ended up sounding creaky without clear vowel articulation. This phenomenon presumably weakens recognition of these vowels, as well as affecting the training phase.

As an example of error patterns, Appendix B shows a list of vowel-specific errors for subject E1. For instance, when the correct vowel to be imitated was /a/, LeVI's imitation was not annotated as /a/ in total six times. Out of those six times, three times LeVI imitated the vowel /a/ using a LAC lying approximately in the region for /o/. It can be seen in the appendix that many of LeVI's erroneous imitations are located in areas acoustically close to the imitated vowel, except in case of /æ/

where errors are more widely spread. It is thus possible that even if the absolute vowel imitated by LeVI were interpreted as a different vowel category, the pronunciation of a word using these articulatory configurations might be more easily understood by CG, since the context, or for example surrounding consonant sounds, would probably bias CG to interpret the "erroneous" vowel as the correct one.

Table 1. Detailed imitation accuracy measures for evaluation (left) and development (right) sets, shown separately for vowels located in the first (top) or last (bottom) syllable of the CVCV word.

	Evaluation set						**Development set**							
	E1	**E2**	**E3**	**E4**	**Mean**		**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**Mean**
Mean all	75.6	80.6	80.3	83.8	**80.1**		76.9	73.1	79.4	68.4	80.0	60.9	72.2	**73.0**
Mean, word set #1	77.5	80.6	76.9	85.6	**80.2**		75.0	73.8	80.6	69.4	80.0	62.5	76.9	**74.0**
Mean, word set #2	73.8	80.6	83.8	81.9	**80.0**		78.8	72.5	78.1	67.5	80.0	59.4	67.5	**72.0**
First vowel, word sets #1+#2														
a	82.6	56.5	82.6	100.0	**80.4**		82.6	0.0	87.0	60.9	56.5	87.0	78.3	**64.6**
e	80.0	96.0	84.0	72.0	**83.0**		80.0	88.0	88.0	36.0	92.0	28.0	84.0	**70.9**
i	77.3	86.4	81.8	77.3	**80.7**		86.4	86.4	90.9	72.7	86.4	95.5	90.9	**87.0**
o	88.2	76.5	100.0	94.1	**89.7**		88.2	88.2	64.7	100.0	88.2	47.1	47.1	**74.8**
u	78.9	84.2	57.9	84.2	**76.3**		89.5	84.2	94.7	73.7	89.5	94.7	78.9	**86.5**
y	60.0	100.0	93.3	100.0	**88.3**		86.7	86.7	100.0	86.7	80.0	100.0	100.0	**91.4**
æ	48.1	37.0	51.9	85.2	**55.6**		70.4	55.6	0.0	40.7	55.6	48.1	44.4	**45.0**
ø	91.7	100.0	75.0	58.3	**81.3**		50.0	33.3	100.0	66.7	100.0	66.7	50.0	**66.7**
Vowel mean	75.9	79.6	78.3	83.9	**79.4**		79.2	65.3	78.2	67.2	81.0	70.9	71.7	**73.3**
Second vowel, word sets #1+#2														
a	86.7	80.0	80.0	100.0	**86.7**		93.3	0.0	100.0	80.0	0.0	80.0	80.0	**61.9**
e	73.9	82.6	91.3	82.6	**82.6**		60.9	95.7	95.7	56.5	78.3	21.7	91.3	**71.4**
i	71.4	94.3	91.4	77.1	**83.6**		91.4	97.1	100.0	88.6	82.9	62.9	82.9	**86.5**
o	75.0	100.0	85.0	95.0	**88.8**		70.0	95.0	100.0	75.0	90.0	30.0	20.0	**68.6**
u	81.0	81.0	81.0	95.2	**84.5**		81.0	90.5	90.5	76.2	85.7	66.7	90.5	**83.0**
y	80.0	80.0	93.3	66.7	**80.0**		86.7	86.7	100.0	93.3	100.0	73.3	86.7	**89.5**
æ	71.4	57.1	61.9	76.2	**66.7**		28.6	81.0	0.0	47.6	100.0	28.6	38.1	**46.3**
ø	90.0	100.0	80.0	70.0	**85.0**		80.0	40.0	100.0	60.0	100.0	80.0	100.0	**80.0**
Vowel mean	67.4	71.9	73.0	74.1	**71.6**		64.0	68.2	73.3	64.7	67.1	45.4	61.2	**63.4**

4.2 Final LeVI auditory categories

During the 1000 babble-response pairs of the training stage, LeVI formed on average 21.72 LACs, varying between 19 and 24 depending on the participant. The number is higher than the number of Finnish vowel categories, but this does not pose a problem in the communicative context of speech: even if LeVI were to use several LACs, resulting in several different articulations for each vowel, communication succeeds as long as the caregiver interprets these as belonging to the same vowel category. As mentioned in the introduction, Chung et al. (2012) found that 2-year-old children had larger amounts of variability than 5-year-olds or adults in their /i/, /u/ and /a/ vowel

productions when they were asked to reproduce native words including the corresponding vowels. It is difficult to establish whether infants indeed have several auditory categories for each vowel type, but based on the study of Chung et al. (2012) it is clear that infants use multiple articulatory variants during imitation even when imitating the same vowel sound spoken by an adult speaker. The occurring variation may later be compressed into a more compact set of sound categories. This may take place for example with the help of word meanings (e.g. the child finally converges in the minimum set of vowels needed in the native language to discriminate between the words of the language), articulation efficiency constraints or acoustic characteristics (e.g. more resonant vowels could be preferred), or learning of more refined motor control. However, it has to be noted that phonemes may not be the unit of perception when humans listen to speech (see, e.g., Mitterer, Scharenborg & McQueen, 2013) indicating that we should not necessarily expect the perceptual categories of the learner to ever converge into linguistically motivated phonemes, but simply to learn proper acoustic contrasts in different contexts. This study indicates that when using a similar learning strategy to the one proposed, infants can converge into a functional communication system without the need to know the exact number of native vowel categories.

Even though the final number of LACs in this study is reasonably large, LeVI still ends up preferring certain LACs over others. This is simply due to the fact that when a LAC is created inside the participant's perceptual category - opposed to a boundary region of the participants categories - the resulting babbles are consistently annotated in a single vowel category, and the acoustic models end up capturing the acoustic properties of the corresponding vowel sounds more accurately. When recognizing the words from the test set, the less noisy models activate more strongly during the corresponding vowel sounds, and are thus more likely to become the model with which LeVI imitates the sound. For example, for subject E1 during the imitation of the two word sets in the final test phase after 1000 interactions, the number of imitations according to all 21 LACs were (35, 15, 7, 7, 0, 14, 1, 13, 21, 18, 35, 14, 9, 37, 12, 22, 30, 8, 10, 5, 7). The numbers are illustrated visually in Figure 8, where each circle represents a LAC centroid and the circle's size is proportional to the amount it was used in imitation. The circles are plotted over all LeVI's training babbles for E1, colored according to the annotation by E1. It can be seen that the LACs in between E1's vowel categories are used less for imitation, and categories inside E1's vowel categories are used rather frequently.

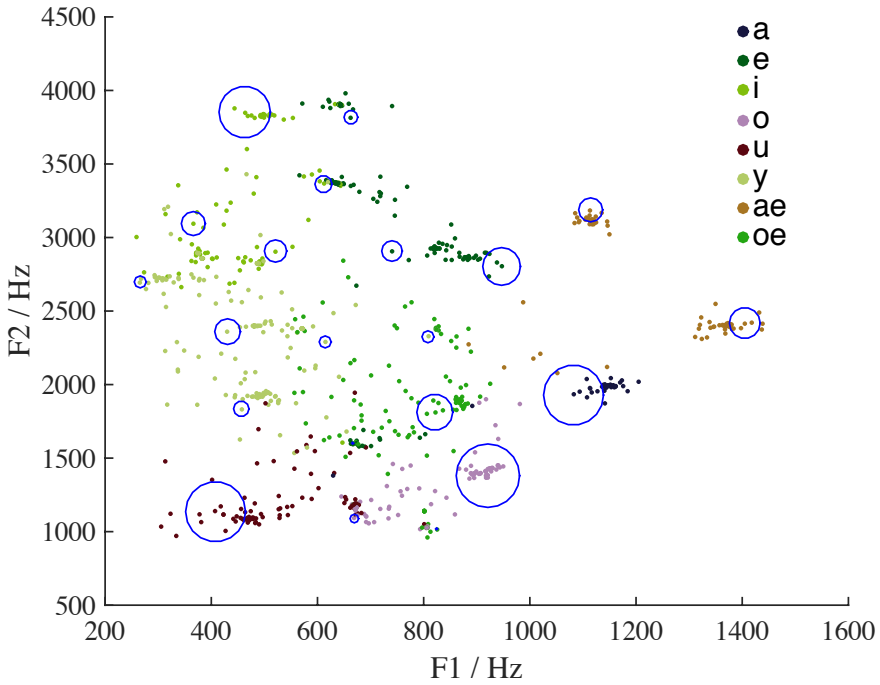

Figure 8. The circles show the locations and frequencies of LACs when imitating the vowels in both word sets. The bigger the circle area, the more frequently the corresponding LAC was used by LeVI in imitation. Circles are drawn over all training phase babbles for E1, colored according to the annotation by E1.

4.3 Reaction times in babble annotation

One of the original hypotheses was that participants would be more consistent in annotating LeVI's babbles as a single vowel if the corresponding LAC would lie inside the participant's perceptual vowel category. In order to get an idea of how confident the participants were when they annotated heard vocalic babbles in different regions in the acoustic domain, we measured the reaction time (RT) of the participant in terms of the time difference between the offset of the babbled sound and the moment when the participant presses a key for vowel annotation. The hypothesis is that longer RTs would be inversely correlated with the quality of the exemplars with respect to Finnish vowel categories, and that the participant should therefore require more time to decide which vowel should be assigned to the babble.

In order to illustrate RTs, we divided the F1-F2 space into 20 x 60 Hz sized rectangles and calculated the mean annotation time inside each of them across the combined development and evaluation sets. Figure 9 shows the average annotation time for each square, colored according to the time in seconds, as explained in the color bar. The black polygons indicate approximate Finnish vowel categories, calculated using all babbles by LeVI, annotated in Finnish vowel categories by the participants (i.e. from data shown in Figure 6). The black polygons correspond to the "bags" of a bagplot (a bivariate version of a boxplot) calculated for each vowel category, surrounding the *depth median* of the category samples (see Rousseeuw, Ruts & Tukey, 1999 for details). The bag polygon includes 50% of the category samples. The bag for vowel /e/ shows the effect of some samples having a weak formant peak at a lower frequency than is typical for the second formant of /e/, discussed in section 4. It appears that in the regions of Finnish vowel categories (compare also with Figure 6), the mean annotation time is generally less than one second, whereas in regions between categories longer annotation times are recorded.

Based on the visual illustration of annotation times, it seems that participants are faster in identifying vowel identities when LeVI's babbles fall clearly inside the participants' perceptual categories. In further models, the reaction time of the caregiver could be presumably used as a cue to indicate the quality of the infant's babble and as an additional factor to reinforce certain babbles more than others.

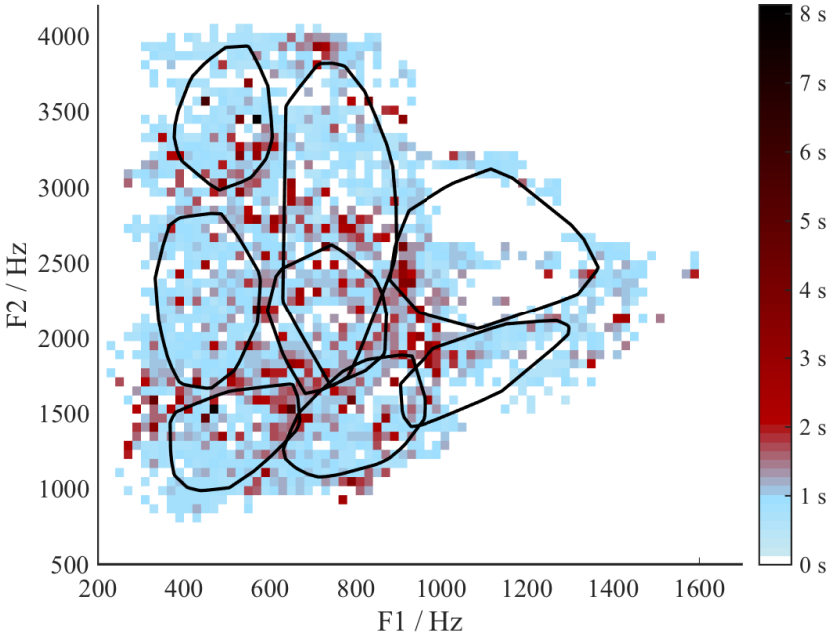

Figure 9. The reaction time (in seconds) for annotating a babble in a Finnish vowel category in different parts of the vowel triangle. The red areas, where annotation takes more than one second, are seen to lie approximately between the participants' perceptual vowel categories. Approximate vowel category boundaries are calculated by using the data shown in Figure 6 and marked with black polygons (see text for details).

4.4 Detailed analysis of model components

We also analyzed the importance of the chosen model components towards learning performance. Since it was infeasible to perform the analysis using human subjects for all the model variants, we simulated human imitative behavior by using a k-nearest neighbor (*k*NN) classifier to assign the babbles into vowel categories while still using the real recorded speech of the evaluation set to probe the final performance.

We used a *k*NN classifier ($k = 20$) and 1000 LeVI babbles, manually annotated by the first author, as the training data for the classifier. The 1000 babbles were obtained with the same babbling procedure as used in the experiments described earlier. An analysis showed that the classifier agreed with 69% accuracy with the human categorization decisions on the development and evaluation sets. However, it has to be noted that variability may exist between participants' classifications of the babbles – participants' vowel categories and category boundaries may not be equal for all participants, and different participants may thus have classified a certain babble consistently in different vowel categories. A classifier trained with the annotations of one participant may thus not agree well with the classification of another participant. The most important aspect for training and testing is that the classifier stays consistent during both phases, eventually leading to participant specific imitation performances. When the babbles for participants E1-E4 and D1-D7 were classified by using 999 annotated samples from the same participant and testing with the remaining one, 80.6% agreement was reached. Thus the classifier's participant-specific performance

can be considered rather reliable. When a single classifier is used during LeVI's training and testing, LeVI is trained to share the vowel system known by the classifier, and we can expect results that are close to individual participants' performances.

The basic model introduced in this study ("Basic method DCM" in Figure 10) was compared to four variants, where the basic method was modified by changing one of its components at a time:

Variant 1) The importance of LeVI evaluating LACs online was studied using a method where all LACs have an equal probability to be babbled on each training trial, i.e. $p(x) = 1/N_c$ in equation (A-3), so that LeVI's selection of the babbled sound does not depend on which LACs are activated when listening to CG's speech or how many times each LAC has been previously babbled ("Equal weights").

Variant 2) The importance of LeVI listening to CG's speech in order to bias babbling towards the most activated LACs was studied by comparing the basic model to a variant where the numerators in equation (A-3) were set to 1, i.e., acoustic model activations in CG's speech did not affect learning, but LeVI aimed to have an equal number of babbles in all LACs ("No listening bias"). The difference between variants 1 and 2 is that in variant 2, LeVI actively tried to equalize the number of babbles assigned to LACs, where as in variant 1, the difference in the number of babbles between LACs may have become larger due to uniform sampling.

Variant 3) The importance of using the Dynamic CM method, i.e. recognizing every training word by CG and using the information to weight learning on promising acoustic regions in the CG responses was studied by using an alternative method where $a = 1$ in equation (A-12), i.e. all acoustic information in CG's responses was treated equally in the acoustic model updates ("CM").

Variant 4) Finally, we also study the effect of the articulatory precision component, simulating learning when $c_{acc} = 0$ in equation (A-5), i.e. causing LeVI to have infinite accuracy when reproducing babbles ("Infinite accuracy").

With each model variant the learning for the four participants of the evaluation set was simulated 16 times in order to get corresponding variance measures. Variability between runs is caused by LeVI's random articulatory exploration as well as nondeterministic selection of CG's responses. The vowels in word set #2 (see Section 3.1) were imitated by LeVI and automatically classified for LeVI's imitation performance results every 60 babbles. It can be seen from Figure 10 that using automatic classification of vowels leads to a slightly better imitation performance than when human participants perform the classification subjectively (82.8% vs. 80.1%). This is due to the fact that the automatic classifier is more consistent with its annotations than human subjects who might annotate even the same babble in different categories on different listenings.

We further analyzed model performances by using Wilcoxon rank-sum tests on the simulated data results for 960 babble-imitation pairs. Based on Figure 10, it appears that the variant with an infinite babbling accuracy outperformed the basic model. The learning rate with infinite accuracy appears considerably faster, however after 960 training samples the difference is not significant ($W = 4339$, $p = 0.316$, $r = 0.177$). This variant may be considered as cognitively less plausible because of the evidence for variability in children's articulatory targets for vowels (e.g. Lee, Potamianos & Narayanan, 1999; Chung et al., 2012). From the other variants, the average performance after 960 babbles is the best for the learning model used in this study. The statistical test shows a significant improvement for the final model when compared to "Equal weights" ($W = 4622$, $p = 0.019$, $r = 0.416$). LeVI should thus try

to guide its reproductive babbles so that each LAC would end up having an approximately equal number of productions. If each articulatory target category had the same probability of being produced, variability in the articulatory production of the targets would lead to an uneven distribution of sounds in the perceptual domain as variation in the articulatory realizations have different perceptual consequences in different parts of the articulatory space. Marginal improvement is also noticed when the basic model is compared against the variant "CM" (W = 4550, p = 0.044, r = 0.355). LeVI thus seems to benefit slightly from hypothesizing the location of the babbled vowel in CG's imitative utterance. Comparison with "No heard bias" is not significant (W = 4272, p = 0.494, r = 0.121), and using the current learning model it seems that LeVI does not benefit from listening to CG's speech and guiding its babbling towards the most activated LACs. We hypothesize that optimizing the weights in equation (A-3) or improving LeVI's recognition of CG's speech's vocalic parts in the listening phase (see section A.3) might make a difference in the learning rate.

Note that the statistical tests performed after the last training sample do not necessarily indicate the methods' differences in the learning *rate*, that appear as consistent differences between the curves in Figure 10. In addition, the differences in performance may change if more training data were available, as performance of some of the variants has not saturated after 960 iterations. Thus the statistical tests should be taken only as approximate indications of final imitation accuracies after training with the present experimental paradigm. For example the basic method might reach the imitation rate of the "infinite accuracy" method if more training was applied. Articulatory inaccuracies in babbling would in this case only slow down learning.

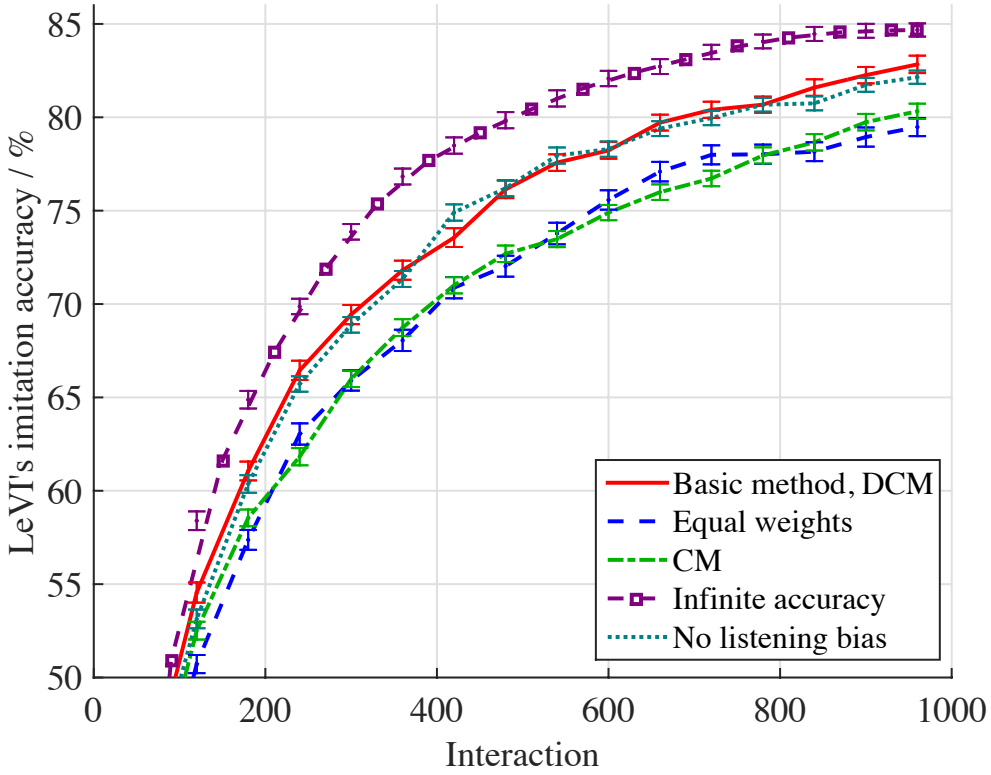

Figure 10. Performance of the final method and its variants in the machine-annotated trials. Mean performance of 16 runs for all four participants is shown with standard error bars.

5 Conclusions

In this paper, we presented a novel computational model of infants' vowel imitation learning. We created a model capable of learning in online interaction with real humans acting as the model's caregivers in realistic and individually ambiguous communicative situations. Previous computational models of vowel learning have used several simplifications, such as synthesized training signals lacking the variability of normal speech, pre-defined sets or numbers of either the infant's or the caregiver's vowel categories, sets of vocal primitives, or bypassing the normalization problem (i.e. infants' and caregivers' vocal tract morphologies and thus acoustic productions differ, and direct comparison of their vowels' acoustic characteristics is not always possible), some of which we have attempted to relax.

Our study builds upon the findings of the Asada group (e.g. Miura et al., 2008) and Howard and Messum (2011, 2014), expanding towards the ability to cope with more natural caregiver-child interactions and the ambiguities involved in them. Our learning virtual infant, LeVI, equipped with an infant sized vocal tract, started with only knowledge of the ranges of its own articulatory parameters and then heard an amount of continuous speech from its caregiver in response to its own babbles. We tied articulatory exploration (present in Howard & Messum, 2014, but not in Miura et al., 2008) and ambiguity in caregivers' responses to babbles (present in Miura et al., 2008; In Howard & Messum, 2014 94% of responses were reformulations) together in an online vowel learning model, resulting in 70–80% vowel imitation accuracy for the eight Finnish vowels present in CVCV words spoken by human caregivers. The imitation accuracy was evaluated by classifying the imitations into Finnish vowel categories by the same participants that acted as LeVI's caregivers, without knowing the original word or vowels that the learner aimed to imitate. Also, none of the imitations or vocalic productions by LeVI were ever *confirmed* or accepted to be good productions by human participants – the learning was solely based on associative learning without strong reinforcement signals (e.g., in the work of Howard & Messum, 2014, participants were asked to accept or reject the proposed imitations). Robust vowel imitation learning thus seems to be possible with similar general associative learning techniques that are assumed to play a central role in infant word learning under the name of cross-situational learning (e.g., Smith & Yu, 2008; see also Räsänen & Rasilo, 2015, for an overview), and we expect that positive/negative type reinforcement signals would bias the learner towards even better vowel representations (see Goldstein & Schwade, 2008).

In the current study, the infant's vocal exploration and interaction with the caregiver were intertwined in one learning phase. Learning is significantly faster when the infant aims to associate equally many caregiver's imitative responses in all of its explored sound categories. LeVI also tracks the activations of LACs related to caregiver's speech, and guides its own babbling behavior towards the most activated LACs, although detailed analysis revealed that this did not improve the model performance. To our knowledge, our model is the first fully online vowel imitation learning model, where vocal exploration and interaction are intertwined, evaluated using real speech signals of human caregivers.

When evaluated after 1000 babble-response pairs for four Finnish participants with good quality recordings, the model reached an average of 80.1% weighted and 75.5% unweighted imitation accuracy across the vowel categories. For the seven participants of the development set whose word recordings contained additional noise due to a recording issue, the corresponding results were 73.0% and 68.4%, respectively.

With our analysis of separate model components we showed that it is beneficial for LeVI's learning to recognize each of its caregiver's imitative responses with current LACs and to update the acoustic models more strongly on those parts of speech, where the babbled model is the most activated one. This can be thought of as a form of self-supervised learning: the supervision signal used to time-align the heard speech with the babbled model is created by LeVI itself based on previous learning. A similar mechanism may presumably be used later to update LeVI's recognizers even when no babbles are produced – when caregiver's speech is heard and recognized, LeVI may keep on updating its acoustic models as soon as a decision of their locations in the heard speech can be reliably made.

5.1 Relation to behavioral research on language acquisition

For practical reasons concerning the effort for the human participants, we wanted to show vowel learning ability in a rather short time-scale when compared to real learning situations. Previous experimental studies shed some light on the issue of how much imitative interaction is available for real speech learning infants. In a study by Kokkinaki and Vitalaki (2013), about 3.35 imitative interactions were observed in 10 minute sessions with the infants, out of which about 72.5% were initialized by the infant and imitated by the mother (averages of Group 1 and Group 2 of infant-mother pairs). For example Molemans (2011) reports that 12 month-old infants produce on average 226 (SD = 62) utterances per 20 minutes when the children were vocally active. Hsu, Foger and Messinger (2001) report babbling frequencies of about 1.3 vocalizations per minute at 6 months and Chen, Lee and Kuo (2011) about 2.3 vocalizations per minute at 12 months of age. A rough estimate from these studies is then that infants would babble about five times a minute, on average, out of which 0.335 babbles are related to imitative interactions, of which 72.5% are imitated by the parent, i.e. only about five percent of all infant vocalizations are imitated by their caregivers. Simply associating all babbled vocalizations in all subsequently heard speech from the parents would probably not lead to sufficient consistency for learning, but other cues such as pitch patterns and comparing some more general acoustic features might help the infant to extract the corresponding vowels or syllables from caregiver responses (see, e.g., Hörnstein, Gustavsson, Santos-Victor & Lacerda, 2008). However, if the imitation detection does not function perfectly, a cross-situational learning mechanism (c.f., our weakly supervised acoustic model) guarantees vowel learning under uncertainty in response alignments. The exact noise tolerance of the learning model was not measured in this work, but in principle, higher probability of co-occurrence of the same phonetic content in babbles and imitative responses, as well as larger diversity concerning all other phonemes in the responses should facilitate learning (see, e.g., Kachergis, Yu & Shiffrin, 2009, for factors that affect cross-situational learning of words).

If we make another rough calculation, estimating that infants would be involved with active interaction with their parents about two hours a day, having 0.335×0.725 imitations by a parent per minute on average, interaction would result in approximately 29 imitative parental responses per day and approximately 11,000 imitative responses in a year. If the infants were able to detect parental imitation reliably enough, it would be plausible that the total number of imitative interactions would suffice to learn vowel imitation following a similar learning strategy to the one proposed in this study.

While it is clear that parents' interaction with infants' vocalizations does affect vocal learning (e.g. Goldstein & Schwade, 2008; Goldstein et al., 2003) and that parents provide imitative feedback to infants' babbles (e.g. Kokkinaki & Kugiumutzakis, 2000; Vigil et al., 2005; Kokkinaki & Vitalaki, 2013), it is more difficult to study whether imitative feedback is really a *necessary* condition for vocal learning. In a study by Schiff (1979, as cited in Sachs, Bard & Johnson, 1981) children with deaf parents (but using both oral and sign language) were found to have normal language development when they interacted at least 5 hours per week with hearing speakers. When comparing two-year old children with normal language development and children with language delay, Vigil et al. (2005) found that the parents of normally developing children responded more to their children's vocalizations and used significantly more expansions to the children's utterances when compared to the parents of language-delayed children.

In some rare cases infants have been raised with minimal amount of vocal interaction with other people. One such child, "Jim" (Sachs, Bard & Johnson, 1981), was raised by two deaf parents and had little interaction with hearing adults, but was exposed to English language through watching TV. At the age of three years and 9 months of age Jim had a "*severe articulation problem with some utterances being unintelligible*" (p.39), but he learned language fast after the conversation sessions with an adult started. Although anecdotal, the example at least shows the importance of dynamic interaction between the learner and a proficient language user in the acquisition of speech. In general, based on the above studies, it seems that even a relatively small amount of interaction with hearing people can lead to normal language development whereas complete lack of interaction seems to lead to severe articulation and development problems.

5.2 Limitations of the proposed learning model

One of the limitations of our proposed model is that we assume that infants' auditory domain for their own speech is represented in terms the first two formant frequencies of their babbles. This translates the concept of an "auditory perceptual distance" to a conceptually simple Euclidean distance in the F1-F2 space, but is only a proxy for some underlying perceptual representation for different speech sounds that infants are actually using. We also have experimented with using more general MFCC-vectors of the babbled sounds to cluster LeVI's productions, but it appears that creation of hyperspherical LACs in the multidimensional MFCC domain results in less "pure" native sound categories when they are evaluated by humans, likely because vowel categories in the MFCC-space are not hyperspherical–not even after any simple normalization technique such as mean and variance normalization. The creation of MFCC-based LACs that would get consistent vowel classifications by participants would therefore require more sophisticated clustering methods (e.g., Gaussian mixture models) with a parameter estimation technique that would find the correct shape (covariance matrix) of the auditory clusters in the high-dimensional feature space, a difficult problem without any additional constraints or cues to the vowel categories. Another limitation is that, even though we include ambiguity in all caregivers' imitative responses, this far the model performance has been tested with relatively little ambiguity when compared to real infants' learning environments.

Also, even though LeVI is not provided with a predefined set or number of vowel categories (or LACs), the threshold parameter for category creation (See

Appendix A.1) affects the amount of LACs found. Using a large threshold value leads to a smaller number of LACs covering larger areas of the acoustic space. Using too large a threshold may prevent LeVI from learning a native distinction between some adjacent vowels (e.g. /e/, /i/), since the babbles from both regions may end up in the same LAC. Lowering the threshold increases the number of LACs and acoustic resolution, but also requires more babble-response pairs in total in order to learn robust acoustic models. In general, lower threshold values should be preferred towards larger values, since a large number of LACs does not necessarily degrade imitation performance – LeVI's babbles that do not fall in regions of typical native vowels will end up getting more ambiguous responses and are thus less likely to be used in imitation. Again, this should be seen primarily as a technical consideration since it is unlikely that infants would actually categorize speech input in terms of discrete and disjoint cluster identities. Instead, they are more likely to have a perceptual mechanism that has adapted to the distributional properties of the speech input and where different speech sounds have different but potentially overlapping distributed representations in the brain. The present categorization method was chosen for its reasonable computational cost and because it provides a conceptually simple way to describe how the learner might differentially represent similar versus different sounds and articulatory configurations and use them as a basis for the associations between the two (see Section 5.3. for suggestions for more sophisticated acoustic perceptual organization methods).

In this study, we have this far only discussed the learning of native vowels. Learning of consonants was left out of this study due to a number of practical computational reasons. First, production of realistic consonant sounds with our articulatory model requires the use of an additional set of dynamic parameters controlling the movements of the articulators. In earlier synthesis experiments it was noticed that the parameters controlling the temporal dynamics of the articulation (e.g., closure duration and release time, voice onset time etc.) have an important effect on the perceptual quality of the consonant sounds. Also the values given for these dynamic parameters are dependent on the articulator, typically articulators with higher masses moving more slowly (see Rasilo, 2012 for a detailed description). If LeVI was to explore its whole range of possible articulations, consisting of combinations of positions and dynamic properties, the dimensionality of the articulatory space would vastly increase, and LeVI's task of exploring proper articulations would become nearly impossible. Since our articulatory model is a rough approximation, lacking real physical constraints related to muscle masses or possible synchronies between articulators, the real-world task of infants' articulatory exploration might be significantly easier – with a natural vocal tract it should be easier to find speech sounds that sound natural to the caregivers. Also using non-vocalic positive feedback on babbles of good quality might reinforce infants towards realistic babbles (see Goldstein & Schwade, 2008, as well as the simulations in Rasilo et al., 2013).

Another challenge when learning consonants concerns the acoustic model used for the speech sounds. Whereas vowel sounds are usually characterized by relatively long lasting and steady acoustic properties, stop consonants, for instance, are short in duration and characterized by rapid changes in the acoustic domain. Since the current weakly supervised approach uses vector quantization of the MFCC features for caregiver speech, the quality of the resulting representation is not detailed enough to provide detailed differentiation of different consonant sounds. Although good performance for weakly supervised learning of synthetic consonants was shown in Rasilo et al. (2013), the consonants in real speech have proven to be much more

problematic due to greater variation in their acoustic characteristics than when working with synthesized signals. In order to model the acoustic effects of real spoken consonant sounds, we should presumably work with shorter time windows as well as VQ-codebooks biased towards the subtle changes occurring during consonant sounds.

If the exploration and acoustic details of consonant recognition were refined, we hypothesize that a similar associative learning account may also underlie consonant learning. When babbling canonically, infants' babbles consist of both a consonant and a vowel sound, and may invoke imitations by the caregiver including both heard phonemes. The proposed associative learning mechanism may update several associations on a single babble-response pair, and thus the total number of babbles needed to learn acoustic models for consonants and vowels would not necessarily increase.

Even though the simulations of the present study evaluate the feasibility of the associative learning paradigm in vowel learning, we do not claim that infants learn vowels as independent and discrete speech units during the early stages of speech acquisition. Since babbling generally consists of consonant-vowel combinations or more dynamic vocalizations than stable vowel sounds, it is possible that infants associate these dynamic constructions to imitative responses. Such learning of more holistic structures could explain the acquisition of e.g. progressive phonological idioms (see Messum & Howard, 2015). In reality, associative learning is likely to operate on multiple representational levels. It is possible that associations are learned between speech patterns of various lengths and their corresponding motor commands. However, in order for this kind of multi-level associative learning to work, the learning mechanism still needs to be able to account for the ambiguity in imitation-response pairs, such as exemplified by this study.

5.3 Summary and discussion of the model's implications

We hypothesized that the learning infant would benefit from listening to the caregivers' speech and weighting its reproductive babbles towards activated acoustic perceptual vowel categories. During the training phase, LeVI weights its reproductive babbles towards LACs that are activated often when listening to several utterances in CG's speech (see Appendix A.3). However, closer analysis revealed that this component did not increase the final learning performance. Based on these experiments, it would thus seem that LeVI should only focus on its own babbles – exploring the articulatory and acoustic spaces – rather than use CG's speech to guide babbling. CG's responses are merely used for acquiring (noisy) correspondence information for the babbled sound. However, this effect may also be an artifact caused by the methodology used. During the training phase, LeVI does not aim to imitate CG's individual utterances. When CG's utterances are imitated in the testing phase, we observed that the imitations are biased towards native language categories (see Figure 8). Consequently, weighting LeVI's babbles towards imitations of CG's individual utterances slowly over time would presumably lead to LeVI's babbling to shift towards more native-language like vowels, as well as speed up learning as less time would be spent on productions that do not seem to frequently appear in CG's speech. More research is needed to explore this idea further.

Because of the ambiguity in caregivers' responses and inaccuracies in infants' babbles, in order for cross-situational learning to work it is necessary for the infant to group its babbles somehow into distinct representations ("categories") (see section 2.2

for explanation). In order to have a maximal correspondence (and thus a maximal cross-situational learning rate) between the infant's interpretation of its own vowels and the caregiver's interpretation of the infant's vowels, it is beneficial that the caregiver and the infant interpret vowel sounds in the same domain and based on the same features. Since caregivers interpret infants' babbles in the acoustic domain (articulations are not visible), it appears that it is beneficial for the infant to perform perceptual grouping of its babbles in the acoustic domain as well (as opposed to categorization in the articulatory domain, see section 2.2) and use that grouping as the basis for further babbling. It is beneficial to explore new acoustic regions and aim at repeating previous acoustic productions, because acoustic characteristics are directly available to the caregiver.

In general, we found that the following components to be critical in achieving successful imitation capability if no innate knowledge is available:

1) The learner has the capability to babble sounds via articulatory exploration.
2) The learner is capable to perform some type of categorization of its own vocalizations so that sets of similar babbles can be associated with sets of caregiver responses (the entities analyzed during the associative learning).
3) There is contingent (linguistically aligned) caregiver feedback following babbling, i.e., responses in which certain adult speech sounds occur more frequently after certain type of babbling. This can range from strict speech sound imitation to above-chance recurrence of the same sounds in the context of larger words or phrases.
4) The learner has a statistical associative learning mechanism that can keep track of the distributional characteristics of the babble-response pairs, thereby learning the sound mapping between the learner and the caregiver.

In addition to these basic principles, our model incorporates several technical solutions that were found useful in practical implementation of the model, and that could be improved even further in the future work:

1) It may be beneficial for the learner to aim to explore the full range of its possible *acoustic* productions, in order to find vowel sounds that may lie in the borders of the full vocal range and possibly have a small region in the articulatory domain for their production. We found that for example finding a pure Finnish vowel sound /i/ by uniformly random articulatory exploration was slow because corresponding articulatory configurations were limited to a very small proportion of the complete articulatory domain, where several articulators take extreme values from their corresponding ranges. Since there is an initially unknown relation between articulatory configurations and acoustic outputs, the learner cannot directly explore the *acoustic* space. In this study we have sped up acoustic exploration by giving the learner a bias to try out articulatory configurations that lie far away from previously produced ones. This is seen to lead to faster exploration of the acoustic space.

   It is currently unclear how infants actually succeed in vocal exploration, but due to a number of different biomechanical constraints and covariances between different articulatory movements (not implemented in the present model), it is likely that the space of possible articulatory gestures is constrained enough to be explored within a reasonable time in real life. In addition, factors such as novelty/reward signals from somatosensory feedback

may bias articulatory gestures towards extreme values of the articulator positions – regions where small articulatory changes are typically associated with large acoustic changes. Finally, real canonical or variegated babbling is dynamic instead of static. Instead of resulting in one articulatory configuration and "one sound" for each babble, there is actually a trajectory of articulatory positions and corresponding auditory outputs for each voiced sound segment between the surrounding consonants in CVCV-type babbling. This may speed up exploration significantly as long as the infants are capable of tracking the articulatory-to-sound correspondences along different sections of these trajectories.

2) The current model categorizes LeVI's productions into discrete and disjoint categories using the Euclidean distance between the babbled vowel sounds in the F1-F2 domain. Although sufficient for proof-of-concept, future work could consider more sophisticated clustering methods together with more general acoustic features such as Mel-spectrum. However, there are some challenges related to the use of non-formant acoustic features since the sound categories tend to be more spherical in the F1-F2 domain than in, e.g., MFCC domain. If Gaussian mixtures are used to define vowel categories, a good prior on feature covariances or a proper way of estimating covariance from the data is needed. Estimation of a covariance matrix for a multidimensional acoustic category requires multiple samples from the category, but in the learning scenario the identity of the category of a produced babble is not known to the infant. One option would be to first store a number of babble-response pairs before attempting to learn a mixture model for the data in a batch or mini-batch-like mode. Alternatively, more advanced methods such as on-line variants of non-parametric Bayesian mixture models (e.g., Hoffman, Blei, Wang & Paisley, 2013) could be used to solve the auditory clustering task with various feature types without imposing definite constraints on the shape and size of the clusters.

   Also, additional constraints on the clustering could be utilized besides purely auditory cues. For instance, the acoustic variability of speech sounds in caregiver responses could be reduced by using contextual information. For example, if the infant babbles a different variant of a sound resembling the phoneme /ɒ/ on two different occasions and the caregiver responses both with the word "dog" where the vowel varies slightly on the two occasions, the infant can use the referential information to deduce that both of the caregiver's phoneme variants belong to the same category and perhaps both babbled sounds have the same category identity as well. This aspect has not been considered in the current study but should be explored in further research.

3) In our simulations, participants respond to LeVI's vocalic babbles with CVCV utterances containing two different vowel sounds, one matching with the participant's interpretation of the babble. We showed that ambiguity in the caregivers' responses can be reduced by predicting the acoustic outcome in the response after babbling, and associating the predicted portion of the response to the babbled sound with increased weight. This is seen to speed up imitation learning. In this study this is implemented by recognizing caregivers' responses using LeVI's current category representations. If the babbled vowel category is recognized in the response, the category's recognizer parameters are updated with more weight at the location identified.

4) We have used an associative learning algorithm that accumulates statistics of acoustic features occurring in caregivers' responses into a category that corresponds to the acoustic characteristics of LeVI's babbles. Future work should consider associative learning of consonants as well. The CM algorithm used in this study has not been successfully used in learning acoustic models for consonants due to rapid changes in the consonant acoustics. The generic MFCCs extracted from 25-ms time windows and vector quantized to discrete categories lose information that would be important for consonant discrimination. The use of phoneme-specific Hidden Markov Models (HMMs) with continuous observation vectors would probably bring additional power in order to discriminate between consonant categories, but training standard generative HMMs is challenging in a weakly supervised manner where only a subset of the data is representative of the class to be learned. Ideally, there would be at least a small number of representative samples to initialize the training (see e.g. Sun, Van hamme & Zhang, 2014).

In order to also learn consonants, a more realistic vocal tract model would presumably facilitate learning. Implementing realistic constraints and dynamics on articulatory movements, possible correlations between articulators as well as sophisticated articulatory-to-acoustic synthesis would lead to realistic babbling patterns and presumably facilitate the exploration effort by the infant as well as the evaluation by participants.

Finally, our model provides a number of predictions regarding child language learning. First of all, the present work suggests that the development of speech production is tightly coupled to the amount of contingent feedback from the parents, higher levels of phonetic and lexical alignment in parental responses leading to faster learning. However, parents do not have to imitate their children exactly as associative statistical learning can overcome ambiguities in the parental responses through accumulation of cross-situational evidence. This also means that less consistent parental alignment can be compensated with larger amount of interaction experience, making the mechanism robust to variation in infant-caregiver dynamics. Expansions, rather than exact imitations, may play a role in learning about the acoustic variation of phonemes in different contexts and thereby improve the robustness of their perception.

In addition, the present study suggests that learners can acquire several different articulatory configurations to produce the same speech sounds, as long as the alternatives are accepted as valid speech sounds by the caregivers in the given communicative context. The model also predicts that infants first learn to imitate speech sounds that are part of their own babbling repertoire. Thus, the infant cannot imitate adult speech sounds whose infant counterparts it has never produced before. However, if the infant were able to gradually align its own acoustic productions to adult productions, as well as its articulatory configurations to their acoustic counterparts (i.e. learn in more and more detail about the non-linear articulatory-to-acoustic mapping based on increasing babbling experience), it might be able to interpolate between known productions and hypothesize likely ways to imitate sounds that it has not previously produced. If this is the case, interpolation should be easier in close proximity to known productions or in articulatory regions where the articulatory-acoustic mapping is more linear. Such a gradual alignment process has not been implemented in this study and is left for future research.

A third prediction is that the vocal exploration phase and imitation learning do not have to take place sequentially, but that the awareness of the triad consisting of articulatory gestures, acoustic sounds corresponding to these gestures, and adult sounds corresponding to the same gestures and sounds can start to develop jointly as soon as the infant becomes engaged in babbling in social contexts.

5.4 Concluding remarks

With this computational study of vowel imitation learning we would especially like to direct modelers' attention to the non-ideal learning conditions that human infants face in the same task. Speech learners have to cope with large amounts of variation in the caregivers' utterances as well as their own articulatory productions. When associating vocalic productions to caregivers' responses, we have shown that weakly supervised associative learning may bring robustness to acoustic variation in the responses (a similar technique was used in Miura et al., 2008). However, the productions themselves may have variation due to inaccuracy in articulation, or alternative articulations for similar acoustic sounds may be found during vocal exploration. We have dealt with these problems by allowing the infant to create auditory categories based on the acoustic outcomes of babbles (in Miura et al., 2008, pre-defined vowel categories were used). If vocal exploration is also included in the learning process, effective exploration techniques are needed so that both exploration and associative learning are possible with a limited amount of interaction. We have introduced a method that helps to explore the acoustic space of the infant's vocalizations faster than uniform sampling from the infant's articulatory domain.

It may be difficult to find out if learning components similar to those suggested here are present in real human speech learning, but the vowel learning framework presented aims to restrict all the information used by LeVI to information that a real human infant would have an access to as well. As well as some previous studies (Huckvale & Sharma, 2013; Howard & Messum, 2014), we show that vowel learning is possible without learning a normalizing mapping between the caregiver's and the infant's voices – the learning occurs based on a general associative learning mechanism that simply associates speech features of the caregiver's speech with vocal productions by the infant.

6 Acknowledgements

This research was funded by the Academy of Finland project titled "Computational Modeling of Language Acquisition", the ERC starting grant project ABACUS, grant number 283435, and the Finnish Cultural Foundation. The authors would like to thank Bart de Boer and Unto K. Laine for helpful discussions as well as Bill Thompson and Hannah Little for general comments on the manuscript.

Appendix A. Model implementation details

A.1. Creation of LeVI auditory categories

LeVI creates perceptual auditory categories, *LACs*, in the acoustic domain based on the two first formant frequencies of the babbled sounds. The formants are extracted using linear predictive coding (LPC) with autocorrelation method on the impulse response of LeVI's vocal tract. The impulse response is first resampled to 12000 Hz sampling frequency and LPC of order 10 is used. By taking into account the approximate maximal values of F1 and F2 produced by LeVI (see Figure A-1), we normalize all formant vectors by dividing F1 values by 1500 and F2 values by 4500, in order to form approximately spherical LACs. The normalized formant values form a formant vector $\mathbf{f}_n$.

The Euclidean distance of the babbled $\mathbf{f}_n$ is calculated to the normalized formant values of all N_c centroids of the existing LACs, $\mathbf{f}_n^c$.

$$D(c) = d(\mathbf{f}_n, \mathbf{f}_n^c), \quad c = 1 \ldots N_c \tag{A-1}$$

If for any centroid, D falls below a preset threshold of 0.10, the new babble is assigned to the LAC given by $\arg\min_c(D(c))$, otherwise a new LAC is created at location $\mathbf{f}_n$.

A.2. Vocal exploration

When LeVI babbles an exploratory babble, it aims to extend the acoustic space covered by its previous acoustic productions by babbling an articulatory configuration far from already babbled configurations. The articulatory space is divided into a grid of articulatory parameter combinations. We define 25 sampling points from the two-dimensional regions for the tongue tip and tongue body parameters, and five points uniformly for jaw angle, hyoid x-coordinate, lip protrusion and lip opening parameters (see Rasilo, 2012). Velum is sampled either fully open or fully closed. After rejecting configurations that have a vocal tract constriction smaller than 0.1 cm^2 in cross-sectional area, we are left with 238,080 possible articulatory configurations in a set **P**, which LeVI uses as a base for its exploration. In order to illustrate the F1-F2 range of the productions, Figure A-1 shows the acoustic outputs of all the sampled parameters in the F1-F2 domain.

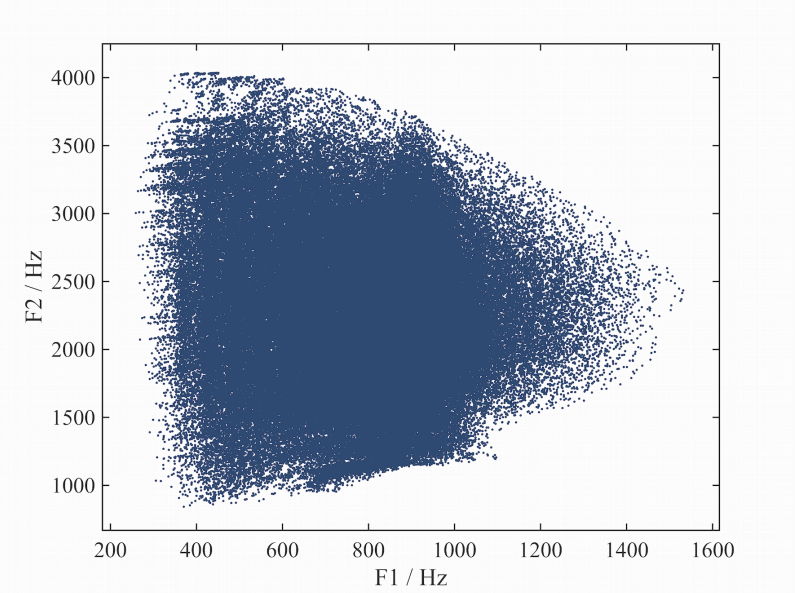

Figure A-1. Formant values of all the sampled vocal tract configurations.

When LeVI decides to explore and create a new babble, it compares the N articulatory parameter vectors it has already produced to a subset of **P**. LeVI chooses to babble an open configuration from an articulatory region that has a small density of neighboring articulatory parameters stored during previous babbles. In practice, we calculate the inverse squared distance, or "gravity" G of all N articulatory configurations babbled this far (known articulatory vectors $\mathbf{k}=[\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_n]$) to a random subset of 20,000 possible configurations (**S**) drawn randomly from the set **P** (the whole set is not used for the sake of faster computation in the following distance calculation), by using a formula

$$G(s) = \sum_{n=1}^{N} \frac{1}{d(\mathbf{k}_n, \mathbf{S}_s)^2}, \quad s = 1 \ldots 20{,}000 \tag{A-1}$$

where d is the Euclidean distance between articulatory vectors. The selected exploratory babble is chosen to be the open configuration with the least gravity from all known configurations $\mathbf{b} = \mathbf{S}_{\text{argmin}(G)}$.

Performance of the selective sampling method, when compared to uniform sampling from all possible open configurations, is illustrated in Figure A-2. On this run, only for illustrative reasons, LeVI samples a new exploratory configuration at 200 consecutive babbles, and the process is run 10 times in total. On each run, the area covered by the convex hull of the F1-F2 values of the babbles produced this far is compared to the area covered by the convex hull of the F1-F2 values of the set **P**. The selective sampling method has constantly a larger proportion of the acoustic domain covered, indicating that the selective sampling method produces articulatory vectors that lead into more spread out acoustic outputs than the uniform sampling. Taking into account the fact that in many languages important vowel categories lie in the borders of the acoustic space, the selective sampling thus helps LeVI to find meaningful vowel categories – as interpreted by human listeners – better than just sampling uniformly from its articulatory range.

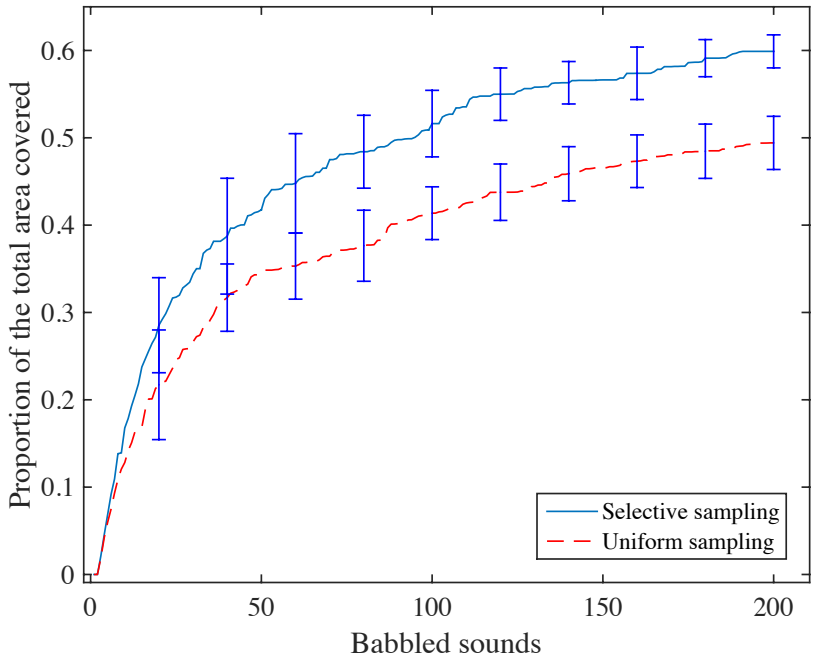

Figure A-2. Mean and standard deviation of the proportion of the acoustic area (in F1-F2 domain) covered by the discovered babbles towards the total area that can be produced with the articulatory model. 10 runs where LeVI produces up to 200 exploratory babbles are averaged.

A.3. Repeating existing LACs

As explained in Section 2.4., LACs that fall inside CG's perceptual categories, and thus end up getting consistent feedback from CG, end up having acoustic models that are more refined towards the corresponding CG's vowel category. When recognizing CG's speech, these models thus provide higher likelihoods than the more noisy recognizers that fall in between CG's perceptual categories. In order to try to bias LeVI's babble towards the pure LACs we use the following procedure: Every 20 interactions, LeVI listens to and recognizes a number (here 80) of spoken words by CG (randomly drawn from the training set, defined in section 3.1), detects local energy envelope maxima[4] with a minimum distance of 40 windows (200 ms) between consequent maxima, and sees which LACs have the highest activation at the given locations. From the set of 80 words, LAC activation frequencies are counted in a vector $\mathbf{a} = [a_1, \ a_2, \ \cdots, \ a_{N_c}]$ each element a_x indicating how many times LAC x was activated. However, if some elements of the vector get a value of 0, we increase it to one in order not to completely forget the LAC.

We similarly count the number each LAC has been babbled this far in total into a frequency vector $\mathbf{f} = [f_1, \ f_2, \ \cdots, \ f_{N_c}]$. Since we want LeVI to babble LACs with low frequency more, and LACs with high activation in CG's speech more, we calculate the probability of LeVI choosing to babble LAC x as:

[4] The aim of this technique is to work as a really simple detection mechanism for vowel locations. In further versions, voice activity detection or other more sophisticated methods could be used for improved accuracy.

$$p(x) = \frac{(a_x)^{0.3}}{f_x} \Big/ \sum_n \frac{(a_n)^{0.3}}{f_n} \quad \text{(A-3)}$$

where the latter division normalizes the sum of probabilities to one, and the exponential component of 0.3 reduces the contribution of the activation component relative to the frequency component.

When LeVI chooses LAC *c* to be babbled based on the calculated probabilities, it aims to babble again the mean of the articulatory parameter vectors that are stored in the biggest articulatory cluster related to LAC *c* (see Figure 3).

A.4. Accuracy of LeVI's reproductions

When LeVI creates a reproductive babble, we introduce an inaccuracy term so that LeVI cannot reach the intended articulation exactly but with certain accuracy depending on babbling experience on the related articulatory region. Note that in the reproduction phase babbled articulatory vectors are not anymore tied to the previously simulated open configurations (set **P** in Appendix A.1.) but can have arbitrary values from their related ranges.

In our rough model of the effect of articulatory experience, LeVI's babbling accuracy in the articulatory domain depends on the number of babbles it has previously produced in a limited hyperspherical region surrounding its intended articulatory vector **i** = [i_1, i_2, … . i_9]. In this work, we calculate the number of previously produced babbles, $N_{\text{neighbours}}$ that are on a distance of less than 0.2 units from the intended target, when calculated as an Euclidean distance between the 9-dimensional articulatory vectors, whose each element is normalized linearly to lie in the range [0,1] (using the allowed ranges explained in Rasilo, 2012). An articulatory accuracy coefficient c_{acc} is then calculated as

$$c_{acc} = 10^{-\frac{N_{neighbours}-1}{20}} \quad \text{(A-4)}$$

and the final babbled articulatory vector **b** = [b_1, b_2, … , b] after implementing the inaccuracy is

$$b_x = i_x + U\left[-0.1, 0.1\right] \cdot c_{acc} \quad for\ all, x = 1...9 \quad \text{(A-5)}$$

where U[-0.1, 0.1] is uniform random noise drawn from the given range. Here, because of the added noise, the final articulatory configuration is not guaranteed to be open, or the articulatory parameter values to lie on allowed ranges. For practical reasons, we round parameter values that lie outside their allowed ranges to the nearest allowed values (for tongue tip and tongue body parameters, to the nearest point in the given polygon borders, see Rasilo, 2012). If the resulting area function has a cross-sectional area of less than 0.1 cm^2 at any tube section, the randomization process (equation A-5) is repeated until the openness condition is met, in order to have an audible, vocalic, output sound.

In order to illustrate the effect of the inaccuracy component in the formant domain, we have given LeVI an intended articulatory target vector in the regions for Finnish vowels /e/ and /i/ and created 20 babbles for both targets with the proposed mechanism. Figure A-3 shows the two first formant values of all the babbled vowels. The color scheme changes from bright red to blue from babble one towards babble 20. It can be seen that the inaccuracy between the target (black crosses) and actual

babbles is rather large during the first babbles, but decreases when the number of babbles increases. Inaccuracy in the babbling process also may help to expand the vowel region of LeVI and possibly to discover new LACs especially in the border regions of the vowel space, where small articulatory inaccuracy may lead to big enough change in the acoustic domain - because of articulatory inaccuracy the acoustic distance threshold for creating new LACs (see Section A.1) may be exceeded.

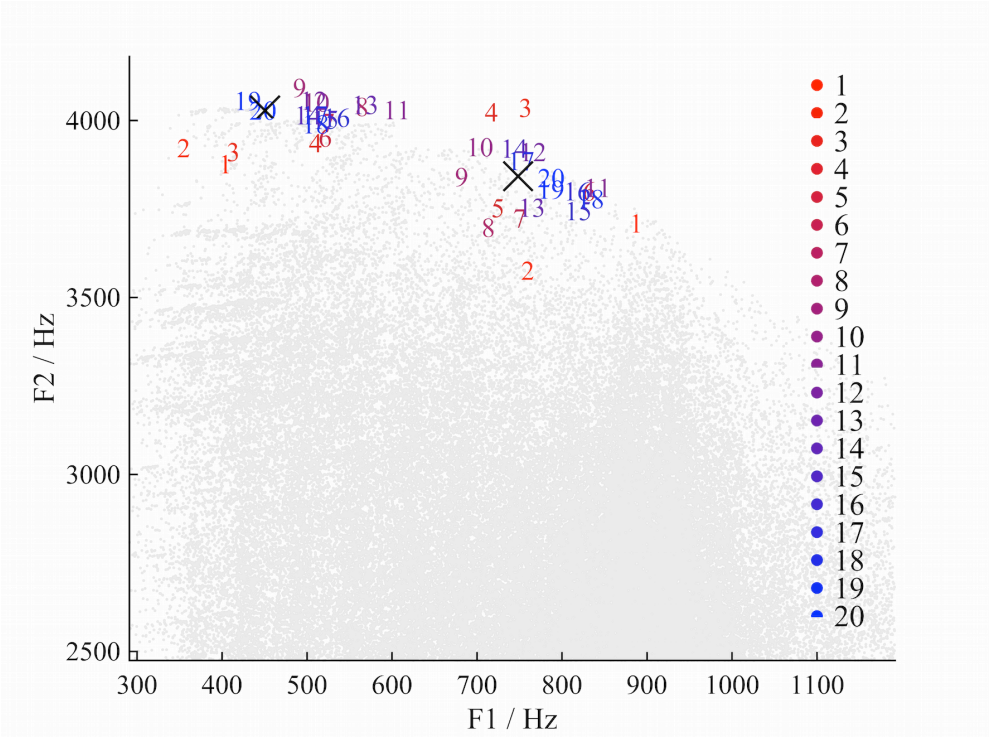

Figure A-3. Illustration of the effect of the articulatory inaccuracy component. LeVI tries to reach articulations underlying the acoustic targets marked as black crosses. When approaching 20 trials, the accuracy of reaching the targets is seen to increase.

If LeVI creates a LAC that is difficult to reproduce, for example by finding an articulatory vector in a highly nonlinear articulatory region where close by babbles end up always a long acoustic distance away, it is better for LeVI to forget it and concentrate on more beneficial articulations. We delete a LAC if LeVI's success rate of reproducing it falls below 10%. Success rate of LAC *c* is measured by the number of intended babbles from *c* that end up in the acoustic region corresponding to *c* divided by the total number of intended babbles for *c*. Note that the numerator is not always equal to the total number of times LAC *c* has been babbled, since also random exploration or noise when trying to reproduce other LACs may lead to an acoustic output in the region of *c*.

A.5. Associating CG's responses to LeVI's babbles

The acoustic characteristics of CG's imitative responses are associated with LACs using a weakly supervised learning algorithm, the Concept Matrix algorithm (Räsänen & Laine, 2012) as a framework. In this study, a dynamic adaptation (DCM) of the algorithm is used.

A.5.1 Pre-processing of participants' speech signals

The CM-algorithm works with Vector Quantized (VQ) speech features, where each spectral slice of the speech signal is represented as an integer number. The speech signals are compressed in an unsupervised manner into sequences of integers from which bi-gram statistics are calculated. For each participant, recorded speech signals are first pre-emphasized with a first order high-pass filter with $\alpha = 0.95$. Then MFCC features are extracted from the complete training speech data in windows of 25 ms of length with a 5 ms overlap. The first MFCC coefficient (spectral tilt) is discarded, and the remaining 11 MFCC coefficients are used as the basis for VQ. Out of all obtained feature vectors, a set of 50,000 is selected and clustered with standard *k*-means algorithm into 150 clusters. Now all training and testing data are transformed into integer values lying between 1 and 150, according to the closest cluster centroid in the obtained codebook.

VQ represents the unsupervised reorganization of the infant's auditory perceptual system to perceive some fundamental units in CG's speech. In practice we thus assume that prior to learning LeVI is exposed to an amount of CG's speech for this reorganization to take place.

A.5.2 The DCM algorithm

Every LAC *c* has an acoustic model that describes the probability of the LAC, given a sequence of acoustic observations. The basis for the model is a tensor $\mathbf{F}^c$ of size $N \times N \times L$, where $N = 150$ is the alphabet size for the discrete observations (VQ-indices) and L is the number of used lags $\boldsymbol{l} = \{l_1, l_2, \ldots, l_L\}$ at which pairs of observations are analyzed. The matrix stores frequencies of transitions between VQ-indices at different lags. When a sequence of VQ-indices, $V = [v_1, \ldots, v_{t-1}, v_t, v_{t+1}, \ldots, v_{t+m}]$, corresponding to a spoken word by CG, is observed after babbling the LAC *c*, the elements in the matrix $\mathbf{F}^c$ are updated on every time instant for every lag as

$$f^c_{v_t, v_{t+l}, l} \leftarrow f^c_{v_t, v_{t+l}, l} + a \tag{A-6}$$

for all values of *l*. Here the values inside the brackets represent matrix elements for notational simplicity. In this work, lags $\boldsymbol{l} = \{-10, -9, \ldots, -1, 1, 2, \ldots, 10\}$ were used. In the basic CM algorithm *a* would always be one, $\mathbf{F}_c$ corresponding to a frequency matrix of transitions so that all information in the input signal would be assigned to the babbled LAC with equal weight. In the DCM method used in this work, the value of *a* depends on the result of recognizing the utterance as described below.

In order to use co-occurrence statistics for recognition, the statistics need to be normalized into conditional probabilities. In contrast to work described in Rasilo & Räsänen (2015), preliminary simulations with vowel learning revealed that conditional probabilities of (lagged) element pairs lead to better performance than the use of transition probability statistics. The joint probability of a pair of VQ-indices observed in the context of production *c* at a given lag *l* becomes

$$P\left(v_i, v_j \middle| l, c\right) \equiv f^c_{v_i, v_j, l} \Big/ \sum_{j=1}^{N} \sum_{i=1}^{N} f^c_{v_i, v_j, l} \tag{A-7}$$

Assuming that all articulatory productions *c* are equally likely, the conditional probability of concept *c* when a given transition is observed becomes:

$$P\left(c \middle| v_i, v_j, l\right) \propto P\left(v_i, v_j \middle| l, c\right) \Big/ \sum\nolimits_c P\left(v_i, v_j \middle| l, c\right) \tag{A-8}$$

When recognizing a given VQ-sequence, an instantaneous activation value for each concept at every time instant is acquired by

$$A(c,t) = \sum_{l \in \mathbf{l}} P\left(c \middle| v_t, v_{t+l}, l\right) \tag{A-9}$$

Finally, the activation sequences are smoothed by summing activations in a sliding window of $N = 20$ time steps (200 ms)

$$A_{smooth}(c,t) = \frac{1}{L_t} \sum_{k=-N/2+1}^{N/2} A(c,t+k) \tag{A-10}$$

where L_t is the total number of lags that could be used on time instant t. L_t may be smaller than L in the beginning and end of the sequence where not all lags can be used.

From the smoothed activation values, we can choose the winning LAC as the most activated one for each time window as

$$winner(t) = \text{argmax}_c(A_{\text{smooth}}(c,t)) \tag{A-11}$$

In this work, when LeVI is to imitate a speech sound produced by CG, LeVI imitates using the articulations related to the winning LAC on the given time instant.

As we use the dynamic version of the CM algorithm in this work, during the training phase, we recognize each CG's imitative response with LACs acquired this far, and see if the babbled category is the winner at any time instant. If the babbled LAC is detected in the utterance, the recognizer for the babbled LAC c is updated more strongly surrounding the winning locations of this model (see Rasilo & Räsänen, 2015 for more details). This helps LeVI's recognizers to converge to the corresponding acoustic transitions more rapidly, and leads to less noisy recognizers in the end of training (see section 4.4 for proof of the effect). Based on this dynamic update procedure, the update term in equation (A-6) becomes

$$a = \begin{cases} 2, & if\ c \in \{winner(t-s), ..., winner(t+s)\} \\ 1, & otherwise \end{cases} \tag{A-12}$$

Where s is a predefined spreading term. In this work we use $s = 10$, meaning that the babbled LAC is updated with double activation at the time instants where the LAC wins as well as up to ten windows before and after the winning time instants.

An example of LACs' smoothed activation curves for a Finnish word "vino", taken from the final experiment for one participant after 1000 babble-response pairs, is shown in Figure A-4. The figure shows the activation value of each of the 23 discovered LACs over time. It can be seen that some LACs (presumably lying in the acoustic regions for /i/ and /o/) have high activation during the vowel sounds in the speech signal.

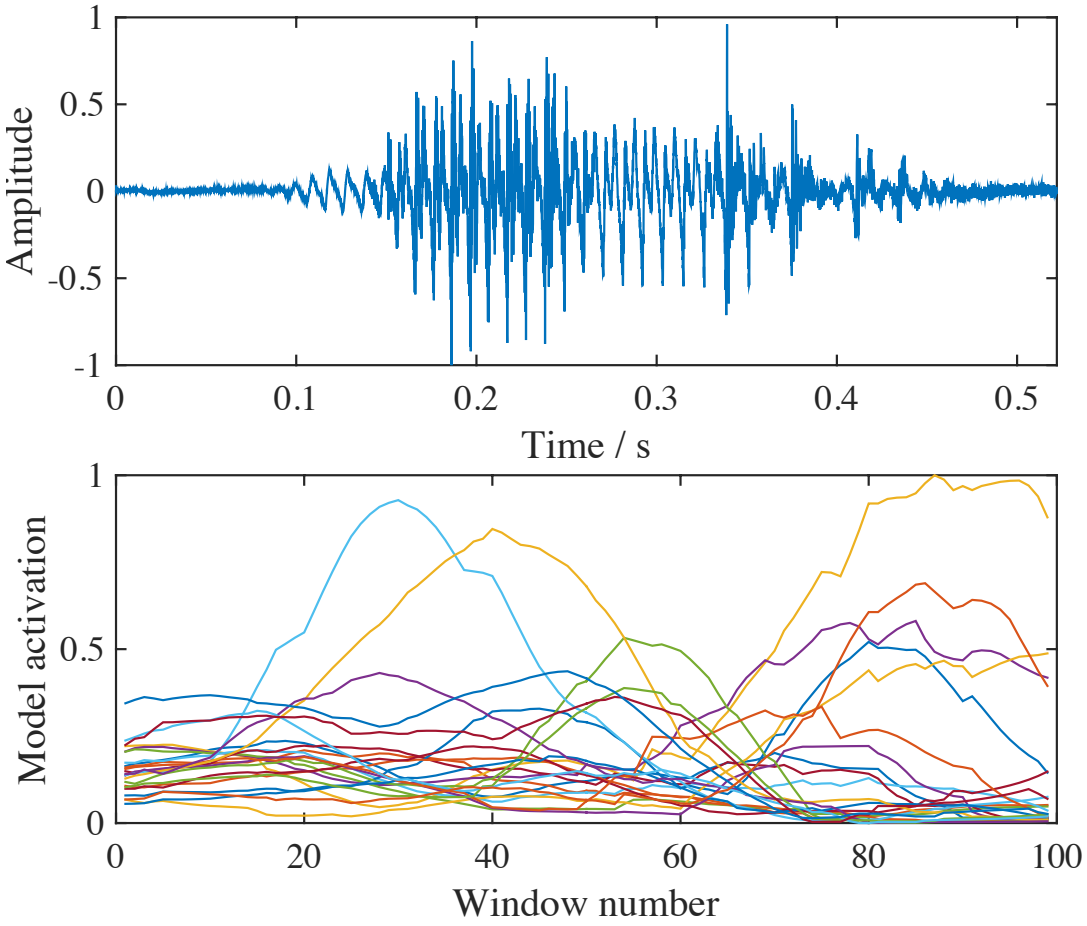

Figure A-4. Original waveform for the Finnish word “vino” (above), and the activations of LACs during the word (below). Each color represents one LAC.

Appendix B. Analysis of errors made by P1

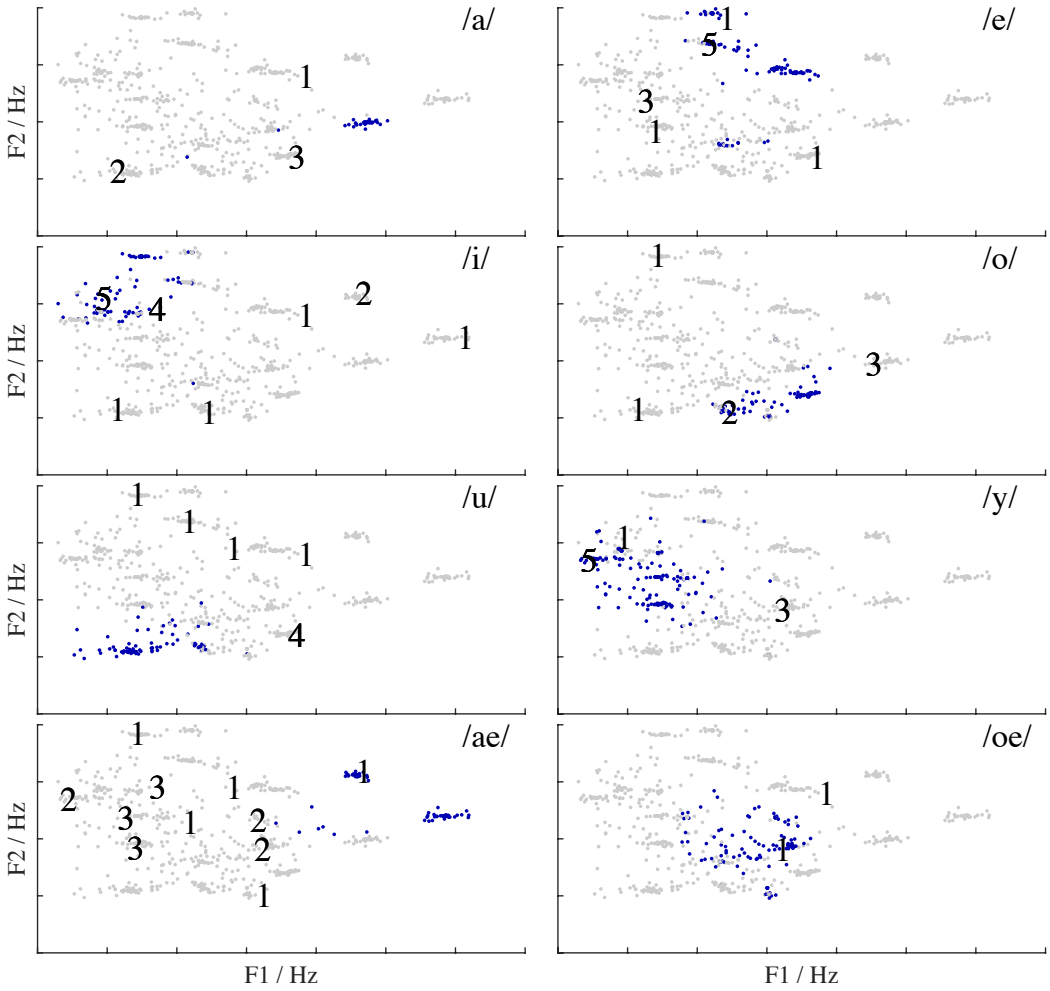

Figure B-1. Erroneous imitations by LeVI located in the F1-F2 domain when evaluated by participant P1. The index of the small image shows the vowel by CG that LeVI intended to imitate. The number shows how many times and in which area the LAC with which LeVI imitated the vowel was located, if CG did not annotate this imitation as the original vowel. It can be seen that errors are often made by imitating with a sound that is acoustically close to the vowel sound to be imitated. The colors show all the babbles in the training phase annotated as the corresponding vowel.

Appendix C. Word lists

Word set #1 (13 tokens of each for training, 1 for testing)		Word set #2 (testing)	
mitä	jyvä	kytö	home
lasi	puhe	koje	menu
näky	haju	meri	täry
kova	pipo	taru	täti
puna	talo	tosi	repo
setä	lisä	toki	hile
halu	käsi	lime	kesy
väri	muki	vilu	pora
lelu	väsy	savu	näre
runo	sade	juro	sulo
melu	sika	räsy	viro
jäte	tipu	kyse	vety
käpy	nenä	tina	köli
jänö	minä	kajo	väli
susi	lohi	muro	köhä
kisu	kani	joku	kumi
kesä	kipu	vino	hiha
vale	side	täpö	pako
kuti	särö	hitu	keto
höpö	piha	karu	pore
tykö	pesu	kela	pöpi
säde	syli	heti	töni
lumi	sinä	lupa	taso
syvä	koti	pyhä	mäti
köhä	hepo	käte	jänö
kone	näkö	vanu	jako
kuje	joki	vähä	sitä
kylä	pöty	väre	kulo
läpi	veli	latu	kota
hyvä	kato	loru	säle
vesi	katu	möly	tykö
pöly	romu	kymi	mökä
levä	söpö	mesi	tavu
levy	mato	väki	tuli
kynä	meno	sysi	hovi
hame	sose	mikä	hyve
peli	hely	kate	peti
kuka	käki	hake	vävy
möly	lepo	tupa	hämy
kuva	juna	risa	jyrä

7 References

Ananthakrishnan, G., & Salvi, G., (2011). *Using Imitation to learn Infant-Adult Acoustic Mappings*. In Interspeech 2011 (765–768), Florence, Italy.

Anisfeld, M. (1996). Only tongue protrusion modeling is matched by neonates. *Developmental Review*, *16*(2), 149-161.

Brass, M., & Heyes, C. (2005). Imitation: is cognitive neuroscience solving the correspondence problem? *Trends in cognitive sciences*, *9*(10), pp. 489-495.

Chen, L. M., Lee, C. C., & Kuo, T. W. (2011). *Measures of early phonetic development: A longitudinal analysis.* In *IEEE International Symposium on Multimedia* (524-529), Laguna Cliffs Marriott Resort & Spa Dana Point, California, USA.

Chung, H., Kong, E. J., Edwards, J., Weismer, G., Fourakis, M., & Hwang, Y. (2012). Cross-linguistic studies of children's and adults' vowel spaces. *The Journal of the Acoustical Society of America*, *131*(1), 442-454.

Goldstein, M. H., King, A. P., & West, M. J. (2003). Social interaction shapes babbling: Testing parallels between birdsong and speech. *Proceedings of the National Academy of Sciences*, *100*(13), pp. 8030-8035.

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, *19*(5), pp. 515-523.

Gros-Louis, J., West, M. J., Goldstein, M. H., & King, A. P. (2006). Mothers provide differential feedback to infants' prelinguistic sounds. *International Journal of Behavioral Development, 30*(6), pp. 509–516.

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological review*, *102*(3), pp. 594-621.

Heyes, C. M., & Ray, E. D. (2000). What is the significance of imitation in animals? *Advances in the Study Behavior, 29,* pp. 215–245.

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic Variational Inference. *Journal of Machine Learning Research*, 14, 1303–1347.

Howard, I. S., & Messum, P. (2011). Modeling the development of pronunciation in infant speech acquisition. *Motor Control*, *15*(1), pp. 85-117.

Howard, I. S., & Messum, P. (2014) Learning to pronounce first words in three languages: an investigation of caregiver and infant behavior using a computational model of an infant. *PLoS ONE 9*(10).

Hsu, H. C., Fogel, A., & Messinger, D. S. (2001). Infant non-distress vocalization during mother-infant face-to-face interaction: Factors associated with quantitative and qualitative differences. *Infant behavior and development*, *24*(1), pp. 107-128.

Huckvale, M. & Sharma, A. (2013). *Learning to imitate adult speech with the KLAIR virtual infant*, In Interspeech 2013 (582-586), Lyon, France.

Hörnstein, J. (2013). *Developmental approach to early language learning in humanoid robots* (Doctoral dissertation). Universidade Técnica de Lisboa, Instituto Superior Técnico.

Hörnstein, J., Gustavsson, L., Santos-Victor, J., & Lacerda, F. (2008). *Modeling speech imitation.* In IROS-2008 Workshop-From motor to interaction learning in robots, Nice, France.

Hörnstein, J., & Santos-Victor, J. (2007). *A unified approach to speech production and recognition based on articulatory motor representations*. In International Conference on Intelligent Robots and Systems (3442-3447), San Diego, California, USA.

Hörnstein, J., Soares, C., Santos-Victor, J., & Bernardino, A. (2007), *Early Speech Development of a Humanoid Robot using Babbling and Lip Tracking*. In Symposium on Language and Robots, Aveiro, Portugal.

Ishihara, H., Yoshikawa, Y., Miura, K., & Asada, M. (2008). Caregiver's sensorimotor magnets lead infant's vowel acquisition through auto mirroring. In *7th IEEE International Conference on Development and Learning* (49-54), Monterey, CA, USA.

Jones, S. S. (2006). Exploration or imitation? The effect of music on 4-week-old infants' tongue protrusions. *Infant Behavior and Development*, *29*(1), 126-130.

Kachergis, G., Yu, C., and Shiffrin, R. M. (2009). *Frequency and contextual diversity effects in cross-situational word learning*. In proceedings of the 31st annual meeting of the cognitive science society (2220–2225).

Kelly, J., & Lochbaum, C. (1962). *Speech Synthesis.* International Congress on Acoustics (1-4), Copenhagen, Denmark.

Kokkinaki, T., & Kugiumutzakis, G. (2000). Basic aspects of vocal imitation in infant-parent interaction during the first 6 months. *Journal of reproductive and infant psychology*, *18*(3), pp. 173-187.

Kokkinaki, T., & Vitalaki, E. (2013). Exploring spontaneous imitation in infancy: A three generation inter-familial study. *Europe's Journal of Psychology*, *9*(2), pp. 259-275.

Kröger, B. J., Kannampuzha, J., & Neuschaefer-Rube, C. (2009). Towards a neurocomputational model of speech production and perception. *Speech Communication*, *51*(9), pp. 793-809.

Kuhl, P. K., & Meltzoff, A. N. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. *The journal of the Acoustical Society of America*, *100*(4), 2425-2438.

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, *105*(3), pp. 1455-1468.

MacWhinney, B. (2000). The childes project: Tools for analyzing talk. 3rd edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Markey, K. L. (1994), *The sensorimotor foundations of phonology: a computational model of early childhood articulatory and phonetic development* (Doctoral thesis), University of Colorado, Boulder, 1994.

Masur, E. F., & Rodemaker, J. E. (1999). Mothers' and infants' spontaneous vocal, verbal, and action imitation during the second year. *Merrill-Palmer Quarterly,* pp. 392-412.

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, *198*(4312), 75-78.

Mermelstein, P. (1973), Articulatory model for the study of speech production, *J. Acoust. Soc. Am*., *53*(4), pp. 1070-1082.

Messum, P., & Howard, I. S. (2015). Creating the cognitive form of phonological units: The speech sound correspondence problem in infancy could be solved by mirrored vocal interactions rather than by imitation. *Journal of Phonetics*, *53*, 125-140.

Mitterer, H., Scharenborg, O., & McQueen, J. M. (2013). Phonological abstraction without phonemes in speech perception. *Cognition*, *129*(2), 356-361.

Miura, K., Yoshikawa, Y., & Asada, M. (2007). Unconscious anchoring in maternal imitation that helps find the correspondence of a caregiver's vowel categories. *Advanced Robotics 21*(13), pp. 1583-1600.

Miura, K., Yoshikawa, Y., & Asada, M. (2008). *Realizing being imitated: Vowel mapping with clearer articulation*. In 7th IEEE International Conference on Development and Learning (262-267), Monterey, CA, USA.

Molemans, I. (2011). *Sounds Like Babbling: A Longitudinal Investigation of Aspects of the Prelexical Speech Repertoire in Young Children Acquiring Dutch: Normally Hearing Children and Hearing-impaired Children with a Cochlear Implant* (Doctoral thesis), Universiteit Antwerpen, Faculteit Letteren en Wijsbegeerte, Departement Taalkunde.

Murakami, M., Kröger, B. J., Birkholz, P., & Triesch, J. (2015). *Seeing [u] aids vocal learning: babbling and imitation of vowels using a 3D vocal tract model,*

reinforcement learning, and reservoir computing. In 5th International Conference on Development and Learning and on Epigenetic Robotics (208-213), Rhode Island, USA.

Oostenbroek, J., Suddendorf, T., Nielsen, M., Redshaw, J., Kennedy-Costantini, S., Davis, J., Clark, S. & Slaughter, V. (2016). Comprehensive longitudinal study challenges the existence of neonatal imitation in humans. *Current Biology*, *26*(10), 1334-1338.

Pawlby, S. (1977). Imitative interaction. In H.R. Schaffer (Ed.), *Studies in mother-infant interaction* (pp. 203–223). London: Academic Press Inc.

Plummer, A. R. (2012), *Aligning manifolds to model the earliest phonological abstraction in infant-caretaker vocal imitation*, In proceedings of Interspeech 2012 (2482-2485), Portland, Oregon, USA.

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), pp. 421-435.

Rasilo, H. (2012) *Articulatory model for synthesizing sequences of arbitrary speech sounds or pre-programmed Finnish phonemes*, work report, available on http://users.spa.aalto.fi/hrasilo/.

Rasilo H., & Räsänen O. (2015) *Weakly-supervised word learning is improved by an active online algorithm*. In proceedings of Interspeech'2015, Dresden, Germany.

Rasilo, H., Räsänen, O., & Laine, U. K. (2013). Feedback and imitation by a caregiver guides a virtual infant to learn native phonemes and the skill of speech inversion. *Speech Communication*, *55*(9), pp. 909-931.

Rousseeuw, P. J., Ruts, I., & Tukey, J. W. (1999). The bagplot: a bivariate boxplot. *The American Statistician*, *53*(4), 382-387.

Räsänen, O. (2012). Computational modeling of phonetic and lexical learning in early language acquisition: existing models and future directions. *Speech Communication*, *54*(9), 975-997.

Räsänen, O., & Laine, U. K. (2012), A method for noise-robust context-aware pattern discovery and recognition from categorical sequences. *Pattern Recognition*, *45*, pp. 606–616.

Räsänen, O. & Rasilo, H. (2015). A joint model of word segmentation and meaning acquisition through cross-situational learning. *Psychological Review*, *122*(4), 792–829

Sachs, J., Bard, B., & Johnson, M. L. (1981). Language learning with restricted input: Case studies of two hearing children of deaf parents. *Applied Psycholinguistics*, *2*(1), 33-54.

Schiff, N. B. (1979). The influence of deviant maternal input on the development of language during the preschool years. *Journal of Speech, Language, and Hearing Research*, *22*(3), 581-603.

Smith, A., & Zelaznik, H. N. (2004). Development of functional synergies for speech motor coordination in childhood and adolescence. *Developmental psychobiology*, *45*(1), pp. 22-33.

Smith, L., & Yu, C. (2008). Infants rapidly learn word- referent mappings via cross-situational statistics. *Cognition*, *106*(3), pp. 1558–1568.

Sun, M., Van hamme & Zhang, X. (2014). *Weakly supervised hmm learning for spoken word acquisition in human computer interaction with little manual effort*. In 12th International Conference on Signal Processing (ICSP), pp. 1341-1345.

Tamis-LeMonda, C. S., Bornstein, M. H., & Baumwell, L. (2001). Maternal responsiveness and children's achievement of language milestones. *Child Development, 72*, pp. 748–767.

Vaz, M. J. L. R. M. (2009), *Developmentally inspired computational framework for embodied speech imitation* (Doctoral thesis), Universidade do Minho, Escola de Engenharia.

Vigil, D. C., Hodges, J., & Klee, T. (2005). Quantity and quality of parental language input to late-talking toddlers during play. *Child Language Teaching and Therapy*, *21*(2), 107-122.

Vorperian, H. K., Kent, R. D., Lindstrom, M. J., Kalina, C. M., Gentry, L. R., & Yandell, B. S. (2005). Development of vocal tract length during early childhood: A magnetic resonance imaging study. *The Journal of the Acoustical Society of America*, *117*(1), pp. 338-350.

Vorperian, H. K., Wang, S., Chung, M. K., Schimek, E. M., Durtschi, R. B., Kent, R. D., Ziegert, A. L. & Gentry, L. R. (2009). Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging study. *The Journal of the Acoustical Society of America*, *125*(3), pp. 1666–1678.

Walsh, B., Smith, A., & Weber-Fox, C. (2006). Short-term plasticity in children's speech motor systems. *Developmental psychobiology*, *48*(8), 660-674.

Westermann, G., & Miranda, E. R. (2004). A new model of sensorimotor coupling in the development of speech. *Brain and language*, *89*(2), pp. 393-400.

Yoshikawa, Y., Koga, J., Asada, M., & Hosoda, K. (2003), "A Constructive Model of Mother-Infant Interaction towards Infant's Vowel Articulation", In Proceedings of the 3rd International Workshop on Epigenetic Robotics, pp. 139-146.

Yurovsky, D., Doyle, G., & Frank, M. C. (2016). "Linguistic input is tuned to children's developmental level", In Proceedings of the 38th Annual Conference of the Cognitive Science Society (2093–2098), Philadelphia, PA.