Feedback and imitation by a caregiver guides a virtual infant to learn native phonemes and the skill of speech inversion

Abstract: Despite large-scale research, development of robust machines for imitation and inversion of human speech into articulatory movements has remained an unsolved problem. We propose a set of principles that can partially explain real infants’ speech acquisition processes and the emergence of imitation skills and demonstrate a simulation where a learning virtual infant (LeVI) learns to invert and imitate a virtual caregiver’s speech. Based on recent findings in infants’ language acquisition, LeVI learns the phonemes of his native language in a babbling phase using only caregiver’s feedback as guidance and to map acoustically differing caregiver’s speech into its own articulation in a phase where LeVI is imitated by the caregiver with similar, but not exact, utterances. After the learning stage, LeVI is able to recognize vowels from the virtual caregiver’s VCVC utterances perfectly and all 25 Finnish phonemes with an average accuracy of 88.42%. The place of articulation of consonants is recognized with an accuracy of 96.81%. LeVI is also able to imitate the caregiver’s speech since the recognition occurs directly in the domain of articulatory programs for phonemes. The learned imitation ability (speech inversion) is strongly language dependent since it is based on the phonemic programs learned from the caregiver. The findings suggest that caregivers’ feedback can act as an important signal in guiding infants’ articulatory learning, and that the speech inversion problem can be effectively approached from the perspective of early speech acquisition.

Keywords: Language acquisition; speech inversion; articulatory modeling; imitation; phonetic learning; caregiver feedback

* Used abbreviations: ATP = articulatory target position, CM = concept matrix, CG = caregiver, DM = direct mapping, IM = indirect mapping, LeVI = Learning Virtual Infant, MFCC = Mel-frequency cepstral coefficients, VQ = vector quantization.
1 Introduction

Speech is an insurmountable tool of passing on messages and emotions in human communication. The path from an initial idea to a fully articulated utterance includes many complex procedures: transforming a message into a linguistically meaningful utterance, an utterance into muscular movements controlling the articulators, and finally into pressure variation in air. The task of the hearer is to decode and recognize the original message. Due to its extremely complex nature and importance in everyday tasks, it is no wonder that speech acquisition and recognition related questions are researched in various fields of study, including engineering, linguistics and psychology.

Speech inversion from acoustic signals into corresponding articulatory movements has interested researchers for several decades. Would it be possible to build a machine that could understand the articulatory gestures behind the spoken message? Such a system would be able to understand and imitate\(^1\) human speech despite the different acoustic characteristics of the speech production systems across different talkers, and it could be used, for example, to aid in the training of language learners and articulatory rehabilitation of impairments. However, the inversion problem has proved to be extremely complicated and still remains unsolved. At the same time, a research field concentrating on computational modeling of language acquisition has emerged in order to better understand how humans learn spoken language (see, e.g. [1]). These models attempt to explain how human-like speech processing capabilities can emerge from experience and interaction with the surrounding environment, in the same way as human infants learn the language effortlessly in the absence of explicit language training.

One long-term basic research topic and hypotheses in the speech communication studies has been - and still is - the motor theory of speech perception [2]. It emphasizes the active role of articulatory movements and corresponding neural processes of the motor cortex in human speech perception. Also, the McGurk effect demonstrates the power of the seen articulatory movements over the perceived sounds [3]. Based on these theories and observations, the long-term goal of our research is to integrate auditory and articulatory mechanisms (including rough processing of visual information) into an agent that learns to interact and communicate with its environment by voice. The simulations with this system could gain new knowledge about the main factors involved in human language acquisition and communication skills, especially related to the role of the articulatory mechanisms and, representations of the speech signals on the articulatory domain. Also, the relative roles of different communication modalities and their associations can be simulated and studied [4].

In our current study, we attempt to link speech inversion and language acquisition together by proposing a methodology for infant speech acquisition, and noticing how the speech inversion skill emerges during the speech acquisition process. Our learning virtual infant, LeVI, learns to babble language dependent speech sounds, or phonemes\(^2\), with the help of a virtual caregiver. After the phonemic learning phase, LeVI learns to imitate speech by learning a mapping between the caregiver’s vocalizations and its own articulations. The learning of the mapping is based on the statistical correspondences between LeVI’s own articulatory gestures throughout this work, the term imitation is considered as replicating the functionally important characteristics of another person. In case of speech, imitation thus means creating an utterance that has functionally the same interpretation as the original one, but it can be created with a different vocal tract size, pitch, and even slightly different articulation, like infants are known to do. Exact characteristics of voice color thus do not have to be imitated, as opposed to what mimics would do.

\(^1\) We use the term phoneme loosely when referring to the categories of speech sounds in the articulatory domain as well as the perceptually distinct units giving meaningful contrasts in utterances. The vocal tract model aims to produce the same units that will be categorized acoustically into phonemes, but since some articulatory parameters may get arbitrary values depending on the surrounding phonemes, different phones are formed.
and the caregiver’s imitative responses with similar linguistic content. LeVI and the caregiver have different vocal tract sizes and fundamental frequencies, and therefore differing acoustic outputs. We show that the used methods result in LeVI learning the same amount of consonant and vowel categories as the caregiver, nearly exact articulation of the two and recognizing phonemes in the caregiver’s VCVC utterances in terms of its own articulatory gestures almost perfectly in case of vowels and with 96.81% accuracy in case of place of articulation of consonants. Since the recognition takes place in the domain of the articulatory gestures, imitation by LeVI is immediately possible. In general, our simulations show how the two agents converge to a shared set of speech sound categories through interaction, suggesting that similar principles may govern the speech acquisition of real-life infants.

1.1 Speech inversion

Speech inversion aims to find the movements of the articulators that have produced a given acoustic speech signal. The task is easy for humans when listening to their native language articulated in a correct way (allowing for imitation for example). Furthermore, the Motor theory of speech perception suggests that speech inversion and the resulting articulatory representation of speech is used as an aid in speech perception [2]. Listening to speech has been shown to activate motor areas of the brain, supporting the idea that the motor system is used to find the phonetic representations for acoustic inputs (e.g. [5]). Similarly, stimuli to the motor cortex controlling lip or tongue movements have been shown to affect speech sound discrimination [6]. Although it is unclear whether the motor cortex is really necessary for successful speech perception, it is clear from the experimental evidence that speech perception and production processes are connected to each other. Being able to perform speaker-independent, human-like speech inversion in machines would presumably lead to numerous technological innovations, such as efficient speech coding using articulatory parameters, robust, human-like speech recognition, mobile phone bandwidth extension, and tools for language teaching or speech therapy.

Researchers struggling with speech inversion have to cope with the ill-posed nature of the problem. As in many inversion tasks, there is no unique solution available for the given input data. In other words, many vocal tract shapes can lead to the same acoustic output. This is used by ventriloquists to create normal sounding speech using different articulatory configurations than a common speaker of the language would. The nonlinear direct speech production problem from articulation into an acoustic signal is of the form \( A(z) = u \), where \( A \) is a continuous nonlinear operator of the speech production, \( z \) is a vector of articulatory parameters and \( u \) is a vector of resulting acoustic parameters. The direct problem can generally be considered uniquely solvable: certain articulation always leads to a unique acoustic output. The inverse problem of finding the correct articulatory parameters \( z \) from the acoustic parameters \( u \) is one-to-many in nature, meaning that one acoustic vector can be produced by several articulatory configurations. In addition, the same phoneme can be articulated in different articulatory configurations (e.g. due to coarticulation), which leads the mapping of a phoneme to the articulatory vector \( z \) being also of one-to-many nature.

The theory of ill-posed problems claims that the non-uniqueness of the solutions can be facilitated using proper constraints. In articulation these constraints can include limitations in muscle forces controlling the articulators, ranges inside which the articulators can move, structure of the motor commands controlling the articulation, synchrony of the articulatory parameters etc. In fact, the exact optimality criteria used by humans in speech production are not known. When proper constraints are chosen, a stable inversion solution can be searched using iterative methods, such as regularization techniques (e.g. [7] cited in [8]). Because the used vocal tract model is rarely equal to the actual vocal tract of the speaker whose speech is to be
inverted, a certain amount of discrepancy between the measured acoustic signal and synthesized signal must be allowed. The solution will always be a compromise between the match in the acoustic signals and the articulatory parameters taking into account the chosen constraints. The theory of solving speech inversion problems, including the above-mentioned aspects, are described in more detail for example in [8].

Speech inversion has been the subject of numerous studies during the last decades. The non-uniqueness of acoustic-to-articulatory mappings has been first investigated in the works of Atal et al. [9]. Flanagan et al. [10] used an optimization loop to search for articulatory model parameters that would lead to acoustic characteristics similar to the goal speech signal. Codebooks or lookup-tables have been widely used to obtain articulatory correspondents for acoustic features with varying optimization and search strategies [11-14]. The mentioned methods of speech inversion still suffer from inaccuracy of the obtained articulatory trajectories when compared to human speech, due to the possible compensatory ways of producing the known formant trajectories, especially in case of consonants. Some approaches rely on statistical methods, training Gaussian Mixture Models [15] or Hidden Markov Models [16] in order to learn the correspondence between articulatory and acoustic data. However, statistical methods generally require detailed measured articulatory data (nowadays increasingly available, see e.g. [17] and [18]) that is not available to human infants in real learning situation.

In our earlier research, we have used a simple greedy algorithm to guarantee smooth dynamics of the vocal tract area function when inverting vowel transitions using formant frequencies as acoustic criteria [13]. Later, we used an eight-parameter vocal tract model, whose parameters’ dynamics were maintained using a two-pole predictor structure. The result of the inversion was evaluated by clustering a large database of Finnish syllables using the acoustical and thereof derived articulatory parameters. The results showed that the mapping from three formant frequencies into eight articulatory parameters worked systematically, and the syllables were clustered into proper categories slightly more reliably using the articulatory trajectories than the original formant frequencies [14], but the inversion reliability remained far from human capabilities.

Even though humans seem to be able to effortlessly imitate other people, even we are not infallible masters of speech inversion, especially when generalized across different languages or speaker groups – probably every language student has noticed this when trying to speak or imitate a foreign language with a completely different phonological system from our native tongues. Without practice, some sounds in foreign languages may not even be perceived correctly due to the assimilation of arbitrary speech sounds to the correspondents in the native language, and without careful training, pronunciation of some sounds may be very challenging. Similarly, people with articulation disorders, habitually pronouncing some speech sounds incorrectly, do not normally imitate the corresponding correct pronunciations by others even in their native speech using the correct pronunciation. Intuitively, it seems that the humans’ ability to invert speech must be linked with the process of language acquisition: a certain way of pronouncing language-specific speech sounds is learned early in childhood, presumably with the guidance from surrounding people, and these reinforced representations are used in producing and imitating speech sounds later in life.

In this study we approach the speech inversion problem from the viewpoint of native language acquisition. Our virtual infant first learns to produce native speech sounds in a certain manner guided by a virtual caregiver, and later uses these learned categories when recognizing and inverting the caregiver’s speech. This approach is especially interesting since it is motivated by natural phenomena considering speech acquisition, and the resulting set of possible solutions for the recognition and inversion problems becomes language dependent. The sounds present in the caregiver’s speech end up falling easily in the correct phoneme categories of the infant, tying the production and recognition of speech sounds closely together: the units (phonemes) that are
to be recognized in speech have gained their meanings and labels in the articulatory domain of the recognizer. Since the same phoneme categories are used in both recognition and production, imitation of the sounds is immediately possible by the infant. No measured articulatory-acoustic data is needed since the necessary information is passed on from the caregiver to the infant during the learning process. Also, heavy computation usually related with frame-by-frame optimization methods at the inversion phase is avoided, since the inversion is performed from acoustics into articulatory gesture activations.

1.2 Phonemic learning and imitation

Imitation and speech inversion of the native language may become easily solvable for humans during the language acquisition process. The language and speech acquisition skills of children still amaze researchers, and engineers have yet to reproduce them in their machines despite several promising works and hypotheses. One of the important questions is, how do children learn to speak in a similar way to their caregivers despite the drastic difference between their vocal tracts and therefore the acoustic signals. Here, an overview of the studies of children’s language acquisition is given, and the findings are used as hypotheses in our simulations.

Newborn infants are known to be able to discriminate between phonetic contrasts appearing in all languages without previous language experience [19,20]. During the first months of life, exposure to the native language causes the speech sound perception to adapt for enhanced discrimination of native contrasts [19,21] and as a result, adults are reported to discriminate sounds with equal acoustic distance better when they belong to different phonetic categories of the native language, than when they belong to the same phonetic category [22-23]. This quality of speech sound perception is called categorical perception. Perceptual magnet effect refers to the tendency of human listeners to assimilate speech sounds inside a phonetic category to the category prototype [24]. This effect causes for example Japanese speakers to assimilate English /r/ and /l/ sounds to the same Japanese phonetic prototype /r/, while English speakers separate the same sounds in two categories [23]. Perceptual magnet effect has an important role in the current study, having already been developed for the caregiver and developing for LeVI during the simulation.

During the first year of their lives, normal infants’ vocalizations develop dramatically towards adult-like speech, until producing the first words at about 12 months [25]. Canonical babbling stage of children is reported to start around 5-10 months of age and consist of reduplicated sequences of vowels and consonants such as “bababa” or “dadada” (e.g. [26] cited in [27-29]). Our work uses reduplicated babble as LeVI’s early speech utterances.

Relevant to the current work, the language learning does not take place in a vacuum but the learner’s interaction with the caregiver and the surrounding environment is also essential. Research shows that parents’ contingent feedback to babbling guides children to learn important vocalizations: Goldstein [30] has shown that non-vocalic, non-imitative (smiling, touching, moving closer) parental responses occurring immediately after 8-month-old infants’ vocalizations (contingent group) lead into more complex and mature vocal behavior than non-contingent caregiver behavior with manipulated response timing. Gros-Louis et al. [31] showed that mothers regulate their contingent feedback depending on the quality of the child’s vocalization; mothers expanded more on children’s consonant-vowel vocalizations than vowel-like sounds. Mothers also imitated consonant-vowel clusters, significantly, eight times more often than vowel-like sounds. The results suggest that by giving contingent feedback mothers encouraged infants to produce more speech-like sounds. Similar results are described in [32], where it is found that 3-month-olds producing intonational syllabic sounds (as opposed to vocalic sounds) were rated by adults as more communicative and socially favorable.
Kokkinaki and Kugiumutzakis [33] have investigated vocal imitation in infant-parent interactions at 15-day intervals from two to six months of age. Parents were concluded to imitate their children significantly more than vice versa. Jones [34] examined infants’ imitation of eight behaviors including two vocalizations and concluded: “No reproduction of these motor acts—that is, no mimicry—was observed at 6 months. Mimicry appeared to develop slowly through most of the 2nd year, emerging at different ages for different behaviors”.

On the other hand, infants are reported to be able to mimic certain qualities in adult speech. Kuhl and Meltzoff [28] report a test situation where infants from 12 to 20 weeks of age watched a previously recorded female speaker pronounce /a/, /u/ and /i/ vowels on a video. The infants responded with vocalizations resembling the vowel they were shown. It was not analyzed in which proportions the successful vowel matching by the child was caused by the visual and the auditory modality. Thus, it remains possible that the match was mainly caused by the visual image of mouth positions. It is mentioned that when the face was presented with tones not identifiable as speech, children did not produce speech sounds.

Goldstein & Schwade [35] have shown that 9.5-month-old infants reformed their babbling according to the contingent feedback by parents. Half of the parents giving contingent feedback responded by speaking fully resonant vowels, whereas the other half responded with consonant-vowel alterations. Infants in the contingent feedback groups modified their babbling to better correspond the phonological structures (resonant vowels / CV alterations) in the parental feedback, but they did not imitate the sounds using exactly the same phonemes although their vocal repertoire included the vowel sounds produced by the mothers. The research suggests that children are able to imitate some more abstract qualities in adult speech.

Our simulated interaction between LeVI and the caregiver follows the above findings. LeVI cannot imitate the caregiver before it learns to link its articulations with the caregiver’s speech sounds. The caregiver gives a constant positive feedback signal from babble that resembles its own speech sounds and LeVI reforms its phoneme vocabulary based on the feedback. The correct amount of phoneme categories is not known by LeVI but emerges during learning. Finally, when LeVI is able to produce the learned native phonemes, the caregiver imitates LeVI with utterances that do not need to be equal than LeVI’s original productions. In this imitation phase, LeVI learns the associations between its articulation and the caregiver’s speech sounds.

### 1.3 Previous computational models of speech acquisition

Computational models can be used in order to test hypotheses and find mechanisms that underlie real-life language acquisition. When the components present in a normal language acquisition situation are accurately modeled, the normal phases and characteristics of the process may be replicated in computer simulations, possibly verifying the existence of certain learning mechanisms. Several studies have been carried out to model infant language acquisition using a variety of different hypothesis, methodologies and initial assumptions. In this section previous models of speech acquisition are reviewed.

In some studies the emergence of infants’ skill of imitation [36,37], or speech motor control [27,38,39] were studied without considering the correspondence problem, i.e. the infant and the caregiver having vocal tracts of differing sizes and thus differing acoustic sound characteristics. Imitation or guidance by the caregiver has been used as a basis for the infant to tackle the correspondence problem in [40-50]. In [40,41,43,44,46,47,49,50] the imitative utterance by the caregiver’s imitative answer did not include additional phonemes to the infant’s original intended phoneme, whereas in [42,45] additional phonemes were allowed. Learning of
only vowel systems was considered in [36,40,41,42,43,45,46,47,50], whereas in [27,37,38,39,44,49] also consonantal sounds were studied.

Markey [37] has developed HABLAR (Hierarchical Articulatory Based Language Acquisition by Reinforcement learning) as a computational model for early childhood language acquisition. The auditory perception of HABLAR learns to categorize speech sounds into phonetic events using acoustic cues, namely regions of static and dynamic spectra. Categories are learned from synthesized speech signals using soft competitive learning. A phonological controller links these phonetic events with articulatory events, and articulatory controllers select corresponding articulatory movements. Controllers were trained by reinforcement learning, where the reward signals were calculated in proportion to the difference between the target utterance’s phonetic events and the phonetic events present in the feedback after the model tries to articulate the same speech signal. In other words, HABLAR holds on to an acoustic representation of an utterance and learns articulatory control policies based on the difference in its productions and the signal to be imitated. It is assumed that the voices and the vocal tract sizes resulting in the target and the feedback signals are similar, since the acoustic and thus phonetic events are synthesized with the same vocal tract model. Markey mentions that “Its [HABLAR’s] greatest weakness is its inability to reconfigure the dimensions of the vocal tract to conform to a child’s anatomy” (p. 95 in [37]). In our simulations, the vocal tract sizes of the caregiver and the infant are different, and the correct amount and qualities of the speech sound categories are acquired based on caregiver’s feedback without comparing the two acoustic productions.

Guenther has studied speech acquisition and production with a cognitively plausible neural network model, DIVA [27,38,39]. The model learns spatiotemporal auditory target regions as well as somatosensory target regions for different speech sounds stored in a speech sound map. When attempting to articulate a speech sound, auditory and somatosensory targets are activated, giving the sound a sensory expectation. When the speech sound is articulated, the targets are compared to the current auditory and somatosensory states, and the possible errors are used to correct the articulation. The model can replicate several speech-related phenomena, including motor equivalence, coarticulation and speech-rate effects as well as account for some neuroimaging data. However, Guenther’s work does not cope with the problem of dissimilar bodies. It simply assumes that the auditory targets for sounds can be learned by monitoring speech of other people, and in one learning experiment the formants extracted from the human speech were “modified slightly to form an auditory target that better matched the vocal tract characteristics of the Maeda synthesizer” (p. 289 in [38]). In our current work, the infant learns to match his productions into the productions by others with different acoustic sound qualities, and thus aims to answer questions on how others’ speech acoustics can be recognized as speech sound activations – i.e. how the skill of imitation is learned.

Howard and Messum [49] have used similar caregiver-child interaction hypotheses to simulate language acquisition in a very interesting setup. Their virtual child, Elija, first learns speech motor patterns corresponding to speech sounds by exploring its vocalic abilities and storing the most rewarding vocalizations based on acoustic and sensory qualities of the actions. Later, human caregiver’s imitative responses to the vocalizations were used to remove non-speech actions and store the associations between adult speech sounds and Elija’s motor actions. The imitated sound segments were stored and used as templates in the speech recognition phase. Caregiver’s speech could be recognized and imitated by classifying it into segments, selecting the closest stored audio templates using a dynamic time warping algorithm and selecting the stored motor actions corresponding to them. Maeda’s articulatory model [51] was used, but the quality of nasal sounds was reported to be poor and the model lacked a possibility for an animated face that could make the interaction with the model more pleasant.
In Howard’s and Messum’s study the imitative response of the human caregiver was a reformulation of the infant’s utterance, and the reformulations were used in parsing caregiver’s speech in the recognition phase. They refer to Meltzoff’s work [52], when pointing out that “infants know when they are being imitated” (p. 91 in [49]). Meltzoff [52] writes: “...infants directed more visual attention and smiled more at the person who was imitating them. They preferred an adult who was playing a matching game” (p. 256 in [52]), referring to [53], but this kind of action was reported in a study of non-verbal interaction as opposed to vocal imitation. It seems likely, that in the beginning the child may not be able to parse the exact imitated utterance from the caregiver’s imitative response, and the imitation may not be exactly equal in the first place. In our study some deviation in caregiver’s imitative utterances are allowed.

In [36], relations between articulator positions and formant frequencies are learned by strengthening Hebbian connections between corresponding parts in auditory and motor maps during a babbling phase. Connections between the most linear parts of the domains (small articulatory change results in small auditory change) end up having the biggest weights. In another experiment the model is exposed to vowels extracted from ambient languages and the babbling adapts towards the target language because of frequent activation of language-specific units in the auditory map. Again, the ambient language sounds (and consequently their mappings) are assumed to match the model’s productions and the correspondence problem is not taken into account.

Ishihara et al. [46] have simulated infants’ vowel category learning hypothesizing imitative turn-takings by the caregiver and the child, where the two alternately imitate each other’s vowel production. The child is provided with a fixed amount of vowel categories to be learned. The caregiver uses sensorimotor magnets to bias perceived infant’s utterances towards his own vowel prototypes, as well as automirroring bias to assume that the infant’s utterance resembles the caregiver’s previous vocalization. The results imply the importance of the caregiver’s participation in the vowel learning process.

In [40] a physical vocal tract robot was used to create vowel sounds and a human caregiver imitated them by using the exactly same vowel if a human-like vowel (/a/, /i/, /u/, or /e/) was recognized. A criterion based on articulatory effort (toil) was used to reduce the many-to-one mapping problem. In, [41] it is assumed that the learning robot knows the desired categories of the caregiver’s vowels and has a rough estimate of the mapping function between the acoustic domain of the imitating caregiver and the robot. When the caregiver imitates the robot, the robot measures the distance between the formant frequencies of the correct caregiver’s vowel category and the imitated utterance, and corrects his own babbling in the same direction using the rough mapping function. Human caregiver imitates again using only one, the correct, vowel sound. In Vaz’s [47] work a learning robot utters a predefined set of vocal primitives, i.e. spectral vectors extracted from normal speech, and a human caregiver imitates them using the same phonetic content.

Miura et al. [42] have conducted a study where the caregiver and the learning robot have dissimilar bodies and the caregiver does not have to imitate the learner with exactly similar utterances. The learner develops imitation detectors to detect if its utterance was present in the caregiver’s answer, and updates these detectors during the learning using a method called ADA-boost. In the training phase a human caregiver imitates the learner’s vowels with different probabilities. The learner is given 15 vowel primitives (three for each of the five caregiver’s vowel categories) in the beginning, each having a probability of being recognized as the corresponding vowel. The learner succeeds in learning in fact above chance level (P = 0.2) probability of being imitated.

In Hörnstein’s work [43] a sound-motor map (a neural network) is trained to a robot first in a babbling phase in order to find an initial mapping between Maeda’s vocal tract model [51] parameters and their synthesized resulting sound features. In a second phase the robot learned
native Swedish or Portuguese vowels when trying to imitate a caregiver’s vowel. The caregiver had to adjust his or her voice to help the robot to find a correct imitation for the current vowel because of the correspondence problem, and when the robot managed to imitate, a reinforcement signal was given on a keyboard and the corresponding vowel sound was stored. The robot was also trained to recognize vowels produced by humans by recording words spoken by Portuguese speakers, extracting the vowel segments and training the robot by using the extracted vowel samples as imitations of the robot’s learned vowel productions. After training with exactly the same vowel than what the robot uttered, about 58% recognition accuracy of Portuguese vowels spoken by humans was reported. In [44], a visual component representing the lips of the caregiver was integrated in the simulations resulting in an increase of the vowel recognition rate to 63.3%. In addition, the robot was reported to learn the consonants /b/, /d/ and /g/ when the caregiver imitated the robot’s CVCV utterances, and the inverse mapping of the consonant was based on the caregiver’s sound feature just before the closure and the lip position at the same moment. Again, when the authors were happy with the robot’s production, based on investigating the vocal tract geometry, a reinforcement signal was given and the articulatory position was stored. Finally, in [45], the importance of detecting the imitative part in the caregiver’s utterance was shown, when training with additional vowels in the imitative utterance degraded the recognition accuracy.

In contrast to Hörnstein’s work [43-45], the methodology of our work differs in several ways: LeVI’s phoneme targets are reinforced without LeVI having to try to imitate the caregiver, and the targets gradually approach the correct native ones when the caregiver demands better performance from LeVI. Instead of using a neural network we use a straightforward statistical mapping from acoustic speech features into already learned phonemic gestures. Also, in the imitation phase LeVI babbles CVCV utterances instead of isolated speech frames, and the caregiver imitates in a similar fashion, but the order of the C and the V can be changed and additional phonemes can be added.

In [48], an infant with a non-linearly scaled vocal tract when compared to an adult, was set to learn a topological mapping between the two differing acoustic domains by clustering the two acoustic spaces separately using Self Organizing Maps with the same set of parameters for the infant’s and adult’s acoustic productions. The most suitable set of parameters is found when the infant tries to imitate caregiver’s utterances using different sets of parameters and the caregiver gives an overall rating of the success of the imitation. The most rewarded clustering consisted of 15 clusters in the two acoustic spaces, and the inter-domain similarity was confirmed by a phonetically trained listener. Being promising in its results, and providing a rational approach to the speaker normalization problem, the work did not yet offer a solution to the acquisition of the exact number of phoneme categories. Also the productions of the adult and the infant were generated with the same babbling algorithm instead of the caregiver’s speech being biased towards already learned native phonemes, as would be expected to happen in real interactions. The work did not consider the role of caregivers’ imitative responses.

Plummer [50] offers a solution to the correspondence problem by arguing that infants can represent vowels of individual speakers in a speaker-independent mediating space. The infant is hypothesized to create perceptual manifolds for the acoustic features of the caretaker’s and his own vowel productions, and align the manifolds by mapping the points of the manifolds into a common “latent space”. The aligning occurs based on imitative interactions between the infant and the caregiver. Vowel data for the caregiver and the infant are synthesized, and imitation data considering five vowel sounds are drawn manually for both speakers, but it is assumed that the infant imitates the caregiver’s vowel production and receives contingent feedback to approve the imitation success. While Plummer offers an interesting solution for the vowel normalization problem, our focus in the current study is more on phoneme acquisition and making the imitative interaction situation more natural and open for variation in the imitative utterances.
Similarly to some of the mentioned studies, our work also takes the approach of imitation and feedback by caregiver as the experimental studies of language acquisition imply. The main differences when compared to the mentioned previous studies are:

- Our method does not require automatic speech segmentation or dynamic time warping methods.
- Articulatory gestures corresponding to the phonemes of the native language are learned without imitative interaction or acoustic-to-articulatory inversion, but using only feedback from the caregiver, when the caregiver interprets infant’s speech sounds based on the perceptual magnet effect.
- LeVI learns all Finnish vowels and consonants, and the number of correct vowel or consonant categories are not given a priori for LeVI but are learned during the interaction.
- Imitation by the caregiver happens only after the learning of the articulatory targets. In the imitation phase LeVI learns the correct mappings between the caregiver’s speech acoustics and its own articulatory gestures, based on simple probability calculus.
- The caregiver does not have to imitate LeVI precisely, but above chance level similarity in linguistic content of the caregiver’s reformulations is sufficient for learning.
- The reliability of the learned mappings is evaluated in a recognition test, where LeVI recognizes phonemes spoken by the virtual caregiver.

In our learning system, we use articulatory synthesis based on target positions in the articulatory domain for Finnish phonemes and trajectories towards/through the targets using minimum-jerk trajectories, and language-dependent timing and excitation parameters. Vowels and consonants (including voiced and voiceless stops, nasals, fricatives, liquids, and a trill /r/), coarticulation, vocal tract length differences and variation in fundamental frequency, can be modeled with the synthesizer, and are used in the speech acquisition experiments in order to provide natural richness and variation in the experiments. The synthesizer, programmed to be able to produce chiefly Finnish phonemes and speech from text, is used as the virtual caregiver, and a synthesizer with a linearly scaled-down vocal tract length, higher fundamental frequency and without language-specific knowledge is used as the learning virtual infant, LeVI. We hypothesize that infants learn to speak and imitate adult speech using the following principles that correspond to the findings in language acquisition described above:

- Infants are unable to match their own sounds directly with the caregiver’s sounds in the auditory domain and are thus unable to imitate parents’ acoustic outputs as such. The appropriate mapping from the acoustic (and visual) properties of adult vocalization to the articulatory domain of the infant is a premise for imitation that must be fulfilled by learning.
- The infant improves the linguistic quality of his/her babbling based on positive feedback from an adult caregiver.
- The caregiver is able to understand and imitate (invert) the infant’s speech, and categorizes infant’s babbling into his own phonemic categories according to the perceptual magnet effect.
- The caregiver rewards more adult-like babbles with positive feedback and finally starts to imitate consonant-vowel clusters produced by the infant according to the perceived sounds.

3 /b/, /f/, /g/ or /z/ do not naturally appear in Finnish speech, but they are also modeled and experimented with in this work.
• The infant learns the mapping between the caregiver’s speech acoustics and its own speech acoustics in a phase where the caregiver imitates the infant. Statistically frequent correspondences between the infant’s and the caregiver’s acoustic features are learned when the caregiver imitates utterances of canonical babble.

• Similarly to the previous point, the infant learns what sound events in his/her own acoustic output, as well as in the caregiver’s acoustic output, are likely to be caused by the articulation of the learned phonemic targets. This information is used to map acoustics into articulation (i.e. language-dependent inversion).

2 Articulatory synthesizer

The simulations of the current work were performed using an articulatory synthesizer by Rasilo (see [54] for a detailed description), but the methodology presented in this work should be universal and work with any vocal tract model. The synthesizer was created from scratch to allow for easy future modifications of different model components or development. This chapter introduces some of the details of the used model.

The model is a geometrical model based on the description of Mermelstein [55], but with certain modifications so that the model could produce all Finnish vowel and consonant sounds. Synthesis is based on phonemes, which are intended to be reached at certain time instants. Phonemes consist of articulatory target positions (ATP), as well as timing and excitation parameters that capture important information about the articulatory gestures related to the actual target positions. The ATPs themselves are defined for nine model parameters related to the positions of elementary articulators (tongue base, tongue tip, hyoid, velum, jaw and lips). All the properties defining a phoneme are listed in Table 1. Figure 1 illustrates the mid-sagittal view of the vocal tract as well as the vowel triangle related to it. Animations of articulatory speech synthesis can be produced with the vocal tract model (a video of babbling is provided as supplementary material).

Flash & Hogan [56] have found that human arms’ point-to-point movements follow a minimum-jerk trajectory, and similar dynamics of point-to-point movements are assumed to be innate for the elementary articulators, and thus do not have to be learned. Nevertheless, the time frame during which the point-to-point movement happens is considered to be related to the physical or language dependent characteristics, and is thus phoneme-dependent. For example, in our model, tongue body movement in /k/ takes more time than a tongue tip movement during /t/ due to its higher mass, but both of the movements follow a minimum-jerk trajectory.

Differences in the timing parameters were noticed to be crucial to consonant identification purposes in unreported listening tests by the first author. All the timing parameters, namely the approach, release and closure times, as well as excitation types, voice onset and voice offset times are considered phoneme-, and thus language-, specific, and they have to be learned by the infant. In reality some of these parameters, for example approach and release durations may be dependent only on the masses of the articulators, but since this would be a strong assumption, they are defined as phoneme-dependent qualities in our work.

The synthesizer reproduces a coarticulation effect, where for example a vowel’s ATP may not be reached exactly before an approach period corresponding to another vowel or consonant target begins. Consonants only affect a subset of the nine articulatory position parameters, so that some parameters can vary freely during a consonant sound, creating a context dependent vocal tract configuration. White Gaussian noise (SNR = 50 dB) is added to the final synthesized signals in order to diminish the inaudible sound waves that keep reflecting in the anterior part of the vocal tract during closures. Without the noise, the feature extractor would
provide a clear acoustic feature where it normally could not be heard, providing an unrealistically clear indication of the closure place.

A detailed description of the model can be found in the supplementary material [54]. What is important for this work is that the model calculates articulatory trajectories and corresponding acoustic outputs for a string of arbitrary phonemes defined as the parameter set in Table 1, placed at arbitrary time instants. The virtual caregiver in this work uses pre-defined phoneme characteristics of Finnish, whereas in the beginning, LeVI does not know any of the parameters corresponding to the Finnish phonemes but learns them by interacting with the caregiver. The learning mechanisms are language independent; if the phonetic system of any other language were to be programmed to the virtual caregiver, LeVI would learn the new system in a similar fashion. The next section describes the methods that are used in the learning process of LeVI, as well as the principles underlying LeVI’s phoneme recognition and imitation capacity.

Table 1. List of all properties needed to define a phoneme in our work.

<table>
<thead>
<tr>
<th>Property of the phoneme</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Articulatory target position (ATP)</td>
<td>The nine vocal tract parameters defining the target position of the phoneme in the articulatory domain (tongue body x-coordinate, tongue body y-coordinate, tongue tip x-coordinate, tongue tip y-coordinate, jaw angle, hyoid x-coordinate, lip protrusion, lip opening and velum opening)</td>
</tr>
<tr>
<td>Approach duration (or target lookahead-time)</td>
<td>Defines, how long before the target instant the approach towards it is started.</td>
</tr>
<tr>
<td>Hold duration (or closure duration for consonants)</td>
<td>Duration of holding the articulators in the target position when achieved.</td>
</tr>
<tr>
<td>Release duration</td>
<td>Duration of the target release towards the next target after the hold period.</td>
</tr>
<tr>
<td>Voice onset time, $V_{on}$ (Only for consonants)</td>
<td>The duration between target release and voicing.</td>
</tr>
<tr>
<td>Voice offset time, $V_{off}$ (Only for consonants)</td>
<td>Defines, how long before the target voicing ends.</td>
</tr>
<tr>
<td>Excitation</td>
<td>Excitation during the target, $0 = $ silence, $1 = $ voicing, $2 = $ hiss (used in /h/).</td>
</tr>
<tr>
<td>Priority (consonant / vowel)</td>
<td>Defines if the target overrules the current target (generally, consonants replace all targets, but vowels do not replace consonants).</td>
</tr>
</tbody>
</table>
Figure 1. Mid-sagittal image of the vocal tract model (left) and the vowel triangle generated by random sampling through possible non-nasalized vocal tract shapes. The green diamonds mark the Finnish vowels as defined into the caregiver’s vocal tract model. The red circles mark the positions of typical Finnish vowel sounds [57].

3 Methods

The articulatory synthesizer described above was created in order to synthesize intelligible human-like Finnish speech, or alternatively to create utterances using arbitrary phoneme characteristics. The language-dependent, pre-defined, Finnish phonemes are used as the “brain” of the virtual caregiver (CG); he knows how to speak and what the correct language-dependent articulatory targets and their parameters are, and he can perceive and measure the quality of LeVI’s speech productions and finally invert LeVI’s speech into his own phoneme categories. This section describes the mechanisms that lead in LeVI learning the skill of imitation in three phases. In the first phase, LeVI learns the Finnish phonemes, initially only mastered by CG, in an online supervised learning phase based on positive/neutral-type direct feedback on LeVI’s babbles given by CG. In the second phase, LeVI’s auditory perception adapts towards sounds present in its own and CG’s speech, and in the third phase, LeVI learns to recognize CG’s speech in terms of the learned phonemes, based on CG’s imitations of the phonemes uttered by LeVI.

In the first phase we simulate phoneme learning by assuming that caregivers have expectations on how the correct infants’ phoneme productions should sound like, and are sensitive to errors in infants’ productions when compared to the expectations. Due to the inferiority of all known speech recognition algorithms to the human ear, we are forced to simulate CG’s perception by providing LeVI’s articulatory parameters directly to CG, which are then compared to CG’s known correct articulations (see also discussion in section 5). In the third phase, LeVI babbles using the learned phonemes, and CG recognizes and inverts them in terms of his own phonemes and imitates LeVI, helping LeVI to associate its phonemes to the correct acoustic features in CG’s speech. LeVI extracts features from the acoustic speech signals based on commonly occurring features learned in the second phase.

The vocal tract used for LeVI is a length-wise linearly scaled down version of CG’s vocal tract. The exact anatomical differences between infants’ and adults’ vocal tracts (for example
infants’ relatively shorter vertical part of the vocal tract than the horizontal part, when compared to adults (see e.g. [58]) are out of the scope of this work and not modeled. Modeling the differences more exactly would affect the first phonemic learning phase: presumably the infant would learn different articulatory configurations when compared to those of the caregiver for the phonemes that the caregiver rewards with positive feedback. In case of non-linear differences in the morphologies of the adult’s and the infant’s vocal tracts, the infant’s vocal tract parameters could not be interpreted as such by the caregiver and the caregiver’s inversion in this phase should be implemented based on the acoustic output of the infant as happens in real-life situations. It is likely that real caregivers cannot invert their children’s productions in exact articulatory representations because of the already mentioned many-to-one property, complicating the phonemic learning phase of these experiments. The corresponding problems and possible solutions are discussed in more detail in section 5. However, in these experiments we assume that the caregiver can invert and interpret the infant’s babble in terms of his own phonemes whatever the morphology of the vocal tract.

3.1 Learning of language dependent phonemes based on online supervised learning

LeVI uses the method described briefly in section 2 and in more detail in [54] for articulatory trajectory calculation, as we assume it to be an innate, language-independent skill of muscle control. The task left for LeVI is to learn the 16 parameters defining the language-dependent phonemes without being able to imitate the caregiver. LeVI is hypothesized to learn the correct phonemes using the following principles:

- LeVI babbles canonical reduplicated CV utterances.
- LeVI wants to communicate with the caregiver and wants positive feedback on its actions, causing it to produce articulations that have previously resulted in positive feedback.
- The caregiver gives more positive feedback on more adult-like articulations by LeVI (CG gets more critical when learning happens).
- The caregiver categorizes LeVI’s speech sounds based on the phonemic qualities according to his categorical perception.
- Based on the feedback by the caregiver, LeVI adds new phonemes to its vocabulary, or adjusts the properties of the already existing ones.

LeVI babbles reduplicated utterances by choosing a vocal tract configuration for one vowel and one consonant. LeVI is assumed to differentiate vowels from consonants so that first a static vocal tract configuration consisting of all 16 articulatory parameters is selected as the vowel, and then a dynamic consonantal gesture consisting of a subset of the parameters is superposed on the vowel configuration, resulting in reduplicated babbling. The subset of the vocal tract parameters that can be used to create a consonantal configuration are defined manually to be similar to those of CG (all possible configurations from the parameter table in [54, p. 15], that have one or more non-defined values). We thus assume that the consonants in reduplicated babble have constrictions in similar positions than the phonemes of the native language (or that the phonemes of the native language have been developed from patterns present in babbling), generally in labial, alveolar or velar positions. In our simulations the infant thus does not experiment with gestures that would have for example a velar and labial closure at the same time. Such cases would probably be possible in real life and they could be simulated in our model, but in such cases the learning time using the current method would increase and the simplification is thus used. For example, when LeVI is experimenting with an alveolar consonant, the affected parameters are set to be tongue tip x- and y-coordinates, tongue body x-coordinate, hyoid, velum, jaw and the 7 timing / excitation parameters. As can be seen in the parameter values listed in [54],
/s/ does affect also the jaw parameter but does not affect the hyoid, and this subset is used as a different possibility for LeVI to experiment. The values of the affected parameters remain unknown and subject to learning – in these simulations we do not for example assume that a consonantal gesture should have a complete closure of the vocal tract.

LeVI is set to experiment with one randomly chosen consonant configuration at a time (e.g. bilabial, alveolar, velar, bilabial-nasal, alveolar-fricative etc.), since repeating CV syllables with the same consonantal place of articulation is a typical characteristic of reduplicated canonical babbling (e.g. [59]). In real-life, some of the parameters, such as voice onset time ($V_{\text{on}}$), voice offset time ($V_{\text{off}}$), approach duration or release duration, may be partly defined by the physiological factors of the sound production mechanisms that are not modeled in our articulatory model (see section 2.1.1), and thus may not be free parameters having to be learned by human infants. For simplicity, in our study we decided to teach all the parameters to LeVI using the same algorithm. The used distance measure is Euclidean distance in the 16-dimensional parameter space, when all the parameters are scaled to lie in the range $[0,1]$. In the next description the phonemes are referred to as targets, referring to the target phonemes of the language in question that LeVI aims to master. The learning process goes on as follows:

- LeVI has lists of his vowel and consonant targets that have resulted in positive feedback, $A_v^l$ and $A_c^l$ (hereon called positive targets). At the first iteration both of the lists are empty.

- At every step, LeVI chooses a positive target with a probability of 4/5 or randomizes a new target from the parameter space with a probability of 1/5 for the vowel and the consonant separately. The new target is randomized in a distance of minimum of $R_{\text{new}}$ from any positive target. To start the iteration at the first learning step, a random target is chosen. LeVI is not able to reproduce the attempted target exactly but within a radius of $R_{\text{acc}}$ in the 16-dimensional parameter space from the actual target point. This enables explorative behavior, and the approaching of CG’s targets when positive feedback signals are obtained. The exact targets that LeVI finally pronounced are $a_v^l$ and $a_c^l$ for the vowel and the consonant targets correspondingly.

- CG carries a list of the distances of the best target productions of LeVI to his own vowel and consonant articulatory targets. The distances are $D_v^{ CG}(j)$ and $D_c^{ CG}(k)$, $j = 1 \ldots V$ and $k = 1 \ldots C$, where $V$ is the amount of vowel targets and $C$ the amount of consonant targets mastered by CG.

- CG perceives the articulatory targets uttered by LeVI, $a_v^l$ and $a_c^l$, and calculates the distances to his own inventory of articulatory targets $d(a_v^l, A_v^{ CG})$ and $d(a_c^l, A_c^{ CG})$.

- CG selects the closest target from his inventory, i.e., due to the categorical perception of CG, LeVI’s targets are perceived as: $P_v = \text{argmin } d(a_v^l, A_v^{ CG})$ and $P_c = \text{argmin } d(a_c^l, A_c^{ CG})$.

- If the whole utterance surpasses the expectations of CG, i.e., both the vowel target and consonant target are closer to the recognized CG’s targets than the best distances in the list: $d(a_v^l, A_v^{ CG}) < D_v^{ CG}(P_V)$ and $d(a_c^l, A_c^{ CG}) < D_c^{ CG}(P_C)$, the distance lists are updated $D_v^{ CG}(P_V) = d(a_v^l, A_v^{ CG})$, $D_c^{ CG}(P_C) = d(a_c^l, A_c^{ CG})$ and positive feedback is given to LeVI.

- LeVI analyzes the feedback so that if a positive target was chosen, and the babbling resulted in positive feedback, the positive target is shifted to the new
location inside the radius $R_{acc}$ where the babbling happened, and the lifetime of the target is set to $L \in \mathbb{Z}^*$. If a positive target was chosen but positive feedback was not gained, the positive target does not change but its lifetime is reduced by one.

- If a new target was randomized and it led into positive feedback, the target is saved as a positive target and it is given a lifetime of $L$. If the new target did not gain positive feedback, it is forgotten immediately (not saved).
- LeVI deletes positive targets with a lifetime of zero from the positive target lists. LeVI thus forgets non-meaningful targets.
- If CG does not perceive LeVI producing a phoneme corresponding to some of CG’s phoneme categories during $M \in \mathbb{Z}^*$ babbles, CG sets the distance of the best LeVI’s production to the corresponding target ($D_{CV}^C$ or $D_{CC}^C$) to infinity. CG thus forgets that LeVI has produced a phoneme corresponding to the target of CG if it has not been uttered in a long time, allowing for re-learning of the target in rare occasions where too many consecutive productions of the target outside the positive radius (see below) causes LeVI to forget the target. Without this property, CG would continue expecting a good production of a target already forgotten by LeVI, making it really difficult for LeVI to re-learn.

We use a positive radius $R_p$ around the correct target so that when LeVI reaches the correct pronunciation of the target, rewarding continues and the targets are not easily forgotten. That is - if the Euclidean distance between LeVI’s target and CG’s target are less than $R_p$, positive feedback is given. Otherwise the lifetime would gradually decay and the targets be forgotten because improvement cannot happen anymore at this level. In real life, the reinforcement of the correct pronunciation may continue at higher levels such as after production of meaningful syllable like units or words that are further rewarded by other people.

The described online supervised learning algorithm allows LeVI to learn language dependent articulatory target parameters by only using constant positive feedback from its caregiver. The caregiver follows the rule of giving more positive feedback on phonations that are close to the correct production of the sounds of the language known by the adult. In real-world interactions the feedback is likely to be more complicated than of the pure positive/neutral form. For example, the infant could pay attention to some phonological structures in the parental feedback, e.g. fully resonant vowels or consonant-vowel alterations [35], and get cues about which phonemes and their properties in the utterance should be adjusted.

### 3.2 Learning of audio-codebooks

In our experiments, LeVI learns to map commonly occurring acoustic characteristics of CG’s speech either into its own speech acoustics or into its articulatory gestures. First, LeVI learns codebooks that store the acoustic features that are known to be common for a particular speaker. The speech signals can be later quantized using these codebooks to obtain a compressed representation of the speech signals for further processing. This codebook learning can be interpreted as the development of the infant’s auditory perception towards the qualities present in its own or its caregiver’s speech. Note that the use of codebooks does not imply that a real child would quantize acoustic input into discrete categories before phonemic perception, but is simply a convenient way to perform statistical learning on the speech signals. If desired, continuous valued models can be approximated to a desired degree using semi-continuous vector quantization (e.g. [60]).
We do not use any speech segmentation methods to parse unique acoustic features from speech streams, but use simply the spectral qualities of 25–ms speech frames. Vector-quantized (VQ) MFCC features (Mel-Frequency Cepstral Coefficients) are used as the acoustic representation. Vector quantization transforms the speech signal into a sequence of discrete symbols using a codebook to choose the correct symbol for each speech frame based on the extracted MFCC coefficients. The VQ codebook is created by extracting MFCC vectors from a large amount of speech data and clustering them into a certain amount of clusters. Each cluster centroid is assigned with a symbol (integer numbers in our work).

In the current experiment LeVI learns separate codebooks for the auditory perceptual features of CG’s speech sounds and its own speech sounds. The codebook for CG’s sounds is learned while listening to a large amount of speech by the caregiver. CG’s speech data is generated by synthesizing 250 grammatically correct but random Finnish sentences from a set of 34 words. MFCC features are extracted from all the output data using a Hamming window of a length of 25 ms and a step size of 5 ms. The first MFCC coefficient \(c_0\) is discarded and 11 coefficients \(c_1\)…\(c_{11}\) are used as the MFCC vector representation. All MFCC vectors are normalized to a length of one. 50,000 MFCC vectors are randomly chosen from the speech data and quantized into a codebook using a standard k-means algorithm with 150 cluster centroids. Each cluster is assigned with a label, that is, an integer number running from 1 to 150. The same clustering approach is used to produce a separate codebook for LeVI’s own speech sounds in a babbling-phase where he babbles 1000 CVCV vocalizations using the learned phonemes. A vocal tract length of 4/7 of the caregiver’s vocal tract length (10 cm / 17.5 cm) and a fundamental frequency of 250 Hz (as opposed to 120 Hz by the caregiver) are used, corresponding to a typical infant’s voice. Due to the articulatory synthesizer, the fundamental frequency varies slightly around the given value as described in [54].

The number of cluster centroids has not been specifically optimized for the task and does not have a clear relation to the amount of phonemes in Finnish language, but in general the number should be large enough so that all important or distinct characteristics of speech have at least one feature representing them. Too low a number might leave some phonemes without a representation in the codebook and too large a number might result in so many representations for a single phoneme (because of variation caused by background noise or differences in fundamental frequency, for example) that the generalizing ability of the recognition model might suffer, and training with more exemplars would be needed. For comparison, the simulations were run also with codebook sizes of 100 and 200 with no remarkable changes in recognition accuracy (see also the results section). We propose that the codebook creation could also be integrated in the imitation phase to work online, using for example a self-learning vector quantization method (see e.g. [61]), but this is left for future research.

3.3 Learning the mappings between the caregiver’s and infant’s features using imitation by caregiver

After the learning of articulatory targets, LeVI should be able to produce highly intelligible canonical babble that resembles the qualities of Finnish language. As already mentioned in the introduction, mothers are likely to imitate CV patterns of their infants. We hypothesize that the infant learns to associate the qualities of the two differing acoustic productions or between the caregiver’s acoustic productions and the infant’s articulation, based on this parental imitation. Because the imitated utterance cannot be considered exactly equal in duration, order of phonemes, or even in exact content, the mapping has to be learned statistically: the acoustic characteristics of the infant’s and parent’s utterances that occur together most frequently are considered to have the same meaning. A real-world example of an imitative interaction could be for example [31]:
Infant says: “ba-ba”
Mother answers: “Ma-ma. Yes, and da-da is working”

This section describes how such imitative responses can be used by the infant to finally achieve the links between the acoustic caregiver’s speech signal and the infant’s articulatory gestures, resulting in the skill of imitation.

LeVI learns to link the acoustic characteristics of CG’s speech with its own vocalizations by associating VQ indices extracted from CG’s speech directly with the phonemes babbled by LeVI itself (from hereon direct mapping, DM) or by associating VQ indices of CG’s speech with LeVI’s own VQ indices and also LeVI’s VQ indices with LeVI’s phoneme labels (from hereon indirect mapping, IM). This phase can be interpreted as the training phase of a phoneme recognition task where the number and the characteristics of the phonemes have been learned in the earlier learning phase. By babbling, LeVI requests an acoustic correspondent for his phonemes from CG, and aims thus to learn the mappings between the two.

The indirect and direct mappings are first experimented in a task, where temporal information or ordering of acoustic events are not taken into account. The direct mapping is also performed with a concept matrix [62] based approach, saving the important temporal information and finally leading to the best performance of our task.

The learning takes place during a period when LeVI babbles 1000 canonical VCVC utterances that are meaningless to itself, but are interpreted as communicative, meaningful sounds by CG that then reproduces them in a linguistically correct form. LeVI does not know the temporal spacing between the phonemes used by CG, and the phonemes in LeVI’s VCVC utterance are separated by randomly drawn values from the set of \{80, 90, ..., 200\} ms randomized for each phoneme. CG’s reproduction consists of the same phonemes but in a randomized order resulting in a VCVC or CVCV utterance, e.g. a LeVI’s utterance “amam” may be responded by CG as “mama”.

### 3.3.1 Learning mappings without temporal information

The first experiment aims to clarify the difference between the direct and indirect mapping by using a simple recognition method where temporal information present in CG’s utterances is discarded. The method results in a distribution of LeVI’s phoneme activations for CG’s individual acoustic features.

MFCC vectors are extracted from LeVI’s babble and from the reformulation by CG and quantized according to the learned codebooks. This leads to an integer number sequence corresponding to the VQ labels for each utterance. Only the unordered set of unique elements in the representations is processed further, so that differences in speech rate or utterance duration would not play a role in learning. An example VQ index sequence of [3 3 3 9 8 7 4 8 4 4 4] is thus converted into [3 4 7 8 9].

A co-occurrence frequency matrix $F$ stores the frequencies of LeVI’s and CG’s VQ indices that have co-occurred in the interactions (indirect mapping). $F$ is an $I \times C$ matrix, where $I$ is the number of VQ indices of LeVI and $C$ is the number of VQ indices of CG. In our main experiment $F$ is thus a $150 \times 150$ matrix. In principle all the matrix elements are zero, but at every step of the imitative interaction, the values of the simultaneously occurring elements indicated by the VQ indices of LeVI’s and CG’s utterances are increased by one. In the end of the period of interaction, the frequency matrix can be transformed into a probability matrix $P_{CGVQ \rightarrow I_{VQ}}$ by normalizing each column so that their values sum to one. $P_{CGVQ \rightarrow I_{VQ}}$ thus tells the probability of each LeVI’s VQ index given a certain CG’s VQ index.

Similarly, probability matrices between the CG’s acoustic features and LeVI’s phoneme models $P_{CGVQ \rightarrow IPH}$ and between LeVI’s acoustic features and LeVI’s phoneme models $P_{IVQ \rightarrow IPH}$
are learned during the process. If the frequency matrices’ rows correspond to the labels of the phonemes babbled by LeVI, after similar training the resulting probability matrices tell the probabilities of LeVI’s phonemes given a certain CG’s (or LeVI’s) VQ index. In our case of reduplicated babbling, all the unique VQ indices of CG’s (LeVI’s) utterance are thus linked to the two LeVI’s phoneme labels after every interaction. After the learning period, first the sums of each row of the frequency matrices $F_{IVQ\rightarrow IPH}$ and $F_{CGVQ\rightarrow IPH}$ are normalized to one reducing the effect of the domination of more frequently occurred targets in training data. Secondly, the sum of each column is normalized to one to obtain the final phoneme activation probabilities for each VQ index.

### 3.3.2 Learning mappings with temporal information

If temporal information is to be captured in the mapping from CG’s acoustic units to LeVI’s phoneme categories, the relative ordering of acoustic VQ indices must be taken into account. Here, we apply the weakly supervised Concept Matrix (CM) algorithm [62] that models the temporal evolution of sequential data as a mixture of bigrams measured from different temporal distances, or lags $k$. The input to the CM are the VQ sequences $X_{CG,VQ}$ corresponding to CG’s speech and LeVI’s phonemes $I_{PH}$ during the same dialogue. As a result, the algorithm learns the maximum-likelihood mapping $X_{CG,VQ} \rightarrow I_{PH}$ for each phonetic target present in the training data. The temporal dependencies in the VQ data are modeled up to 120 ms ($k = \{-9, -8, \ldots, +10\}$), approximately spanning across the duration of a single phoneme or a transition from a phoneme to another. See Appendix A for more detail.

### 3.4 Phoneme recognition and imitation using the learned mappings

The learned mappings can now be used to recognize phonemes in CG’s speech. Since LeVI’s phonemes have already acquired their meanings in articulatory domain in the supervised learning phase, LeVI is immediately able to imitate the phonemes that are recognized in CG’s speech. LeVI is thus able to invert CG’s speech acoustics into articulation with its own vocal tract, related to the phoneme gestures it has learned previously. This section describes the details of the phoneme recognition task using the indirect mapping and the direct mapping approaches.

#### 3.4.1 Recognition without temporal information

A test utterance by CG is transformed into a corresponding VQ sequence using the codebook that LeVI learned for CG’s speech sounds. In indirect mapping each CG’s VQ index corresponds to LeVI’s VQ indices with certain probabilities, and each LeVI’s VQ index corresponds to LeVI’s phoneme models with certain probabilities. The activation of each LeVI’s phoneme model $I_{PH}$ at time $t$ can be calculated indirectly as

$$A(I_{PH}, t) = \sum_{I_{VQ}=1}^{150} P_{I_{VQ}\rightarrow I_{PH}}(I_{PH}, I_{VQ}) \cdot P_{CGVQ\rightarrow I_{VQ}}(I_{VQ}, CGVQ(t))$$

where $CGVQ(t)$ is the VQ index of CG’s utterance at time $t$ and $I_{VQ}$ is a VQ index in LeVI’s codebook. In the direct mapping, the probabilities of LeVI’s phoneme activations are directly obtainable from the probability matrix $P_{CGVQ\rightarrow I_{PH}}$ for each CG’s VQ index.
In order to smoothen the model activation curves, a sliding window of length of 10 time steps (70 ms) is used to sum up the model activations of adjacent windows. The total model activation at time \( t \) is thus:

\[
A_{\text{tot}}(I_{PH}, t) = \sum_{i=t-9}^{t} A(I_{PH}, i)
\]  

(2)

3.4.2 Recognition with temporal information

In the CM approach, context activations at every time instant \( t \) are obtained by summing the learned context activations for the VQ label transitions found in CG’s utterance at different lags. The recognition method is described with more detail in Appendix A. Again, the resulting activations were smoothed using equation 2.

4 Experiments and results

The complete learning simulation is run through 10 times, and the learning of LeVI is investigated. All the 10 simulation consist of the four main phases discussed in section 3:

1. LeVI’s learning of language dependent phonemes based on online supervised learning by CG

2. Learning of auditory perceptual codebooks for LeVI’s and CG’s speech

3. Learning mappings between auditory features and between auditory features and articulatory gestures

4. LeVI’s recognition of CG’s phonemes in VCVC utterances

4.1 Experiments in phoneme learning

All the phoneme parameter values in our experiments were scaled linearly to lie in range \([0,1]\).

The parameter values considering the learning methods were \( L = 100, M = 500, R_{\text{new}} = 0.02, R_{\text{acc}} = 0.01, R_p = 0.05 \). CG’s phoneme vocabulary was set to consist of the pre-programmed phonemes /a/, /e/, /i/, /o/, /y/, /æ/, /œ/, /b/, /d/, /t/, /z/, /ŋ/, /j/. The proposed online learning method reaches the correct targets in about 60,000 iterations (equals the amount of utterances produced by LeVI). The choice of the used parameters affect the rate of learning, and use of additional qualities such as rewarding tactile feedback (in case of closures) or spectral qualities of LeVI’s own productions (in case of fricatives for example) probably speed up the learning process of real-life infants. These properties are not yet implemented in our model. Also, taking possible interdependencies of the phoneme parameters into account in a more physiologically plausible vocal tract model would reduce the degrees of freedom of the system and decrease the duration of the learning.

Figure 2 illustrates the progression of each of the phonemic learning processes. It can be seen that LeVI reaches the correct number of 8 vowel and 17 consonant targets with minimal error in about 60,000 iterations. The consonant targets are learned about 10,000 iterations later than the vowel targets. The amount of positive targets can initially have larger values than there are phonemes in the native language, because sometimes when a new random target is created and babble, it may be a better representation of CG’s target than what CG has heard previously from LeVI. Thus, LeVI gets positive feedback and saves the target, but the old target(s) corresponding to the same speech sound category (as interpreted by CG) will not be forgotten by LeVI before their lifetime decreases to zero. At this point, LeVI does not know if the new rewarded target is a better exemplar of one of already known targets or if it corresponds to a new CG’s speech sound. It is questionable if real infants learn a large set of possible phonemes before
converging to the correct amount of phones in their native language. Children are known to babble a wide variety of sounds, which are usually interpreted and transcribed in phonetic categories - defined using adult speech - in babbling related research (see e.g. [63,64]), possibly losing information about small variations or inaccuracy in infants’ babbles. However, for example Kent & Murray [58] have extracted formant frequencies of infants’ babbles in the ages of 3, 6 and 9 months, indicating that at least the number of babbled vocalic sounds can grow to exceed the amount of vowels in English language.

In all the 10 runs, LeVI’s final targets have a very small error when compared to CG’s targets. The errors do not decrease exactly to zero due to the small but ongoing reinforcement of LeVI’s targets around the category center. In a separate simulation, the phonemic learning experiment was run 100 times, and only one of the 100 runs produced 7 vowel targets for the infant, being one less than in CG’s vocabulary. The number of consonant targets was the expected 17 in all the runs. The final amount and error of LeVI’s phonemes seem to be the most sensitive to LeVI’s targets’ lifetime parameter $L$ and CG’s forgetting parameter $M$. Using our choice of parameters, the learning algorithm seems to act in an adequately stable manner. A video about LeVI’s variegated babble using the learned phonemes of one language acquisition run is provided in the supplementary material.

Figure 3 shows the two first formant frequencies of all the positively reinforced vowel targets at every 20th iteration during a single run of 60,000 iterations. For the sake of this visualization, the articulatory targets of LeVI are synthesized with the caregiver’s vocal tract to allow comparison with a typical adult F1-F2 charts. The drift towards the rewarded targets can be seen, and in the end of the training, only 8 positive targets corresponding to Finnish vowels remain.

![Figure 2](image2.png)

Figure 2. Number of LeVI’s consonant (blue dashed line) and vowel (red line) targets (left) and their RMS distances to CG’s targets (right) during 10 runs of the online supervised phonemic learning algorithm.
4.2 Experiments in LeVI’s phoneme recognition

After LeVI has learned the Finnish phonemes in the previous phase, the mappings between the articulatory and acoustic domains are trained as discussed in section 3.3. In the recognition experiments, VCVC-utterances are synthesized with the vocal tract corresponding to CG, and the activations of the phonemes at every time instant are obtained as described in section 3.4. Also, the energy of each window is calculated in the MFCC feature extraction phase, and the final activation curves are multiplied with the energy envelope. This results in the activation values during closures being almost zero, roughly estimating weaker recognition accuracy where the signal energy is low.

Only the most active phoneme model is chosen for every time frame from the resulting activation curves, resulting in an activation envelope. The total activation for each phoneme present in the activation envelope is calculated by integrating the corresponding activation values through the whole utterance. In these experiments, simply the most activated vowel and the most activated consonant are chosen as the recognized phonemes at the utterance. A detector for the exact moment of the phoneme activation has not yet been constructed. Examples of LeVI’s model activations during an utterance “oto” by CG are shown in Figure 4. The activation envelope is labeled with the maximal phoneme activations.
Figure 4. LeVI’s phoneme model activations using indirect mapping, direct mapping and the CM approach (from top to bottom) during an utterance “oto” by CG. The most activated models (activation envelope) are indicated with corresponding phoneme symbols. It can be seen that indirect mapping gives the worst result, direct mapping confuses the unvoiced alveolar consonant /t/ into a voiced alveolar fricative /z/ and alveolar nasal /n/ (with the same place of articulation). CM-method gives strong model activations for /t/ both in the closure and opening phase of the /t/ gesture. During the closure, the background noise results in MFCC features that are assimilated with /g/ or /h/, but the model activation amplitudes are almost zero. Phoneme activations are scaled linearly so that the found maximum activation gets an amplitude of one.

During each of the 10 simulations, CG speaks all combinations of vowel targets (/a/, /e/, /i/, /o/, /u/, /y/, /ae/, /oe/) and consonant targets (/b/, /d/, /f/, /g/, /h/, /k/, /l/, /m/, /n/, /p/, /r/, /s/, /t/, /v/, /z/, /ŋ/) in VCVC utterances ten times to average over the possible differences due to slightly varying fundamental frequency in CG’s synthesis. Phoneme-specific and place-of-articulation-specific recognition scores are calculated using the described direct and indirect mapping methods as well as the CM approach. With each method, the confusion matrix \( C \), where the rows define the phonemes uttered by CG’s and the columns define the recognized phonemes by LeVI, is calculated during recognition. Phoneme-specific recognition scores are obtained simply by dividing the number of the correctly recognized phoneme \( p \) with the total amount of occurrences of the phoneme \( p \) in caregiver’s utterances:

\[
S(p) = \frac{C(i_{CG}^p, i_t^p)}{\sum_{k=1}^N C(i_{CG}^p, i_t^k)}, \quad p = 1 \ldots N
\]

where \( N \) is the total number of phonemes. Place-of-articulation-specific scores are calculated because consonants may easily be confused with other consonants with the same place of articulation due to the similarity of the spectral properties around to the closure and the insensitivity of the two recognition methods towards timing or order of spectral events. Thus /k/
may be easily confused with /g/, /s/ with /t/ etc. Group-specific correct recognitions are calculated by summing the intersections of the rows and columns corresponding to /b/, /f/, /v/ and /p/ to form the bilabial and labiodental group, /l/, /t/, /d/, /s/, /z/ and /r/ to form the alveolar group and /k/, /g/ and /ŋ/ to form the velar group and dividing by the total number of occurrences in the rows corresponding to the consonant categories.

The phoneme specific recognition scores and their standard deviations are listed in Table 2, when the size of the VQ codebook was set to 150 for LeVI and CG. The standard deviations represent the deviation between different runs of the whole acquisition procedure. The complete language-acquisition procedure consists of several random components (LeVI’s utterances in the codebook creation phase, creation of the two codebooks by k-means, varying accuracy in LeVI’s learned phoneme targets, variation in LeVI’s utterances in the training phase, variation in LeVI’s and CG’s fundamental frequencies) resulting in variation between different runs of the simulation. The standard deviation in phoneme-specific recognition accuracy is rather large between different runs of the acquisition procedure for consonants, mainly because of confusion between consonants with the same places of articulation. For example, in the case of CM, when comparing the first and the second run, /t/ is recognized with accuracies of 91 and 73 per cent correspondingly, when /d/ is recognized with accuracies of 73 and 81 per cent correspondingly. Nevertheless, the variation between the average phoneme recognition scores in different runs is very small (SD next to the averages in Table 2). The averages and the standard deviations of the place-of-articulation-specific recognition scores of the 10 runs are listed in Table 3.

Figure 5 illustrates the confusion matrices in all the three methods. Since the phonemes with the same place of articulation are placed close to each other in the graph, the confusion between these phonemes can be seen as regions of activations around the diagonal. In the confusion matrices it can be seen that in indirect and direct mapping several CG’s consonants are recognized as the glide /j/. This occurs mostly in cases where the vowel in CG’s utterances is /i/, having similar parameter values and thus acoustic qualities as /j/. Therefore, several consonants before and after the closure are recognized as /j/ before /i/ is activated.
Table 2. LeVI’s average phoneme recognition scores (in percentages) of 10 runs of speech acquisition simulation using indirect mapping, direct mapping and CM, when the VQ codebook size is 150 for CG and LeVI. The standard deviations represent the differences between the 10 complete simulation runs.

<table>
<thead>
<tr>
<th>Phoneme Indirect mapping</th>
<th>Direct mapping</th>
<th>CM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Vowels</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/a/</td>
<td>100.00</td>
<td>0.00</td>
</tr>
<tr>
<td>/e/</td>
<td>100.00</td>
<td>0.00</td>
</tr>
<tr>
<td>/i/</td>
<td>100.00</td>
<td>0.00</td>
</tr>
<tr>
<td>/o/</td>
<td>100.00</td>
<td>0.00</td>
</tr>
<tr>
<td>/u/</td>
<td>99.88</td>
<td>0.37</td>
</tr>
<tr>
<td>/y/</td>
<td>99.94</td>
<td>0.19</td>
</tr>
<tr>
<td>/ae/</td>
<td>100.00</td>
<td>0.00</td>
</tr>
<tr>
<td>/oe/</td>
<td>99.18</td>
<td>1.65</td>
</tr>
<tr>
<td>/j/</td>
<td>99.88</td>
<td>0.37</td>
</tr>
<tr>
<td>/h/</td>
<td>96.50</td>
<td>11.07</td>
</tr>
<tr>
<td>Velars</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/g/</td>
<td>44.25</td>
<td>29.56</td>
</tr>
<tr>
<td>/k/</td>
<td>60.13</td>
<td>21.94</td>
</tr>
<tr>
<td>/ŋ/</td>
<td>94.13</td>
<td>4.97</td>
</tr>
<tr>
<td>Alveolars</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/n/</td>
<td>90.88</td>
<td>8.82</td>
</tr>
<tr>
<td>/l/</td>
<td>43.13</td>
<td>14.07</td>
</tr>
<tr>
<td>/r/</td>
<td>17.25</td>
<td>28.93</td>
</tr>
<tr>
<td>/s/</td>
<td>49.00</td>
<td>14.75</td>
</tr>
<tr>
<td>/z/</td>
<td>16.25</td>
<td>21.70</td>
</tr>
<tr>
<td>/t/</td>
<td>22.25</td>
<td>15.90</td>
</tr>
<tr>
<td>/d/</td>
<td>7.75</td>
<td>7.16</td>
</tr>
<tr>
<td>Bilabials / labio-dentals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/m/</td>
<td>13.38</td>
<td>11.41</td>
</tr>
<tr>
<td>/v/</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>/f/</td>
<td>20.75</td>
<td>20.17</td>
</tr>
<tr>
<td>/p/</td>
<td>61.38</td>
<td>21.49</td>
</tr>
<tr>
<td>/b/</td>
<td>3.50</td>
<td>7.63</td>
</tr>
<tr>
<td>Avg.</td>
<td>57.58</td>
<td>2.16</td>
</tr>
</tbody>
</table>

Table 3. Place-of-articulation-specific recognition scores (in percentages) for consonants using indirect mapping, direct mapping and CM. Average and standard deviation after 10 complete language acquisition runs.

<table>
<thead>
<tr>
<th>Phoneme category</th>
<th>Indirect mapping</th>
<th>Direct mapping</th>
<th>CM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Velar</td>
<td>81.79 (SD=6.66)</td>
<td>60.08 (SD=3.19)</td>
<td>96.29 (SD=1.12)</td>
</tr>
<tr>
<td>Alveolar</td>
<td>49.31 (SD=11.06)</td>
<td>78.90 (SD=3.01)</td>
<td>98.48 (SD=1.34)</td>
</tr>
<tr>
<td>Bilabial / labio-dental</td>
<td>71.78 (SD=12.29)</td>
<td>84.34 (SD=2.49)</td>
<td>95.66 (SD=1.51)</td>
</tr>
<tr>
<td>Avg.</td>
<td>67.63</td>
<td>74.44</td>
<td>96.81</td>
</tr>
</tbody>
</table>
4.3 Discussion of the experiments

The results show that LeVI learns to recognize and reproduce CG’s vowels with a very good accuracy using the proposed methods. The direct mapping leads to higher phoneme recognition accuracy than the indirect method, but the temporal information preserving CM approach is clearly the most powerful method in the task.

The bad recognition score for /h/ is due to the fact that in most of the cases /h/ is recognized as the surrounding vowel sound, and no consonant gets activation during the utterances. This is due to the fact that in our vocal tract model /h/ is defined as only a change in the glottal excitation, and MFCC coefficients as the feature representation give more value to the vocal tract resonances than the glottal source.
The recognition scores calculated for the consonant categories based on the place of articulation are on average 67.63 percent for the indirect mapping and 74.77 percent for the direct mapping. The correct phoneme class is more easily confused than the place of articulation probably due to the similarity in the spectral transitions before and after the consonant produced with the same constriction. In these methods of recognition the transitions of longer duration dominate the quick effects of $V_{on}$ or $V_{off}$ for example. Also, in indirect mapping, in case of some consonants, inaccuracy in the learned targets may cause a large difference in the vocalization. For example if LeVI does not learn a complete closure for consonant /t/, turbulence occurs, and the following feature vectors are similar to /s/. This may lead to LeVI linking CG’s phoneme /t/ with its spectral qualities corresponding to /s/. Taking more accurate physical phenomena, such as tactile feedback, into account in the articulatory model, would presumably not only speed up the learning process, but as well avoid such confusion in phoneme recognition.

The better success of the direct mapping methods proposes that infants may learn to associate their articulation directly with caregivers’ speech and so called vocal tract normalization or methods for normalizing acoustic qualities of different speakers may not be necessary by the infant in order to imitate. That is, the caregiver’s voice does not have to be linked with the infant’s voice when the caregiver’s speech is directly interpreted in the articulatory domain of the infant. In the indirect case, the additional acoustic-to-acoustic mapping causes extra distortion to the signal that is to be mapped to the articulatory domain, resulting in worse phoneme recognition. The direct CM-approach, mapping CG’s acoustic features in LeVI’s phoneme models using also temporal information present in the utterances, results in very high phoneme-specific recognition scores, at least when recognizing the calmly paced CG’s VCVC utterances in our experiment.

In order to compare recognition with different amount of clusters in the VQ codebooks, similar simulations were run with codebook sizes of 100 and 200. The average phoneme recognition scores were correspondingly 54.00% (SD=1.34) and 58.42% (SD=2.50) for IM, 58.71% (SD=1.78) and 64.96% (SD=1.02) for DM, 86.25% (SD=1.03) and 90.13% (SD=1.03) for CM. Using the selected amount of training data it thus seems that increasing the amount of VQ elements to 200 increases the recognition accuracy somewhat. It is expected that if the codebook sizes would be increased too much, the generalizing ability of the recognizer would suffer because of specializing too much to the exact feature vectors present in the training data.

The quality of the speech produced by the developed articulatory model has not yet been evaluated by human listeners or a standard speech recognizer, but is left for future work. However, in this study the primary goal has been to demonstrate possible principles governing the spoken language learning ability of infants even though the quality of the produced speech is strongly “distorted” when compared to a typical human speech. However, based on informal listening tests, the vowels can easily be classified correctly, and consonant distinctions can also be perceived. Lacking the accurate prosody, the synthetic speech could be thus classified as a sort of a dialect, or a very personal way of speaking Finnish. An example of LeVI’s babbling using the learned Finnish phonemes is made available in the supplementary material.

5 Conclusions

In this work we have simulated a language acquisition experiment using a learning virtual infant (LeVI) with no preceding language knowledge and a virtual caregiver (CG) whose speech parameters and vocabulary are manually defined to approximate spoken Finnish. LeVI acquires speech in three phases: (1) learning language-dependent articulatory gestures, or phonemes, (2) perceptual development towards commonly occurring acoustic features in CG’s and his own speech, and (3) learning the mapping between LeVI’s articulation and CG’s speech. An
articulatory synthesizer including lip radiation losses, nasal sounds, the effect of lip lengthening, and offering a possibility of animated speech synthesis, was created for the purpose.

The learning procedure follows few basic principles described in related literature. Children are known to adapt their vocalizations due to contingent feedback by parents [30] and parents are reported to subconsciously give more positive feedback on closely correct articulations [31]. Parents are also reported to imitate their children more than vice versa [33], presumably due to the lack of vocalic imitation skills of young infants. In our work, it is assumed that the caregiver can invert and imitate infant’s speech, and rewards advanced vocalizations with more positive feedback. LeVI learns language-specific articulatory gestures corresponding to phonemes using only positive feedback as guidance. After learning, LeVI babbles reduplicated consonant-vowel alterations, which CG interprets as communicative, and answers to the vocalizations by imitation. This is the step where LeVI learns to link its articulations with CG’s acoustic productions and thus learns the skill of imitation of articulatory sequences of ever increasing complexity.

The learned mappings were tested using a simple phoneme recognition test, where LeVI tries to find correct correspondents from its phoneme vocabulary for CG’s utterances. The best mapping method recognizes CG’s vowel sounds perfectly, and the consonants’ place of articulation with an average accuracy of 96.81%. Average phoneme-specific recognition score is 88.42% when averaged over 10 independent language acquisition simulations. Direct mapping (CG’s acoustic features into LeVI’s articulation) results in the highest scores when compared to indirect mapping (CG’s acoustic features into LeVI’s acoustic features and further into LeVI’s articulation), indicating that the additional acoustic-to-acoustic transformation causes unnecessary distortion to the signal that is to be interpreted in the articulatory domain. The apparent stability of the method and the good recognition scores may indicate that similar principles govern a real speech acquisition process.

Speaker or vocal tract normalization is not needed at the learning phase, where LeVI learns to map his articulations in acoustic features present in CG’s speech directly. In the indirect mapping experiment, the mapping between CG’s and LeVI’s acoustic features may be considered a non-linear vocal tract normalization mapping, but the best learning results were obtained with the direct methods.

Our proposed method of speech inversion or imitation does not need articulatory-acoustic codebooks or frame-by-frame optimization of articulatory trajectories. When compared to the study by Howard & Messum [49], our method does not require dynamic time warping in order to recognize caregiver’s speech and the imitative responses by the caregiver do not have to be exactly equal to LeVI’s utterances during the learning process. Because of the simple working principles of the simulation, the methodology is clear and easily replicated with whichever vocal tract model or articulatory synthesizer. In addition, our caregiver is also simulated (as opposed to a human caregiver), all the phonemes present in Finnish language are modeled and recognition results are reported in a large-scale evaluation test.

In our current experiment CG imitates LeVI by vocalizing the same phonemes as LeVI but in arbitrary order, i.e. the imitation does not have to be exactly similar to LeVI’s utterance. This leads to the fact that at every imitative interaction LeVI associates his uttered phonemes always with the correct phonemes, but also with incorrect ones, by CG. With enough training data the correct representation will dominate in the activation frequency matrices and correct recognition can still be performed. The imitative utterances could include moderate amounts of additional random information or phonemes at the cost of slowing down of the learning process.

The biggest drawback of the current learning method is the lack of accurate speech recognition mechanism for CG and thus the provision of LeVI’s exact articulatory parameters to CG. In real life the caregiver would interpret infants’ sounds based on the acoustic signal and he may not be able to tell a difference between articulatory configurations that lead into acoustic
signals that fall into the same phonemic category known by the caregiver. If this would be the case, in the end of the phoneme learning phase, LeVI would presumably be left with a number of phonemes corresponding to a single CG’s phonemic category, and in the imitation phase LeVI would associate CG’s phonemes into a larger number of possible phoneme categories when compared to CG’s phoneme vocabulary. However, LeVI’s phonemes could presumably be categorized into a correct number of phonemes in the native language, since after learning the function (i.e. role in CG’s acoustic domain) of each phoneme, phonemes having the same function could be clustered together. This issue has been studied more closely in [65]. Even though the phonemic learning phase in the current work is a simplification of a real phoneme acquisition process, we propose that a similar mechanism may help real infants to acquire at least rough estimations of the caregiver’s phoneme categories.

Since during this simulation, LeVI learns to map his own acoustic features, as well as CG’s acoustic features, into his own articulatory phoneme categories, we propose that the learning mechanism could allow for passing on of the phonetic system from generation to generation: LeVI could be used as the new caregiver with a lengthened vocal tract, and a new oblivious infant with original LeVI’s vocal tract dimensions could be introduced to be trained. The similar learning mechanism would lead the new infant to learn almost exact phonemes to LeVI, and the new infant would learn to recognize LeVI’s phonemes with similar accuracy to these simulations. The phonemes would drift from their original positions in the course of time because of the small allowed discrepancy between the infant’s and the caregiver’s phoneme productions. However, for the generation-wise learning to happen, LeVI would still have to obtain direct information about the new infant’s articulatory parameters as was done in the current work, since LeVI learns only to recognize the exact articulation of phonemes with the obtained 88% accuracy. LeVI would thus not be able to give clear feedback signals on the new infant’s babbles that are not close enough to the correct articulations in the exploration phase due to the limitations of the used phoneme recognition method.

A real-life language acquisition process is presumably much more complex in nature as the simulated situation. In real language acquisition all the phases present in the proposed method are likely to overlap: possible mappings may be learned at the same time with possible phonemic targets, the auditory receptive fields may be updated in real time during the whole process, and some articulations may remain unlearned when some caregiver’s speech can already be imitated. The caregiver may not be able to extract exact articulations from infant’s speech, complicating the phonemic learning phase. The perception of speech sounds begins probably with perception of some more abstract surface features when compared to phonemic sounds [35], and this may also guide the infant to distinguish which of the caregiver’s utterances was the actual imitative one, or if imitation happened in the first place. Tactile feedback in the articulatory system probably guides more accurate articulatory target learning, and provides important information about reaching closures for instance. Inclusion of some of the missing details might make the speech acquisition easier for LeVI, and some might make it more complex, but the objective of this study was to draw a general outline of the speech acquisition situation, that will be advanced further in future research.

Many acoustic-to-articulatory inversion methods concentrate on sampling the articulatory space of the vocal tract model into a codebook, and estimating articulatory trajectories using certain optimality criteria for the movements of the articulators and match between acoustic features (e.g. [8,11,12]). It seems that the speech inversion problem becomes much more easily solvable if language-dependent criteria are taken into account. When the speech sounds of the native language in question are normally produced by a certain learned set of articulatory gestures, the problems caused by the many-to-one property faced in the frame-by-frame inversion approaches are largely diminished. The acoustic features or even the articulations of the infant’s speech sounds do not have to match the caregiver’s ones exactly; they just have to be
good enough to enable success in communication as evaluated by the caregiver. The resulting articulatory trajectories are not only based on the physiology of the speech production mechanism but very importantly also on the native language of the caregiver - the evaluator and the guide of the language learning.

In future, recognition tests will be made when a new talker with a differing voice is introduced. Infants are known to distinguish words more accurately when they occur in speech produced by similar voice to the voice the words were originally familiarized with. For example Houston and Jusczyk [66] report that 7.5-month-old infants showed capacity to generalize words across two speakers of same gender, but not across speakers of different sex. By 10.5 months generalization over the opposite sexes could be done. This is supposedly the case also in our virtual infant, since the MFCC features of the new talker may differ drastically from the original virtual caregiver. Our idea to learn the new mapping into articulation is to use the context information available in the situation of interaction. If the new voice for example says “mama”, and the child is aware that the physical object (mother) is in question, he can link the new observed acoustic features with the articulatory targets he has previously used for the same object. Thus, the new acoustic signal “mama” could be mapped into infant’s articulatory targets /mama/ using the context as a medium (see, e.g., [67], for context based merging of acoustic variants of a word in infant’s lexicon).

The real-life interactive situation between an infant and a parent is very complex, including touching, facial expressions, eye contact, infant-directed speech, differences in word stress, word-repetition, etc., and thus a realistic interaction with a virtual infant is very difficult to reproduce in computer simulations. A real person could be used as the caregiver in the phonemic learning process but currently the limitations and the unnaturalness of the interaction with a computer would make it too an exhausting experience. The use of human caregiver in the imitative learning process (as has been done in [49]) would be more plausible and will be left for future experiments. However, with more physiologically realistic models for the vocal tract, hearing and intelligent memory architectures, a virtual infant capable of human-like language acquisition and speech recognition might be closer to reality.

6 Acknowledgements

This study was supported by the graduate school of Electronics, Telecommunications and Automation (ETA), Finnish Foundation for Technology Promotion (TES), KAUTE foundation and the Nokia foundation. The authors would like to thank Bart de Boer for his valuable comments on the manuscript.

7 References


[65] H. Rasilo, B. de Boer, Virtual infant’s online acquisition of vowel categories and their mapping between dissimilar bodies, Submitted to Interspeech 2013.


Appendix A - The concept matrix algorithm

The temporal information preserving recognition algorithm used in this work is a variant of the concept matrix (CM) algorithm by Räsänen & Laine [62]. This appendix describes in more detail how the algorithm is used in our experiments.

Training phase

Given a sequence of the caregiver’s VQ labels $X=[a_t, a_{t+1}, \ldots, a_{t+m}]$ corresponding to the caregiver’s imitative VCVC or CVCV utterance, and LeVI’s “context tags” $C=\{c_1, c_2, \ldots, c_{N_C}\}$ that are activated concurrently with $X$, the training algorithm counts the occurrences of element pairs in $X$ at lags $k = \{k_1, k_2, \ldots, k_K\}$ into frequency matrices $F_{k,c}$ so that a transition at lag $k$ corresponds to a transition from $a_t$ to $a_{t+k}$. Every matrix $F_{k,c}$ is a $150 \times 150$ matrix due to the number of VQ indices in the codebook for the caregiver, and there are a total of $N_C \times K$ of such matrices to account for all values of $c$ and $k$. The contexts used in our experiments are the phoneme categories corresponding to the vowel and consonant gestures babbled by LeVI, so that during each iteration the same set of transitions will be stored to the matrices of the two contexts $c_i$ and $c_j$ corresponding the vowel gesture $i$ and consonant gesture $j$.

Recognition phase

When the training phase has been completed, the frequency matrices corresponding to all gestures $c$ and lags $k$ are normalized to represent the joint probabilities of element pairs:

$$P(v_i, v_j, c, k) = \frac{F_{k,c}(v_i, v_j)}{\sum_{i=1}^{N_A} \sum_{j=1}^{N_A} F_{k,c}(v_i, v_j)} \quad (A.1)$$

Where $N_A$ represents the alphabet size (150 in our experiments), and $v_i$ and $v_j$ refer to the elements in the frequency matrices. Next, the obtained probability matrices are normalized over all gestures resulting in the activation probability of gesture $c_n$ when transition $i \rightarrow j$ occurs at lag $k$:

$$P^c(c_n | v_i, v_j, k) = \frac{P(v_i, v_j, c_n, k)}{\sum_{n=1}^{N_C} P(v_i, v_j, c_n, k)} \quad (A.2)$$

where $N_C$ represents the total number of gestures (the 25 phonemes in our experiments). This provides the maximum likelihood estimate that each lagged element pair occurs in a gesture $c_n$ under a uniform prior assumption for $c$. Now, when a new input signal $X = [a_1, a_2, \ldots, a_M]$ is to be recognized, the activation of each gesture is calculated at every time instant using

$$A(c_n, t) = \frac{1}{K_{total}} \sum_{m=1}^{K} P^c(c_n | a_t, a_{t+k_m}, k_m) \quad (A.3)$$

with the constraint $t + k_m > 0$ and $t + k_m \leq M$. $K_{total}$ corresponds to the number of lags that were used in total taking into account the mentioned constraint, and limits the activation values at each time window between zero and one.
Captions

Figures

Figure 1. Mid-sagittal image of the vocal tract model (left) and the vowel triangle generated by random sampling through possible non-nasalized vocal tract shapes. The green diamonds mark the Finnish vowels as defined into the caregiver’s vocal tract model. The red circles mark the positions of typical Finnish vowel sounds [57].

Figure 2. Number of LeVI’s consonant (blue dashed line) and vowel (red line) targets (left) and their RMS distances to CG’s targets (right) during 10 runs of the online supervised phonemic learning algorithm.

Figure 3. Formants from all positive vowel targets at every 20th iteration during a run of 60,000 iterations of the phonemic learning are shown with small blue dots. The plus signs indicate the formant frequencies of the targets produced by CG. Xs indicate the (scaled) formant frequencies of LeVI after the learning process. For this illustration LeVI’s synthesis output is produced with a normalized vocal tract length to allow visual comparison to CG’s vocabulary. It can be seen that LeVI has explored the articulatory space and finally drifted towards CG’s targets.

Figure 4. LeVI’s phoneme model activations using indirect mapping, direct mapping and the CM-approach (from top to bottom) during an utterance “oto” by CG. The most activated models (activation envelope) are indicated with corresponding phoneme symbols. It can be seen that indirect mapping gives the worst result, direct mapping confuses the unvoiced alveolar consonant /t/ into a voiced alveolar fricative /z/ and alveolar nasal /n/ (with the same place of articulation). CM-method gives strong model activations for /t/ both in the closure and opening phase of the /t/ gesture. During the closure, the background noise results in MFCC features that are assimilated with /g/ or /h/, but the model activation amplitudes are almost zero. Phoneme activations are scaled linearly so that the found maximum activation gets an amplitude of one.

Figure 5. Confusion matrices of all three recognition methods. The areas of the black squares are related to the frequency of corresponding caregiver’s uttered phoneme - LeVI’s recognition pair: the bigger the area, the more matches in the element. The three bordered squares are the category boundaries for the consonants grouped due to their place of articulation. Number 2 refers to /æ/, number 3 to /œ/ and number 4 to /ŋ/.

Tables

Table 1. List of all properties needed to define a phoneme in our work.

Table 2. LeVI’s average phoneme recognition scores (in percentages) of 10 runs of speech acquisition simulation using indirect mapping, direct mapping and CM, when the VQ codebook size is 150 for CG and LeVI. The standard deviations represent the differences between the 10 complete simulation runs.
Table 3. Place-of-articulation-specific recognition scores (in percentages) for consonants using indirect mapping, direct mapping and CM. Average and standard deviation after 10 complete language acquisition runs.