Multimodal language acquisition based on motor learning and interaction

Jonas Hörnstein, Lisa Gustavsson, José Santos-Victor, Francisco Lacerda

1 Introduction

In this work we propose a methodology for language acquisition in humanoid robots that mimics that in children. Language acquisition is a complex process that involves mastering several different tasks, such as producing speech sounds, learning how to group different sounds into a consistent and manageable number of classes or speech units, grounding speech, and recognizing the speech sounds when uttered by other persons. While it is not known to which extent those abilities are learned or written in our genetic code, this work aims at two intertwined goals: (i) to investigate how much of linguistic structure that can be derived directly from the speech signal directed to infants by (ii) designing, building and testing biological plausible models for language acquisition in a humanoid robot. We have therefore chosen to avoid implementing any pre-programmed linguistic knowledge, such as phonemes, into these models. Instead we rely on general methods such as pattern matching and hierarchical clustering techniques, and show that it is possible to acquire important linguistic structures directly from the speech signal through the interaction with a caregiver. We also show that this process can be facilitated through the use of motor learning.

The interaction between an adult caregiver and a human infant is very different from the interaction between two adults. Speech directed to infants is highly structured and characterized by what seems like physically motivated tricks to maintain the communicative connection to the infant, actions that at the same time also may enhance linguistically relevant important aspects of the signal. Also, whereas

Jonas Hörnstein and José Santos-Victor
Institute for System and Robotics (ISR), Instituto Superior Técnico, Lisbon, Portugal, e-mail: {jhornstein,jasv}@isr.ist.utl.pt

Lisa Gustavsson and Francisco Lacerda
Department of Linguistics, Stockholm University, Stockholm, Sweden e-mail: {lisag, frasse}@ling.su.se

communication between adults usually is about exchanging information, speech directed to infants is of a more referential nature. The adult refers to objects, people and events in the world surrounding the infant [33]. Because of this, the sound sequences the infant hears are very likely to co-occur with actual objects or events in the infant's visual field. The expanded intonation contours, the repetitive structure of IDS and the modulation of the sentence intensity are likely to play an important role in helping the infant establishing an implicit and plausible word-object link. This kind of structuring might very well be one of the first steps in speech processing, a coarse segmenting of the continuous signal in chunks that stand out because of some recurrent pattern the infant learns to recognize. Infants are very sensitive to the characteristic qualities of this typical IDS style, and a number of studies indicate that infants use this kind of information to find implicit structure in acoustic signals [27] [11] [32] [6] [46]. Some evidence on the usefulness of these co-ocurring events can also be found in robotics. In the CELL [45], Cross-channel Early Lexical Learning, architectures for processing multisensory data is developed and implemented in a robot called Toco the Toucan. The robot is able to acquire words from untranscribed acoustic and video input and represent them in terms of associations between acoustic and visual sensory experience. Compared to conventional ASR systems that maps speech signal to human specified labels, this is an important step towards creating more ecological models. However, important shortcuts are still taken, such as the use of a predefined phoneme-model where a set of 40 phonemes are used and the transition probabilities are trained off-line on large scale database. In [51], no external database is used. Instead the the transition probabilities are trained online only taking into account utterances that have been presented to the system at the specific instance in time. While this make the model more plausible from a cognitive perspective, infants may not rely on linguistic concepts as phonemes at all during these early stages of language development. In this work we have instead chosen a more direct approach and map the auditory impression of the word as a whole to the object. Underlying concepts like phonemes instead are seen as emergent consequences imposed by increasing representation needs [44] [35].

In this work we have chosen to represent those underlying structures, i.e. pseudo-phonemes, in the form of target position in motor space, rather than as auditory goals. The rationale for this can be found in the motor theory of speech perception [37], which hypothesizes that we recognize speech sound by first mapping the sound to our motor capabilities. This statement is also supported by more recent work in neuroscience that demonstrates an increased activity in the tongue muscles when listening to words that requires large tongue movements [9]. This leads to believe that the motor area in the brain is involved not only in the task of production, but also in that of recognition. Earlier works including neurophysiologic studies of monkeys have shown a similar relationship between visual stimulation and the activation of premotor neurons [14]. Those neurons, usually referred to as mirror neurons, fire both when executing a motor command and when being presented with an action that involves the same motor command. To learn the audio-motor maps and finding key target positions, interactions again play an important role in the form of imitation games.

Already during the first year a child slowly starts to mix what can be perceived as spontaneous babbling and some early imitations of the caregivers. During the second year as much as one third of the utterances produced by the child can be perceived as imitations, [54]. Broadly speaking, these "imitation games" where the child may play either the role of demonstrator or that of the imitator, serve two different purposes: (i) learning elementary speech units or speech vocabulary and (ii) gaining inter-speaker invariance for recognition.

One frequent situation during the child's language acquisition process is when the caretaker repeats certain utterances, words or sounds that the child tries to imitate. During this type of interaction it has been shown that caregivers actively change their voices in order to facilitate the task for the child [8]. When the child eventually succeeds in producing an "acceptable" imitation, the adult provides some sort of reinforcement signal. Through this reinforcement, the child identifies elementary speech units (e.g. phonemes or words) that play a role in the communication process and will form a (motor) speech vocabulary that can be used in later times. In addition, the child can learn how to better produce those specific terms in the vocabulary.

Another frequent situation is when the adult speaker imitates the sounds produced by the child. This may help gaining inter-speaker invariance by affording the child with the possibility of perceiving the same word (or phoneme or utterance) produced by the child's own vocal tract and that of an adult speaker.

Hence, the main hypothesis for this work are that (i) the infant or robot rely on the interaction with a caregiver in order to acquire their language, and (ii) underlying linguistic structures such as phonemes do not need to be pre-programmed but can emerge as a result of the interaction and can be represented in the form of vocal tract target positions.

Here we first take a closer look at the interaction between the infant and a caregiver and discuss how these interactions can guide the infant's language acquisition. We then discuss the learning architecture and necessary resources that has to be implemented in a robot for it to be able to acquire early language structures through the described interactions. The architecture described is an extended version of the architecture presented in [23]. A similar architecture can also be found in [28], and some earlier work can be found in the DIVA model [19]. Finally we show how a humanoid can use the architecture and interaction scenarios to learn word-object relations and extract target positions for a number of vowels.

2 The role of interaction

In this section we take a closer look at how interaction can facilitate language acquisition, more specifically we look at the interaction between a caregiver and the child (or robot). When interacting with a child, adults tend to adapt their speech signal in a way that is attractive to the infant and in the same time can be helpful for learning the language. It is therefore natural to start this section with an explanation of some of the typical aspects of Infant Directed Speech.

This interaction it typcially multimodal in its nature. The child does not only hear the caregiver, but receives mutlimodal information such as combinations of speech, vision, and tactile input. In this section we discuss how this multimodal information can be used to ground speech. This kind of interaction, based on shared attention, starts even before the infant can produce speech itself. However, as the infant starts to explore its capacity to produce sound, verbal interaction will become more and more important. Initial verbal interaction is mainly based on imitations. Here we separate between two types of imitation scenarios, one where the infant imitates its caregiver, and one where the caregiver imitates the infants. Both have been found equally common in adult-infant interactions, but they serve two different goals (i) to learn which sounds that are useful for communication (the motor vocabulary of speech gestures like phonemes or pseudo-phonemes) and (ii) to learn to map human sound to the articulatory positions of the robot.

2.1 Infant directed speech

An important portion of the physical signal in the ambient language of almost every infant is in the form of Infant Directed Speech (IDS), a typical speech style used by adults when communicating with infants. IDS is found in most languages [2] [31] [10] and is characterized by long pauses, repetitions, high fundamental frequency, exaggerated fundamental frequency contours [11] and hyperarticulated vowels [31]. A very similar speech style is found in speech directed to pets [22] [4], and to some degree also in speech directed to humanoid robots [17], and pet robots [3].

The function of IDS seem to change in accordance with the infant's developmental stages, phonetic characteristics in the adult's speech are adjusted to accommodate the communicative functions between the parents and their infants, for example a gradual change in consonant specifications associated with the infants communicative development was found in a study by Sundberg and Lacerda [49]. In longitudinal studies it has been shown that parents is adapting their speech to their infants linguistic and social development the first post-natal year. On the whole they use higher fundamental frequency, greater frequency range, shorter utterance duration, longer syllable duration, and less number of syllables per utterance when speaking to their infants as compared to speaking to adults. Sundberg [50] suggests that these phonetic modifications might be an intuitive strategy adults use automatically that is both attractive and functional for the infant.

In a study more directly related to infants word learning, Fernald and Mazzie [13] found that target words in infant directed speech were typically highlighted using focal stress and utterance-final position. In their study 18 mothers of 14-month-old infants were asked to tell a story from a picture book called Kelly's New Clothes, both to their infants and to an adult listener. Each page of the book introduced a new piece of clothes that was designated as a target word. When telling the story to the infants target words were stressed in 76% of the instances, and placed in utterance-final position in 75% of the instances. For adult speech the same values were 40% and 53% respectively.

Albin [1] found that an even larger portion of the target words (87% - 100% depending of subject) occured in final position when the subjects were asked to present a number of items to an infant.

2.2 Multimodal interaction

Whereas communication between adults usually is about exchanging information, speech directed to infants is of a more referential nature. The adult refers to objects, people and events in the world surrounding the infant [33]. When for example playing with a toy, the name of the toy is therefore likely to be mentioned several times during the interactions. Finding such recurrent patterns in the sound stream coming from the caregiver can help the infant to extract potential word candidates that can be linked to the visual representation of the object. On the other hand, also words that are not directly related to the object may be mentioned repeatably to the same extent or even more often than the target word itself. By linking all recurrent sound patterns to the most salient object in the visual field we are likely to end up with a large number of false word-object links. However, if the same wordlike pattern consistantly appears when, and only when, a certain object is observed it is hightly likely that it is actually related to that object. It is therefore necessary to look for cross-modal regularities over a relatively long time.

The CELL-model [45] therefore make use of both a short-term memory that is searched for recurrent patterns and a long-term memory that is searched for cross-modal regularities in order to form word-object associations.

2.3 Babbling and imitation

One of the first key observations regarding a child's language development is the use of babbling [36], an exploration process that allows the child to explore different actuations of the vocal-tract and the corresponding acoustic consequences. This process allows to to build sensorimotor maps that associate the articulatory parameters of the vocal tract and the produced sounds. In other words, through babbling the child (or the robot) learns the physics of the vocal tract and how to produce sounds. While babbling was first seen as an isolated process, it has later been shown to have a continuous importance for the vocal development [53]. It has also been shown that in order to babble normally, children need to be able not only to hear both to themselves and other con-specifics [48], but also to establish visual contact with others [42]. Speech babbling could therefore be seen as an early type of interaction, instead of just a self exploration task. When a caregiver is present, he or she is likely to imitate the sound of the infant, giving the infant the possibility to also create a map between its own utterances and that of the caregiver. Imitation studies performed at Stockholm University has shown that in about 20% of the cases, an eventual response from the caregiver is seen as an imitation of the infant's utterance when judged by other adult listeners. We have also shown that it is possible to get a good estimation of when there is an imitation or not by comparing a number of prosodic features [25].

By creating speech sound and detecting possible imitations, the child or robot can overcome differences between its own voice and that of the caregiver, and by repeating this kind of "imitation game" with several different caregivers it is also possible to overcome inter-speaker variations.

As stated in the introduction of this section, there are two different goals with the imitation games. Apart from the goal to overcome the inter-speaker differences and allow the robot to correctly map sounds from different speakers to its own motor positions in order to reproduce or recognize those sounds, the second goal is to help the robot to separate between sounds that are useful for communication (thus becoming part of a motor vocabulary of speech gestures) and sounds that can be considered as noise and should be forgotten.

Given that the child or robot is already able to correctly map speech sound from the caregiver to its vocal tract, it can use statistical methods to find useful positions. However, while the map is still incomplete it may need feedback or reinforcement provided by the caregiver. This is possible by letting the robot imitate utterances from the caregiver. The robot uses its sensor-motor maps to calculate the corresponding vocal tract positions and then reproduces the same utterance with its own voice. Depending on how well the robot is able to map the voice of the caregiver to its own vocal tract positions the reproduced sound may or may not be perceived as an imitation by the caregiver. If a correct imitation or, at least, as a useful speech utterance is perceived, the robot receives positive feedback and stores the vocal tract positions in the speech motor vocabulary. Alternatively, if the reproduced sound is not validated by the caregiver, he may try to change his/her voice in order to guide the robot towards the correct sound. This behaviour can also be found in the interaction between a child and its parents and has been studied in [8].

3 Embodiment

In order to interact with the caregiver as explained in the previous section, the robot must be able to see, hear, and also to produce sounds. To mimic the way human infants interact with it is of course an advantage to have a robot that looks and acts as would be expect from an infant. However, the embodiment is important not only to evoke emotions and make it more attractive for interaction, but also from a learning perspective. Having a similar body structure facilitates imitations since it allow for a more direct mapping between the demonstrator and the imitator while at the same time limiting the search space and hence the risk of incorrect mappings.

In this section we describe the architecutre and the system components that are implemented in the humanoid robot in order to allow the robot to engage in the interaction tasks necessary to acquire early language structures. The described architecture is an extension of the work in described in [23]. It includes a speech production module, a sensing unit, a sensorimotor coordination module and a memory unit, as shown in 1.

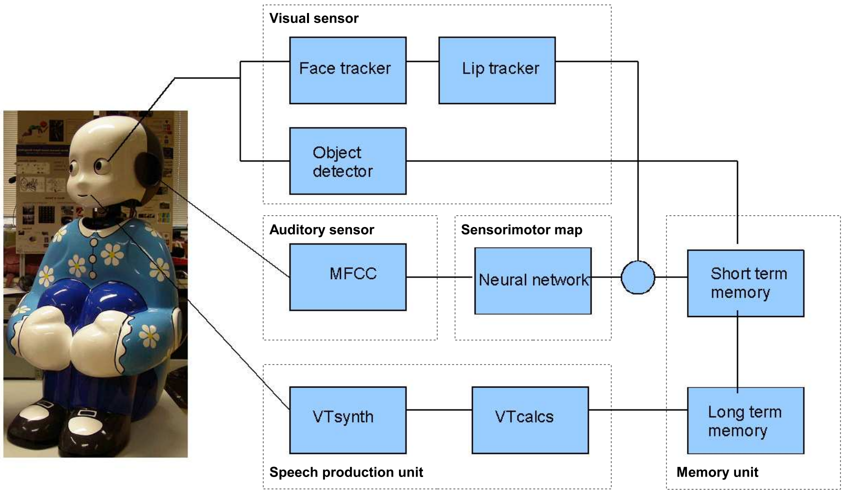

Fig. 1 System architecture for language acquisition

3.1 Speech production unit

The speech production unit consists of an articulatory model of the human vocal tract and a position generator. There have been several attempts to build mechanical models of the vocal tract [21] [18]. While these can produce some human like sounds they are still rather limited and there are no commercially available mechanical solutions. The alternative is to simulate the vocal tract with a computer model. Such simulators are typically based on the tube model [40] where the vocal tract is considered to be a number of concatenated tubes with variable diameter. On top of the tube model an articulator is used that calculates the diameter of the tubes for different configurations of the vocalization units. In this work we have chosen to simulate the vocal tract by using VTcals developed by Maeda [41]. This model has been developed by studying x-rays from two women articulating French words, and has six parameters that can be used to control the movements of the vocal tract. One parameter is used for controlling the position of the yaw, one for the extrusion of the lips, one for lip opening, and three parameters for controlling the position of the tongue. A synthesizer converts the vocal tract positions into sound. While the synthesizer works well for vowel-like sounds, it is unable to produce fricatives sounds and can hence only produce a limited set of consonants.

Apart from creating sound, the lip parameter is also used to control a number of Light Emitting Diods (LEDs) in the robot's face in order to simulate the movements of the lips. The current lip model is very simple and only show the mouth as either open or closed.

3.2 Sensing units

As explained in the previous section, not only ears but also other sensing modalities are useful when learning to speak. Here we have implemented two sensing modalities, an auditory sensor unit and a visual sensor unit that extract features from the acoustic and visual spaces respectively.

The auditory sensor consist of two microphones and artificial pinnas [26]. To get a more compact representation of the sound signal it is transformed into a tonotopic sound representation (sound features). There exist various representations that can be used for this. For production and recognition of vowels, formants are commonly used [56]. However, formants only contain information useful for vowels so its application is rather narrow. In other related work, LPC has been used [30] [43]. LPC are more generally applicable than formants, but still require rather stationary signals to perform well. Here we use Mel frequency cepstral coefficients (MFCC) [7] as our speech features since these do not require a stationary signal. To calculate the MFCC each utterance is first windowed using 25 ms windows with 50% overlap between the windows, and MFCC are then calculated for each window.

As visual sensors the robot is equipped with two cameras with pan and tilt. However, for the sake of language acquisition only one camera is used. The visual sensor is used both for finding objects during multimodal interaction, and for tracking faces and lipmovements to provide information on the lip-opening when trying to imitate the caregiver.

Starting with the object detector, the robot takes a snapshot of the camera's view and segment the image in order and look for the object closest to the center of the image. The segmentation is done by background subtraction followed by morphological dilation. Using the silhouette of the object we create a representation of its shape by taking the distance between the center of mass and the perimeter of the silhouette. This is done for each degree of rotation creating a vector with 360 columns. The tranformation of an image to the object representation is illustrated in Figure 2.

To estimate the opening of the mouth, the visual sensor takes a video sequence of the speaker's face as input and calculates the openness of the mouth in each frame. The openness is defined as the distance between the upper and lower lip, normalized with the height of the head. We use a face detection algorithm, based on [55] and [39], to calculate the size of the head and the initial estimate of the position of the lips. Within the estimated position we use colour segmentation methods to select candidate pixels for the lips based on their redness. While there are methods to find the exact contour of the lips, like snakes or active contour methods [29], we are only interested in the openness of the mouth and have chosen to represent the lips with an ellipse. To fit the ellipse to the lip pixels we use a least square method described in [16]. Finally we calculate a new estimate for the lip position in the next frame of the video sequence by using the method in [38].

3.3 Sensorimotor maps

The sensorimotor maps are responsible for retrieving the vocal tract position from the given auditory and visual features. We use two separate neural networks to model the sound-motor and visiuomotor maps. The sound-motor map is the more complicated of the two, mapping the 12 cepstral coefficients back to the 6 articulatory parameters of the vocal tract model. The added difficulty of the problem lies in the fact that several positions of the vocal tract may result in the same sound, hence giving several possible solutions for a given set of input features. These ambiguities have to be solved through the interaction with a caregiver. For the sound-motor map we use an artificial neural network with 20 hidden neurons.

The vision-motor map is a very simple unit, performing a linear mapping from the mouth opening to the lip height parameter of the synthesizer.

Since the output from both the sound-motor map and the vision-motor map consist of vocal tract positions, the integration of those sensor outputs becomes very simple. Here we simply use a weighted average of the lip height calculated from the two maps. The weight is currently set by hand, but should preferably be set automatically according to the quality and intensity of the visual and auditory stimuli.

3.4 Short term memory

The short-term memory receives a continuous stream of visual and auditory data that are stored during a fixed time (here we have used 10-20 s).

The auditory sound stream is sequenced into utterances. This is done automatically when the sound level is under a certain threshold value for at least 200 ms. Each utterance within the short term memory at a given time is compared pair-wise with all other utterences in the memory in order to find recurrent pattern. For each utterance-pair we first make sure that the utterances have the same length by padding the shortest utterance. The utterances are then aligned in time and we calculate the sum of differences between their mel coefficients creating a vector with the acoustic distance between the two utterances at each window. The second utterance is then shifted forward and backward in time and for each step a new distance vector is calculated. These vectors are averaged over 15 windows, i.e. 200 ms, and combined into a distance matrix as illustrated in Figure 3. By averaging over 200 ms

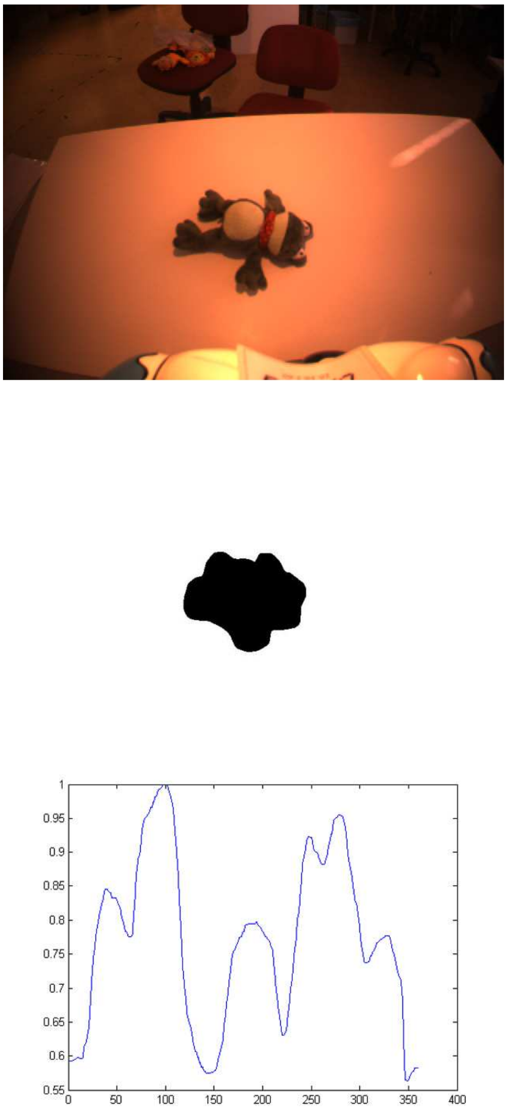

Fig. 2 Original image (top), silhouette image after background substraction and morphologic operations (center), and the silhouette perimeter in polar coordinates (bottom).

we exclude local matches that are too short and can find word candidates by simply looking for minimas in the distance matrix. Starting from a minima we find the start and the end points for the word candidate by moving left and right in the matrix while making sure that the distance metric at each point is always below a certain critical threshold.

In order to take advantage of the structure of infant directed speech and to mimic infants' apparent bias towards target words in utterance-final position and focal stress, we also check for these features. For a word candidate to be considered to have utterance-final position we simply check that the end of the candidate is less than 15 windows from the end of the utterance. To find the focal stress of an utterance we look for the F0-peak. While there are many ways for adults to stress words (e.g. pitch, intensity, length) it has been found that F0-peaks are mainly used in infant directed speech [13]. If the F0-peak of the utterance as a whole is within the boundaries of the word candidate, the word candidate is considered to be stressed. If a word candidates is not stressed and in utterance-final position we may reject it with a specified probability.

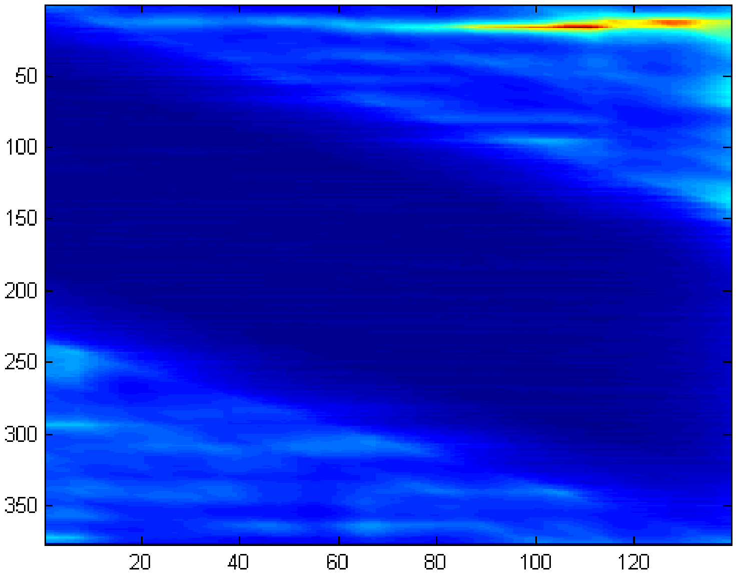

Fig. 3 Finding word candidates in sentences "titta här är den söta dappan" and "se på lilla dappan". The best match is found by shifting the sentences 15 windows.

The same technique for pattern matching is also be used to compare visual objects. When comparing two object representations with each other we first normalize the vectors and then perform a pattern maching, much in the same way as for the auditory representations, by shifting the vectors one step at a time. By doing this we get a measurement of the visual similarity between objects that is invariant to both scale and rotation.

When we find both a word candidate and a visual object we pair these representations and send them to the long term memory.

3.5 Long term memory

The long term memory is used to store both word candidates, visual objects, and vocal tract positions that are found interesting during the interaction with the caregiver. To organize the information we use an hierarchical clustering algorithm [20]. Word candidates, visual objects, and vocal tract positions are organized independently into three different tree clusters. The algorithm starts by creating one cluster for each item. It then iteratively joins the two clusters that has the smallest average distance between their items until only one cluster remains.

While the algorithm is the same for all three trees, the distance measure varies slightly between them. The distance between the visual objects is measured directly through the pattern matching explained above. For the acoustic similarity we use Dynamic Time Warp (DTW) [47] to measure the distance between different word candidates. The reason to use DTW instead of directly applying the pattern matching described earlier is to be less sensitive to how fast the word candidate is pronounced. For the vocal tract position we simply use the Euclidean distance to measure the distance between each target position.

The resulting hierarchical trees can now be analyzed in a second step to determine the correct number of clusters. For each level of the clustering process, we have different relationships between data groupings. The question is then to find the "natural" grouping for this dataset. To estimate the adequate number of clusters in the dataset we have used the Gap statistic [52]. This function compares the within-cluster dispersion of our data with that obtained by clustering a reference uniform distribution. This is to compare the gain of raising the cluster number in a structured data with that arising from adding another cluster to a non-informative and not structured set of points. To find the optimal number of clusters we look for the first maximum in the Gap. Each position within the same cluster is considered to be part of the same phoneme or pseudo-phoneme in the motor speech vocabulary.

When we have interconnected multimodal representations, which is the case for the word candidates and visual objects that assumingly refers to the same object we can make use of these connections, not only to create associations, but also to find where we should cut the trees in order to get a good representations of the words and the objects. In order to find which branch in the word candidate tree that should be associated with which branch in the object tree we use the mutual information criterion [5]. In the general form this can be written as

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log \left(\frac{p(x,y)}{p_1(x)p_2(y)} \right) \tag{1}$$

Where $p(x,y)$ is the joint probability distribution function of X and Y, and $p_1(x)$ and $p_2(y)$ are the marginal probability distribution functions of X and Y respectively.

We want to calculate $I(X;Y)$ for all combinations of clusters and objects in order to find the best word representations. For a specific word cluster A and visual cluster V we define the binary variables X and Y as

$$X = \{1 \; if \; observation \in A; \; 0 \; otherwise\}$$
$$Y = \{1 \; if \; observation \in V; \; 0 \; otherwise\}$$

The probability functions are estimated using the relative frequencies of all observations in the long-term memory, i.e. $p_1(x)$ is estimated by taking the number of observations within the cluster A and dividing with the total number of observations in the long-term memory. In the same way $p_2(y)$ is estimated by taking the number of observations in the cluster V and again dividing with the total number of observations. The joint probability is found by counting how many of the observations in cluster A that is paired with an observation in cluster V and dividing by the total number of observations.

4 Humanoid robot experiments

In this section we examplify how the interaction strategies and the described architecture works in practise by showing the results from two different experiments.

In the first experiment we show how the robot can extract useful word-candidates and associate those to visual object. This is done both by interacting directly with the robot and by using recordings of real Infant Directed Speech taken from adult-infant interactions.

In the second experiment we teach the robot nine Portuguese vowels and see how well the robot can recognize the same vowels when pronounced by different human speakers. Here we especially look at the effect that the different developmental stages of babbling and interaction have for the recognition rate and the role played by the visual input with respect to the recognition rate.

4.1 Experiment 1: Word learning

In this experiment the robot makes use of multimodal information in order to learn word-object associations when interacting with the caregiver. The experimental setup is shown in Figure 4.

The caregiver shows a number of toys for the robot and, at the same time, talks about these objects in an infant directed speech style. The objects that were used during the experiment were one ball and two dolls named "Pudde" and "Siffy". The experiment was performed by demonstrating one object at a time by placing it in front of the robot for approximately 20 s, while talking to the robot about the object by saying things like "Look at the nice ball!" and "Do you want to play with the ball?". Each utterance contained a reference to a target word and we made sure that the target word has always stressed and in utterance-final position. For the dolls we referred to them both by using their individual names and the swedish word for doll, "docka". The ball were always referred to using the swedish word "bollen".

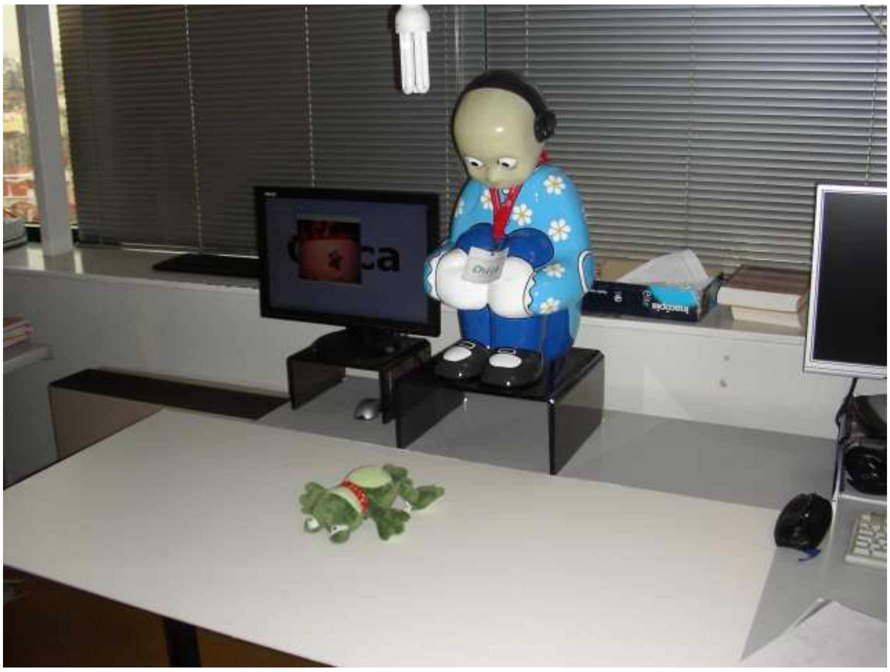

Fig. 4 Experimental setup for robot test

During the length of one demonstration, sound and images are continuously stored in the short-term memory. The sound is then segmented by simply looking for periods of silence between the utterances and each utterance is then compared to the others as explained in the previous section. Each word candidate, i.e. matching sound pattern, in the short-term memory is paired with the visual representation of the object and sent to the long-term memory. After having demonstrated all three objects we repeat the procedure once more, but this time with the objects in slightly different orientations in front of the robot. This is done in order to verify that the clustering of the visual objects is able to find similarities in the shape despite differences in the orientation of the objects.

When word candidates have been extracted from all six demonstrations, the hierarchical clustering algorithm is used to group word candidates in the long-term memory that are acoustically close. The result from the hierarchical clustering of the word candidates can be seen in Figure 5. The same is done for the visual objects, Figure 6. The numbers at each leaf shows the unique identifier that allows us to see which of the word candidates that was paired with which of the visual objects.

Looking only at the hierarchical tree for the word candidates it is not obvious where the tree should be cut in order to find good word representations. By listening to the word candidates we notice that the cluster containing candidates (25 26 19 20 2 6 18 14 16 1) represent the word "dockan", the cluster (3 7 4 9 5 8 10 12 15 11 13 17) represent the word "Pudde", the cluster (21 22 23 27 28 29 24 31 30) represent the word "Siffy", and the cluster (32 33 34 36 35) represent the word "bollen". The hierarchical tree for the visual objects may look more simple and it is tempting to select the five clusters in the bottom as our objects. However, it is actually the clusters one level up that represents our visual objects, but of course the robot does not know that at this point.

To find out which branch in the respective tree that should be associated with which branch in the other we calculate the mutual information criterion. Calculating the mutual information criterion for all pair of branches shows that we get the highest score for associating the word candidates (32-36) with the same visual objects (32-36). This is what we could expect since all visual observations of "bollen" were also paired with a correct word candidate. In the case of the objects "Pudde" and "Siffy" part of the observations are not paired with the object name, but instead with the word "docka". Still we get the second and third heighest scores by associating word candidates for the word "Pudde" with object Pudde and the word "Siffy" with object Siffy respectively. We can also find that the branch above the visual representations of Pudde and Siffy receives the heighest score for being associated branch containing word candidates for "dockan".

The experiment was repeated without putting any bias on word candidates that were stressed and in utterance-final position. This resulted in four false word candidates for the object Pudde and one for object Siffy. However, this did not affect the word-object associations as these candidates were found in separate branches in the word candidate tree and only received low scores by the mutual information criterion.

A second experiment was performed using recordings of interactions between parents and their infants. The recordings were made under controlled forms at the Department of Linguistics, Stockholm University. A lot of care were taken to create a natural interactions. The room was equipped with several toys, among those two dolls called "Kuckan" and "Siffy". The parents were not given any information of the aim of the recordings but were simply introduced to the toys and then left alone with their infants. In this study we have only used a single recording of a mother interacting with her 8 month old infant. The total duration of the recording is around 10 minutes. The audio recording has been segmented by hand to exclude sound coming from the infant. In total the material consists of 132 utterances with time stamps and also object references in those case that an object were present. In 33 of

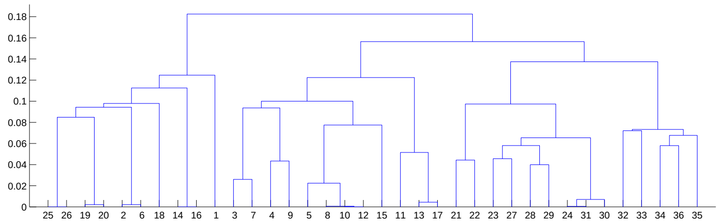

Fig. 5 Clusters of the extracted word candidates during the robot experiment. Word candidates 1-17 are paired with object Pudde, nr 18-29 with object Siffy, and 32-36 with object bollen.

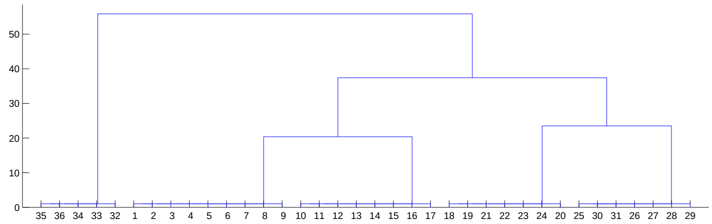

Fig. 6 Clusters of the extracted visual objects during the robot experiment. Objects 1-17 corresponds to object Pudde, nr 18-29 to object Siffy, and 32-36 to object bollen.

these the doll "Kuckan" was present and in 13 of them the doll "Siffy". In total the word "Kuckan" is mentioned 15 times and "Siffy" is mentioned 6 times.

In this experiment we limit the short-term memory to 10 s. The utterances enter in the short-term memory one at a time and any utterance older than 10 s is erased from the memory. Word candidates that also have an assigned object label are transferred into the long-term memory.

After searching all utterances for word candidates we cluster all the candidates in the long-term memory. The result can be found in Figure 7. Here we don't have any hierarchical tree for the visual objects. Instead we use the labels assigned by hand that can be used for calculating the mutual information criterion. Doing so gives us that the object Kuckan is best represented by word candidates (5 6 3 9 11 10 22 4 7 17 18) and Siffy by (32 33 34). Listening to the word candidates confirms that they represent the names of the dolls, but the segmentation is not as clear as in the humanoid experiment and there are a few outliers. Among the word candidates associated with Kuckan, nr 22 was unhearable and nr 17 and 18 were non-words but with a prosodic resembly of the word "Kuckan". For the word candidates associated with Siffy all contained parts of initial words.

When repeating the experiment without bias on focal stress and utterance-final position, the number of word candidates grew significantly resulting in lots of outliers being associated with both the objects. In the case of Kuckan it even caused the correct word candidates to be excluded from the branch that was associated with the object. However, it should be stated that the experiment was very small in order to draw any general conclusions.

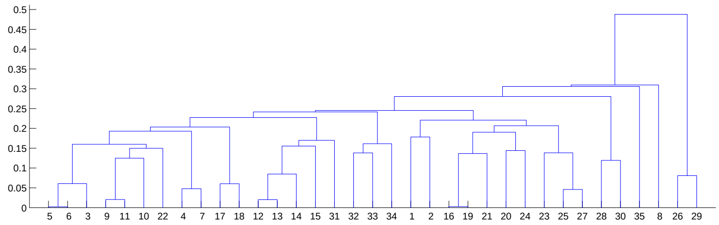

Fig. 7 Cluster formations from word candidates taken from infant directed speech. Word candidates between 1 and 30 are paired with object Kuckan and word candidates between 31 and 35 are paired with Siffy. Using the mutual information criterion, cluster (32 33 34) gets associated with Siffy and cluster (5 6 3 9 11 10 22 4 7 17 18) gets associated with Kucka.

4.2 Learning target positions

The objective of the second experiment is to show how the robot can learn target positions for a number of Portuguese vowels.

To learn vowels the robot first has to create an initial sound-motor map. Using the initial map it can then try to imitate the caregiver in order to get some first estimated motor configurations that represent vowels in the speech motor vocabulary. Local babbling is used to explore the neighbourhood of the terms in the vocabulary, while the caregiver gives feedback on the result. Finally, the clustering algorithm is used to group all positions learned into a feasible number of elements in the vocabulary.

The initial sound-motor map is created through random babbling. We generated 10000 random positions vectors for this phase. Each vector contains information about the position of the 6 articulators used in Maeda's model. These configurations are used by the speech production unit to calculate the resulting sound, which is coded in MFCC by the auditory unit. The sound-motor-map then tries to map the MFCC back to the original articulator positions that originated the sound. The error resulting from the comparison with the correct motor configuration given by the random articulator generator is used with a back-propagation algorithm to update the map. Repeating this will create an initial map between sound and the articulator positions used to create this sound.

The second step can be seen as a parroting behaviour where the robot tries to imitate the caregiver using the previously learned map. Since the map at this stage is only trained with the robot's own voice, it will not generalize very well to different voices. This may force the caregiver to change his/her own voice in order to direct the robot. There can also be a need to over-articulate, i.e. exaggerate the positions of the articulators in order to overcome flat areas in the maps that are a result of the inversion problem. When two or more articulator positions give the same sound the initial maps tends to be an average of those. However, for vowels the articulator positions are usually naturally biased towards the correct position as the sound is more stable around the correct positions than around the alternative positions. For most of the vowels it was not necessary to adapt the voice too much. Typically between one and ten attempts were enough to obtain a satisfying result. When the caregiver is happy with the sound produced by the robot it gives positive feedback which causes the robot to store the current articulator positions in its speech motor vocabulary. Using this method the caregiver was able to teach the robot prototype positions for nine Portuguese vowels. Visual inspection of the learned articulator positions showed that the positions used by robot are similar to those used by a human speaker, Figure 3.

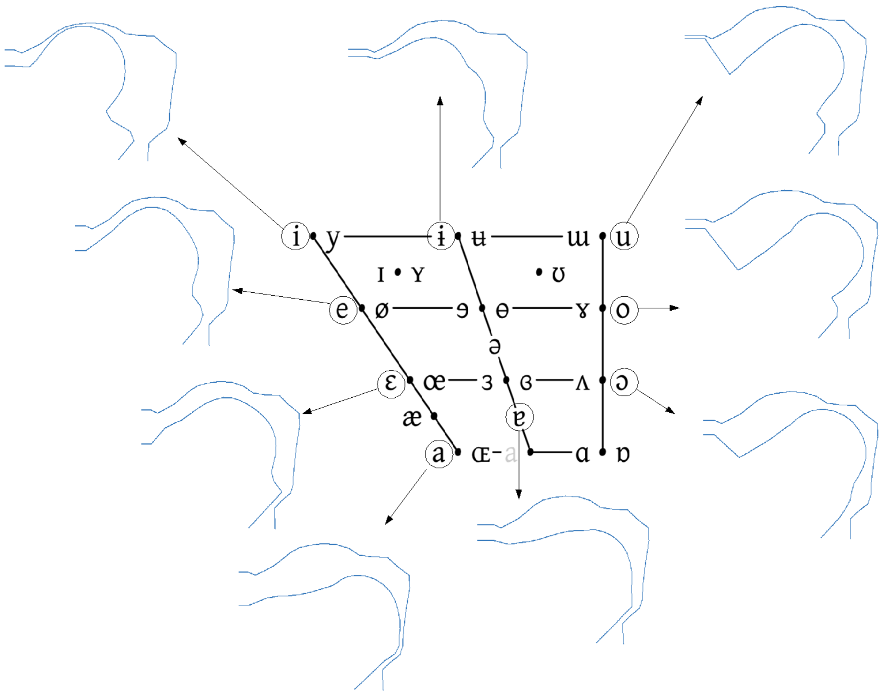

Fig. 8 Articulator positions used by the robot for the Portuguese vowels. In the center we show the positions of the vowels in the International Phonetic Alphabet (IPA). The vertical axis in the IPA corresponds to the vertical position of the tongue and the horizontal axis to the front-back position when the vowel is pronounced by a human speaker. For the simulated articulator positions used by the robot the upper line corresponds to the soft palate and the lower line to the tongue. There is a strong correlation between how the robot and a human articulate the vowels.

Since the vowel positions were learned under controlled forms where only one position was stored for each vowel sound we did not do any clustering of the target positions, but simply let each position represent a pseudo-phoneme. In [24] we did a larger scale experiment where 281 target positions were learned, each representing one of the nine vowels above. We then used the hierarchical clustering algorithm together with GAP statistics to group the target positions into a number of pseudo-phonemes. This showed that the robot automatically would group the positions into nine pseudo-phonemes corresponding to the Portuguese vowels.

Here we instead study how well the robot is able to recognize the learned vowels when those are pronounced by human speakers. We especially look at how the recognition rate is improved as a result of the different stages of babbling and interaction. To study this training and test data were collected with 14 speakers (seven males and seven females) reading words that included the nine Portuguese vowels above. We used the vowels from seven speakers for training and the other seven for testing. Each speaker read the words several times, and the vowels were hand labelled with a number 1 to 9. The amplitude of the sound was normalized and each vowel was then divided into 25 ms windows with 50% overlap. Each window was then treated as individual data which resulted in a training set of 2428 samples, and a test set of 1694 samples.

During training, we simulated the interaction where the humans' imitate the robot by having the robot pronouncing one of its vowels at the time, and then present the robot with the same vowel from one of the humans in the training set. In this step we used both auditory and visual input. The auditory input consisted of a single window of 25 ms sound, and the visual input is an image showing the face of the human at the same instant of time. The robot then mapped these inputs to its vocal tract positions, compared the result with the position used by the robot to create the same sound, and used the error to update the both the auditory-motor map and the vision-motor map.

For testing, we let the robot listen to the vowels pronounced by the speakers in the test set, i.e. speakers previously unknown to the robot. The input was mapped to the robot's vocal tract positions and the mapped positions were compared to the vowel positions stored in the speech motor vocabulary. Based on the minimum Euclidean distance, each position was classified as one of the stored vowel positions.

We performed this test several times using the maps obtained at each of the different stages of babbling and interaction. First we tested how well the robot was able to map the human vowels using maps that had only been trained using the robot's own voice, i.e. after the initial random babbling. As expected at this stage, the estimated positions were relatively far from the correct ones and it was not possible to recognize more than 18% of the human vowels. This is mainly due to the difference between the voice of the robot and the voices of the the human adults in the test set, and it is because of this that the human caregiver may need to adapt his or her voice during the early interaction with the robot.

When the robot has already had some human interaction, through the people in the training set, we noticed a significant increase in the performance. The distance between the vocal tract positions estimated from the human utterances in the test set, and the positions used by the robot to create the same utterance, decreased, and the recognition rate improved. Using only sound as input, the recognition rate became close to 58%, and using both sound and visual data the recognition rate reached 63%. A summary of the results is shown in Table 1.

Table 1 Recognition rates at the different stages of development

Training data	Sum of square distance	recognition rate
Only babbling	9.75	18%
Using interaction	0.52	58%
Using interaction with vision	0.47	63%

5 Conclusions

In this work we have described a method for language acquisition in humanoid robots that mimics the process of language acquisition in children. Our approach relies on two main axes. Firstly, we adopt a developmental and ecological approach driving the robot through sensorimotor coordination stages (babbling) and mediated by the interaction with the caregivers, to acquire a speech vocabulary. Secondly, we use motor representations both for speech production and recognition as suggested by the studies in mirror neurons.

We discuss how different interaction scenarios between the robot and its caregiver(s) can help the robot to find linguistic structures such as words and phonemes without the need for pre-programmed linguistic knowledge. In this discussion we identify a number of enabling capabilities that must be embodied in the robot for it to be able to participate in the interactions. Especially we find the need for multimodal sensors, and the need for the robot to not only retrieve speech sounds, but also to be able to produce them. We show how these capabilities can be implemented in a humanoid robot.

Through the described interaction scenarios, the robot is able to learn both word-like structures that are associated with objects in the robot's visual field, and also underlying structures in the form of target positions for a number of vowels.

Acknowledgements This work was partially supported by EU Project CONTACT and by the Fundação para a Ciência e a Tecnologia (ISR/IST pluriannual funding) through the POS Conhecimento Program that includes FEDER funds.

References

1. Albin, D. D., and Echols, C. H., "Stressed and word-final syllables in infant-directed speech", Infant Behavior and Development, 19, pp 401-418, 1996
2. Andruski, J. E., Kuhl, O. K., Hayashi, A., "Point vowels in Japanese mothers' speech to infants and adults", The Journal of the Acoustical Society of America, 105, pp 1095-1096, 1999
3. Batliner, A., Biersack, S., Steidl, S., "'The Prosody of Pet Robot Directed Speech: Evidence from Children", Proc. of Speech Prosody 2006, Dresden, pp 1-4, 2006
4. Burnham, D., "'What's new pussycat? On talking to babies and animnals", Science, 296, p 1435, 2002
5. Cover, T. M., Thomas, J. A., "Elements of information theory", Wiley, July 2006
6. Crystal, D. "Non-segmental phonology in language acquisition: A review of the issues", Lingua, 32, 1-45, 1973
7. Davis, S. B., Mermelstein, P., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Transactions on Acoustics, speech, and signal processing, Vol. ASSP-28, no. 4, August 1980
8. de Boer, B. (2005)., "Infant directed speech and evolution of language", In Evolutionary Prerequisites for Language, Oxford: Oxford University Press, 2005, pp. 100Ű121
9. Fadiga, L., Craighero, L., Buccino, G., and Rizzolatti, G., "Speech listening specifically modulates the excitability of tongue muscles: a TMS study", European Journal of Neuroscience, Vol 15, pp. 399-402, 2002
10. Ferguson, C. A., "Baby talk in six languages", American Anthropologist, 66, pp 103-114, 1964
11. Fernald, Al, "The perceptual and affective salience of mothers' speech to infants", In The origins and growth of communication, Norwood, N.J, Ablex., 1984
12. Fernald, A., "Four-month-old infants prefer to listen to Motherese", Infant Behavior and Development, 8, pp 181-195, 1985
13. Fernald, A., and Mazzie, C., "Prosody and focus in speech to infants and adults", Developmental Psychology, 27, pp. 209-221, 1991
14. Gallese, V. and Fadiga, L. and Fogassi, L. and Rizzolatti, G. "Action Recognition in the Premotor Cortex", Brain, 199:593-609, 1996
15. Gustavsson, L., Sundberg, U., Klintfors, E., Marklund, E., Lagerkvist, L., Lacerda, F., "Integration of audio-visual information in 8-months-old infants", in Proceedings of the Fourth Internation Workshop on Epigenetic Robotics Lund University Cognitive Studies, 117, pp 143-144, 2004
16. Fitzgibbon, A., Pilu, M., and Risher, R. B., "Direct least square fitting of ellipses", Tern Analysis and Machine Intelligence, 21., 1999
17. Fitzpatrick, P., Varchavskaia, P., Breazeal, C., "Characterizing and processing robotdirected speech", In Proceedings of the International IEEE/RSJ Conference on Humanoid Robotics, 2001
18. Fukui, K., Nishikawa, K., Kuwae, T., Takanobu, H., Mochida, T., Honda, M., and Takanishi, A., "Development of a New Humanlike Talking Robot for Human Vocal Mimicry", in proc. International Conference on Robotics and Automation, Barcelona, Spain, pp 1437-1442, April 2005
19. Guenther, F. H., Ghosh, S. S., and Tourville, J. A., "Neural modeling and imaging of the cortical interactions underlying syllable production", Brain and Language, 96 (3), pp. 280-301
20. Hastie, T., "The elements of statistical learning data mining inference and prediction", Springer, 2001
21. Higashimoto, T. and Sawanda, H., "Speech Production by a Mechanical Model: Construction of a Vocal Tract and Its Control by Neural Network" in proc. International Conference on Robotics and Automation, Washington DC, pp 3858-3863, May 2002

22. Hirsh-Pasek, K., "Doggerel: motherese in a new context", Journal of Child Language, 9, pp. 229-237, 1982
23. Hörnstein, J. and Santos-Victor, J., "A Unified Approach to Speech Production and Recognition Based on Articulatory Motor Representations", 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Diego, USA, October 2007
24. Hörnstien, J., Soares, C., Santos-Victor, J., Bernardino, A., "Early Speech Development of a Humanoid Robot using Babbling and Lip Tracking", Symposium on Language and Robots, Aveiro, Portugal, December 2007
25. Hörnstein, J., Gustavsson, L., Santos-Victor, J., Lacerda, F., "Modeling Speech imitation", IROS-2008 Workshop - From motor to interaction learning in robots, Nice, France, September 2008.
26. Hörnstein, J., Lopes, M., Santos-Victor, J., Lacerda, F., "Sound localization for humanoid robots - building audio-motor maps based on the HRTF", IEEE/RSJ International Conference on intelligent Robots and Systems, Beijing, China, Oct. 9-15, 2006
27. Jusczyk, P., Kemler Nelson, D. G., Hirsh-Pasek, K., Kennedy, L., Woodward, A., Piwoz, J., "Perception of acoustic correlates of major phrasal units by young infants", Cognitive Psychology, 24, pp 252-293, 1992
28. Kanda, H. and Ogata, T., "Vocal imitation using physical vocal tract model", 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Diego, USA, October 2007, pp. 1846Ű1851
29. Kass,M., Witkin, A., and Terzopoulus, D., "Snakes: Active contour models", International Journal of Computer Vision., 1987
30. Krstulovic, S., "LPC modeling with speech production constraints", in proc. 5th speech production seminar, 2000
31. Kuhl, P., Andruski, J. E., Christovich, I. A., Christovich, L. A., Kozhevnikova, E. V., Ryskina, V. L. et al., "Cross-language analysis of Phonetic units in language addressed to infants", Science, 277, pp. 684-686, 1997
32. Kuhl, P. and Miller, J., "Discrimination of auditory target dimensions in the presence or absence of variation in a second dimension by infants", Perception and Psychophysics, 31, 279-292, 1982
33. Lacerda, F., Marklund, E., Lagerkvist, L., Gustavsson, L., Klintfors, E., Sundberg, U., "On the linguistic implications of context-bound adult-infant interactions", In Genova: Epirob 2004, 2004
34. Lacerda, F., Klintfors, E., Gustavsson, L., Lagerkvist, L., Marklund, E., Sundberg, U., "Ecological Theory of Language Acquisition", In Genova: Epirob 2004, 2004
35. Lacerda, F., "Phonology: An emergent consequence of memory constraints and sonsory input", Reading and Writing: An Interdisciplinary Journal, 16, pp 41-59, 2003
36. Lenneberg, E., Biological Foundations of Language, New York: Wiley., 1967
37. Liberman, A. and Mattingly, I., "The motor theory of speech perception revisited"Š, Cognition, 21:1-36, 1985
38. Lien, J. J.-J., Kanade, T., Cohn, J., and Li, C.-C., "Detection, tracking, and classification of action units in facial expression", Journal of Robotics and Autonomous Systems., 1999
39. Lienhart, R. and Maydt, J. , "An extended set of haar-like features for rapid object detection", IEEE ICIP, 2002, pp. 900Ű903
40. Liljencrants, J. and Fant, G., "Computer program for VT-resonance frequency calculations", STL-QPSR, pp. 15-20, 1975
41. Maeda, S., "Compensatory articulation during speech: evidence from the analysis and synthesis of vocat-tract shapes using an articulatory model", in Speech production and speech modelling (W. J. Hardcastle and A. Marchal, esd.), pp. 131-149. Boston: Kluwer Academic Publishers
42. Mulford, R., "First words of the blind child", In Smith M D & Locke J L (eds): The emergent lexicon: The child’s development of a linguisticvocabulary. New York:Academic Press., 1988
43. Nakamura, M. and Sawada, H., "Talking Robot and the Analysis of Autonomous Voice Acquisition"Š in proc. International Conference on Intelligent Robots and Systems, Beijing, China, pp 4684-4689, October 2006

44. Nowak, M. A., Plotkin, J. B., Jansen, V. A. A., "The evolution of syntactic communication", Nature, 404, pp 495-498, 2000
45. Roy, D. and Pentland, A., "Learning words from sights and sounds: A computational model", Cognitive Science, 2002, vol 26, pp 113-146
46. Saffran, J. R., Johnson, E. K., Aslin, R. N., Newport, E., "Statistical learning of tone sequences by human infants and adults", Cognition, 70, 27-52, 1999
47. Sakoe, H. and Chiba, S., "Dynamic programming algorithm optimization for spoken word recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, 26(1) pp. 43-49, 1978, ISSN: 0096-3518
48. Stoel-Gammon, C., "Prelinguistic vocalizations of hearing-impaired and normally hearing subjects: a comparison of consonantal inventories", J Speech Hear Disord 53(3), 1988, pp 302-15.
49. Sundberg, U, and Lacerda, F., "Voice onset time in speech to infants and adults", Phonetica, 56, pp 186-199, 1999
50. Sundberg, U., "Mother tongue Ű Phonetic aspects of infant-directed speech", Department of Linguistics, Stockholm University, 1998
51. ten Bosch, L., Van hamme, H., Boves, L., "A computational model of language acquisition: focus on word discovery", In Interspeech 2008, Brisbane, 2008
52. Tibshirani, R., Walther, G., and Hastie, T., "Estimating the number of clusters in a data set via the gap statistic", Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2)., 2001
53. Vihman, M. M., Phonological development, Blackwell:Oxford., 1996
54. Vihman, M and McCune, L., "When is a word a word?", Journal of Child Language, 21, 1994, pp. 517-42
55. Viola, P. and Jones, M. J., "Rapid object detection using a boosted cascade of simple features", IEEE CVPR., 2001
56. Yoshikawa, Y., Koga, J., Asada, M., Hosoda, K., "Primary Vowel Imitation between Agents with Different Articulation Parameters by Parrot-like Teaching", in proc. Int. Conference on Intelligent Robots and Systems, Las Vegas, Nevada, pp. 149-154, October 2003