Virtual infant's online acquisition of vowel categories and their mapping between dissimilar bodies

Heikki Rasilo [1,2], *Okko Räsänen* [1], *Bart de Boer* [2]

[1] Aalto University, School of Electrical Engineering, Department of Signal Processing and Acoustics
[2] Vrije Universiteit Brussel, The Artificial Intelligence Laboratory
heikki.rasilo@aalto.fi, okko.rasanen@aalto.fi, bart.de.boer@ai.vub.ac.be

Abstract

In order to understand how humans learn speech imitation without access to detailed articulatory data of other talkers, simulated speech acquisition experiments between two virtual agents were carried out with the goal of maintaining the interaction between the two as natural as possible. As an outcome, a novel model of infants' vowel acquisition is presented. In the experimental setup, a virtual infant learns vowels in interaction with a virtual caregiver: it babbles vowels randomly, the caregiver answers every babble with an utterance that contains the vowel uttered by the infant in addition to other vocalic content, and the infant associates its own productions to the caregiver's responses. The infant and the caregiver have different vocal tract sizes, and hence the acoustic qualities of the same vowel differ between the infant and the caregiver. The infant learns on line to map acoustic qualities of its caregiver's speech onto its own vowel articulations, allowing for instant imitation of the caregiver's vowel sounds when recognized. As opposed to previous computational studies of vowel acquisition, the infant does not need initial mappings, initial vowel primitives, or knowledge of the caregiver's vowel categories.

Index Terms: speech acquisition, vowel learning, imitation

1. Introduction

Speech inversion techniques aim to find underlying articulatory trajectories from acoustic speech signals. Humans generally perform well when inverting their native language, which also shows as an ability to imitate speech. Since human infants learn the skill of speech imitation at a young age without access to exact information about articulation, we have decided to approach speech inversion related issues naturally from the point of view of infant speech acquisition. In this work, by *speech inversion* we mean human-like reproduction of heard speech, possibly with a different vocal tract (= imitation), instead of finding detailed articulations of individual speakers.

In [1], we showed how an initially naive Learning Virtual Infant, LeVI, equipped with an articulatory model could learn to produce and recognize native phonemes in interaction with a virtual caregiver, CG, who had already mastered the Finnish phonetic system. Importantly, the learned phonemes were learned during CG-supervised babbling phase, and thus the phoneme categories, in terms of which the caregiver's speech was later recognized, were primarily articulatory by nature. Speech inversion thus happened as a by-product – when LeVI recognized a phoneme in CG's speech, LeVI had immediate access to the underlying articulatory phonetic gesture.

In this work, we describe a novel scenario of vowel acquisition where the interaction between CG and LeVI is made more natural for improved cognitive plausibility. Our results could have repercussions for adaptive speech recognition, but are also meant to provide insight into the cognitive processes of speech production, perception, and acquisition.

1.1. Models of articulatory learning and automatic speech recognition

Modern speech recognition systems use certain features and representational units extracted from acoustic speech signals as a basis for word recognition. However, the used features are mainly chosen due to their performance in the prevailing speech recognition architectures so that they perform well in speech-to-text conversion, most often leading to the use of triphone HMM-models together with MFCCs, RASTA-PLP, LPC, or other fixed-frame spectral features [2]. These systems work properly as long as they have narrowly defined goals, but their limitations become clear when dealing with more complex levels or aspects of language. How humans learn the important features in speech signals, what the sub-word level representations that humans use to code speech are (c.f., [3]), and what the role of speech production mechanisms in human speech recognition is, are often given less attention.

It is likely that human infants start learning longer, word-like sequences from speech signals, but later words can be parsed as shorter *building blocks,* such as syllables, diphones or phones, on which combinatorial speech structures can also be produced [4]. Since speech communication is not only hearing and understanding, but also speaking, it is possible that children gradually learn to analyze acoustic speech signals for features that most robustly serve *all* the aspects of speech, including combinatorial production and grammatical requirements. Since speech production explicitly requires discrete choices between different motor actions, it may also play a role in the organization of the sub-word level perception of speech, and provide means to associate acoustically varying percepts into linguistically meaningful phonetic categories.

When considering recognition of purely acoustic speech (i.e., not having access to the visual modality that importantly affects recognition as well), we do not believe that inverting speech and using the resulting articulatory representation brings additional gain to speech recognition when compared to recognition based on the acoustic signals, *as long as the correct acoustic features are used in recognition*: performing reliable inversion requires a proper set of acoustic features, and these features could probably be used directly to train speech recognizers. If the acoustic speech signals did not contain all the necessary information needed for recognition, humans would not be able to recognize speech robustly either. Instead, understanding and modeling the articulatory development through speech learning simulations using articulatory models may provide the necessary constraints for finding the best possible acoustic level descriptors for spoken language, used also by humans, as well as shed light on human speech learning mechanisms. In our current work, as well as in [1], phones are learned due to speech production, and acoustic signals are parsed for features that best discriminate between found phone classes.

1.2. Previous research on infant speech acquisition

When infants acquire speech, they find themselves facing a correspondence problem: which speech sounds created by their small and underdeveloped vocal tracts correspond to the sounds uttered by their caregivers? The acoustic qualities of the same speech sounds in these two differing bodies may vary drastically. Previous research (e.g. [5–11]) has shown that the caregiver can help the infant to draw the links between the two representations in imitative interactions.

Research on infant language acquisition has shown that infants cannot imitate their caregivers' speech sounds before at least six months of age ([11,12]). According to [13], both infants' and mothers' verbal imitation was almost nonexistent at 10 months of age but exceeded all other imitative actions during the second year. Mothers regularly imitated the infant more (vocally and verbally) than vice versa (see also [11]). Parents also regulate their feedback on infants' babbles based on the quality of their vocalizations [14] and infants are shown to regulate their babbles based on parental feedback [15].

Several studies have investigated the mechanisms governing infants' speech acquisition using computational modeling. However, in most of the studies, the learning situation is heavily simplified from real-world speech learning situations, diminishing the cognitive plausibility of the models. The following paragraphs list previous research on the subject, and their main drawbacks. Our work aims to simulate vowel acquisition without the initial assumptions used in previous research, as concluded in the last paragraph of this section.

A number of papers provide possible methods of how speech of other speakers can be imitated without dealing with the problem of *dissimilar bodies* (e.g. [5,16]), i.e., assuming that the sound to be imitated is created with a vocal tract similar to the one of the imitator, which is not the case in real infants' language acquisition. In our work the vocal tract sizes of the interactors differ.[6–8,10] used dissimilar bodies and imitation by the caregiver to approach the correspondence problem. In [7–8,10], the imitative utterance of the caregiver had the same phonetic content as the learner's utterance. In [6] it is reported that the human caregiver imitated the infant with reproductions if the utterances sounded native, otherwise they were ignored. In addition, [10] assumes that the language learning robot knows a predefined set of vocal primitives whereas [8] assumes that the robot knows the desired caregiver's vowel categories and has a rough estimate of the mapping function between the two acoustic domains.

In [9], the caregiver and the learning robot have dissimilar bodies and the caregiver does not have to imitate the learner with exactly the same utterances. In training, a human caregiver imitates the learner's vowels with different probabilities. The learner is given 15 vowel primitives in the beginning, each having a probability of being recognized as the corresponding vowel. The learner succeeds in learning at and above chance level of being imitated.

In our previous work [1], LeVI learned to map acoustic characteristics of Finnish phonemes (including consonants), spoken by CG, to its own articulations in a dissimilar body in an imitation scene where erroneous associations were also allowed to a certain extent. Before the imitation phase, LeVI had already learned the articulatory gestures corresponding to the native phones in a supervised learning phase, where CG awarded more adult-like babbles. However, learning of the exact articulation of phones using only reinforcement is a strong assumption – human caregivers are not likely to distinguish, and maybe do not even pay attention to, very small differences in the infant's articulation of phones if the acoustic output is sufficiently close to a known phonemic category. In our work this was possible because CG recognized LeVI's speech based on directly provided articulatory data. In the current study, CG's recognition is based on LeVI's acoustic speech output and the supervised learning phase is discarded. Similarly in [17], a robot learns to map vowel sounds of human caregivers, presented in utterances with additional vowels, onto articulatory vowel categories that were previously learned in a reinforcement learning phase.

In an elegant study, [18] proposes that an infant can learn a mapping between the differing adult's and infant's acoustic spaces by clustering the two spaces separately and using the topological correspondence between them. Non-imitative adult feedback on the infant's attempted imitations is used to find the best topology for the clustering. The clustering is performed on synthesized sets of speech data created in a similar babbling procedure for both speakers, leading to the assumption that the distribution of speech sounds is similar for both speakers at the moment of the clustering – i.e., the correct mapping does not evolve gradually "on-line". Also, Plummer [19] has studied vowel normalization by aligning perceptual manifolds created for the caregiver's and infant's vowel spaces based on imitative interactions. The interactions are assumed to consist of the infant's imitations of the caregiver's vowels, confirmed correct by the caregiver, but in practice imitation data is selected manually for the interactors.

In the current study we simulate LeVI's vowel acquisition when: LeVI and CG have vocal tracts of different sizes, CG imitates LeVI's babbles using his native language but not exactly matching phonetic content (imitations include several additional phonemes), no initial vocal primitives or reinforcement learning phases are used but the vowel categories are learned online during the simulation, no initial "rough" mappings are used, LeVI does not know the number of CG's vowel categories nor does it know how to articulate phonemes before learning. The minimal amount of initial assumptions and the speech recognition method, make this work a novel contribution to related research.

2. The method

In this work, every interaction between LeVI and CG consist of the following steps. First, LeVI babbles an open vocal tract configuration, which CG hears and interprets as one of his already known native vowel prototypes (using *the perceptual magnet effect* [20]). LeVI compares the auditory perceptual characteristics of its babble to those of its previous babbles, and based on the perceptual distance, adds the articulatory configuration of the babbled speech sound to an already existing vowel category or a completely new category. CG imitates LeVI using a string of phonemes, of which one is the native vowel as which LeVI's babble was interpreted. LeVI associates the acoustic features found in CG's imitative answer to the vowel category activated during its own production. LeVI will learn several vowel categories, but after multiple interactions, they end up being the most sensitive towards those features in CG's speech that occurred most frequently concurrently with the babbled categories: *CG's knowledge of the vowel domain is thus transferred to LeVI based on CG's imitations of LeVI's initially meaningless babbles.* The method is described with more technical detail in section 2.3.

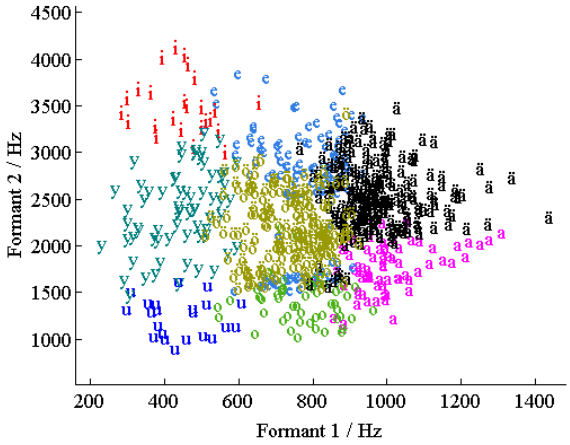

Figure 1. *758 vowels babbled by LeVI and annotated into Finnish vowel categories*

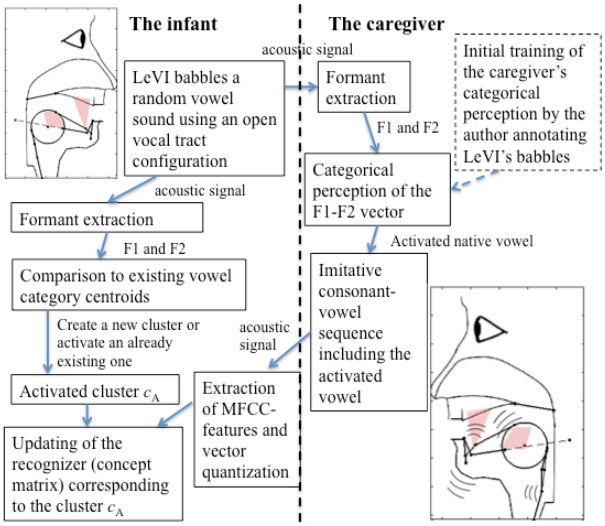

Figure 2. *A flow chart of the interactive learning process*

2.1. Used assumptions and methods

LeVI and CG produce sounds using an articulatory synthesizer developed and described in earlier work [21]. The *look-ahead* model of articulation aims to reach articulatory target positions defined in a 9-dimensional articulatory space, with additional parameters controlling the dynamic characteristics of the articulatory movements. The model is able to produce vowels, nasals and consonants, as well as to synthesize strings of phonemes. In the current experiments LeVI babbles using a vocal tract of 10 cm in length and a fundamental frequency varying slightly around 250 Hz. CG has a vocal tract length of 17.5 cm and a fundamental frequency around 120 Hz.

LeVI learns to recognize CG's speech based on statistical distributions of vector quantized MFCC-features, whereas LeVI's speech sounds are categorized using the two first formant frequencies, extracted with linear predictive analysis from vocal tract impulse responses. In the preliminary experiments, formant representation for LeVI's acoustic domain is used because it allows easy visualization of the emergent vowel space as well as easy visual evaluation of CG's categorical vowel perception. The same MFCC-representation could presumably be used for both parties for better cognitive plausibility, but this is left for future work. MFCC-features are used for the caregiver's speech because we do not want to assume that LeVI is able to extract formant frequencies from continuous caregiver's speech also including consonants. Thus, a more general auditory perceptual representation is preferred. This representation also makes LeVI's speech recognition mechanism more robust for future experiments with real recorded speech signals. The distance between two formant vectors is calculated as Euclidean distance, with weighting factors of 1/1500 and 1/4500, denominators estimating the maximum values of F1 and F2 correspondingly, to compensate for the bigger range of F2 values.

LeVI's recognition of CG's speech is based on a variant of the concept matrix (CM) algorithm [22], a weakly supervised pattern recognition algorithm that approximates temporal structures of higher-order Markov processes. CM is used as its training is one-pass and computationally straightforward. LeVI recognizes CG's speech in terms of activations of clusters $\mathbf{C}=\{c_1,c_2,\ldots,c_C\}$ storing articulatory configurations for vowels learned during the simulation. The acoustic correspondents in CG's speech for each cluster are learned based on CG's imitations and consist of distributions of transitions between acoustic features (vector quantized MFCC-vectors) at different lags. The vector quantization, training and recognition of the CM-recognizers are described with more detail in Appendix A.

2.2. Initial training of the caregiver's perceptual categories

In order to have CG recognize LeVI's vocalizations in a similar way to a human listener, an initial training phase was performed for CG. LeVI was set to babble[1] 1000 CVCV utterances. The first author listened to every utterance and annotated it with the perceived consonants and vowels. In 758 cases out of the 1000, a clear vowel was perceived and annotated. The definition of a "clear" vowel here is of course highly subjective: in almost all cases, a vowel was categorized as one of the Finnish vowels, but some productions sounded like borderline cases between two vowels or simply sounded incorrectly pronounced and thus they were left unannotated. However, the annotation process results in a clear F1-F2 *vowel triangle* with

[1] Throughout this work, a babbled vowel is defined as a random vocal tract configuration in the 9 dimensional articulatory space with a minimum allowed cross-sectional area of 0.1 cm^2 and a babbled consonant as a vocal tract configuration where a complete closure was obtained. The randomization of eight parameters is slightly biased towards the extreme values of the articulators, helping the infant to explore its full range of articulations. One of the nine parameters, the velum, is always set to either fully open or closed.

[2] Categorization of the perceived consonant sounds was also experimented with, but they turned out to be much more difficult to categorize automatically with simple rules into the correct categories in both the articulatory and the acoustic sense. Even though in many cases the place of articulation was a clear indication of a consonant class, in some cases the effect of other gestural parameters, for example velocities of the articulators, can cause the percept to change into a category that would not be obvious looking just at the place of constriction. For example, in some cases, percepts of /k/ could be created with the tongue tip. These effects can be partly caused by lack of natural detail in the articulatory model, but they can also be normal "ventriloquist" effects, just not used in normal Finnish speech. Because we were not yet able to teach the virtual caregiver human-like consonant recognition for LeVI's speech, the current work focuses on vowel acquisition.

reasonably clear boundaries between the vowel categories (see Figure 1). The obtained categorization was used in the further simulations by making CG recognize LeVI's future vowel sounds[2] using a *k*-nearest neighbor (kNN) classifier (k = 6 neighbors used for majority voting) based on these initially annotated babbles, using the Euclidean distance between the weighted two first formant frequencies as a distance measure.

2.3. LeVI's learning of vowel categories and their mapping to the caregiver's vowels

The learning process of LeVI goes on with the following steps (see Figure 2). **1)** LeVI babbles a vowel sound with a probability of 0.5, or tries to reproduce one of the already learned vowel clusters with a probability of 0.5. In the latter case, LeVI always chooses a randomly stored articulatory parameter vector from the selected cluster with the smallest number of possible articulations stored in it. This is done in order to actively attempt to find the matching sound from CG's speech when a new category is learned. In order to introduce a small random error to LeVI's intended reproductions, uniform random noise from a range of [-0.1, 0.1] is added to the 9 parameter values when the parameters' coordinate values are linearly scaled into a range [0, 1]. **2)** LeVI extracts the two first formant frequencies of its production. **3a)** If this is the first production by LeVI, a new cluster centroid is placed in the F1-F2 space in the corresponding location, and the articulatory parameter vector that created the vowel is stored in this cluster and selected as the *cluster centroid*. An empty frequency matrix **F** is created for the cluster centroid for the CM-recognizer (see Appendix A). **3b)** If there already exist cluster centroids in LeVI's memory, the distance of the obtained F1-F2 vector is calculated to all the cluster centroids using the weighted Euclidean distance (see section 2.1.). If the distance to the closest cluster exceeds a threshold of t_c = 0.1, a new cluster is created similarly as in 3a. Otherwise the new articulatory parameter vector is stored in the closest cluster. In both cases, 3a and 3b, the updated cluster is called the *activated cluster* c_A. **4)** CG extracts the two first formant frequencies of LeVI's babbled vowel sound and uses the kNN algorithm to classify it in a vowel category (see section 2.2). This is performed in order to obtain "human-like" perception of the correct vowel category. **5)** CG creates an utterance consisting of P Finnish phones, where phones alternate between consonants from a set of {/k/, /t/, /p/, /g/, /d/, /b/, /s/, /m/, /n/, /ŋ/, /v/, /f/, /l/} and vowels from a set of {/a/, /e/, /i/, /o/, /u/, /y/, /ä/, /ö/}. CG places the recognized vowel of LeVI's babble into one random vowel position in the utterance. Thus, the CG's answer might look like /abik**u**tup/, when P = 8 and the bolded /u/ is the recognized vowel from LeVI's babble. **6)** CG speaks the created utterance. **7)** LeVI hears the acoustic signal of CG's utterance, and extracts vector quantized MFCC-features from it using the method described in the Appendix A. **8)** LeVI updates the frequency matrix corresponding to the cluster c_A with the transitions present in the complete observed VQ-sequence, as described in the training phase in Appendix A.

3. Results

During the simulation, LeVI's success in recognition is measured for every interaction between LeVI and CG. The measurement is made in order to see LeVI's progress in interpreting the CG's vowels during the learning process, and it is performed by a "third-party observer" that does not affect the learning process itself. First, 200 test probes are synthesized with CG's vocal tract, each probe consisting of a sequence of 17 phones alternating between the 8 Finnish vowels and random Finnish consonants. This is done in order to test LeVI's recognition of vowels in varying phonetic contexts depending on the surrounding consonantal gestures. At every iteration of the acquisition process, LeVI's frequency matrices are normalized using the equations (A1) and (A2), and a randomly chosen test probe is recognized using equations (A3) and (A4). This results in activation scores of LeVI's current vowel clusters for the test probe based on how LeVI is trained so far. The most activated LeVI's clusters at the time instants of the 8 vowels, $\mathbf{t}_{vowels}$ (stored for sake of the recognition measurement during the synthesizing of the probes), in the recognized test probe are found as $\mathrm{argmax}_n\, A_{smooth}(c_n, \mathbf{t}_{vowels})$ (see Appendix A) and the formant frequencies corresponding to the cluster centroids are transformed into vowel classes by kNN clustering with the initial annotated data by the author, as described in section 2.2. If the obtained vowel classes match with the vowels in the test probe, it means that if LeVI were to imitate the CG's vowel sounds using the articulations corresponding to the most activated cluster *centroids*, the resulting acoustic signals would again be interpreted as the same vowels by CG. However, if the most activated LeVI's phonemic category crosses a border between two or more of the CG's perceptual categories, not *all* of the articulations stored in the cluster will lead to correct imitation as judged by CG, but generally LeVI's categories of these kinds will have weaker activations for the CG's vowel sounds and are less likely to be the most activated centroids (depending on the range of LeVI's categories defined by the threshold t_c). The score for each interaction is obtained by the number of the matching vowels divided by the total number of vowels (8).

Two simulations were run until 3000 babble-answer interactions were reached. In the first simulation, CG answers using 6 phonemes (P = 6) and in the second one, using 10 phonemes (P = 10). Figure 3 shows the development of LeVI's vowel recognition accuracy during the simulation, as well as the number of clusters obtained by LeVI. The final numbers of clusters are 43 and 39, respectively, and the final recognition accuracy for the vowels present in the probes is about 95% when they are recognized based on the classification of the cluster centroid. Somewhat unexpectedly, there does not seem to be significant difference in the learning rates between the two individual runs. Presumably there is big variance between different runs of the simulation caused by the differing positions of LeVI's phonemic clusters and, on average, we expect

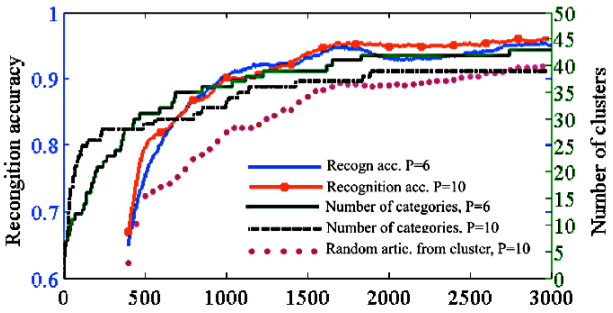

Figure 3. *Running average over 400 interactions, showing LeVI's recognition accuracy of all 8 vowels for P=6 (blue solid line) or P=10 (red dotted line), when LeVI imitates using articulations corresponding to cluster centroids. The lines with step-like shape show the number of obtained categories (P=6 with the green solid line and P=10 with black dashed line). The purple dashed line shows the success of LeVI's imitation when P=10 and LeVI chooses a random articulation from the recognized cluster instead of the cluster centroid.*

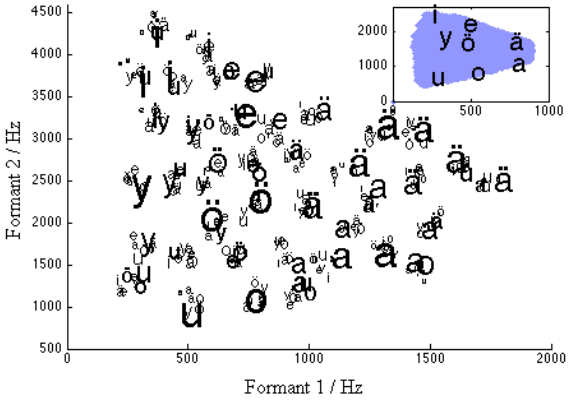

Figure 4. *LeVI's cluster activations for the 8 vowels of the caregiver and the caregiver's prototype vowels, plotted over the caregiver's vowel triangle (top right corner).*

that the learning rate is smaller when the proportion of matching phonemes in CG's answers is larger.

Figure 4 shows the vowel triangle and the vowel prototypes of CG, as well as LeVI's final vowel triangle when P = 10. The activations of LeVI's phonemic clusters are indicated as letters when CG utters the corresponding vowel in isolation. The font size is related to the activation of the corresponding cluster. A small amount of displacement in the F1-F2 dimensions is added for every cluster activation for every CG's vowel in order to show activations for different clusters more clearly. It is clear that the CG's vowels activate LeVI's clusters in corresponding areas, and that several clusters situated close to each other obtain large activations for certain of CG's vowels. This indicates that some clusters could be joined together based on their simultaneous activations, thus reducing the final number of LeVI's phonemic categories.

4. Conclusions

A method allowing a learning virtual infant, LeVI, to acquire a vowel system based on imitative interactions with a virtual caregiver, CG, was introduced. LeVI and CG have different vocal tract sizes, and thus the acoustic vowel spaces of the two are dissimilar. The phonetic content of the imitative answer by CG does not have to match exactly the content of LeVI's initial babble, and LeVI does not have previous knowledge on the number or characteristics of CG's phonemes, nor the characteristics of its own productions. The results suggest that a caregiver simply repeating a vowel, initially babbled by the infant, in a longer utterance may be sufficient for infants to learn an initial set of vowel categories.

Since the infant's vowel categories are learned while *producing* speech sounds, they are articulatory by nature. This means that in this work the entities into which acoustic events in the caregiver's speech are recognized, exist only in the articulatory domain. When the caregiver's speech sounds are recognized, articulatory configurations stored in the activated vowel cluster can be directly used for imitation. If infants learn speech sound categories due to speech production, and learn to parse acoustic signals for the acoustic features that co-occur with these produced categories (be they phones, or underlying gestures of individual articulators, for example), we could expect that hearing these acoustic features activate the articulatory domain. Some studies have shown that listening to speech indeed activates speech production related motor areas in the brain [23].

The proposed method of associating articulations with CG's speech in this study is purely distributional in nature – the infant does not try to detect the imitative part in the caregiver's utterance, but simply assumes that the babbled vowel exists somewhere in the utterance. Although in real language acquisition, other cues such as intonation or rhythmic patterns may help to locate the part imitated by a caregiver (see [17]), making learning of more complex utterances feasible, the current results show that even with very minimal assumptions, the basic sound system of a language can already be acquired.

In these experiments LeVI also explores its acoustic space merely by babbling on its own. In more advanced simulations the infant might use a slowly emerging mapping between the acoustic domains of the infant and the caregiver to make hypotheses on the locations of heard caregiver's vowel sounds in its own acoustic space, and try to produce them systematically, speeding up the articulatory exploration.

In this work, LeVI learns a significantly bigger number of phonemic categories than the caregiver has vowel categories. This is not considered a problem, since when the learning continues, activations in *different* phonemic categories of LeVI that are interpreted as the *same* vowel by the caregiver will eventually get similar activation matrices due to similar content in the imitations (this appears in Figure 4 as several of LeVI's vowel clusters situated next to each other gaining maximal activations for the same caregiver's vowel), and could presumably be reliably clustered together in later phases, thus reducing the final number of LeVI's phonemic categories.

So far we have not analyzed the characteristics of the articulatory parameter vectors stored in the learned phonemic clusters. Due to the many-to-one characteristic of articulatory-acoustic mappings, there exist several possibilities to pronounce each of the learned vowel categories. The current goal was not to find one "correct" way to pronounce each phoneme: people are known to utter the same phonemes slightly differently depending on adjacent phonemes [24]. Allowing a range of different ways to pronounce phonemes may explain several speech related phenomena, such as speech rate effects or compensatory articulation, e.g., in case of constrained articulations using bite blocks [25,26]. Excessive articulatory effort could be used to rule out some improbable articulations belonging to the same phonemic category. Although having some freedom in articulatory parameters therefore seems realistic, we would like to explore the effect of clustering of the obtained articulatory configurations in future work.

This study has shown that a simple distributional learning algorithm may be sufficient for infants to learn the mapping between the vowel productions of two dissimilar vocal tracts with a minimal amount of initial assumptions. In reality, the interactive situation between the infant and the caregiver may contain several additional cues to make the learning process faster and more robust.

Since speech is produced by combining more or less discrete articulatory building blocks, we propose that during the learning of combinatorial speech production in infancy, humans learn to parse continuous acoustic speech signals for features that best categorize them in these discrete, production-conditioned, categories. Current ASR systems use features that are refined to work best in limited tasks when compared to human capabilities. We propose that simulating human-like speech learning processes may help to define acoustic features that work best in robust, human-like speech recognition tasks.

5. Appendix A

Vector quantized MFCC representation. In order to be able to vector quantize CG's speech in discrete labels, LeVI first listens to 250 synthesized grammatically correct, but lexically randomized, Finnish sentences, from a vocabulary of 34 words. MFCC-features (coefficients 1 to 12) are extracted from every sentence with a step size of 5 ms and a window length of 25 ms using Hamming windowing. When the listening process is finished, 50,000 MFCC-feature vectors are randomly chosen from all MFCC-vectors and clustered into 150 clusters using a standard k-means algorithm. Then the MFCC-features of CG's further speech are quantized into these 150 categories and labeled with corresponding integer numbers.

Training of the CM recognizers. Given a sequence of CG's answer's VQ-indices $X = [x_t, x_{t+1}, \ldots, x_{t+m}]$, energy at every analyzed window $E = [e_t, e_{t+1}, \ldots, e_{t+m}]$ (normalized so that the maximum energy in the signal gets a value one) and LeVI's activated cluster c_A, occurrences of element pairs in X at lags $\mathbf{l} = \{l_1, l_2, \ldots, l_L\}$ are counted and added into frequency matrices $\mathbf{F}_{l,c_A}$. A transition at lag l corresponds to a transition from x_t *to* x_{t+l}. In this work the value added to $\mathbf{F}_{l,c_A}(x_t, x_{t+l})$ is equal to e_{t+l} in order to diminish the effect of training on weak parts of the signal. This work uses $L = 20$ lags $\mathbf{l}=\{-9, -8, \ldots, 10\}$, and in the end of the training there will be $L \times C$ matrices of size $N_V \times N_V$, where C is the final number of clusters known by LeVI and $N_V = 150$ the number of VQ labels used.

Recognition with the CM recognizers. When the training of the models is complete, all frequency matrices are normalized to represent the joint probabilities of their element pairs (v_i, v_j):

$$\mathbf{P}(v_i, v_j, c, l) = \frac{\mathbf{F}_{l,c}(v_i v_j)}{\sum_{x=1}^{N_V} \sum_{y=1}^{N_V} \mathbf{F}_{l,c}(v_x, v_y)} \quad \text{(A1)}$$

The obtained probability matrices are normalized over all clusters, providing a maximum likelihood estimate that a lagged element pair (i, j) occurs for cluster c_n, when a uniform prior probability is assumed for $\mathbf{c}$:

$$\mathbf{P}^{\mathbf{C}}(c_n | v_i, v_j, l) = \frac{\mathbf{P}(v_i, v_j, c_n, l)}{\sum_{m=1}^{C} \mathbf{P}(v_i, v_j, c_m, l)} \quad \text{(A2)}$$

A new input VQ-sequence $X = [x_1, x_2, \ldots, x_M]$ provides an activation for each cluster at time instant t by

$$A(c_n, t) = \frac{1}{L_{total}} \sum_{m=1}^{L} \mathbf{P}^{\mathbf{C}}(c_n | x_t, x_{t+l_m}, l_m) \cdot e(t + l_m) \quad \text{(A3)}$$

with $t + l_m > 0$ and $t + l_m \leq M$. L_{total} is the number of lags that could be used in the current window and limits the activation values at each time window between zero and one. The model activation curves are smoothed by summing activations in a sliding window of 20 time steps (120 ms):

$$A_{smooth}(c_n, t) = \sum_{i=t-19}^{t} A(c_n | i) \quad \text{(A4)}$$

6. References

[1] Rasilo H., Räsänen, O. and Laine, U.K., "Feedback and imitation by a caregiver guides a virtual infant to learn native phonemes and the skill of speech inversion", SpeCom, in press, http://dx.doi.org/10.1016/j.specom.2013.05.002.

[2] Huang, X., Acero, A., and Hon, H.W., Spoken language processing: a guide to theory, algorithm, and system development, New Jersey, Prentice Hall PTR, 2001.

[3] Port, R., "How are words stored in memory? Beyond phones and phonemes", New Ideas in Psychology, 25, 143–170, 2007.

[4] Werker J. and Curtin S., "PRIMIR: A developmental framework of infant speech processing", Lang Learn Dev, 1, 197–234, 2005.

[5] Markey, K.L., "The sensorimotor foundations of phonology: a computational model of early childhood articulatory and phonetic development", Ph.D. Thesis, Univ. Colorado, Boulder, 1994.

[6] Howard, I.S. and Messum, P., "Modeling the development of pronunciation in infant speech acquisition", Motor Control 15(1): 85–117, 2011.

[7] Yoshikawa, Y., Asada, M., Hosoda, K., and Koga, J. "A constructivist approach to infants' vowel acquisition through mother–infant interaction", ConSc, 15(4): 245–258, 2003.

[8] Miura, K., Yoshikawa, Y., and Asada, M, "Unconscious anchoring in maternal imitation that helps finding the correspondence of caregiver's vowel categories", Advanced Robotics, 21(13), 1583–1600, 2007.

[9] Miura, K., Yoshikawa, Y., and Asada, M., "Realizing being imitated: Vowel mapping with clearer articulation", 7th IEEE ICDL 2008, 262–267, 2008.

[10] Vaz, M.J.L.R.M., "Developmentally inspired computational framework for embodied speech imitation", PhD Thesis, Universidade do Minho, Escola de Engenharia, 2009.

[11] Kokkinaki, T. and Kugiumutzakis, G., "Basic aspects of vocal imitation in infant–parent interaction during the first 6 months", J Reprod Infant Psyc 18(3): 173–187, 2000.

[12] Jones, S.S., "Imitation in infancy: The development of mimicry", Psychological Science 18: 593–599, 2007.

[13] Masur, E.F., Rodemaker, J.E., "Mothers' and infants' spontaneous vocal, verbal, and action imitation during the second year", Merrill-Palmer Quart. 45: 392–412, 1999.

[14] Gros-Louis, J., West, M.J., Goldstein, M.H. and King, A.P., "Mothers provide differential feedback to infants' prelinguistic sounds", Int J Behav Dev 30(6): 509–516, 2006.

[15] Goldstein, M.H. and Schwade, J.A., "Social Feedback to Infants' Babbling Facilitates Rapid Phonological Learning", Psychological Science 19(5): 515–523, 2008.

[16] Westermann, G., and Reck E.R., "A new model of sensorimotor coupling in the development of speech", Brain and language, 89(2): 393–400, 2004.

[17] Hörnstein, J., Gustavsson, L., Santos-Victor, J. and Lacerda, F., "Modeling speech imitation", IROS-2008 Workshop-From motor to interaction learning in robots, Nice, France, 2008.

[18] Ananthakrishnan, G. and Salvi, G., "Using Imitation to learn Infant-Adult Acoustic Mappings", in Proceedings of Interspeech 2011, 765–768, 2011.

[19] Plummer, A.R., Aligning manifolds to model the earliest phonological abstraction in infant-caretaker vocal imitation, proceedings of Interspeech 2012, Portland, Oregon, USA, 2012.

[20] Kuhl, P.K., "Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not", Percept. Psychophys. 50: 93–107, 1991.

[21] Rasilo, H., "Articulatory model for synthesizing sequences of arbitrary speech sounds or pre-programmed Finnish phonemes", supplementary material for [1], 2013, available at `http://dx.doi.org/10.1016/j.specom.2013.05.002.`

[22] Räsänen, O. and Laine, U., "A method for noise-robust context-aware pattern discovery and recognition from categorical sequences", Pattern Recognition 45: 606–616, 2012.

[23] Wilson, S.M., Saygin, A.P., Sereno, M.I., Iacoboni, M., "Listening to speech activates motor areas involved in speech production", Nature Neuroscience 7: 701–702, 2004.

[24] Öhman, S.E.G., "Coarticulation in VCV utterances: Spectrographic measurements", Journal of the Acoustical Society of America 39: 151–168, 1966.

[25] Guenther, F.H., "Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production", Psychological Review 102(3): 594–621, 1995.

[26] Guenther, F.H., "Cortical interactions underlying the production of speech sounds", Journal of communication disorders, 39(5): 350–365, 2006.

[i] This study was supported by the ETA graduate school of Aalto University, Finnish Foundation of Technology Promotion (TES), KAUTE foundation and the Nokia foundation. The authors would like to thank Unto K. Laine and Hannah Little for valuable comments on the paper.