creating an automatic model of speech imitation

lisa gustavsson
department of linguistics
stockholm university
SSR level II | TMH KTH| spring 2007

1. introduction

The concept of imitation is regarded as one of the cornerstones in infants' early language acquisition but studies concerning word imitation reported in the literature are few and inconsistent and the concept of vocal imitation calls for a more stringent definition. As part of the CONTACT project (language acquisition and motor development in a humanoid robot, http://eris.liralab.it/contact/) a study concerning the role and characterisation of speech imitation in early language acquisition is carried out and the result is used for sketching a model to be used for the humanoid robotic articulatory feedback system. The basic scientific incentive behind the project is to increase theoretical understanding of language acquisition but also to explore ways in which key aspects of human learning can be replicated by computationally controlled machinery. Cognitive models function as wide-ranging tools for studying the fundamental processes in the human species. The challenge is to implement only human motivated components we know about in order to test our hypotheses about processes we lack knowledge of. We do not yet know to what extent the human language ability is written in our genetic code but one way to find out is to investigate how far we can describe language acquisition without incorporating hard-wired knowledge.

As it will be seen it's not trivial to determine exactly what a good imitation should sound like since there are several dimensions of imitative behaviour. For this reason imitative behaviour in infant vocalisation 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. Results are analysed in terms of which acoustic parameters that seemed to guide the listeners in their judgements. These parameters are then assigned different weights according to the results and used as matching criterion for automatic infant vocalisation-adult speech equivalence classification. Evaluation of the system is made by calculating the error rate in relation to results of the listener panel. If the acoustic parameters are weighted correctly the baby robot is expected to perform in agreement with the listeners.

2. background

Because of anatomical and physiological differences between adults and infants there will obviously never occur a perfect imitation in the sense of acoustically similar realisations (see Gustavsson et al, 2006 for an illustration of the infant vs the adult articulatory space). In these terms, 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. Obviously it takes practice to achieve the understanding of the underlying equivalence of the infant's own utterances and the adult's because in the speech material presented in this paper from infants in their very early stages of language acquisition the evidence of spontaneous vocal imitative behaviour is very meagre. This follows the findings in other speech data, in a couple of longitudinal infant studies examined by Vihman and McCune (1994) the number of imitations follows the slow trend of other spontaneous word productions during the first year of life. During the second year imitations add up to a third of the infants' productions reported in one of the studies but only a fifth in the other study. The low number of imitations and other word productions in this study might be explained in that these infants didn't have the same amount of training as the infants in the first study.

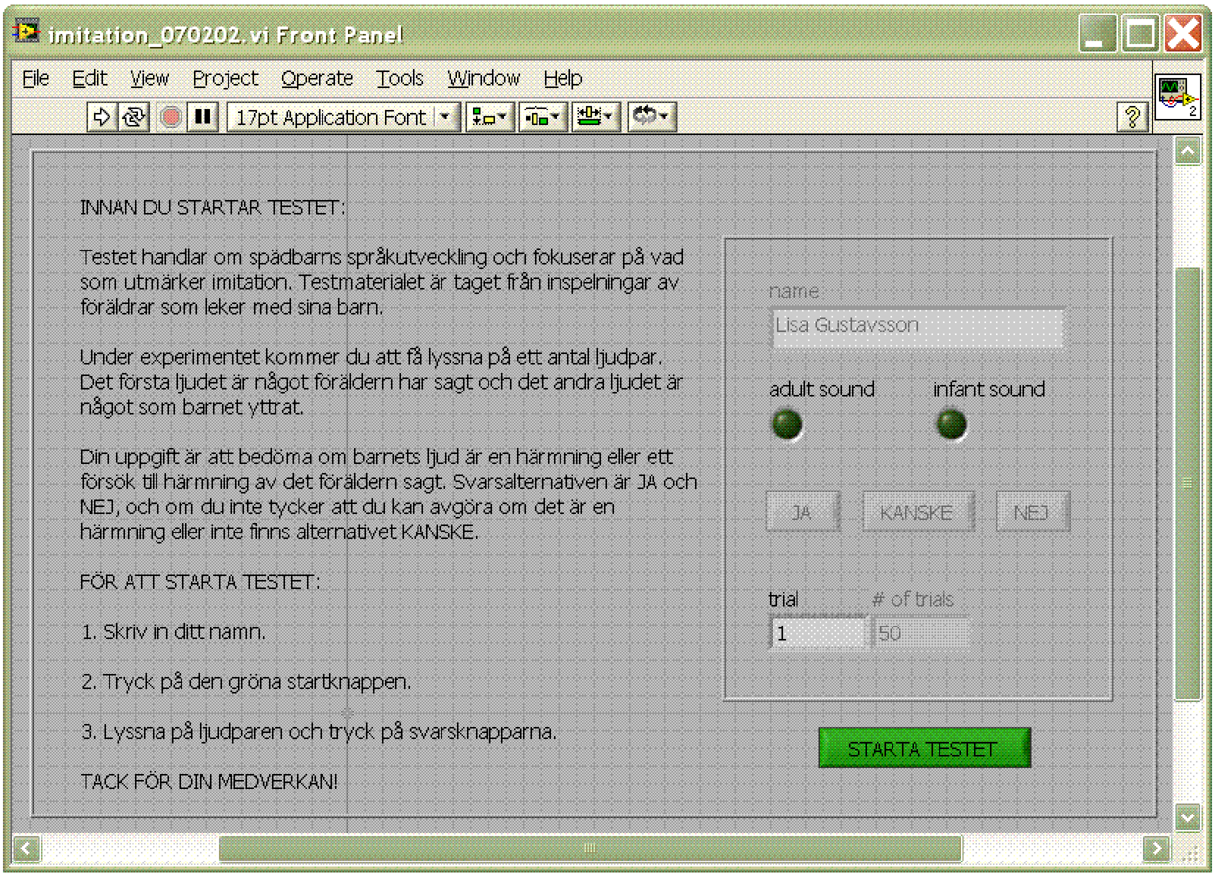
3. imitation judgement experiment

3.1. speech material

The speech data base, used in both the imitation judgement experiment and the imitation modelling, 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 on a vest that the infant wears during the session. 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 behaviour, enabling also the study of natural, interactive behaviour and the possible mutual convergence towards imitation targets. 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). The audio files were subsequently analyzed using the WaveSurfer software (http://www.speech.kth.se). Each recording was labelled in two separate files marking both the infant's and the adult's utterances/vocalisations. In order to obtain short separated utterances to create the speech data base the audio files were split in sequences exactly 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 (the total material consists of approximately 15000 samples, but distorted and noisy samples were excluded, also 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).

3.2. method

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 (Figure 1, LabView program developed together with Ellen Marklund, Phonetics, SU) 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 organised in sets of three different infant age groups, 185-296, 360-457 and 544-628 days, consisting of 50 presentations each. 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.



INNAN DU STARTAR TESTET:

Testet handlar om spädbarns språkutveckling och fokuserar på vad som utmärker imitation. Testmaterialet är taget från inspelningar av föräldrar som leker med sina barn.

Under experimentet kommer du att få lyssna på ett antal ljudpar. Det första ljudet är något föräldern har sagt och det andra ljudet är något som barnet yttrat.

Din uppgift är att bedöma om barnets ljud är en härmning eller ett försök till härmning av det föräldern sagt. Svarsalternativen är JA och NEJ, och om du inte tycker att du kan avgöra om det är en härmning eller inte finns alternativet KANSKE.

FÖR ATT STARTA TESTET:

1. Skriv in ditt namn.

2. Tryck på den gröna startknappen.

3. Lyssna på ljudparen och tryck på svarsknapparna.

TACK FÖR DIN MEDVERKAN!

name: Lisa Gustavsson

trial: 1

# of trials: 50

Figure **1**. The interface of the imitation judgement program.

3.3. results

The extreme pairs, that were considered to be sure imitations and also the pairs that were judged as very unlikely imitations, were chosen for an acoustic analysis. The acoustic parameters that were found in common 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 in seconds and the length of the first, second and the third syllables, relative to the utterance's overall duration were measured. These ratios give some indication on the relative prominence of the syllables they refer to.

The results of a preliminary analysis of the acoustic patterns in these pairs are summarized in figure 2. The number of perceived CV-syllables seems to be a very important factor in the judgment of similarity between the infant's utterances and the adult's. Matching pitch contours also seem to be good indicators of imitation, as well as the phonetic quality of the sounds imitated of course. Other factors that were not considered in the initial acoustic analysis but seemed to be important were intensity and other non-speech characteristics such as laughing or sad voices.

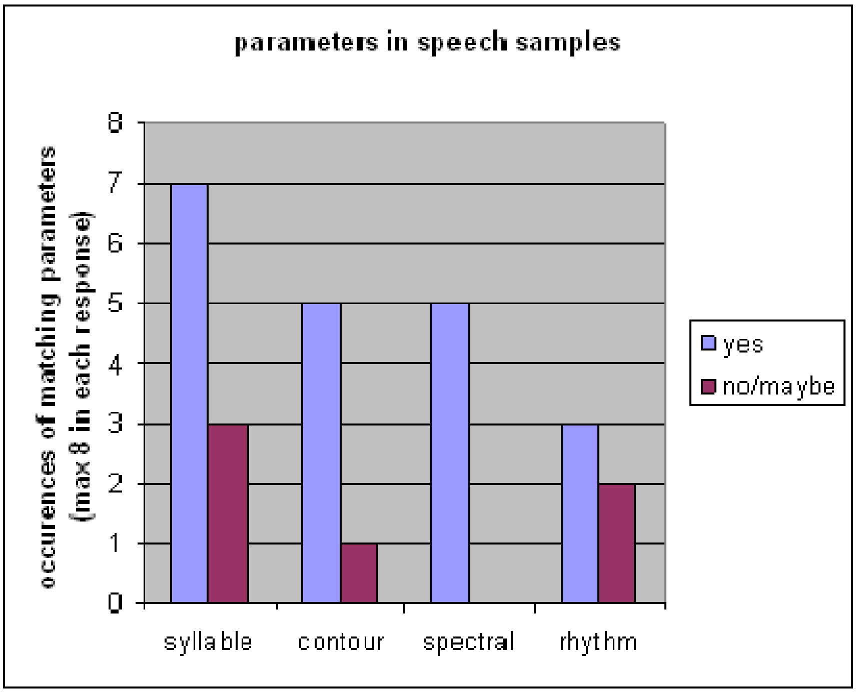


Figure 2. Preliminary results from the acoustic analysis of the imitation judgement answers.

Another 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 judgement 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.

3.4. discussion

Obviously these results indicate how adults judge the infant's imitations of an adult model and reflect the perception test's subject's views of what a successful imitation might be but they are nevertheless very informative because in a naturalistic situation adults are likely to react to an infant's utterances just as the perception test's subjects did. As mentioned earlier, the infant's speech imitation behaviour must deal with substantial and unavoidable differences between the acoustic characteristics of the adult's model utterances and the infant's own utterances. 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 the infant's imitation seemed preferable because it would provide a deeper insight on how adults interacting with young infants are likely to react to the acoustic characteristics of the infant's imitative responses. 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. This is a potentially important result that allows a creation of a realistic algorithm to describe and simulate speech imitation behaviour in young infants. Incidentally, these results provide also a clear coupling to the notions put forward by Lacerda and Lindblom (2006) and MacNeilage and Davis (2000). They consider the potential significance of the initial vocal behaviour 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 behaviours will unavoidable tune the infant's articulatory gestures and vocalisations in the direction of speech. The process of learning to imitate seems to be incremental in its nature, that is initially any vocalisation 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.

4. imitation modelling

The infant vocal tract poses physiological constraints that we have to consider when modelling imitation in spoken language development. Regardless of how much an infant would try, one and the same speech sound produced by an adult and produced by an infant would have drastically different acoustic realisations. As a consequence of this, an infant may not be able to come close the adult model, we have an equivalence problem because of the acoustic properties of an infant's vocal tract of and those of an adult. To avoid this problem imitation would require a model that is able to adopt the physiology of different vocal tract dimensions, as suggested by Engwall (2004). With the model proposed in this paper I will not attempt to solve this equivalence problem, rather shifting the focus to the more general patterns in the speech signal as shown in the imitation judgement experiment. Also the hierarchical judging of imitations according to progress in production has to be taken into account and implementing these weighted parameters in an incremental manner could reveal some the tuning processes towards successful imitations.

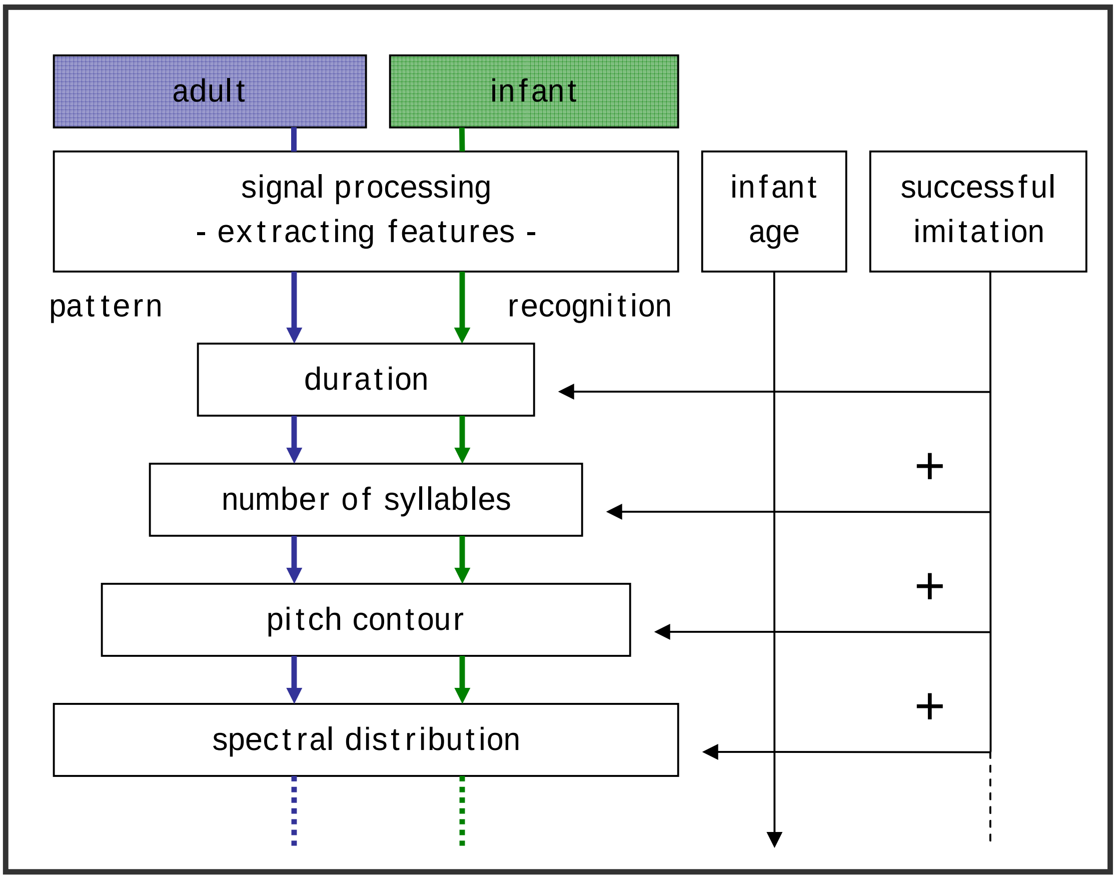


Figure 3. Flow chart of the imitation modelling. The infant utterances are matched against their respective adult utterance with respect to any of the parameters. The number of matching parameters for a successful imitation will increase according to the age of the infant. Perhaps the parameters will also be weighted differently depending on how important they are for a good imitation.

An ambitious sketch of a possible scenario for how infants may be able to imitate and get feedback on adult utterances is displayed in figure 3. The initial feature extraction could be done in different steps depending on what parameters we are interested in but Cepstrum Coefficients should provide a good enough estimation of the features we are investigating. The parameters we are looking at are duration, number of syllables, the overall pitch contour (perhaps autocorrelation methods would be more suitable for extracting the pitch) and crude representations of intensity distributed in the frequency domain. We choose the Mel Frequency Cepstrum Coefficients (MFCC) method where a Mel-scale frequency warp is used for converting the input-signal to what is thought to be its auditory representation. The reason for using the Mel-scale is that the relation between perceived frequency and Hz is non-linear and the Mel-scale is based on the subjective perception of pitch and transforms the signal to correspond to this non-linear human perception of frequency. This is used within ASR to get better results, but in the kind of human perceptual modelling we are dealing with here it is crucial to keep every step as close to human abilities as possible and a Mel-based representation of frequency is very close to the human auditory frequency response. A cepstral processing of the signals will also allow a separation of the source from the filter a task that usually poses problems when dealing with the very high pitch of young infants. The principle is imagining the typical spectrum cross section as a time domain signal, where the rapidly oscillating component due to voice excitation would correspond to high frequency while the slow changes of the vocal tract resonances would be represented in a low frequency domain. Thus the highest order coefficients are roughly equal to the interval between the pitch pulses of the signal when multiplied with sample duration and also the lower order coefficients correspond in the same way to the shape of the filter's transfer function. By this way running another FFT on the original spectrum will usually result in an effective separation of source and filter, however this is only possible if the variation in the spectrum due to the source is faster than the variation due to the filter. As mentioned, perhaps this is not always the case for infants since the harmonics are so far from each other.

In the pattern recognition module a straightforward distance calculation between the infant vocalisations and the adult model involves more than a few difficulties besides the common complications within ASR such as time alignment and end-point detection. The technical difficulties in processing infant utterances raise an even greater challenge than the already difficult processing of high-pitch utterances produced by female speakers or young children. Indeed, not only the fundamental frequency of the utterances produced by an infant is about two or three octaves the typical pitch for a female speaker but there is also a non-linear relationship between the infant’s and the adult’s vocal tracts which introduces additional problems when attempting a comparative study of vocalisations in infants and adults. The commonly used methods for feature transformation used within ASR such as PCA or LDA techniques (Fukunaga, 1990 and Haeb-Umbach, 1999) for example are based on the assumption that the acoustic mapping is linear. The dimensions of the infant vocal tract and the adult are not. Also VTLN or the adaptive-model version VTLA (Blomberg & Elenius, 2007) have proven successful for older children ASR but might have problem with very young infants.

It all comes down to the problem of initially recovering formant values and the high pitch of the infant voice. The harmonics are usually spaced with at least 500 Hz and any possible formant can easily hide between any two of them. Then again, considering the results from the imitation judgement experiment a good imitation based on spectral properties in the signal is not of interest initially, the demands for matching speech sounds seem to come at a later stage once the infant already mastered more basic characteristics such as syllable structure or rhythmical aspects of the vocalisation. Also taking into account the fact that the newborn infant can not, due to immature anatomy and muscle control, produce all the different speech sounds we might be interested in – even if they tried. Consequently, even if we had a technique to cope with a high fundamental and a non-linear acoustic mapping perhaps we would not find what we are looking for. This would in fact also motivate the incremental design of the imitation model. The correct number of syllables for example will have less weight than the correct spectral distribution and is also expected to be mastered earlier than spectral characteristics such as vowels and consonants. The parameters suggested here is of course based on the Swedish speech material used in this study, languages that that use other features for discrimination in speech, such as tone languages for example might have different parameters.

Many of the adult utterances in the adult-infant interaction speech database consist of longer sentences while the infant utterances are shorter. But memory decay is considered in the model hence any imitation pattern matching is done backwards, that is the adult might utter a long sentence such as Have *you seen the nice Dappa!* but the infant attempts to imitate the last segment *…Dappa* rather than the initial segment *Have you seen the nice…*. As revealed in the imitation judgement experiment this is a perfectly good imitation and a right-to-left algorithm will capture this.

5. concluding remarks

Another aspect of the early adult-infant interaction concerning imitation is the adult imitations. Admittedly, in the adult-infant interaction settings in which the recordings were made there are also a great number occurrences of adult imitations of infant utterances but these have not been considered in the current study. However, in further development of this model it is necessary to address exactly that issue when studying the relationship between acoustic and articulatory representations of utterances produced by adults and infants imitating each other. The adult imitations might play a crucial role in guiding the infant in developing its articulatory to acoustical mapping and overcome the equivalence classification problem.

references

Blomberg, M and Elenius, D (2007): Vocal tract length compensation in the signal and model domains in child speaker recognition. In Proc of FONETIK 2007, TMH-QPSR vol 50, KTH, Stockholm.

Engwall, O (2004): Speaker adaptation of a three-dimensional tongue model . In Proc of ICSLP 2004, pp. 465-468. Jeju Island, Korea.

Fukunaga, K (1990): Introduction to Statistical Pattern Recognition, 2nd ed, Orlando, FL, Academic Press.

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

Haed-Umbach, R (1999): Investigations on Inter-Speaker Variability in the Feature Space. IEEE International Conference on Acoustics, Speech and Signal Processing. Phoenix, AZ.

Huang, X, Acero, A and Hon, H-W (2001): Spoken Language Processing. Prentice Hall

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

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

Vihman, M and McCune, L (1994): When is a word a word? Journal of Child Language, 21, 517-42.