interspeech 2008 incorporating SST 08

# Vocal imitation in early language acquisition

*Lisa Gustavsson and Francisco Lacerda*

Department of Linguistics, Stockholm University, Stockholm, Sweden
lisag@ling.su.se, frasse@ling.su.se

## Abstract

This paper presents a study of vocal imitation during the early stages of the language acquisition process. Utterances were extracted from recordings of adult-infant interactions in controlled but naturalistic experimental settings. For each recording session, utterances were used to create pairs of adult-infant samples that were presented to a panel of listeners, whose task was to judge whether the samples in a pair could be considered as imitations of each other or not. The results suggest an age-dependent hierarchy for the impact of different phonetic dimensions on imitation judgments and provide a basis for a quantitative model of vocal imitation.

**Index Terms**: language acquisition, vocal tract, imitation, developmental robot, information processing, equivalence classification

## 1. Introduction

Although the concept of imitation is regarded as one of the cornerstones in infants' early language acquisition, vocal imitation, in particular, calls for a more stringent definition since the few studies on word/vocal imitation reported in the literature also are somewhat inconsistent. Imitation can of course have many different purposes or consequences; adult vocal imitation can be used for impersonating, entertaining or as a social mechanism [1]. Infant vocal imitation is seen in very early communicative turn-taking games [2] but it's not trivial to determine exactly what a good imitation should sound like since there are several dimensions of imitative behavior. For this reason imitative behavior in infant-adult interaction was here assessed perceptually asking a panel of listeners to judge recordings of utterances from dyads of infant-adult interaction as "imitations", "maybe imitations" or "not imitations". The results were analyzed in terms of which acoustic phonetic parameters that seemed to guide the listeners in their judgments. The imitation model based on these phonetic parameters is currently being used to sketch a mathematical model that is evaluated as an articulatory reinforcement component of a developmental humanoid robotic system. A new imitation judgment experiment will be carried out using the robot and its results will be compared with those of human listeners. If we have managed to isolate the key acoustic parameters the robot is expected to perform in agreement with the listeners. The mismatch between the human and the robot results will however allow us to re-assess the parameters and adjust their weights. By using the parameters in this way the robot should be able learn his audio-motor map faster and also in a less controlled environment and reveal some of the tuning processes towards speech.

## 2. Background

It has been shown that infants as young as 12 days can imitate both facial and manual gestures, i.e. infants equate their own unseen behaviors with gestures they see others perform [3]. Also vocal imitation has been documented in infants from 12 to 20 weeks of age [4]. Obviously because of anatomical and physiological differences between adults and infants, a perfect imitation, in the sense of acoustically similar realizations (see [5] for an illustration of the infant vs. the adult articulatory space) is simply not possible. Therefore, the infant's ability to imitate adult utterances seems to require understanding of the underlying equivalence between the infant's own utterances and the adult's utterances rather than just the ability to match physically identical sounds. The assumption behind this study is that this kind of understanding is not necessarily genetically coded in the newborn. Admittedly the infant needs to practice in order to achieve the understanding of the underlying equivalence of its own utterances and the adult's, as suggested by the very meager imitative behavior found in infants in their very early stages of language acquisition among the speech materials used for this paper. [6] reviews several longitudinal infant studies indicating that the number of spontaneous vocal imitations follows the same slow trend of other spontaneous word productions during the first year of life. During the second year imitations add up to about 20-40 percent of the infants' productions. And again, the difficulty for an observer to decide what defines an imitation is obvious.

Another study reveals that it's not only infants that imitate adults; also adults imitate their 2 to 5-month-old infants [7]. In the current study we consider that imitations take place even when the adult imitates the infant, which is in fact even more frequent in our data than infants imitating the adult. This adult behavior can be seen as a way to corroborate the infant's utterances and to provide an efficient updating of the infant's acoustic map for different voices that implicitly conveys information on the underlying equivalence between the infant's utterances and the adult's utterances. Until now the robotic system has been babbling randomly to explore his acoustic space and map sound to his corresponding articulatory movements and positions. The challenge for the robot or any other first language learner is to discover the essential linguistic regularities in the signal to handle the problem of mapping phonetically equivalent speech sounds that may be acoustically very different when uttered by different speakers. To evaluate the features from the listening experiment we will let the robot use these parameters to match his own vocalizations to the reinforcement vocalizations/imitations from different caregivers. Our aim with these studies is thus to explore which are the important acoustic features in early vocalizations that will guide the young infant in converging towards speech.

# 3. Method

## 3.1 Speech material

The speech data base, used in both the imitation judgment experiment and the imitation modeling, was obtained from naturalistic adult-infant interaction situations with seven Swedish infants participating in one, two or three half-our sessions each, altogether 15 sessions at ages ranging from 185 to 628 days. These recordings were made in a comfortable home-like environment, in a recording studio at the Phonetics Laboratory, Stockholm University. The speech signals from the infant and the adult were recorded in separate channels via wireless lavalier microphones clip-mounted on the shirt (adult) and mounted on a vest that the infant wore during the session. Thus the infant and the adult were free to move around in the studio and they were also provided with a number of toys. The sessions were also video filmed from two different angles and recorded on DVD. This naturalistic experimental strategy was adopted in order to increase the probability of observing speech imitation behavior, enabling also the study of natural, interactive behavior and the possible mutual convergence towards imitation targets[i].

The audio files were subsequently annotated using the WaveSurfer software (http://www.speech.kth.se). Each recording was labeled in two separate tracks marking both the infant's and the adult's utterances/vocalizations. In order to obtain short separated utterances to create the speech data base the audio files were split in sequences corresponding to the labels and named according to their relative timing in the audio files. In total, these recordings generated an adult-infant interaction speech data base consisting of 4100 speech samples.[ii]

## 3.2 Procedure

20 subjects were requested to judge whether or not the infant's utterance could be an imitation or an attempt of imitation of the adult's utterance. The subjects listened to the stimuli presented via headphones. The stimuli were presented in pairs, where the first element was an adult utterance and the second element was an infant utterance. There were three possible answers: "Yes" an imitation, "Uncertain" or "No" definitely not an imitation. The subjects responded by clicking buttons on the screen corresponding to the answer they wanted to give. The program[iii] (see figure 1) created pairs of stimuli for presentation by picking at random an utterance from the pool of adult utterances and a random utterance from the pool of infant utterances corresponding to that adult and session. To increase the likelihood of an actual imitation being uttered the infant's utterance was drawn from among the utterances that the infant had produced within five seconds before or after the adult's utterance. The subject's reaction time, the stimuli included in each pair and their order of presentation in the test session were automatically logged by the program, along with the subject's judgment of the pair of stimuli. The listening sessions were organized in sets of three different infant age groups, 185-296, 360-457 and 544-628 days, consisting of 50 presentations each. The subjects were never informed about these three age-groups. In total each subject listened to 150 (adult, infant) pairs of randomly selected stimuli within the age groups from the data base of adult and infant utterances and also meeting the relative timing restrictions described above. Because the stimuli were randomly selected throughout the test session, any given pair could be presented several times within one session.

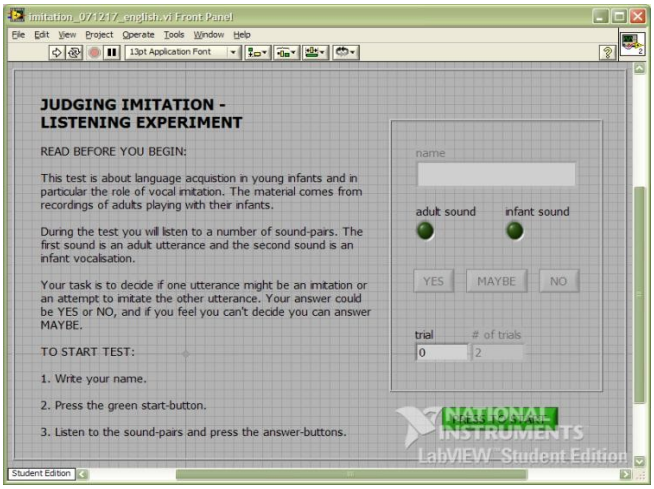


Figure 1: *The interface of the imitation judgment program.*

# 4. Results

The extreme pairs, that were considered to be sure imitations and the pairs that were judged as very unlikely imitations, were chosen for an acoustic analysis. In table 1 a few examples are listed. The utterances followed by "_a" are adult utterances and those followed by "_i" are the infant's. The first column shows the adult utterances and the corresponding infant utterances. "Syllable" stands for the number of CV-syllables observed in the utterance. The "contour" column presents the F0 contour of the utterance, where "L" indicates relative low pitch and "H" stands for relative high pitch. The "duration" column shows the overall duration of the utterances, in seconds. The column "spectral" presents a crude orthographic notation of the utterances, as an indication of how the utterances sound like. Finally, the columns "kvot1", "kvot2" and "kvot3" display the length of the first, second and the third syllables, relative to the utterance's overall duration.

| sound(match) | syllable | contour | duration(sec) | spectral | kvot1 | kvot2 | kvot3 |
|---|---|---|---|---|---|---|---|
| 295385_a | 2 | LH | 0.487 | tetu | 55 | 45 | |
| 298624_i | 2 | LL | 0.482 | tutu | 62 | 38 | |
| 301725_i | 2 | LH | 0.527 | ketu | 74 | 26 | |
| 299806_a | 2 | LH | 0.823 | epu | 52 | 48 | |
| 301725_i | 2 | LH | 0.527 | ketu | 74 | 26 | |
| 796208_a | 1 | LHL | 0.679 | bu | | | |
| 799540_i | 1 | LHL | 0.313 | bu | | | |
| 801950_i | 1 | LHL | 0.509 | bu | | | |
| 825024_a | 3 | LLL | 1.750 | bububu | 38 | 25 | 36 |
| 828186_i | 3 | HLL | 1.530 | bababu | 62 | 23 | 15 |
| 901507_a | 3 | LLH | 0.865 | aeapa | 33 | 56 | 11 |
| 904070_i | 2 | LH | 0.486 | papa | 68 | 32 | |
| 965340_a | 1 | HL | 1.013 | u | | | |
| 966798_i | 1 | HL | 0.948 | u | | | |
| sound(mismatch) | syllable | contour | duration(sec) | spectral | kvot1 | kvot2 | kvot3 |
| 291689_a | 2 | HL | 0.529 | kuka | 73 | 27 | |
| 294122_i | 2 | LH | 0.462 | tedo | 64 | 36 | |
| 302938_a | 3 | LLH | 1.100 | tikuka | 36 | 43 | 21 |
| 310522_i | 1 | HL | 0.273 | ka | | | |
| 811575_a | 2 | LH | 1.091 | upu | 27 | 73 | |
| 821431_i | 2 | LH | 0.496 | auwa | 46 | 54 | |
| 904899_a | 2 | LH | 0.616 | apa | 71 | 29 | |
| 914786_i | 1 | HL | 0.214 | ga | | | |
| 404813_a | 1 | LH | 0.319 | stäng | | | |
| 405918_i | 1 | LHL | 0.160 | daow | | | |
| 412811_i | 2 | HL | 0.469 | köka | 62 | 38 | |
| 410923_a | 3 | HLH | 0.406 | vemede | 34 | 18 | 48 |
| 412811_i | 2 | HL | 0.469 | köka | 62 | 38 | |
| 965340_a | 1 | HL | 1.013 | u | | | |
| 971373_i | 4 | LLHH | 1.202 | tatatata | | | |

Table 1: *Examples of utterances reaching significant scores for imitation (match) or not imitation (mismatch).*

The common acoustic parameters among the good imitations were the number of CV-syllables observed in the utterances, the pitch contours and crude spectral distributions. Also the duration of the utterances and the length of the first, second and the third syllables, relative to the utterance's overall duration, seemed to guide the listeners in their judgments. These duration ratios also give some indication on the relative prominence of the syllables they refer to.

The results of the analysis of the acoustic patterns in pairs like these are summarized as a flowchart in figure 2. An overall similar rhythm or timing seems to govern all parameters in the judgment of similarity between the infant's utterances and the adult's. But the concrete features we will focus on throughout these studies are: The number and length of CV-syllables, pitch contours and spectral/phonetic quality. Other factors that are not considered in the acoustic analysis but seemed to be important were intensity and non-speech characteristics such as laughing or sad voices.

Another very important aspect of judging imitation is the age of the infant (or rather how advanced they are in their production) that seems to affect the listeners judgments in such a way that the older the infant are the higher are the demands on matching parameters. Not only in quantity but also the quality, for example matching spectral characteristics of the adult model may not be required for a six month old infant but maybe so for a 18 month old. To illustrate this we could take the familiar example of [baba] that happily is rewarded as an imitation of both [mama] and [papa] when the infant is very young, but with time the correct syllable structure and vowels are not good enough for an imitation and the infant has to pinpoint the correct consonants to get the same positive feedback from the adult. In table 2 the yes- and no-imitation answers are summarized according to infant age. There were no differences between the age groups regarding yes- or no-imitations. This was a bit of a surprise at first because we thought that imitations would get better as the infants grew older, but then again probably the demands for a successful imitation increases. But in agreement with the longitudinal infant studies examined by [6] just over 20 percent of the production pairs drawn from this database (within time spans of five seconds) are considered to be imitations. In 19 percent of the trials, the listeners couldn't judge if there was an imitation or not.

Table 2. *The amount observed imitations in the listening experiment. All numbers are in percent. Of all the perceived utterance-pairs in the database 22 percent were judged to be imitations. There were no significant differences between the three age-groups (a=185-296, b=360-457 and c=544-628 days).*

| 22 | | | 59 | | |
|---|---|---|---|---|---|
| YES imitation | | | NO imitation | | |
| a | b | c | a | b | c |
| 34 | 32 | 34 | 31 | 33 | 36 |

Indeed the results are imitations in both directions as the panel of listeners was presented with adult-infant utterance-pairs that could be drawn from a situation of either infant vocalizing after the adult or the adult speaking after the infant. In fact 52% of all the utterance-pairs judged to be imitations is the adult imitating the infant.

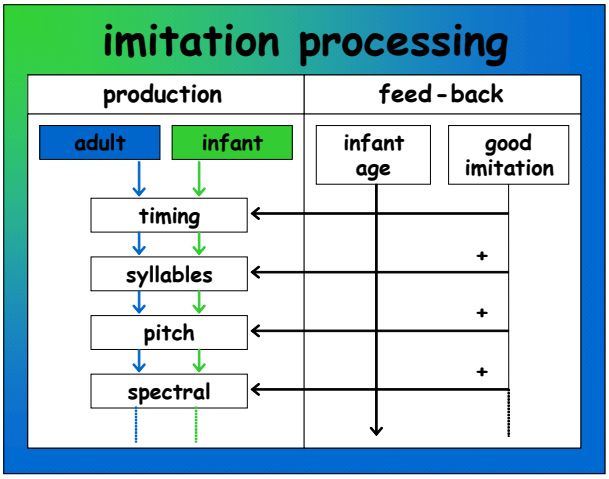


Figure 2: *A schematic summary of the preliminary results from the acoustic analysis of the imitation judgment answers.*

## 5. Discussion

Obviously these results indicate how listeners judge the infant's imitations of an adult model or the adult's imitations of infant vocalizations and reflect the subject's perceptual sense of what a successful imitation might be. In a naturalistic situation adults are likely to react to an infant's utterances just as the panel of listeners did in this study (with 91% agreement in their judgments). In other words, these parameters are valid for an adult reinforcing infant vocalization but the infant perspective on judging imitation is of course nothing we explicitly could assess in this study. Nevertheless considering that these parameters reflect very rudimentary features of speech they might be used as more general sound-processing tools in the pre-linguistic infant. The syllable is in our model described as an opposition between sound and silence and pitch is defined as an opposition between high and low frequency.

As mentioned earlier, speech imitation behavior must deal with substantial and unavoidable differences between the acoustic characteristics of the adult's utterances and the infant's vocalizations. At this point an approach based on an algorithm that would perform a computational evaluation of the similarity between these two very differently sounding speech signals is not available and would have to be calibrated against subject's auditory judgments anyway. For these reasons a listener panel evaluation of imitation seemed preferable because it would provide a deeper insight on how listeners define the important acoustic characteristics of imitation in speech. Under this assumption it can be expected from the present results that adults will tend to provide positive feedback to an infant who for example matches the number of syllables of an adult model and uses a generally adequate pitch contour. In the same way we make the assumption in our model that infants will regard such matches from the adult as equivalent to their own sounds to update their acoustic map.

These are potentially important results that allow a creation of a realistic algorithm to describe and simulate speech imitation behavior in the vocal interaction between young infants and the adult environment. Incidentally, these results provide also a clear coupling to the notions put forward by [8] and [9]. They consider the potential significance of the initial vocal behavior of infants, biological functions such as jaw opening/closing as origin of words. With proper feedback on the right "speech-like" parameters in

the adult sense these initial non-speech vocal behaviors will unavoidably tune the infant's articulatory gestures and vocalizations in the direction of speech. The process of learning to imitate seems to be incremental in its nature, that is initially any vocalization might be considered as an imitation and give enough feedback from the adult to encourage the infant in the imitation game, but once the infant has expanded its articulatory repertoire the demands increases for a successful imitation.

## 6. Conclusions

The acoustic parameters among the good imitations were the number of CV-syllables observed in the utterances, the pitch contours and crude spectral distributions; also an overall similar rhythm or timing pattern seems to govern all parameters. The results also suggest an age-dependent hierarchy (or rather how advanced they are in their production) for the common acoustic parameters among the imitations. At the moment we are evaluating the parameters from the listening experiment in the robotic system by creating a classifier to match his vocalizations to the reinforcement vocalizations/imitations from different caregivers. Our aim with these studies is thus to assess the impact of these acoustic features and explore how they can guide the young infant in converging towards speech.

## 7. Acknowledgements

Research supported by The Bank of Sweden Tercentenary Foundation (MILLE, K2003:0867), grant from The Knut and Alice Wallenberg Foundation (2005.0115) and by EU-NEST (CONTACT, proj. 5010).

## 8. References

[2] Bloom, K. (1988): Quality of adult vocalizations affects the quality of infant vocalisations. *Journal of Child Language* 15:469–80.

[5] Gustavsson, L, Lindblom, B, Lacerda, F and Eir Cortes, E (2006): From movements to sound - Contributions to building the BB speech production system. CONTACT Review Meeting, Genova, November 14-15, 2006 http://eris.liralab.it/contact/docs/discussion-movements-to-sound.pdf

[4] Kuhl, P and Meltzoff, A (1996): Infant vocalizations in response to speech: Vocal imitation and developmental change. *The Journal of the Acoustical Society of America*, Vol. 100, No. 4, 2425-2438

[8] Lacerda, F and Lindblom, B (2006): On bootstrapping BabyBot's speech production. CONTACT Design Meeting, Genova, May 23, 2006 http://eris.liralab.it/contact/reporting-period-1/DELIVERABLES/d0101-review.pdf

[9] MacNeilage, P and Davis, B.L (2000): On the origin of internal structure of word forms. *Science*, 288, 527-531.

[3] Meltzoff, A. & Moore, M. (1977): Imitation of facial and manual gestures by human neonates. *Science*, 198, 75(4312), 74-78.

[7] Papousek, M., & Papousek, H. (1989): Forms and functions of vocal matching in interactions between mothers and their precanonical infants. First Language, 9(6), 137-158.

[6] Vihman, M (1996): Phonological Development: The Origins of Language in the Child. Cambridge, Mass.

[1] Zetterholm, E. (2003): *Voice Imitation. A Phonetic Study of Perceptual Illusions and Acoustic Success.* Doctoral dissertation. Travaux de l'institut de Linguistique de Lund 44, Lund University, Sweden.

[i] Of course from the point of view of a stringent experimental setup, using controlled acoustic stimuli is desirable but a serious drawback is that infants may not engage in spontaneous vocal interaction with for example a loudspeaker emitting target sounds (instead of the caregiver).

[ii] The total material consisted of approximately 15000 samples, but distorted and noisy samples were excluded, also numerous samples in which the infant and the adult are speaking at the same time or when the adult obviously is talking to the experiment leader were excluded.

[iii] Developed together with Ellen Marklund, Department of Linguistics, Stockholm University.