UNIVERSIDADE TÉCNICA DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Developmental approach to early language learning in humanoid robots

Karl Jonas Patrik Hörnstein

Supervisor: Doctor José Alberto Rosado dos Santos Victor

Thesis approved in public session to obtain the PhD Degree in
Electrical and Computer Engineering
Jury final classification: Pass With Merit

Jury

Chairperson:
Chairman of the IST Scientific Board

Members of the Committee:
Doctor Francisco Paulo de Lacerda
Doctor Isabel Maria Martins Trancoso
Doctor Mário Alexandre Teles de Figueiredo
Doctor José Alberto Rosado dos Santos Victor
Doctor Giampiero Salvi

2013

UNIVERSIDADE TÉCNICA DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Developmental approach to early language learning in humanoid robots

Karl Jonas Patrik Hörnstein

Supervisor: Doctor José Alberto Rosado dos Santos Victor

Thesis approved in public session to obtain the PhD Degree in
Electrical and Computer Engineering
Jury final classification: Pass With Merit

Jury
Chairperson:
Chairman of the IST Scientific Board
Members of the Committee:
Doctor Francisco Paulo de Lacerda, Full Professor, University of Stockholm, Sweden
Doctor Isabel Maria Martins Trancoso, Professora Catedrática do Instituto Superior Técnico, da Unversidade Técnica de Lisboa
Doctor Mário Alexandre Teles de Figueiredo, Professor Catedrático do Instituto Superior Técnico, da Unversidade Técnica de Lisboa
Doctor José Alberto Rosado dos Santos Victor, Professor Catedrático do Instituto Superior Técnico, da Unversidade Técnica de Lisboa
Doctor Giampiero Salvi, Assistant Professor, KTH Royal Institute of Technology, Stockholm, Sweden

Funding Institutions
FCT
Contact (European NEST Project 5010)

2013

Abstract

This thesis presents a developmental approach to language learning in humanoid robots. The objectives are both to find more flexible methods for language learning in machines, and to learn more about how infants acquire their language.

The proposed method takes an ecological approach, where language learning does not depend on any pre-programmed linguistic knowledge such as given phonemes or labeled speech data. Instead an initial set of words is learnt through general pattern matching techniques, by mimicking adult-infant interactions, and by taking advantage of the multimodal nature and the inherent structure of infant directed speech (IDS).

In parallel, initial speech units are learnt through imitation games, where the robot and a caregiver take turn imitating each other. These imitation games not only allow the robot to find useful speech sounds, but also to create an audio-motor map. This map works as "mirror neurons", allowing the robot to find the vocal tract positions used to produce a given speech sound, which is useful not only to reproduce the sound, but also to recognize the same.

Finally, initial words and speech units are combined in a statistical model, allowing the robot to overcome the limitations of the initial models.

Keywords: humanoid robots, language acquisition, developmental approach, ecological methods, imitations, mirror neurons

Resumo

Esta tese representa uma abordagem evolutiva da aprendizagem da linguagem em robots humanóides. Os objectivos são: (i) encontrar métodos mais flexíveis que facilitam a aprendizagem da linguagem em maquinas, e (ii) aprender mais sobre o desenvolvimento da linguagem nas crianças.

O método proposto utiliza uma abordagem ecológica onde a aprendizagem de linguagem não depende de conhecimentos linguísticos pré-programados, como fonemas. Em vez disso, um inicial conjunto de palavras é aprendido através de técnicas gerais como reconhecimento de padrões, imitando a interacção entre adultos e crianças e utilizando a natureza multimodal e estructura inerente do sinal dirigido às crianças.

Em paralelo, pseudo-fonemas são aprendidos através de jogos de imitações, onde o robot e o caregiver fazem turnos em imitar um do outro. Além disso, estes jogos servem para criar um mapa entre o som e os posições motoras utilizados para criar o mesmo. Este mapa funciona como neurónios-espelho, que permite o robot encontrar posições do tracto vocal utilizados para produzir um som. Isto é útil, não só para reproduzir o dado som, mas também para reconhecer o mesmo.

Em seguida, as palavras e os pseudo-fonemas são combinados num modelo estatístico, que permite superar as limitações dos modelos iniciais.

Palavras-chave: robots humanóides, abordagem evolutiva, métodos ecológicos, imitações, neurónios-espelho

Agradecimentos

Primeiro quero agradecer ao meu orientador científico, Prof. José Santos-Victor, por me ter convidado a juntar à equipa de Vislab, pela ajuda de elaborar o plano do trabalho que deu origem a esta tese, e pelo apoio na realização da mesma. O trabalho com a tese tem sido uma experiência enriquecedora e ao mesmo tempo sempre divertida, muito graças ao bom ambiente de trabalho no Vislab. Por isso quero estender os meus agradecimentos a todos os colegas e membros do Vislab.

Parte deste trabalho foi realizado no projecto europeu CONTACT. O consórcio deste projecto tem sido fundamental para a abordagem interdisciplinar da tese e muitas ideias nasceram de discussões com os parceiros do projecto. Por isso quero agradecer ao coordenador do projecto, Giulio Sandini, e a todos os parceiros.

Entre todos os colegas e parceiros, quero expressar um agradecimento especial:

Ao Manuel Lopes, pelo apoio científico e pela ajuda de controlar os motores do robot.

Ao Ricardo Beira, pela ajuda de modelar as orelhas.

Ao Ricardo Nunes, pelo desenvolvimento dos amplificadores para as orelhas e pelo apoio técnico sempre que houve um problema com os robots ou os computadores no Vislab.

Ao Francisco Lacerda e à Lisa Gustavsson no Stockholm University, que ofereceram conhecimentos essenciais sobre o desenvolvimento de fala nas crianças, e forneciam exemplos de interacções entre adultos e crianças que serviram para avaliar os métodos desenvolvidos nesta tese.

Finalmente quero agradecer a minha família: a minha mulher Sara Hörnstein pelo apoio pessoal e pela paciência e compreensão quando tive de faltar por causa de viagens ou de trabalho, e o nosso filho Francisco Hörnstein pela inspiração e alegria.

Mais uma vez, obrigado a todos.

Contents

Chapter 1

Introduction

This thesis proposes a developmental approach to language learning in humanoid robots that mimics some of the learning processes found in infants.

Spoken language is a powerful tool for human interactions, and potentially so also for human-machine interfaces. Advances in the area of automatic speech recognition (ASR) have made voice interfaces increasingly popular. When used in a controlled environment or for smaller vocabularies, state-of-the-art ASR-systems can already match or even improve on human word error rates. However, despite of all research effort in the area, current machines are still far from human language capabilities in terms of robustness and flexibility when used in more natural settings. Some of the main challenges are interspeaker differences, coarticulations, and out-of-vocabulary words. The main approach used to address those challenges has been to use more computing power and increase the speech database that is used to train the existing language models. Unfortunately, the difficulties may not be a result of insufficient processing power or insufficient speech samples, but rather a fundamental flaw in the architecture of current models [87]. Compared with current approaches to language learning in machines, infants are able to acquire impressive language skills from very little speech exposure.

A humanoid robot, which needs to work in a more natural environment, may therefore benefit from an alternative approach to language learning with learning capabilities more similar to those of an infant. While infants' learning capabilities are well documented, it is still far from completely understood how they are able to learn those. The European project Contact [17], in which part of this thesis has been developed, has addressed this by studying the coupling between production and perception and by exploring parallels between learning to speak and learning to make hand gestures. This is a multidisciplinary project that includes researchers from different areas such as linguistics, psychology, neuroscience, and engineering. Built on the knowledge in this project, the role of this thesis has been to develop language

acquisition capabilities with biological plausibility in an artificial system. Especially, this thesis combines two lines of research: (i) the ecological or emergent approach to language learning, and (ii) the use of motor primitives for imitations and speech recognition.

According to the ecological and emergent approach to language learning, the infant's linguistic and phonetic knowledge evolve gradually as the infant is forced to deal with an increasing amount of information [73]. It is assumed that no innate linguistic or phonetic knowledge, such as phonemes or grammar, is needed in order to acquire the initial language capabilities. This assumption is debated, and nativists argue that some universal knowledge about language must be present already at birth [16]. It is also a strong contrast to the typical approach used in todays computer-based systems where phoneme and word models are preprogrammed and trained with large databases containing labeled speech examples of individual phonemes and words. Following the ecological approach it is instead necessary to make use of the linguistic cues and structures available directly in the speech signal, and to extract the relevant parts of the signal using general information processing principles, such as detecting recurrent patterns. While this is a challenging route to take in an artificial system, it has the advantage of allowing the system to specifically acquire the language knowledge needed for the tasks and environment where it is deployed.

The second line of research is the use of embodiment and motor primitives for recognition. The hypothesis here is that we recognize a gesture by simulating the same motion and comparing with gestures that we already know how to produce. This is motivated by the discovery of mirror neurons that fire both when performing a gesture and when observing the same gesture performed by another. The existence of mirror neurons was first found during grasping experiments [38], and there is also evidence that the motor area is involved in speech recognition tasks [27]. This means that production and recognition are coupled and that there is a link between what we can do ourselves and how we interpret what others are doing. For speech this means that the auditory signal is first transformed into a speech gesture, i.e. an articulation of the vocal tract, which may then facilitate both reproduction and recognition of the speech sounds.

While there is strong evidence of the existence of both mirror neurons and of linguistic cues in the speech signal directed to infants, it is not clear exactly how infants make use of these when learning their language. By trying to formalize those ideas and implement them in a humanoid robot we aim to create a tool that can be used both to increase the understanding of infant language learning, and to find more flexible methods for language learning in humanoid robots. We focus on humanoid robots both because of the challenges introduced from their need to interact with humans in natural settings such as household environments, and because of the opportunities to use embodiment and richer types of interactions due to their human-

like body structure and rich set of sensors.

In summary, this thesis studies how the ecological perspective of infant language learning, and the concept of "mirror neurons", can be implemented in a humanoid robot. By doing this, it aims at two goals: (i) to find more flexible methods for language learning in humanoid robots, and (ii) to increase the understanding about infant language learning. To delimit this work, only early language learning, i.e. learning single words and speech units, is considered. This corresponds loosely to infants' language learning during their first year of life.

The remainder of this chapter gives an introduction to ecological language acquisition and motor-based learning, and discusses how these theories can be applied to language learning in humanoid robots. Based on this we then give an overview of the approach taken in this thesis and discuss some of the main contributions of this thesis.

1.1 Ecological language acquisition

In this thesis we follow the ecological perspective by not assuming any innate linguistic or phonetic knowledge. Instead we take advantage of the inherent structure available in the speech signal. In order to mimic infants' language acquisition it is important to follow a similar developmental path, and to make sure that we create a similar ecological setting.

This section first gives an introduction to the developmental path of infants' language acquisition, from birth to around one year of age, when infants are able to produce and recognize single words. We then take a more detailed look at the typical speech signal directed to infants and how the structure of this signal may facilitate language learning. We also look at how information coming from other sensor modalities can further be used to guide the language learning.

1.1.1 Developmental path

At birth, infants are able to discriminate phonetic contrasts of all languages. Adults, on the other hand, have very strong preference towards the sounds used in their native language. This is sometimes described as a phonetic magnet that forces sound to be perceived as one of the phonemes that are used in the particular language [70]. Infants start to show signs of such a phonetic magnet for the vowel sounds of their native language already at around six months of age. The same happens for consonants at around 11 months of age. This preference for speech sound utilized in the native language indicates that the infants have been able to acquire language specific phoneme models. It is not known how infants do that or to which extent those phoneme models are already innate. However, whether the acquisition

is done from scratch or by simply tuning a set of innate phoneme models, learning is still necessary. One of the hypotheses in this work is that imitation games, where the infant and the caregiver take turn in imitating each other, may have an important role in the phoneme acquisition. However, to engage in such interaction games, infants need to be able to produce sounds. Interestingly, infants seem to begin producing these sounds shortly before they show a preference for perceiving these same sounds. The first vowel sounds are typically produced between the second and the third month after birth, and canonical babbling that include consonant begins around month seven.

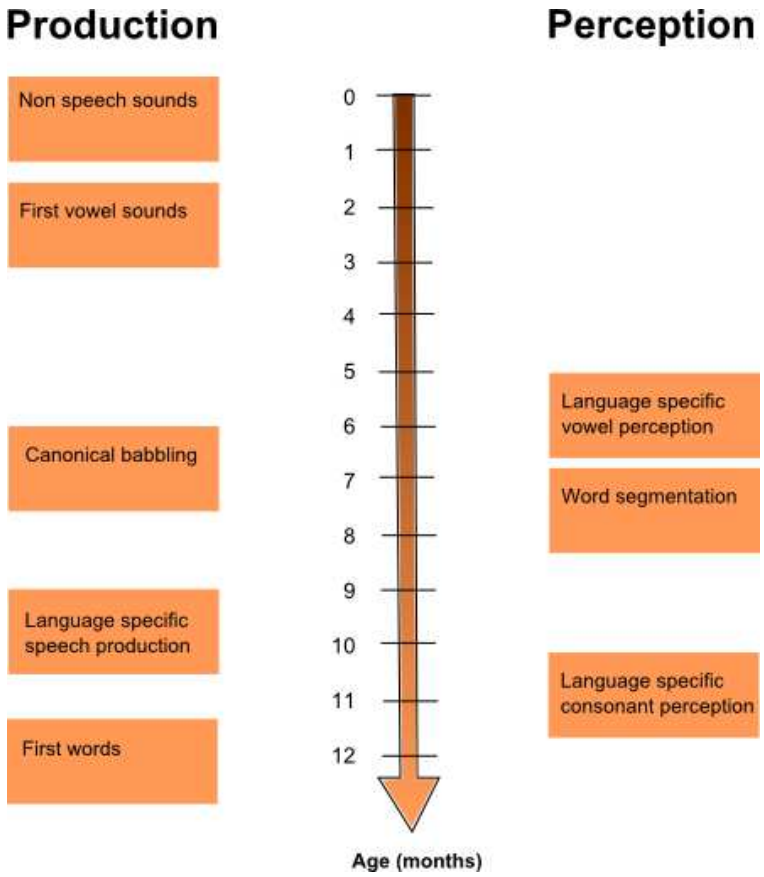

Figure 1.1: Developmental path of an infant's language capabilities. At birth infants can discriminate phonetic contrasts of all languages, but later develops a preference towards the sounds used in their native language. This happens approximately at the same time as they learn to produce those sounds.

At around 6-8 months infants begin to show signs of being able to segment words from fluent speech. This is a difficult task since words are generally not separated by silence, even if it may appear so for an adult listening to its own native language. However, looking at the speech signal it quickly becomes clear that this is not the case, Figure 1.2. Instead the speech signal contains several other cues, based on either prosodic or statistical features, that

may allow infants to do this initial segmentation. At this age infants are still not able to reproduce the words, but they are able to associate the words with visual objects in their surrounding at around 8 months of age. The first words are usually produced around 12 months.

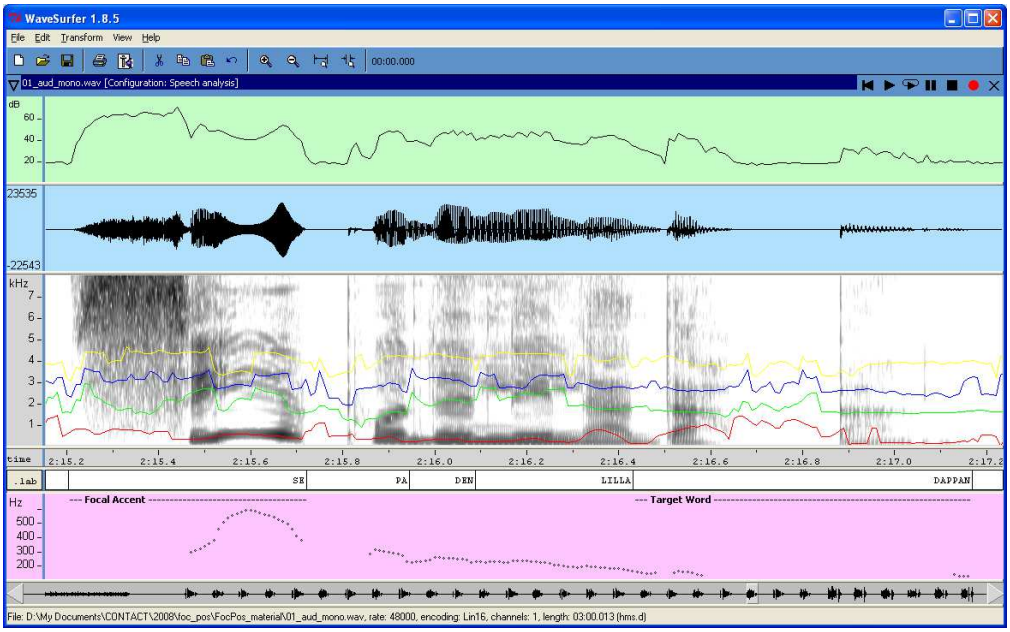

Figure 1.2: Speech signal for the utterance "Se på den lilla Dappan". As seen in the intensity curve (green pane) and waveform (blue pane), there are generally no periods of silence between the words.

1.1.2 Infant directed speech

An important part 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 [6] [70] [29] and is characterized by long pauses, repetitions, high fundamental frequency, exaggerated fundamental frequency contours [30] and hyperarticulated vowels [70]. These phonetic modifications seem to be an intuitive strategy, used automatically by the adults, that is both attractive and functional for the infant [114]. A very similar speech style is found in speech directed to pets [49] [15], and to some degree also in speech directed to humanoid robots [35], and pet robots [9].

The typical speech style and structure of IDS contains several important cues that can facilitate the language acquisition. To make the discussion of those more clear, a separation is done between cues related to the prosody and to statistical features.

Prosodic cues

The prosody is related to the rhythm, stress, and intonation of speech. Acoustically it involves variation in syllable length, intensity, pitch, and the formant frequencies of speech sounds. These are variations of relatively low frequencies compared to the speech signal itself and can therefore be perceived by the infant already inside the mother's belly, which acts as a low-pass filter for the surrounding sound.

It has been found that newborn infants are capable of separating between its native language and languages from other language groups, while they do not seem to separate between languages that have similar prosody [85]. It is therefore likely that some familiarization of the prosody take place even before birth, and that this familiarization helps the infant to focus on speech sound in general, and speech sound from the native language in particular.

This ability to quickly learn and to recognize prosodic patterns may also serve as a basis for learning more complex acoustic patterns. The prosody has been found to be used by infants for word segmentation [65] [84].

In this thesis, prosodic features are used mainly for two tasks, (i) to find target words in fluent speech, and (ii) to detect when a caregiver is imitating the infant.

The first task is relatively well studied. Fernald and Mazzie [32] 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 [2] found that an even larger portion of the target words (87% - 100% depending of subject) occurred in final position when the subjects were asked to present a number of items to an infant. There are indications that infants take advantage of these features and therefore have easier to learn target words with focal stress and utterance final position [44].

The second task has received much less attention from researchers. A study of this has therefore been performed as a part of this thesis together with researchers from the University of Stockholm [55]. This study showed that prosodic features can give important cues to when the caregiver is imitating the infant. While no studies have been performed to verify that infants actually make use of these features, a second study at the University of Stockholm showed that altering the prosody can indeed change adults perception of an utterance as being an imitation or not.

Statistical cues and recurring patterns

The speech signal in general, and IDS in particular, contains lots of repetitions: phonemes are reused and combined to form different words, and the words are then combined to construct different sentences. However, not all combinations are used, and those that are used occur with different frequencies. It has been suggested that 8 months old infants make use of this statistical properties when segmenting words [103]. In that work 24 8-months-old infants each listened to 2 min of continuous speech consisting of a number of three-syllable nonsense word, such as "bidaku", "padoti", and "golabu", repeated in random order. The speech stream was constructed so that the transitional probabilities between syllable pairs within words (e.g. bida, daku) were 1.0, while the probabilities for the syllable pairs between words (e.g. kupa, tigo) were only 0.33. After the exposure, the infants showed a significant difference in listening time for sentences containing the target words, and sentences not containing any of the target words.

The main problem with this approach is that in order to learn the sequential probabilities for the syllable transitions, infants must also have been able to acquire models for the individual syllables. As stated earlier, infants at this age only show signs of grouping vowel sounds into language specific speech units, but not yet consonants. Also, in a later experiment [104], 8-months-old infants show the same tendency to learn statistical transition probabilities for tone sequences.

This leads to believe that there are some alternative explanations as to how infants are able to segment words based on the repetitive nature of the signal. One such explanation could be the use of pattern matching techniques, where the infants look for recurring patterns in the speech signal [73].

While pattern matching can explain how infants are able to segment words without the need for an underlying phoneme model, a statistical model based on phonemes is much more powerful when it comes to separating between words in a large vocabulary of speech. It is therefore likely that such statistical models become effective and necessary as the vocabulary grows and the underlying phoneme model is in place.

1.1.3 Multimodal information

Infant's first words usually refer to persons and objects in the surrounding. To be able to ground those first words infants can make use of multimodal information and especially the combination of audio and vision.

Whereas communication between adults is usually 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 [72]. When for example playing with a toy, the name of the toy is therefore likely to be mentioned several times within a relatively short time as the infant is being introduced to the toy. As discussed, finding such recurring patterns in the sound stream coming from the caregiver can help the infant to extract potential word candidates. These can then be linked to the visual representation of the object.

Infants are very sensitive to such co-occurring events, and a study with 8-months-old infants shows that they are able both to learn names of object and to ground those with visual objects [44]. In this study, two puppets are shown for the infant, one at each time, while a female voice read a number of sentences containing the name of each puppet. After being presented to each of the puppets individually, they are both shown at the same time on the screen while the female voice asks questions about one of the puppets. An eye-tracker measures the time that infant look at each of the puppets and a significant longer looking time was found for the puppet they were talking about.

1.2 Motor-based learning

By studying infant language learning, we get a good overview of the information available to the language learning infant, as well as what part of the signal that is used at each step of the developmental path. This is important information in order to take full advantage of both the information available in the speech signal and to make use of other modalities such as vision. However, these studies do not provide any insights to how the information is processed in the brain, which may be equally important in order to mimic these capabilities in a humanoid robot.

The signal creates sensorial stimulus that may either be used directly to interpret the meaning, or first be transformed in some way in order to facilitate the interpretation. One hypothesis, which is investigated in this thesis, is that the sensorial stimuli is first associated with the articulation that was used to create the original signal, and that this articulation is useful in order to interpret the meaning of the signal.

This was first suggested by the Motor Theory of speech perception [76]. In the original work, it was assumed that this transformation was hard-wired in the brain and thus considered to be innate. However, it has later been argued that this transformation may be learned rather than innate [37]. The latter is motivated by the discovery of "mirror neurons", which is a group of neurons that fire both when a when a specific gesture is executed, and when observing someone that is executing the same gesture.

Mirror neurons were discovered by a group of neurophysiologists when studying grasping movements in the macaque monkey [38]. They found that some of the neurons that were

measured consistently fired both when the monkey picked up a piece of food and when a person performed the same movement. There are some indications that similar mirror neurons exists also for speech. A more recent work in neuroscience demonstrated an increased activity in the tongue muscles when listening to words that require large tongue movements [27].

This type of embodiment, where the sensorial signal is associated with an articulation, has already been found useful for learning to execute and recognize different types of grasps with a humanoid robot [80]. From this work we have identified three developmental steps that allow the robot to take advantage of motor-based learning: (i) creating a sensory-motor map by babbling, (ii) gaining source invariance, and (iii) learning new gestures by imitation, see Figure 1.3.

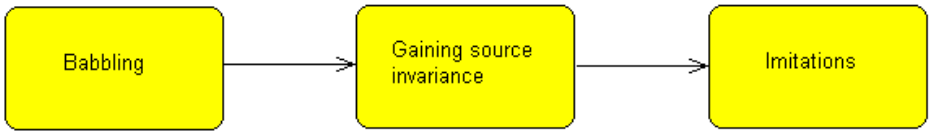

Figure 1.3: Developmental steps for motor-based learning

The first step aims at allowing the robot to associate sensory stimuli with the motoric gesture that produced it. This can be done by a sensory-motor-map that the robot learns through motor babbling. Babbling is a behavior that can be found also in infants, both for learning to control arm movements and when learning to produce speech sounds.

While babbling allows the robot to map the sensory stimuli caused by its own actions to the motor positions used to create them, the same action performed by another may create different sensory stimuli and cause the initial map to fail. The initial map must therefore be extended to compensate for the differences in sensor information depending whether the action is performed by the robot or by another person.

Once the map has been learnt, it can be used in order to imitate the caregiver. This capability has been found to be important, not only in order to learn new gestures, but also in a later stage in order to recognize the gestures that are being performed.

Here we will take a closer look at each of the three steps and identify similarities and differences between hand gestures and speech gestures, as summarized in Figure 1.4.

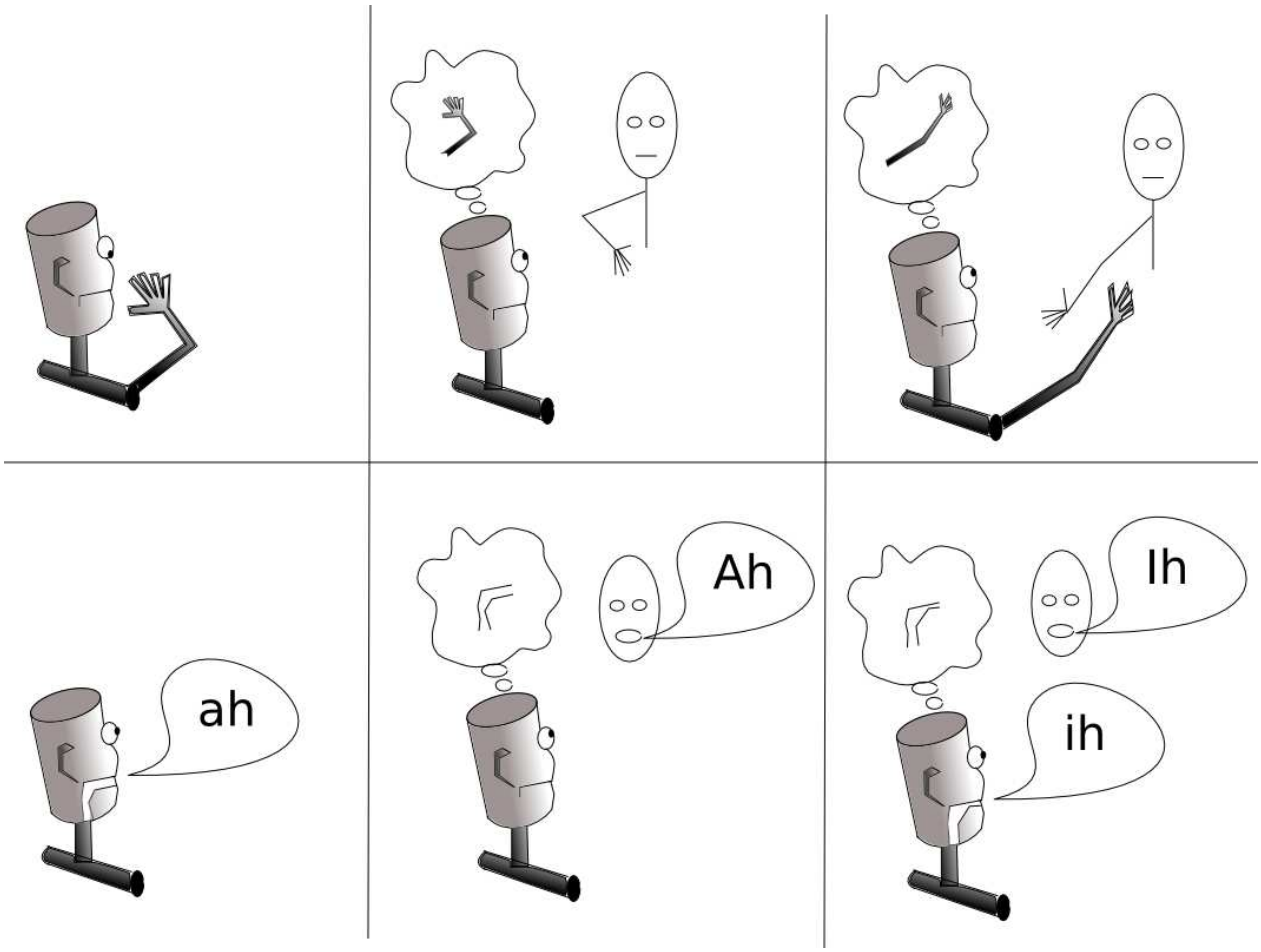

Figure 1.4: Similarities and differences between motor-based learning of hand gestures and speech for each of the developmental steps: babbling (left), gaining source invariance (center), and imitation (right). Apart from the obvious difference between the sensors and actuators used when learning hand gestures and speech, gaining source invariance is more difficult for speech learning due to large interspeaker differences of the vocal tract.

1.2.1 Babbling

The first step towards being able to imitate another person, whether the imitation is used in order to recognize what the person is doing or in order to learn how to produce a new gesture, it is first necessary to learn how to control the own production system. This is done through babbling, an exploration process that allows the child to explore different motor actuations and the corresponding consequences in the sensory system.

The most obvious difference between learning to produce speech sounds and learning to perform arm movements, are the "sensors" and the "actuators". For speech babbling the actuators are the muscles that control the shape of the vocal tract and the ears are our sensors, while for arm movements it is the muscles in the arm that serve as actuators and the eyes are the sensors that provide feedback. As motor-based learning is an embodied approach, where the actual shape of the body is seen as an important part of the cognitive system, it is necessary to have models that provide similar physiological capacities as humans.

During babbling, different actuator positions are experimented and the resulting sensor input is used to train the sensory-motor map. Both in the case of arm movements and speech babbling, the learning is complicated by the fact that several actuator positions can lead to the same or nearly the same sensor result.

The use of speech babbling in infant language learning is described in [75]. While babbling was first seen as an isolated process, it has later been shown to have a continuous importance for the vocal development [123]. 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 conspecifics [111], but also to establish visual contact with others [88].

1.2.2 Gaining source invariance

Having a good sensory-motor map for our own actions may not be sufficient to map the actions of others and find the necessary motor positions to reproduce the action. In the case of arm movements there is a significant difference in the visual stimuli caused by the different viewpoints when watching yourself execute an action compared to watching someone in front of you executing the same action. This can be solved with a view-point transformation (VPT). The VPT can be seen as an additional map that allows the observer to do a "mental rotation" that places the demonstrator's arm in correspondence with the observer's own body. This map can either be learned or pre-programmed.

For spoken language the problem is slightly different. Even though the position of the speaker does affect the sound slightly, it is of little or no significance for how the speech sounds are perceived. A map corresponding to the VPT of the visual input is therefore not

needed for the audio input. On the other hand, there are considerable differences in the vocal tract between different speakers. These differences are often enough to cause the map to fail, especially in the case of adult-infant or human-robot where there are significant differences between the voice used during babbling and the voice of the caregiver. While vocal tract normalization (VTN) [97] can be used to compensate for some differences in length of the vocal tract, speaker variation is still a largely unsolved problem.

An alternative method to obtain generalization is to train the initial sensory-motor map with the voice from several speakers. This can be done in an ecological way by using interactions. Even the initial babbling may actually be more than just a self-exploration task; it may also be the first step towards interaction. When a caregiver is present, he or she is likely to imitate the sound of the infant, giving the infant the possibility to create a map between its own utterances and that of the caregiver. In the previously referred imitation study at Stockholm University [55], it was shown that in about 20% of the cases, the response from the caregiver was seen as an imitation of the infant's utterance when judged by other adult listeners. It was 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.

By producing speech sound and detecting possible imitations, it can therefore be possible overcome the differences between the own voice and that of the caregiver. By repeating this kind of "imitation game" with several different caregivers it is possible for a robot to learn how to map sounds from different speakers to its own motor positions in order to reproduce or recognize those sounds, and thereby overcoming the problem with inter-speaker variations.

1.2.3 Learning by imitation

When the robot has learnt how to map an observed gestures produced by a human to its own motor primitives, it can use this to both facilitate the recognition of the gesture or and to learn new gestures.

If the sensory-motor map is able to fully reconstruct the motor positions used to reproduce the action, this step should be relatively straight forward. However, due to the complexity of this mapping, the sensory-motor map may not give a perfect result. In order to compensate for these shortcomings, the caregiver may try to alter the demonstration. For language learning this means that the caregiver actively changes his or her voice or overarticulates in order to help the infant to correctly pronounce the utterance. This behavior has been studied in [22]. By providing the infant with feedback on how well it is articulating, the infant can locate articulatory target positions that are useful for communication.

1.3 Approach of this work

This thesis takes an ecological approach to language acquisition and applies it to humanoid robots. More specifically, preprogrammed linguistic knowledge such as phone models are avoided and are instead learned through the interaction with a caregiver and by using motor learning. In contrast to the common data-driven approach to language learning, this approach takes advantage of a wider spectrum of features coming from several different modalities.

Another important part is the use of embodiment. The humanoid robot used in the experiments has physical models of human head and outer ears, Figure 1.5, and simulated models of the inner ear and the vocal tract. Creating models that closely match human audition and speech production is important when trying to mimic the human language learning capacity, especially when motor learning is used. The robot is also equipped with eyes in form of cameras, and has motors that can move the eyes and the head.

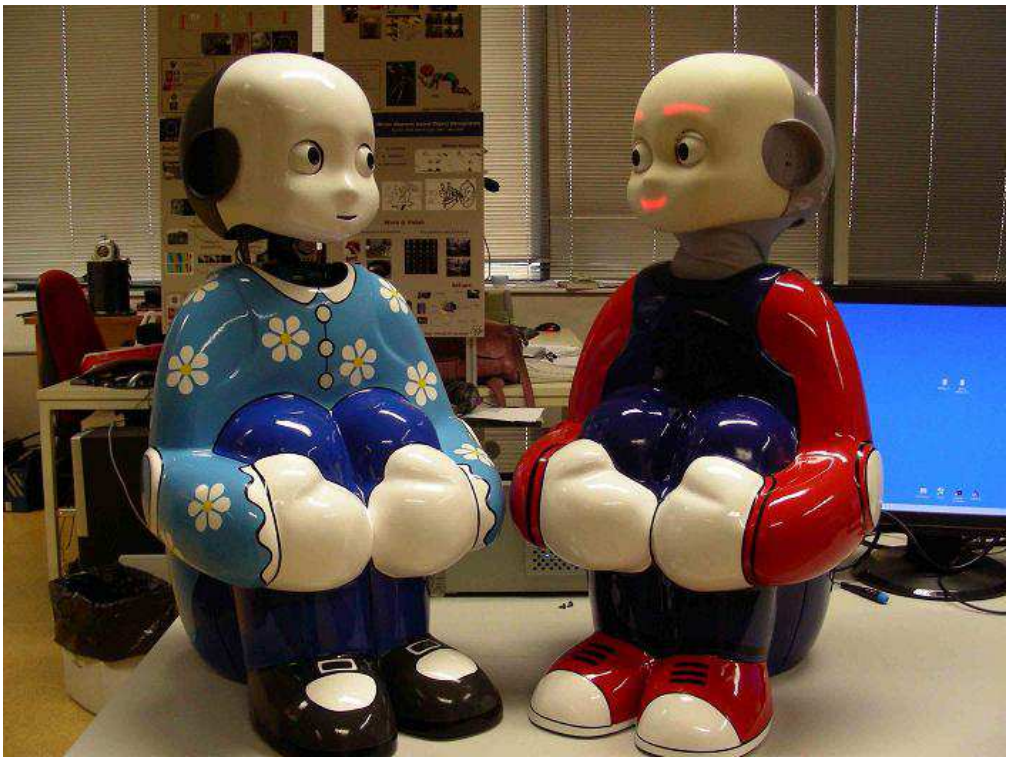

Figure 1.5: Chica and Chico - the humanoid robots used for the experiments in this thesis.

The different steps in the proposed developmental approach to early language learning in humanoid robots are outlined in Figure 1.6. The approach consists of two parallel paths where the robot independently learns an initial word model and initial speech units. These are then combined in a statistical model. We have also included a step for sound localization. Being able to localize speech sounds can facilitate the learning by allowing the robot to turn

towards the speaker and establish visual contact. Having visual contact with the caregiver is important for an infant's babbling and we will show that it can be useful also for robots in order to learn the sensory-motor map.

The sensory-motor map is learnt in the lower of the two parallel paths. This path follows the steps described above for learning through imitations. It is closely related both to how infants learn to produce speech sounds and how they learn to imitate different arm gestures. The goal of this path is both to learn a speaker invariant sensory-motor map that can map speech sounds produced by different speakers to the positions in the motor space of the robot, and to learn initial speech units. The initial speech units correspond to target positions in motor space that are used to create different speech sounds such as vowels.

The initial word learning takes advantage of some of the characteristics of infant directed speech (IDS) in order to extract sound patterns that are consistently used to describe different objects. Several characteristics are used. The most important is the fact that IDS contains lots of repetitions and that utterances typically refer to objects in the surrounding. By using pattern matching techniques such as dynamic time warp (DTW) we can find recurring patterns in the speech signal which may potentially correspond to words. By comparing which speech patterns that consistently co-occur with certain visual objects it is possible to find words that are associated with those objects. We also make use of the fact that the caregiver typically highlights target words by using focal stress and utterance final position.

In the final step, initial words and speech units are combined to create a statistical model. A statistical model such as a hidden Markov model (HMM) provides a very compact description of the words while at the same time taking into account statistical differences. This becomes a necessary step when the vocabulary grows and direct pattern matching is no longer sufficient to separate the words. The initial word learning is still important in order to create a training set for the statistical learning. Also, by delaying the creation of the statistical model until the sensory-motor map has been fully learnt, we can use this map in order to include motor information in the statistical model. Finally the speech units learnt through imitation can serve as a starting point for finding suitable speech units for the statistical model.

1.4 Thesis contributions

This thesis presents a developmental approach to early language learning in humanoid robots. More specifically it provides models and learning methods that allow the robot to acquire speech units and single words. This is done by mimicking infant language learning during their first year of life.

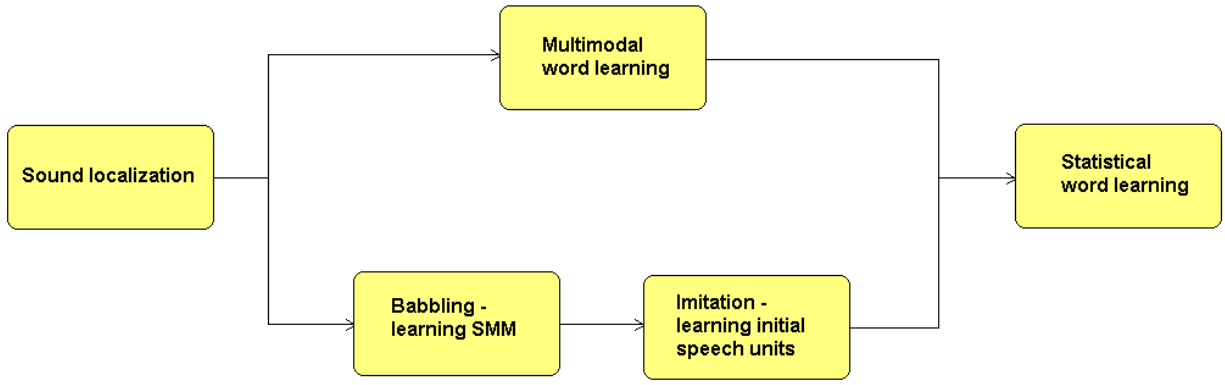

Figure 1.6: Developmental learning approach

On a large scale, the objectives of the work has been to focus on the following 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 ecological computational models for language acquisition in a humanoid robot, and (iii) to investigate similarities and differences between language learning and other learning processes such as grasping.

By doing this, the thesis contributes with insights both on how to create more flexible and human-like methods for language learning in humanoid robots, and to increase the understanding on how infants are able to learn their language.

The main contribution of thesis is probably the developmental learning approach itself, and the way that existing methods are combined in order to allow the robot to learn its language from natural speech and social interactions, without the need for pre-programmed linguistic knowledge such as phonemes. However, specific contributions have also been made both for the embodied models of vocal tract and the ears, as well as for each of the steps in Figure 1.6.

For the articulatory model an open source software, jArticulator [59], has been developed. The software is written in Java and is fully scriptable from either built-in jython console or through Matlab.

For sound localization, a new method for determining the elevation of the sound source has been developed. The method makes use of the fact that reflections in the outer ear or the ”pinna” provides spectral cues that can be used for sound source localization. An artificial pinna was developed that creates spectral notches comparable to those of the human ears. It also makes use of the fact that there are small asymmetries between our ears. The method includes algorithms for finding the spectral notches and learning to map those to the elevation of the sound source.

For the initial word learning we extend existing methods for multimodal word learning

by including features such as word positioning and prosody in order to find suitable target words. This makes the method more robust in natural adult-infant interactions.

For babbling and imitations several contributions have been made, most importantly for expanding the sensory-motor map to account for interspeaker differences. This is done by making use of the fact that caregivers tend to imitate the infant, which allow the infant to collect acoustic examples from several speakers when training the map. A method for detecting when the caregiver is imitating the infant, based on prosodic features, is proposed. Another contribution is the use of visual information in the sensory-motor map, which can improve the recognition rate for the initial speech units.

Finally, the statistical word learning is a novel method in that it only makes use of the information coming from natural interactions and optimizes the number of speech units based on those interactions.

1.5 Structure of the thesis

The first part of this thesis describes the embodiment of the sensor and actuator models. Chapter 2 describes the model of the articulatory system consisting of a tube model that produces synthesized speech and an articulatory model that simulates movements of the jaw, tongue, and lips. Chapter 3 describes the model of the auditory system, which consists of a physical model of the outer ear and computational models for the transformation of the signal in the inner ear. It also derives a number of acoustic features that are used for the different learning methods throughout the thesis. Separate features are derived for sound-source localization, the prosody, and the tonotopic representation.

Chapter 4 describes how the ear model can be used for sound-source localization and making the robot turn towards the speaker, which is important to be able to interact with a caregiver.

Chapter 5 and Chapter 6 are directly related to the motor-based learning described in this introduction. The first describes the babbling and inversion mapping, i.e. the learning of the initial sensory-motor map, and especially describe how to deal with the inversion problem caused by different articulations producing the same or similar speech sounds. The latter explain how to expand the map to include interspeaker variations, and use imitation for learning an initial set of speech units. This chapter also describes how the robot can decide if a given response from the caregiver should be considered as an imitation or not.

Chapter 7 is directly related to the ecological approach and describes how initial words can be learned by looking for recurring patterns in the speech signal and grounding those with objects in the visual field.

Chapter 8 describes how linguistic structure such as speech units can emerge when the initial vocabulary found in chapter 7 increases and direct pattern matching is no longer sufficient to handle the growing vocabulary. It also shows how the initial speech units found during the imitation experiment can help to bootstrap this.

The thesis ends with general discussion and conclusions in Chapter 9, together with some directions for future work.

Chapter 2

Modeling the articulatory system

In order to develop a computer-based system that can acquire language capabilities similar to that of an infant, we must first make sure that the system can produce and perceive speech sounds. In this chapter we will concentrate on how to produce speech sounds and leave audition to the next chapter.

There are several possible methods to develop computer models for speech production. The most straight-forward is to simply record human speech sounds and then reproduce those. In systems with a very limited vocabulary, such as toys, this is often done with words or even complete phrases. Of course, this approach can only be used for very limited speech repertoires and cannot generalize in a simple manner. One way to create more general systems is to concatenate smaller speech units, such as phonemes or diphones that can be reused among combined into many. However, it is difficult to do this concatenation without getting a discontinuity at the boundary. Still these systems are able to produce high quality speech.

An alternative approach is to not use recorded speech at all and instead generate the speech sounds in a completely synthetic way. One way to do this is to use formant synthesis. The formants are spectral peaks of the sound spectrum and by placing those peaks at different frequencies, different speech sounds are generated. Formant synthesis doesn't have the problem with discontinuities, but typically creates less natural and more robotic sounds.

While both of those methods are able to produce speech sounds, none of those really models how humans produce speech. Here we are interested in studying the coupling between production and perception, through the use of motor primitives during recognition. We therefore need models of the vocal organs that humans use to produce speech. Several models already exists and while the resulting speech sounds are still inferior to that produced with the methods above, they are still more interesting for our purpose. There have been several attempts to build mechanical models of the vocal tract [47] [36]. While these can produce

some human like sounds they are still rather limited and there are no commercially available mechanical solutions.

In this thesis we have instead chosen to simulate the vocal organs in software. One software for this purpose is VTcalcs [83]. This has been used for studying syllable production [43] and vocal imitations [66]. One limitation with this model is that the source code is not available so its internals can mainly be viewed as a "black box". This also makes it very difficult to extend or adapt the existing model. Here we have therefore chosen not to rely on third-party software, and instead develop our own open source articulatory speech synthesizer called jArticulator [59]. The jArcticulator offers similar functionality as VTcalcs. A number of parameters make it possible to control the position of the jaw, tongue and lips, and a synthesizer produces the speech signal by simulating the air flow based on these parameters. By being open source it makes it possible to try different parameters and control strategies. There are a few open source initiatives with similar ideas, such as gnuspeech [48] and ArtiSynth [125]. The first provides a ready to use system, which is based on articulatory speech synthesis, but does not include a parametric model of jaw, tongue, and lips. ArtiSynth, on the other hand, does provide a model of the jaw and the tongue, as well as a separate speech synthesizer, but the different parts have not been integrated into a complete system that can be used for studying the inversion mapping. There are also other related software models that are not provided as open source [26] [11].

In this chapter we will have a closer look at the models on which jArticulator is built. We start by giving a short introduction to human speech production. We then describe how the speech synthesis is done using a tube model, and finally describe articulatory model that is used on top of the tube model to simulate human articulations.

2.1 Human speech production

As we want our model to mimic the vocal organs, we will first have a brief look at how humans produce speech sounds. The vocal organs, i.e. the parts of the body that are related to speech production, consist of the lungs, the larynx (containing the vocal cords), the throat or pharynx, the nose and the mouth.

The lungs produce a steady air stream. When we talk, this stream is rapidly turned on and off as our vocal cords vibrate, producing a sequence of short air puffs. The frequency with which the vocal cords vibrate is called the fundamental frequency or the "pitch" of the speech signal. The pitch varies greatly between different people and also for each individual as we alter the tension and the length of the vocal cords and depending on the air pressure from the lungs.

The throat, nose and mouth, produce a tube that is called the vocal tract, Figure 2.1. The character of the air stream, released at the vocal cords, is modified by the acoustic properties of the vocal tract. The properties depend on the shape of the vocal tract. During speech, we continuously alter the shape of the vocal tract by moving the tongue and lips, thereby creating articulated speech.

Next we will look at how to model these vocal organs in the humanoid robot.

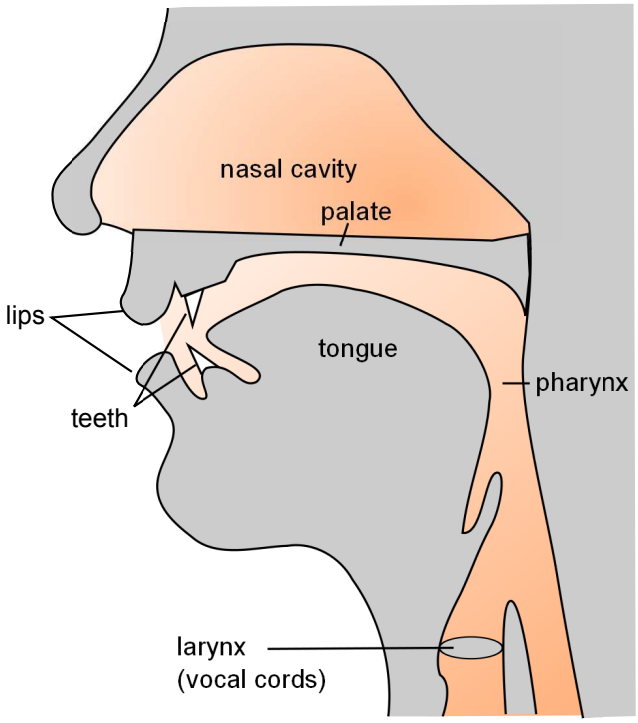

Figure 2.1: Vocal tract

2.2 Tube model

From a technical point of view, the vocal system may be divided into three parts; (i) the speech source, (ii) an acoustic tube between the glottis and the lips, and (iii) the lip radiation.

2.2.1 Glottal source

Starting with the speech source, we need to model the generation of the "air puffs" caused by the vibration of the vocal cords. It is important to notice that this is not a pure sinusoidal vibration, as the resulting sine wave would not contain any harmonics, i.e. the frequency spectra has only a single peak at that the fundamental frequency. If we would use a sine wave as the speech source the vocal tract would therefore not be able to produce resonances

at any other frequencies. Instead of a sinusoidal vibration there is a quite rapid movement once the pressure has built up sufficiently to open the vocal cords. A simple, but functional, way to model this is to use a pulse wave which frequency is set to the desired pitch.

There are also models that more closely fit the glottal source function, such as the LF-model [28].

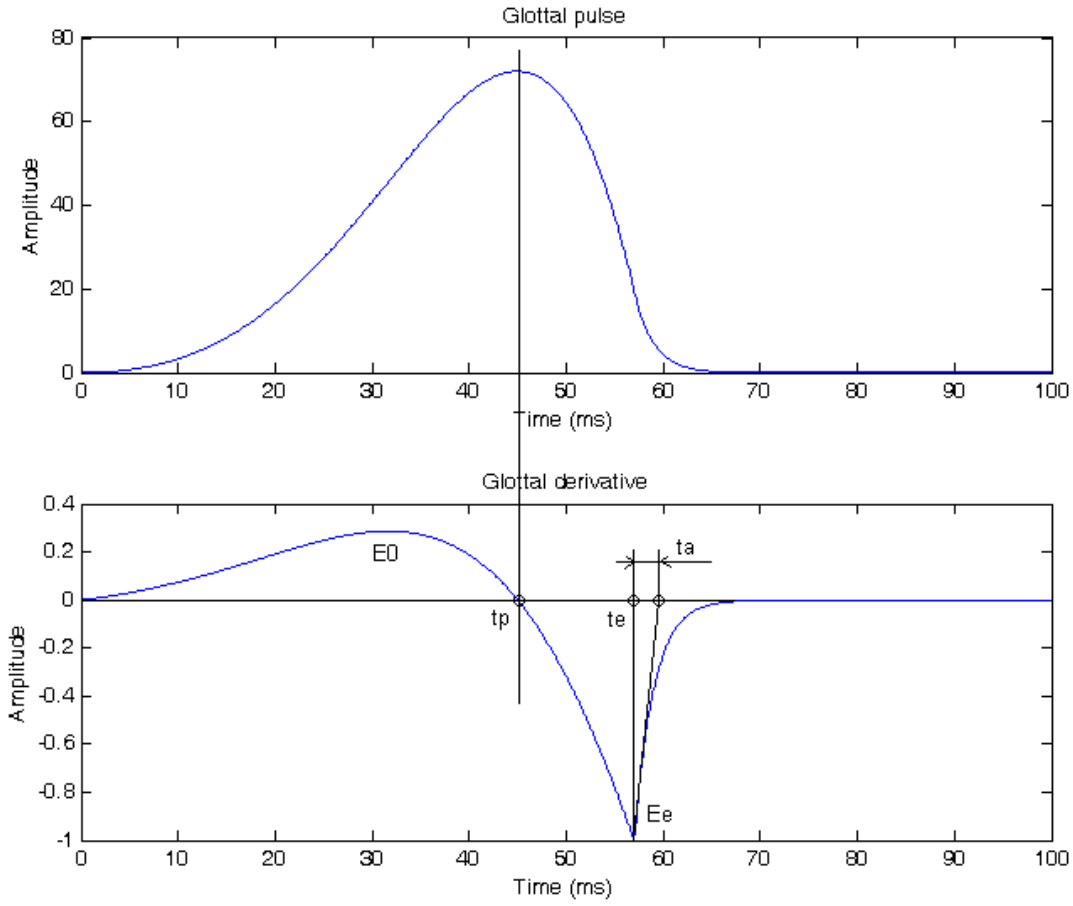

Figure 2.2: A typical shape of glottal flow pulse (above) and its derivative (below) according to the LF model. This model has four parameters: (1) time instant of maximum glottal flow (t_p), (2) time instant of onset of glottal closure and maximum change of glottal flow (t_e), (3) the projection of the derivative of the return phase (t_a), and (4) the negative peak value of the derivative function (E_e).

The LF-model is a four-parameter model represented by the following equation:

$$E(t) = \begin{cases} E_0 e^{\alpha t} sin(\frac{\pi}{t_p} t), & \text{for} \quad 0 \leq t \leq t_e \\ -\frac{E_e}{\epsilon t_a}[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)}], & \text{for} \quad t_e < t \leq t_c \end{cases} \tag{2.1}$$

where the three time points t_e, t_a, t_p, and the amplitude parameter E_e uniquely determine

the pulse. The remaining parameters are the fundamental period t_c, the positive peak value E_0, and the shape controlling parameters α, and ϵ. These can be calculated using a continuity constraint at t_e, and by imposing a requirement of area balance, i.e., zero net gain of flow during a fundamental period. This gives us the following equations for calculating the remaining parameters:

$$-E_e = -\frac{E_e}{\epsilon t_a}[1 - e^{-\epsilon(t_c - t_e)}] \tag{2.2}$$

$$-E_e = E_0 e^{\alpha t_e} sin(\frac{\pi}{t_p} t_e) \tag{2.3}$$

$$\int_0^{t_c} E(t)dt = 0 \tag{2.4}$$

By assuming a small value for t_a, the first equation allows us to directly estimate ϵ as:

$$\epsilon = \frac{1}{t_a} \tag{2.5}$$

Finally we only need to find the value of α, as it directly allows us to calculate the value of E_0. An iteratively search is used to find the value of α. After each iteration the area balance is used to evaluate value of α. If there is a net gain in flow we need to reduce the value of α, and if there is a net loss we instead need to increase the value.

In jArticulator the glottal source can be modeled either with a pulse wave, or with the LF model, see Figure 2.3. We have also added a white noise to produce unvoiced speech sound.

2.2.2 Vocal tract tube

Next we want to model how this wave pulse is modified by the tube formed by our vocal tract. The wave propagation in a tube, whose walls are viewed to have an infinitely high sound impedance, follows Webster's horn equation [126]:

$$\frac{\partial^2 v(x,t)}{\partial x^2} + \frac{1}{A(x)}\frac{dA(x)}{dx}\frac{\partial v(x,t)}{\partial x} = \frac{1}{c^2}\frac{\partial^2 v(x,t)}{\partial t^2} \tag{2.6}$$

where $v(x,t)$ is the sound particle velocity at distance x from the source at time t, c is the speed of sound and $A(x)$ is the area function.

Unfortunately it is not possible to solve the equation for an arbitrary $A(x)$. However, if we assume that $A(x)$ is constant, the equation can be simplified as:

$$\frac{\partial^2 v(x,t)}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 v(x,t)}{\partial t^2} \tag{2.7}$$

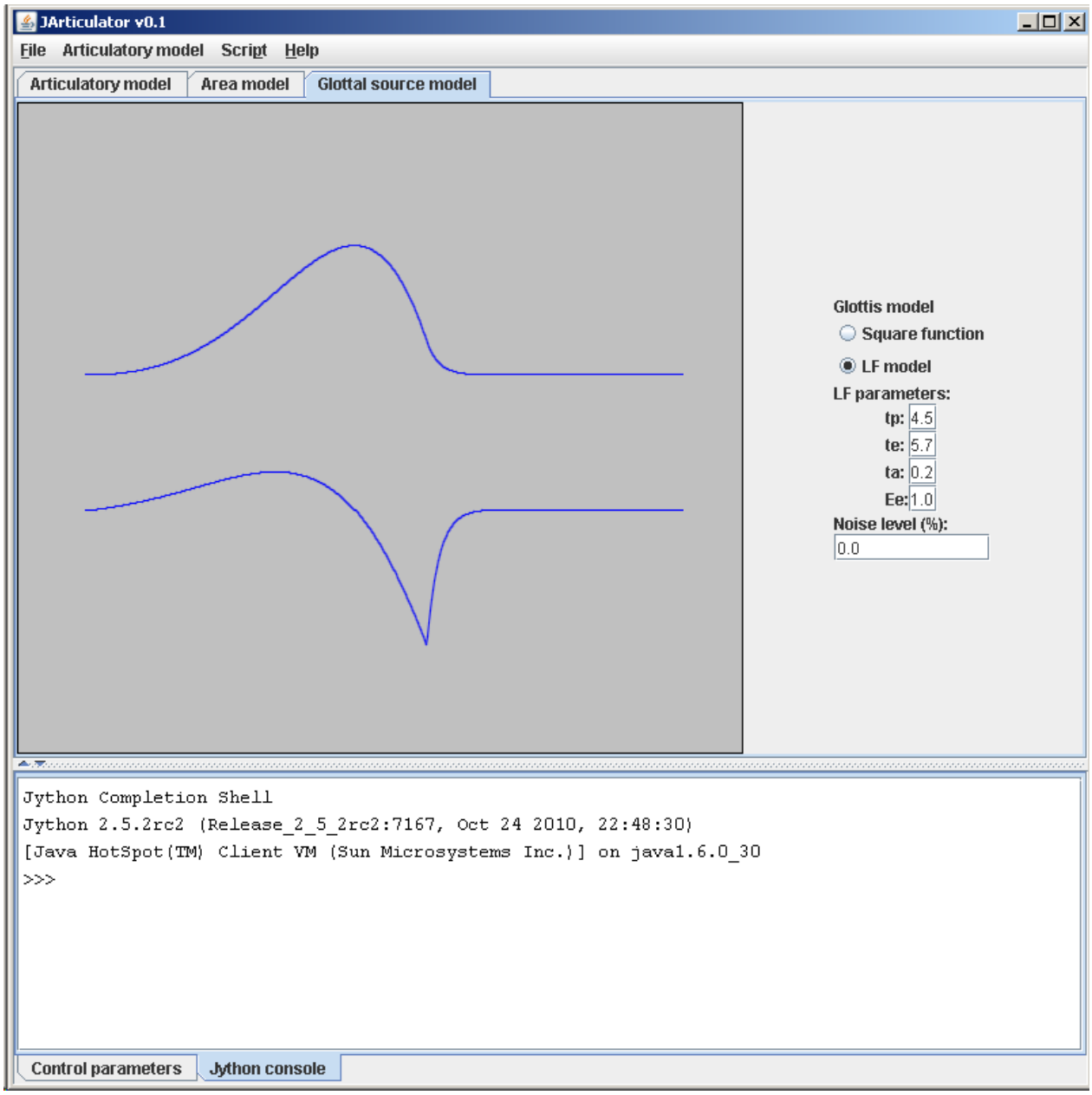

Figure 2.3: The glottal source model implemented in jArticulator

By introducing the volume velocity $u := vA$, a general solution to the equation above can be seen as a combination of two volume velocity waves traveling forward and backward respectively [96]:

$$u(x,t) = u_f\left(t - \frac{x}{c}\right) - u_b\left(t + \frac{x}{c}\right) \tag{2.8}$$

To relax the constraint of constant area, a tube with variable area function can be approximated by concatenating several tubes, where each tube is considered to have a constant area [79]. The sudden change in cross-sectional areas at the tube junctions is equivalent to changes in the acoustic impedances, so that part of the traveling wave is reflected according to the reflection coefficient:

$$r_k = \frac{A_{k-1} - A_k}{A_{k-1} + A_k} \tag{2.9}$$

Each tube segment can be modeled as a time delay. A Simulink model of simple vocal tract model, consisting of two tube segments are shown in Figure 2.4.

By changing the length and the area of each tube segment in the described model, it is possible to produce most vowel sounds. A comparison between vowels produced by a human and by the tube model, are shown in Figure 2.5. As can be seen in the figure, the first formants are at approximately the same frequencies.

In order to model stop consonants a third narrow tube segment, with a small area, is needed at the place where the flow of the vocal tract is stopped. However, instead of using tube segments of variable length it is more convenient to use several short tubes where the length is equal to the distance that sound travel during one sample. In jArticulator we therefore set the length of each tube segment l to:

$$l = \frac{c}{F_s} \tag{2.10}$$

where c is the speed of sound and F_s is the sampling frequency.

2.2.3 Lip radiation

Some extra attention is needed at the end of the last tube. While the most natural might be to consider the free air as an infinite area, this would mean that r_k becomes -1 and no air will be leaving the tube. One solution to this is to choose an arbitrary area for the free air and just make sure that it is larger than the last tube segment. A more detailed model of the lip radiation can be constructed by using an additional high-pass filter for the output signal and a corresponding low-pass filter for the reflected wave [20].

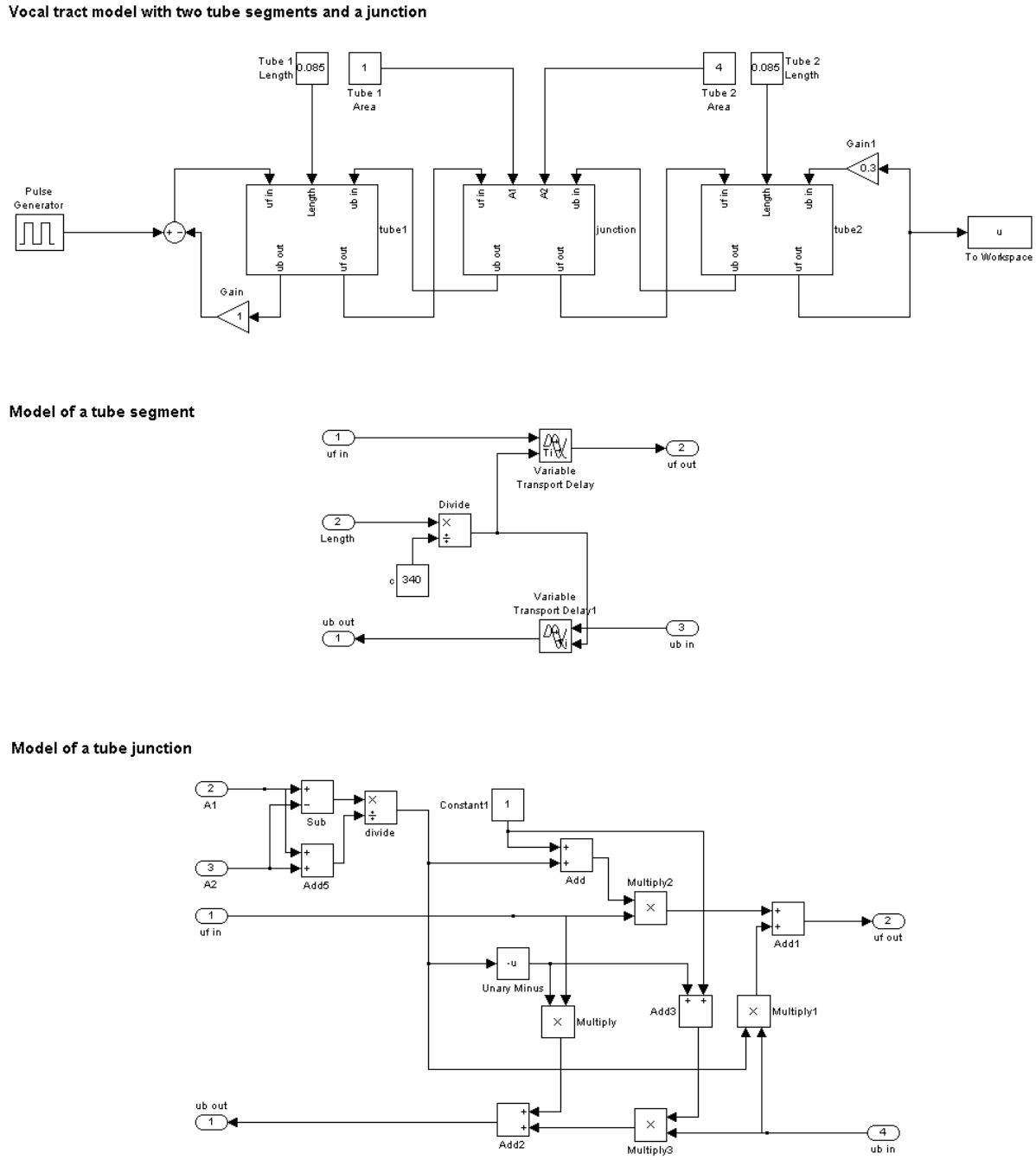

Figure 2.4: Simulink vocal tract model with two tube segments and a junction. The glottal source is modeled with a pulse wave and lip radiation is simulated with a fixed reflection rate (above). The detailed models are shown for a single tube segment (center) and the junction between two segments (below).

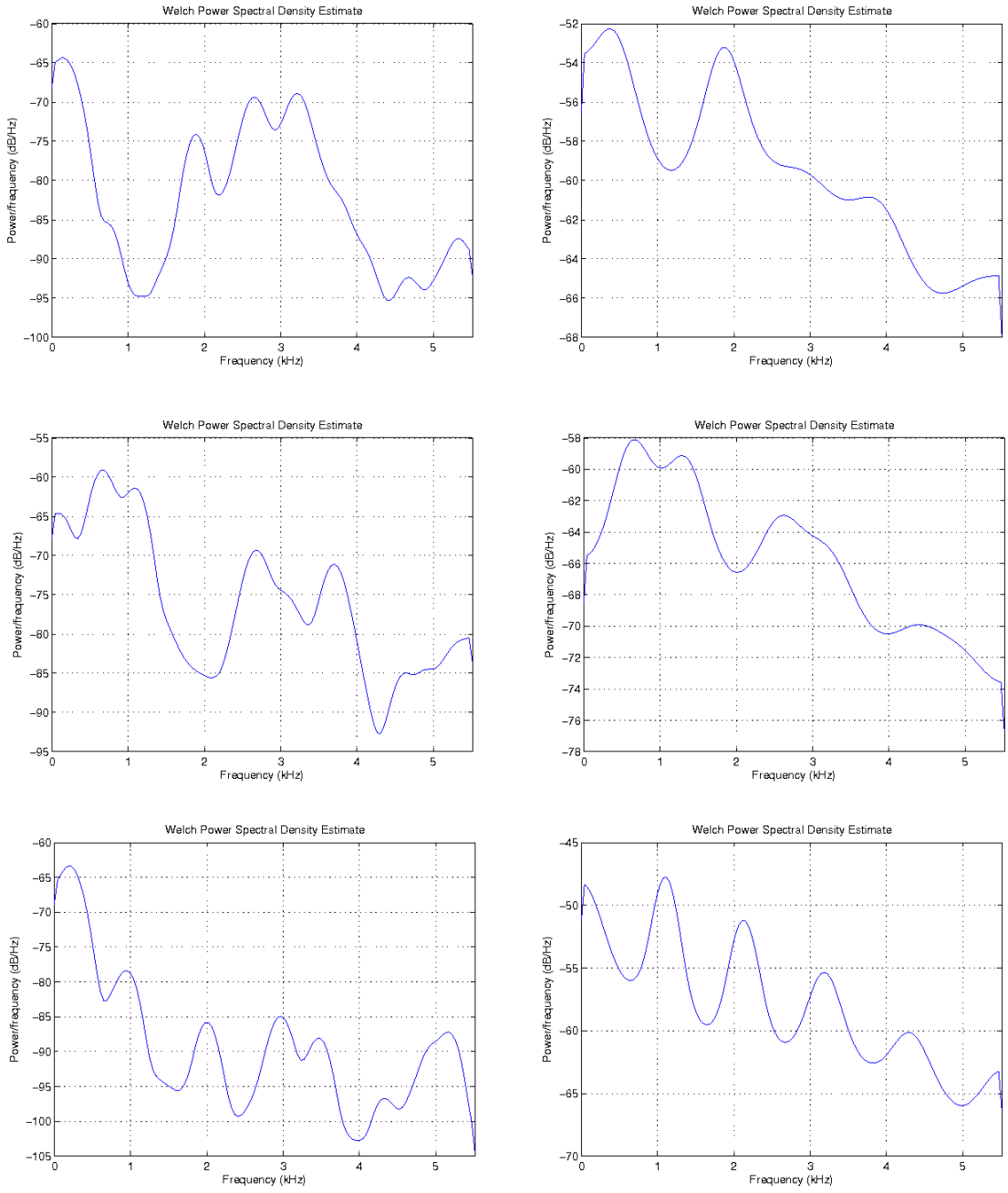

Figure 2.5: Sound spectra for vowels produced by a human (left) and the tube model (right). The vowels used are [i], [a], and [u] (from top to bottom). The first two formant frequencies are around 300 Hz and 1900 Hz for the vowel [i], 800 Hz and 1200 Hz for [a], and around 300 Hz and 1000 Hz for [u].

2.3 Articulatory models

While the tube model allows the area for each tube segment to be set individually, this is not the case for the human vocal tract. As humans move their jaw and tongue in order to articulate, this has coupled effects on the area function of a large part of the vocal tract. Describing the articulations in terms of movements of the tongue, jaw, and lips, instead of a pure area function may therefore help to focus on the relevant vocal tract configuration and avoid anatomically impossible area functions. Even more important, it allows us to get a measurement for how close different articulations are in terms of motor positions.

One of the hypotheses of the CONTACT project, in which part of this thesis has been developed, is that motor space is more invariant than the produced speech sound for many articulations and therefore better for speech recognition tasks.

An advanced equipment for studying articulatory features called "linguometer" was therefore developed by some of the partners in the CONTACT project [41]. The linguometer combined data from a 3D electromagnetic articulograph and an ultrasound system. The articulograph locates the 3D position and orientation of 12 small coils that are glued to the head, soft palate, tongue and lips. The articulograph has high accuracy, but does not provide a complete model of the articulators. The ultra sound system on the other hand provides a good 2D view of the complete tongue profile. By combining data from those sensors, the linguometer is able to produce a good measurement of the configuration of the vocal tract. The linguometer setup is shown in Figure 2.6. However, the direct measurements from the linguometer still describe the positions of certain points in the vocal tract rather than the underlying actuators. By having several subjects articulating a given set of speech sounds, an attempt was made to find the role that different tongue muscles have for the position of the tongue surface. Initial results indicate that those motor features are at least as good as auditory features for classification.

Unfortunately, no complete model of how different muscles affect the shape of the vocal tract was available in time for this work. Instead we have based the articulatory model in jArticulator on the parameters in VTcalcs. The tongue model in VTcalcs was obtained from images of two women articulating French vowels [83]. It was found that three PC's was enough to reconstruct the tongue shape with sufficient precision to separate between the vowels, and were included in the VTcalcs software. These components corresponds to: i) front-back movement of the tongue, ii) bending the tongue around two points, and iii) bending the tongue around a single point, see Figure 2.7. Similar principal components have also been found in other studies [109]. However, in this latter work they also found that a forth PC including three bending points can help to create /s/ and /l/ sounds. While it is

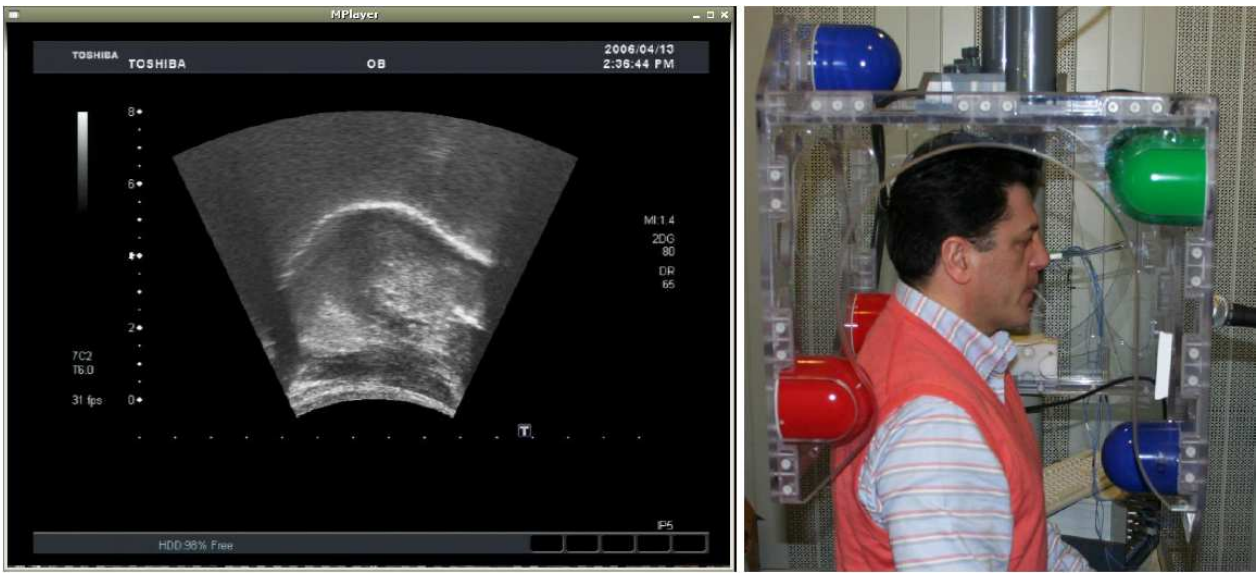

Figure 2.6: The linguometer setup. Left: the tongue, viewed in real-time via an ultrasound machine. Right: an articulograph, which recovers the 3D pose of sensors placed on the tongue and face.

possible to add this and other parameters to the tongue model in jArticulator, the current model only includes three parameters for the tongue, directly inspired by the parameters of VTcalcs.

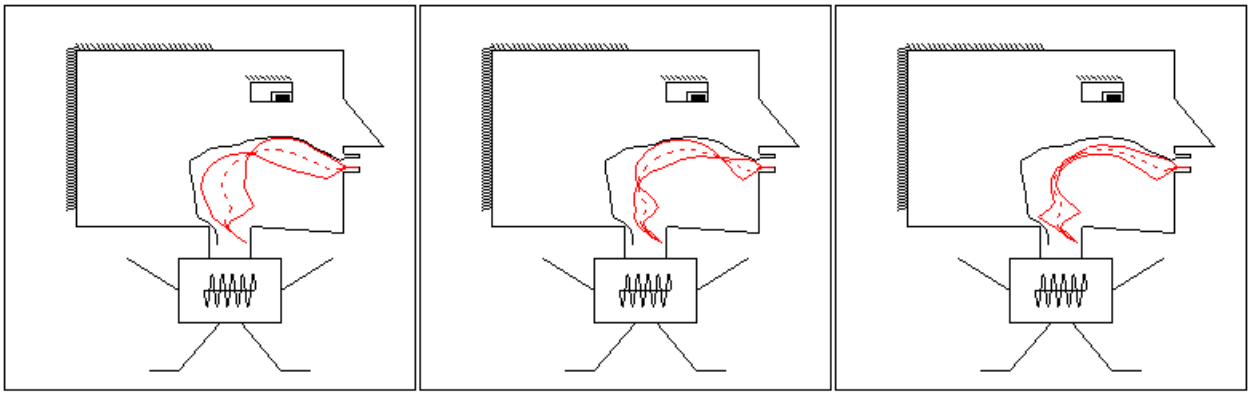

Figure 2.7: Tongue movement in VTcalcs corresponding to the three main PC's as implemented in VTcalcs, i.e. front-back movement, two point bending, and single point bending. The dashed line shows the center line and the upper and lower lines show the respective extreme position.

Apart from the tongue parameters, VTcalcs also includes one parameter for the yaw, one for the extrusion of the lips, one for lip opening, and a parameter for changing the position of the larynx. With the exception of the last parameter, those have all been implemented in jArticulator. The complete set of parameters is summarized in Table 2.1.

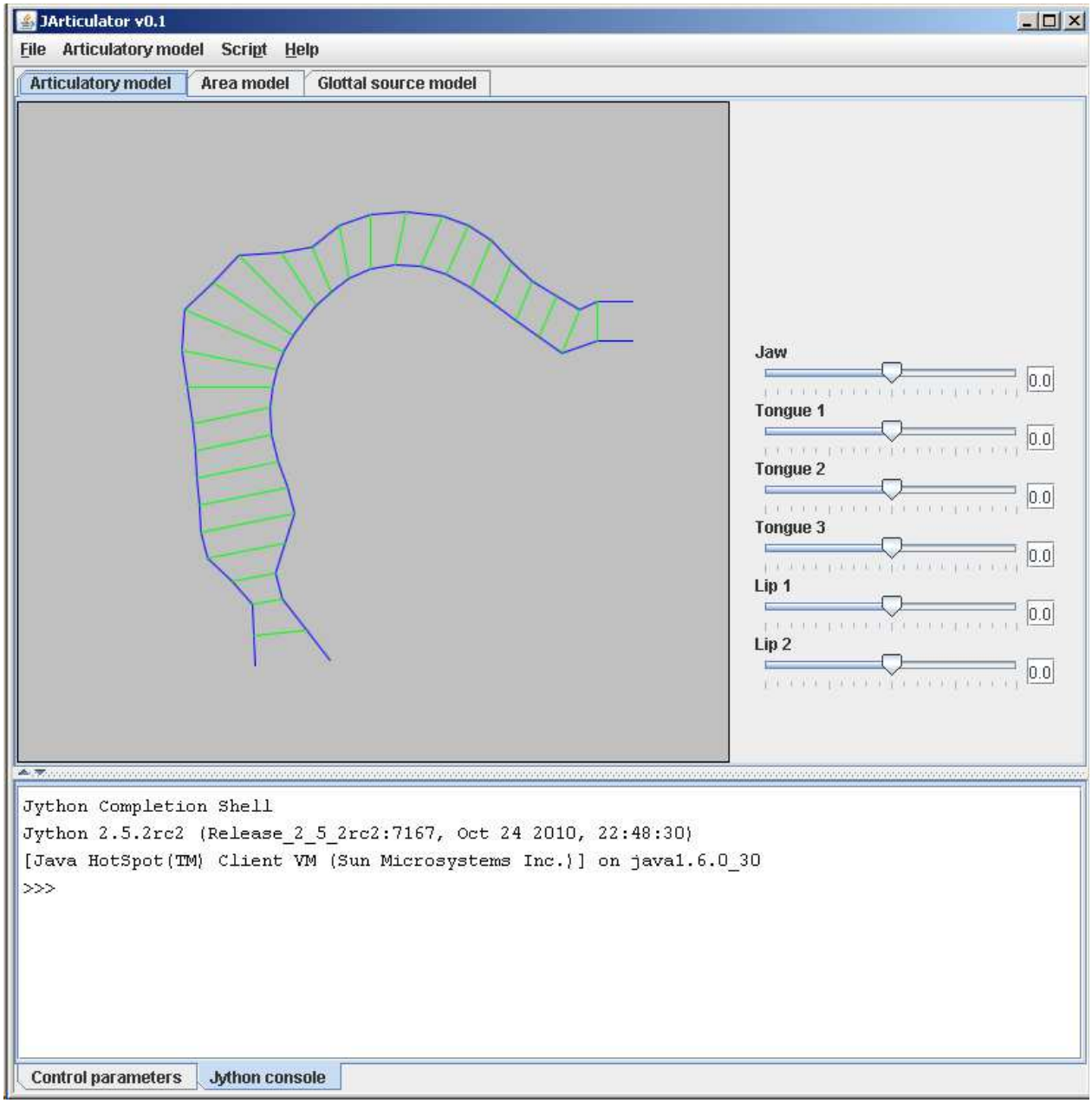

Figure 2.8: Simulation of vocal tract using jArticulator

Table 2.1: *Parameters of the vocal tract model*

Parameter	Name	Explanation
1	Jaw	Jaw opening
2	Tongue	Tongue position front/back
3	Shape	The amount of roundness
4	Apex	The positon of the roundness' apex
5	Lip height	The openess of the lips
6	Lip protrusion	The protrusion of the lips
7	Larynx	Larynx position

2.4 Conclusions

This chapter shows how the vocal organs can be modeled and simulated in an artificial system. While we have not created a physically embodied model, we try to simulate a physical model as close as possible using software.

We have modeled how the vibration of the vocal cords produces a pressure wave, and how this pressure wave is transformed as it passes the vocal tract. The vocal tract is modeled with a number of concatenated tube segments where each tube is considered to have a constant area and to be of a fixed length. This approximation makes it easy simulate how a sound wave propagates through the tube. For demonstration purpose, a simple Simulink model has been derived and implemented. It is shown that the model can produce human-like vowel sounds.

On top of the tube model we have built an articulatory model with six parameters that mimics the movement of the jaw, tongue, and lip, and how this changes the cross section area of each tube segment. The parameters have been chosen so that they are compatible with the VTCalcs software. The advantage with this model is that it is commonly used and that it works very well for producing vowel sounds as well as stop consonants. One disadvantage is that it may not perform very well for other types of speech sounds.

Chapter 3

Modeling the auditory system

In the previous chapter we had a look at how human speech is produced and how to model this in a computer based system. As mentioned, the sound signal is produced from a sequence of air puffs released from the vocal cord and the resonances caused by the form of the vocal tract. This causes a complex wave pattern that is transmitted through the air. In Figure 3.1 we have plotted how the sound level changes over time for a typical speech signal. From this plot we directly see how the sound intensity changes over time. As we mentioned in the introduction, the periods of near silence that occur in the sound signal do not correspond directly to the borders between different words, but rather parts of words such as syllables. In some parts of the signal it is also possible to directly identify recurring spikes caused by the vibration of the vocal cords. This becomes clearer if we look at a shorter time interval. In Figure 3.2 we have a closer look at the speech signal for the interval between 0.26-0.33 s, loosely corresponding to the vowel /i/ in "titta". Here the fundamental period of around 0.01 s, which is equal to a pitch of 100 Hz, is clearly visible. These parts of the sound signal, i.e. sound intensity, syllable length and pitch, and how these changes over time is called the prosody of speech. These are relatively slow alterations of the sound signal compared to the alterations caused by the resonances in the vocal tract. The prosody is mainly related to the rhythm, stress, and intonation of an utterance, while it is the resonances in the vocal tract that provide information about the actual articulation. These resonances are difficult to distinguish directly from plots in Figure 3.1 and 3.2, but become more visible when the signal is decomposed into its frequency components. This is also what happens in the human ear as different places of the inner ear reacts to different frequencies, creating a tonotopic representation of the sound signal.

As we are interested in modeling the auditory system of humans we will first look at what happens when the sound wave hits the human ear, and then at how we can model this in our humanoid robot. We look at both how to simulate the physical properties of the ear, and

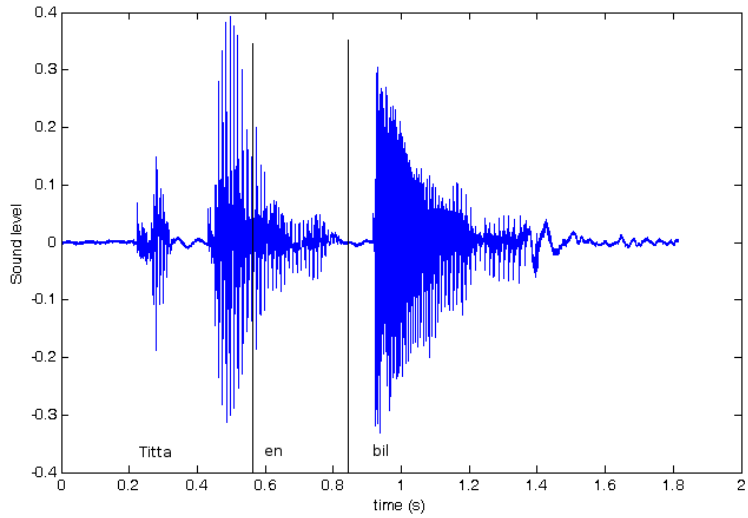

Figure 3.1: Acoustic wave for the phrase "Titta en bil"

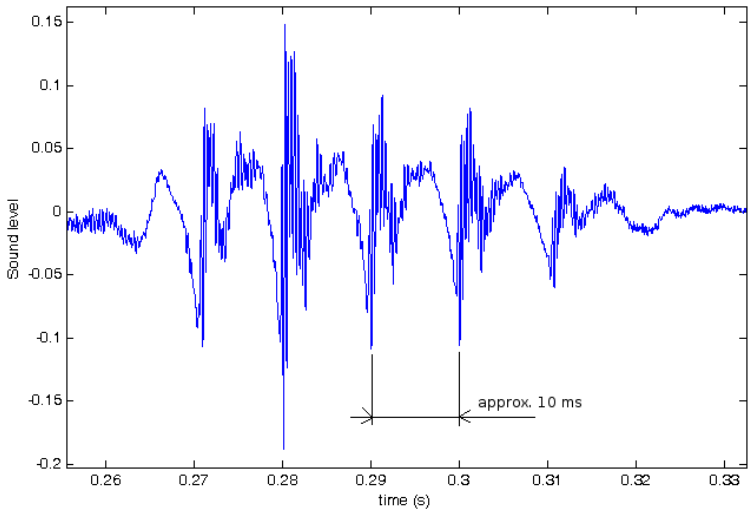

Figure 3.2: Close-up on the acoustic wave for the vowel i

how to derive a number of features that enhance those parts of the auditory signal that are most useful for language learning.

3.1 The auditory system

The acoustic information is primarily processed by our auditory system, shown in Figure 3.3. The sound waves are reflected by the pinna and funneled into the ear canal where they set the ear drum into vibration. This vibration continues through the middle ear and inner ear (the cochlea) where the basilar membrane detects the vibration and passes the information to the brain through the auditory nerve.

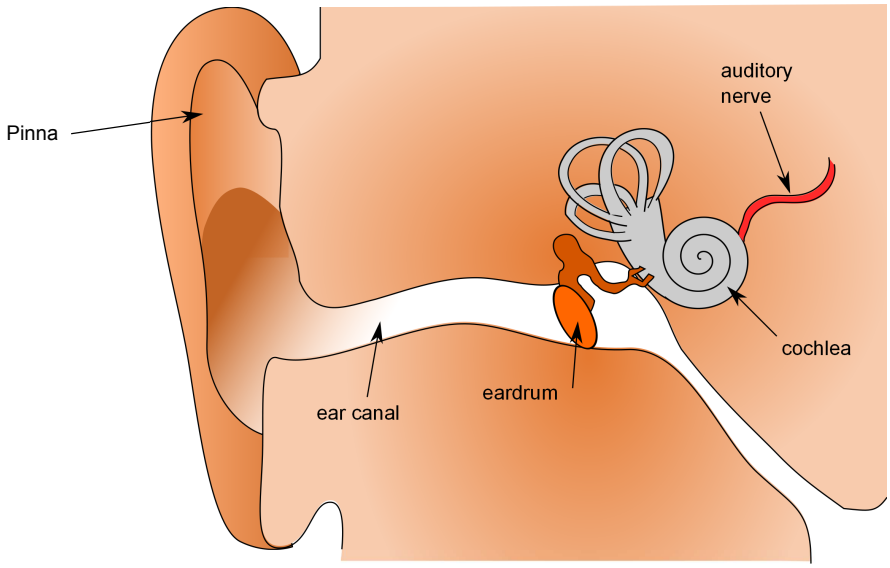

Figure 3.3: The auditory system

The shape of both the pinna and the cochlea has important implications for how we perceive sounds. The reflection of the pinna provides information that is useful for localizing the sound and the cochlea defines how sensitive we are to different frequencies.

While the human ear is a very complex organ we need to create simplified models that are suitable for implementation in a computer model while maintaining some of the main characteristics of the human ear.

3.2 Modeling the cochlea and basilar membrane

The cochlea is a spiral shaped bone structure with a total length of about 3.5 cm. The cochlea contains the basilar membrane, the base for the sensory cells of hearing, the hair

cells. There are approximately 12000 hair cells in each ear [23]. These play a crucial role in the transfer of sound waves to the brain.

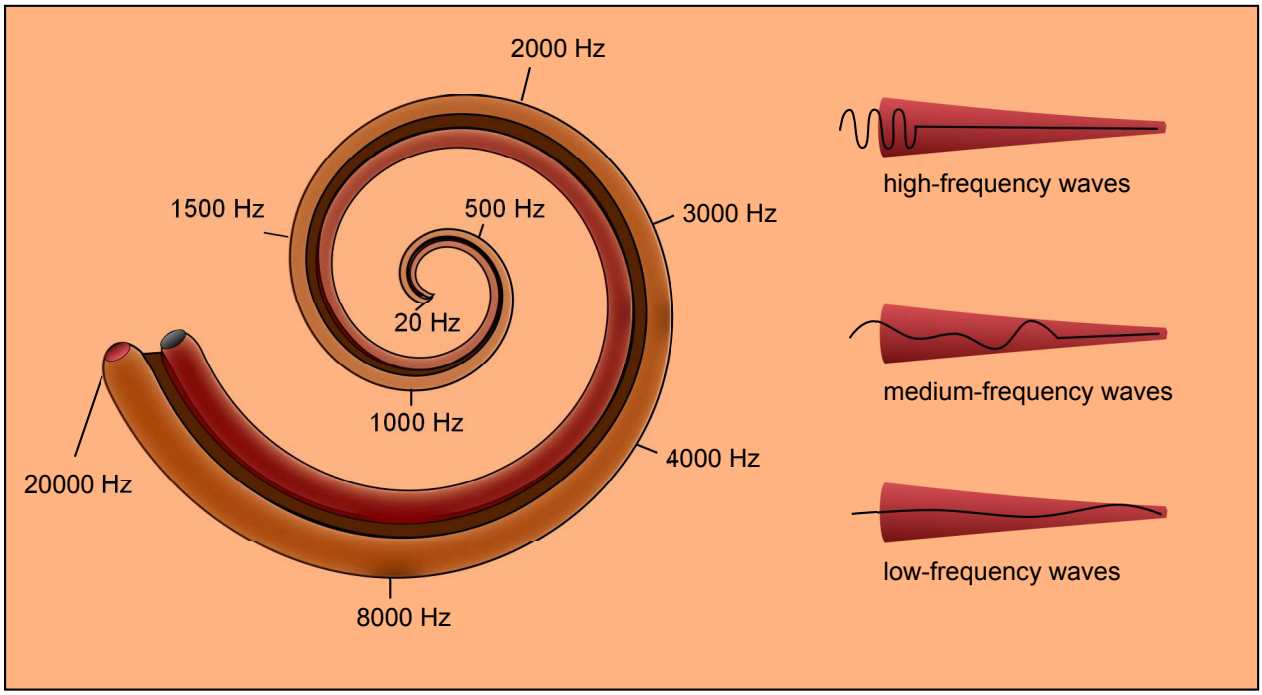

Figure 3.4: Basilar membrane

The cochlea works as a crude mechanical filter that separates the incoming sound into its frequency components, with maximum vibration at the base for high frequencies up to 20 kHz, and maximum vibration for low frequencies around 20 Hz at the apex at the end of the spiral. The cochlea is filled with a fluid with mechanical characteristics similar to water, and the movement of this fluid causes the hair cells at the basilar membrane to move along. As the hair cells are set into vibration, they signal this through the acoustic nerve to the brain.

The basilar membrane imposes limits on the hearable frequency range, but even for the hearable part the basilar membranes response to input frequencies is non-linear: a larger portion of the basilar membrane responds to sounds in the 0-1 kHz range than, for example, in the 10-11 kHz range. As a consequence human listeners are more sensitive to differences in the lower than in the higher frequencies. This sensitivity can be modeled with the mel-scale. The mel scale is based on experiments with pure tones in which listeners adjust the frequency of a test tone to be half as high (or twice as high) as that of a comparison tone, starting at 1000 Hz = 1000 mel. From there 500 mel is defined as the pitch of tone that sounds half as high, and 2000 mel the pitch of tone that sounds twice as high as 1000 Hz.

The mel scale corresponds closely to the Hz scale up to approximately 500 Hz. At higher frequencies, the mel scale is nearly logarithmic. Mels can be interpretable in terms of linear

distance along the basilar membrane.

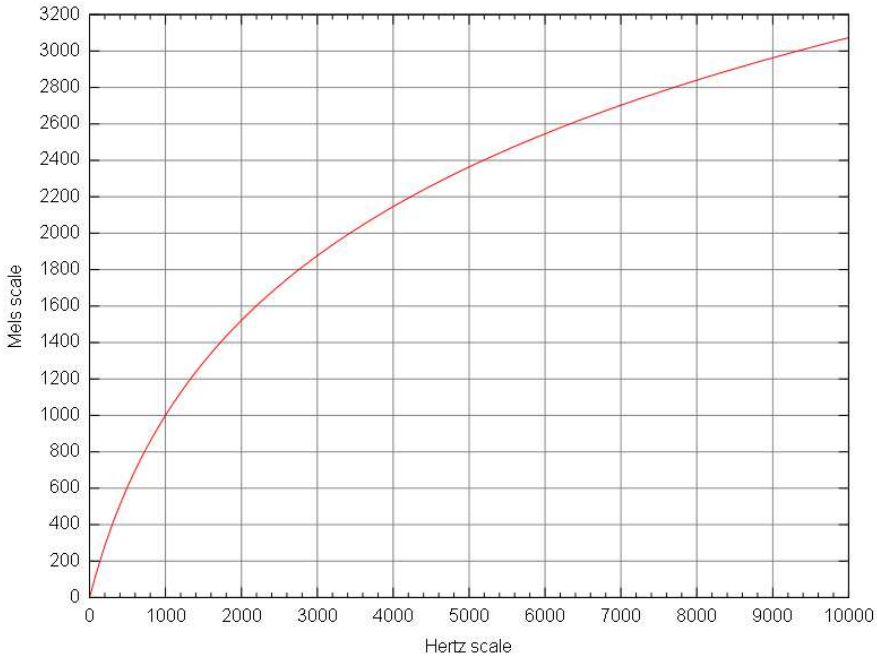

Figure 3.5: Mel scale

3.2.1 Tonotopic representation of speech signal

The purpose of the tonotopic representation is to model how the sound signal is divided into different frequency components as it passes along the basilar membrane, and provide a compact representation of this, which can then be used as a feature vector for the language acquisition. Several tonotopic representations have been proposed to facilitate speech recognition. For production and recognition of vowels, formants are commonly used [129]. Formants are spectral peaks in the sound spectrum caused by resonances in the vocal tract. However, due to difficulties to track the formants for non-stationary signals, they are mostly useful for vowels. In other related work, Linear Predictive Coding (LPC) has been used [68] [89]. LPC is more generally applicable than formants, but still require rather stationary signals to perform well. Here, Mel frequency cepstral coefficients (MFCC) [21] are used as speech features since these are less dependent on having a stationary signal. The steps for calculating MFCC are outlined in Table 3.1.

The first step is to calculate the frequency spectra using FFT. It is not possible to find the frequency components by looking at the signal at a single instant. Instead it is necessary to divide the signal into small segments, or windows, that are long enough to span the wave lengths of the interesting frequencies, but short enough to capture variations of speech sounds.

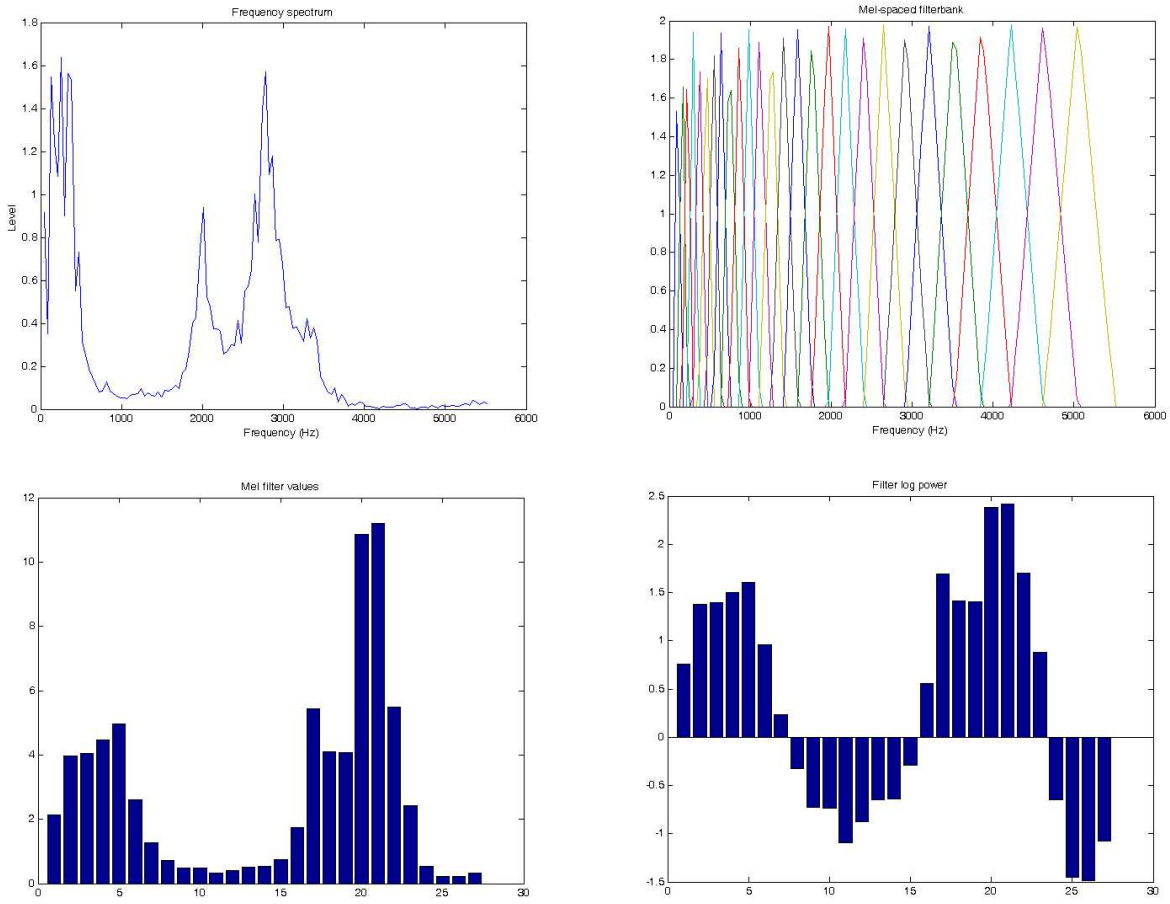

Figure 3.6: Calculation of mel log powers: 1. FFT power spectrum (top left), 2. Mel spaced filter bank (top right), 3. Mel values (bottom left), 4. Mel log powers (bottom right)

Table 3.1: Algorithm for calculating MFCC

1) Make a FFT of the speech signal
2) Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
3) Take the logs of the powers at each of the mel frequencies.
4) Take the discrete cosine transform of the bank of mel log powers.
5) The MFCCs are the amplitudes of the resulting spectrum.

Typically this is done by using 25 ms windows with 50% overlap between the windows, which is sufficient to capture at least two periods of the pitch (and all harmonics).

In the second step the frequency spectrum is mapped onto the mel scale. Instead of making this for each individual frequency, a filter bank is used. This simulates how the individual hair cells are affected not only by a single frequency, but by a wider band of frequencies. Typically 24-32 filter banks are used.

Next we take the logs of the powers at each of the mel frequencies. These first steps are illustrated in Figure 3.6.

The final step is to make a discrete cosine transform of the bank of mel log powers. The components of the resulting spectrum are our MFCC, see Figure 3.7. Both the MFCCs and their derivatives are used as features for speech recognition.

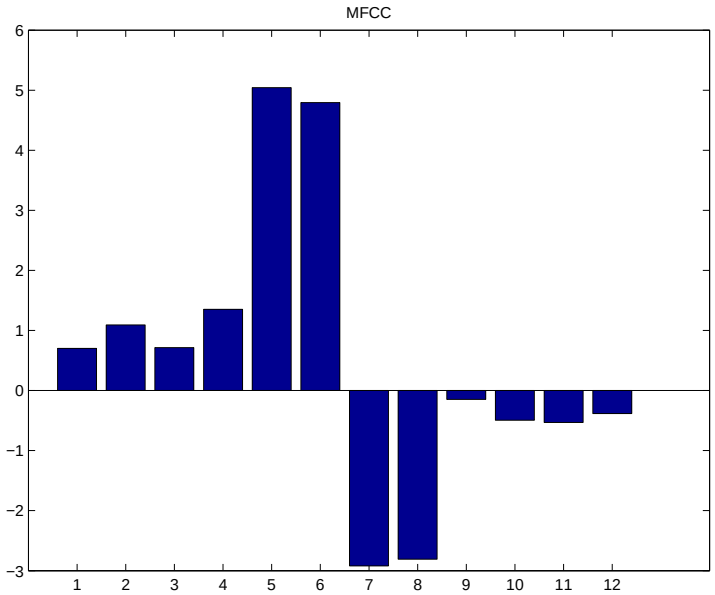

Figure 3.7: MFCCs used as tonotoic representation of the speech signal

3.2.2 Prosodic features

The prosodic features aim at capturing the rhythm, stress, and intonation of an utterance. Some suggestions on prosodic features and tests of their usefulness can be found in [9] [8]. For this thesis, the following prosodic features have been adopted:

1. Number of syllables
2. Length of the last syllable
3. Length of the second last syllable
4. Difference in length between the two last syllables
5. Difference in pitch for the two last syllables

The start of a syllable is detected when the intensity exceeds a certain threshold value, and the end of the same syllable is defined as the point when the intensity drops below a second threshold, set slightly lower than the first.

The pitch, i.e. the fundamental frequency with which the vocal cords vibrate, was tracked over each syllable using RAPT [115]. In order to calculate the difference in pitch between syllables, only the pitch in the center of each syllable was used.

3.3 Modeling the head and pinna

The shape of the head and ear changes the sound as a function of the location of the sound source. This phenomenon is called the head related transfer function (HRTF). This section describes the design of a robotic head and ears that give a human-like HRTF, and especially provides Interaural Time Difference (ITD), Interaural Level Difference (ILD), and frequency notches similar to those observed in humans.

The ITD depends on distance between the ears and the ILD is primarily dependent on the form of the head and to less extent also the form of the ears, while the peaks and notches in the frequency response mainly are related to the form of the ears. For the sake of calculating the HRTF, a human head can be modeled by a spheroid [4] [24]. The head used in this work is the iCUB head which is close enough to a human head to expect the same acoustic properties. The detailed design of the head is described in [10] but here we can simply consider it a sphere with a diameter of 14 cm. The ears are more complex. Each of the robot's ears was built by a microphone placed on the surface of the head and a reflector simulating the pinna/concha, as will be described in detail below. The shape of human ears

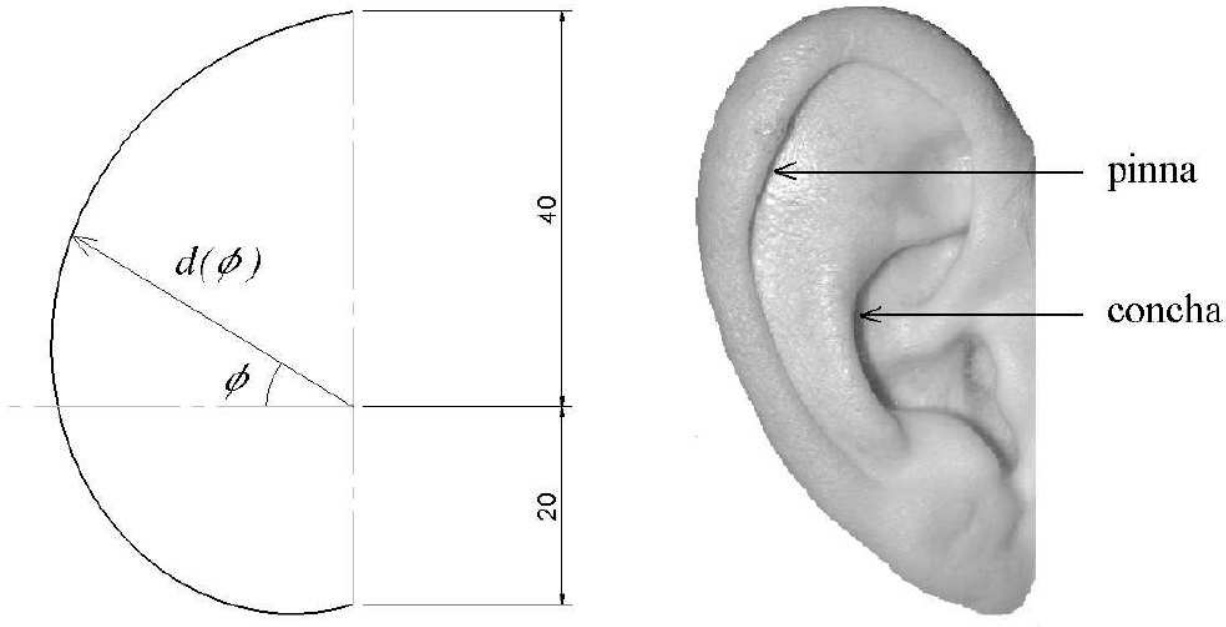

Figure 3.8: Pinna and concha of a human ear (right) and the artificial pinna (left).

differs substantially between individuals, but a database with HRTF for different individuals [5] provides some general information on the frequency spectra created for various positions of the sound source by human ears. Obviously, one way to create ears for a humanoid robot would be to simply copy the shape of a pair of human ears. That way we can assure that they will have similar properties. However we want to find a shape that is easier to model and produce while preserving the main acoustic characteristics of the human ear. The most important property of the pinna/concha, for the purpose of locating the sound source, is to give different frequency responses for different elevation angles. We will be looking for notches in the frequency spectra, created by interferences between the incident waves, reaching directly the microphone, and their reflections by the artificial pinna, and want the notches to be produced at different frequencies for different elevations. A notch is created when a quarter of the wavelength of the sound, λ, (plus any multiple of $\lambda/2$) is equal to the distance, d, between the concha and the microphone:

$$n * \frac{\lambda}{2} + \frac{\lambda}{4} = d \ \ (n = 0, 1, 2, ...) \tag{3.1}$$

For these wavelengths, the sound wave that reaches the microphone directly is cancelled by the wave reflected by the concha. Hence the frequency spectra will have notches for the corresponding frequencies:

$$f = \frac{c}{\lambda} = \frac{(2 * n + 1) * c}{4 * d} \quad (c = \{speed\, of\, sound\} \approx 340 m/s) \tag{3.2}$$

To get the notches at different frequencies for all elevations we want an ear-shape that has different distance between the microphone and the ear for all elevations. Lopez-Poveda and Meddis suggest the use of a spiral shape to model human ears and simulate the HRTF [81]. In a spiral the distance between the microphone, placed in the center of the spiral, and the ear increases linearly with the angle. We can therefore expect the position of the notches in the frequency response to also change linearly with the elevation of the sound source.

We used a spiral with the distance to the center varying from 2 cm below to 4 cm in the top, Figure 3.8. That should give us the first notch at around 2800 Hz for sound coming straight from the front and with the frequency increasing linearly as the elevation angle increases. When the free field sound is white noise as in Figure 3.9, it is easy to find the notches directly in the frequency spectra of either ear. However, sound like spoken language will have its own maxima and minima in the frequency spectra depending on what is said. It is not clear how humans separate what is said from where it is said [63]. One hypothesis is that we perform a binaural comparison of the spectral patterns, as it has also been suggested for owls [91]. Both humans and owls have small asymmetries between the left and right ear that can give excellent cues to vertical localization in the higher frequencies. These small asymmetries that provide different spectral behaviors between the ears should not be confused with the relatively large asymmetries needed to give any substantial difference for the ILD. Here we only need the difference in distance between the microphone and the ear for the right and left ear to be enough to separate the spectral notches. In the ideal case we would like the right ear to give a maximum for the same frequency that the left ear has a notch and hence amplify that notch. This can be done by choosing the distance for the right ear, d_r, as:

$$d_r(\phi) = 2\frac{m_r + 1}{2 * n_l + 1} * d_l(\phi) \tag{3.3}$$

where m_r=maxima number for right ear, n_l=notch number for left ear, and d_l=distance between the microphone and ear for left ear.

If, for example, we want to detect the third notch of the left ear and require the right ear to have its second maxima for that same frequency, we should choose the distance between the microphone and ear for the right ear as:

$$d_r(\phi) = 2\frac{2 + 1}{2 * 3 + 1} * d_l(\phi) = \frac{6}{7} * d_l(\phi) \tag{3.4}$$

In the case of two identical ears we cannot have a maximum of the right ear at the same

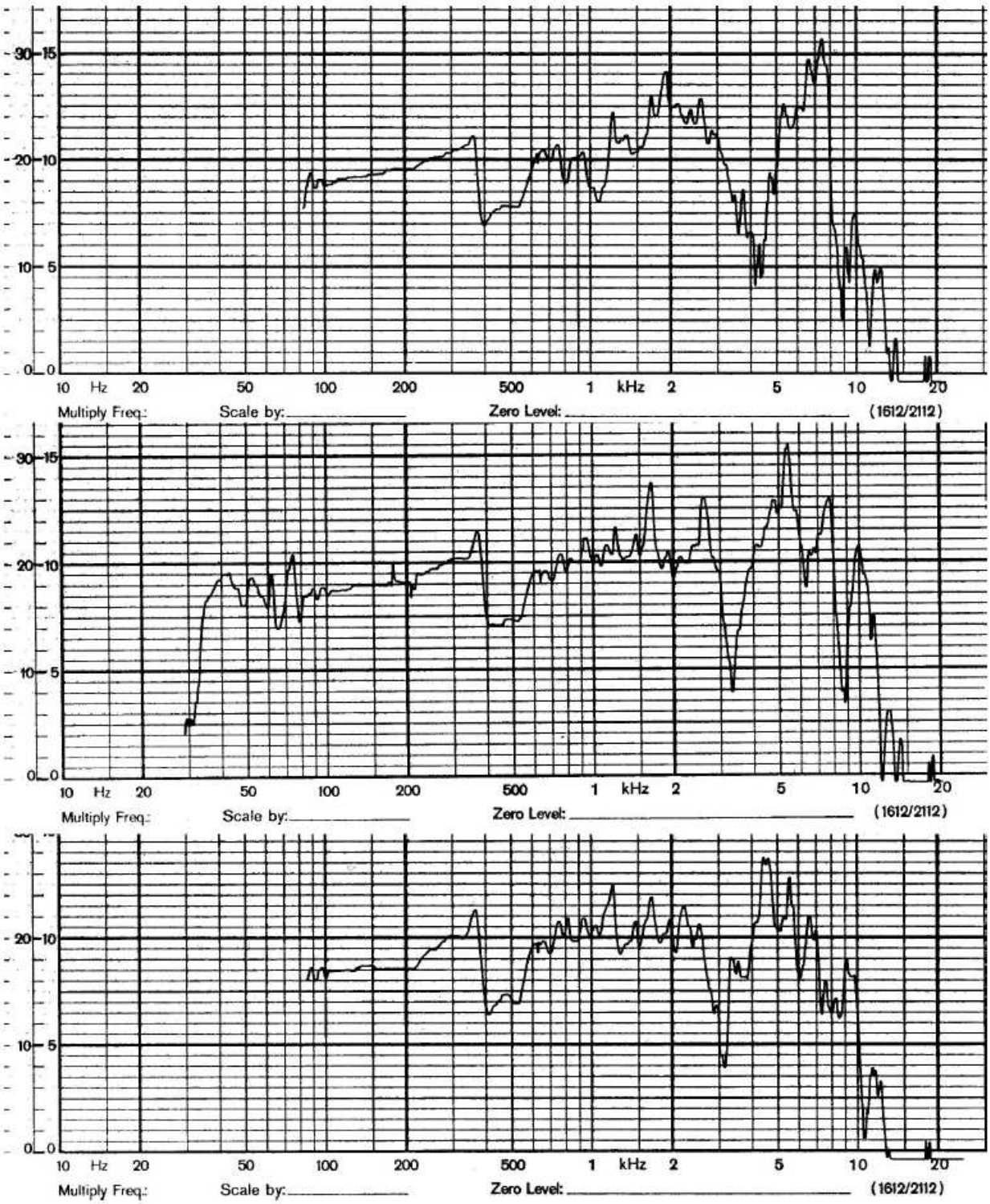

Figure 3.9: Example of the HRTF for a sound source at a) 50 degrees above, b) front, and c) 50 degrees below

place as the left ear has a notch for all elevations. The best we can do is to choose the angle between the ears so that the right ear has a maximum for the wanted notch when the sound comes from the front. In the specific case of the ears in Figure 3.8 the optimal angle becomes 18 degrees, which is the setup used in this work.

3.4 Conclusions

To mimic infants' language learning using a humanoid robot, it is necessary that the signal provided to the robot is similar to that of the infants. This chapter describes how the sound signal is affected by the human ear and how we can model this in the humanoid robot. A physical model of the outer ear is needed to create a human-like HRTF, while the frequency response of the inner ear is modeled using a mel-scale transformation.

A number of features, that provide compact representations of the acoustic information, are also proposed.

For the tonotopic representation we have chosen to use MFCC, as these are the most common features for speech recognition.

Finally we have also suggested a number of prosodic features that provide important additional information that is commonly used in infant-direct speech.

Chapter 4

Sound localization

Sound plays an important role in directing humans' attention to events in their ecological setting. The human ability to locate sound sources in potentially dangerous situations, like an approaching car, or locating and paying attention to a speaker in social interaction settings, is a very important component of human behavior. In designing a humanoid robot that is expected to mimic human behavior, the implementation of a human-like sound location capability as a source of integrated information is therefore an important goal. Humans are able of locating the sound sources in both the horizontal and vertical plane from exploring acoustic information conveyed by the auditory system, but in a robot that uses two simple microphones as ears there is not enough information to do the same. Typically the robot would be able to calculate or learn the positions of the sound source in the plane of the microphones, i.e. the azimuth which usually corresponds to the horizontal plane. This can be done by calculating the difference in time between the signal reaching the left and the right microphone respectively. This is called the interaural time difference (ITD) or the interaural phase difference (IPD) if we have a continuous sound signal and calculate the phase difference of the signal from the two microphones by cross-correlation. However, the ITD/IPD does not give any information about the elevation of the sound source. Furthermore it cannot tell whether a sound comes from the front or the back of the head. In robotics this is usually solved by adding more microphones. The SIG robot [94] [93] has four microphones even though two are mainly used to filter the sound caused by the motors and the tracking is mainly done in the horizontal plane. In [120] eight microphones are used, and in [112] [42] a whole array of microphones is used to estimate the location of the sound.

While adding more microphones simplifies the task of sound localization, humans and other animals manage to localize the sound with only two ears. This comes from the fact that the form of our head and ears change the sound as a function of the location of the sound source, a phenomenon known as the head related transfer function (HRTF). The

HRTF describes how the free field sound is changed before it hits the eardrum, and is a function $H(f, \theta, \phi)$ of the frequency, f, the horizontal angle, θ, and the vertical angle, ϕ, between the ears and sound source. The IPD is one important part of the HRTF. Another important part is that the level of the sound is higher when the sound is directed straight into the ear compared to sound coming from the sides or behind. Many animals, like cats, have the possibility to turn their ears around in order to get a better estimate of the localization of the sound source. Even without turning the ears, it is possible to estimate the location of the sound by calculating the difference in level intensity between the two ears. This is referred to as the interaural level difference (ILD). However, if the ears are positioned on each side of the head as for humans, ILD will mainly give us information about on which side of the head that the sound source is located, i.e. information about the azimuth which we already have from the ITD/IPD. In order to get new information from the ILD we have to create an asymmetry in the vertical plane rather than in the horizontal. This can be done by putting the ears on top of the head and letting one ear be pointing up while the other is pointing forwards as done in [6]. The problem with this approach is that a big asymmetry is needed to get an acceptable precision and ILD of human-like ears does not give sufficient information about the elevation of the sound source.

For humans it has been found that the main cue for estimating the elevation of the sound source comes from resonances and cancellation (notches) of certain frequencies due to the pinna and concha of the ear. This phenomenon has been quite well studied in humans both in neuroscience and in the field of audio reproduction for creating 3D-stereo sound [3] [40] [61] [107] [60] [7] [86], but has often been left out in robotics due to the complex nature of the frequency response and the difficulty to extract the notches.

In this chapter we present an effective, yet relatively easy, way of extracting the notches from the frequency response and we show how a robot can use information about ITD/IPD, ILD, and notches in order to accurately estimate the location of the sound source in both vertical and horizontal space. Knowing the form of the head and ears it is possible to calculate the relationship between the features (ITD, ILD, and the frequencies for the notches) and the position the sound source, or even estimate the complete HRTF. However, here we are only interested in the relationship between the features and the position. Alternatively we can get the relationship by measuring the value of the features for some known positions of the sound source and let the robot learn the maps. Since the HRTF changes if there is some changes to the ears or microphones or if some object like for example a hat is put close to the ears, it is important to be able to update the maps. Indeed, although human ears undergo big changes from birth to adulthood, humans are capable of adapting their auditory maps to compensate for acoustic consequences of the anatomical changes. It has been shown that

vision is an important cue for updating the maps [130], and it can also be used as a mean for the robot to update its maps [90].

4.1 Features for sound localization

Three different features are used for localizing the sound source: ITD, ILD, and the notches in the frequency response of each ear.

4.1.1 ITD

The ITD is calculated by doing a cross-correlation between the signals arriving to the left and right ear/microphone. If the signals have the same shape we can expect to find a peak in the cross-correlation for the number of samples that corresponds to the interaural time difference, i.e. the difference in time at which the signal arrives at the microphones. We can easily find this by searching for the maximum in the cross correlation function. Knowing the sampling frequency F_s and the number of samples n that corresponds to the maximum in the cross-correlation function we can calculate the interaural time difference as:

$$ITD = \frac{n}{F_s} \tag{4.1}$$

If the distance to the sound source is big enough in comparison to the distance between the ears, l, we can approximate the incoming wave front with a straight line and the difference in distance Δl traveled by the wave for the left and right ear can easily be calculated as:

$$\Delta l = l \sin(\Theta) \tag{4.2}$$

where Θ is the horizontal angle between the heads mid sagittal plane and the sound source, Figure 4.1. Knowing that the distance traveled is equal to the time multiplied with the velocity of the sound we can now express the angle directly as a function of the ITD:

$$\Theta = \arcsin(ITD * \frac{c}{l}) \tag{4.3}$$

However for the sake of controlling the robot we are not interested in the exact formula since we want the robot to be able to learn the relationship between the ITD and the angle rather than hard coding this into the robot. The important thing is that there exists a relationship that we will be able to learn. We therefore measured the ITD for a number of different angles in an anechoic room, Figure 4.2.

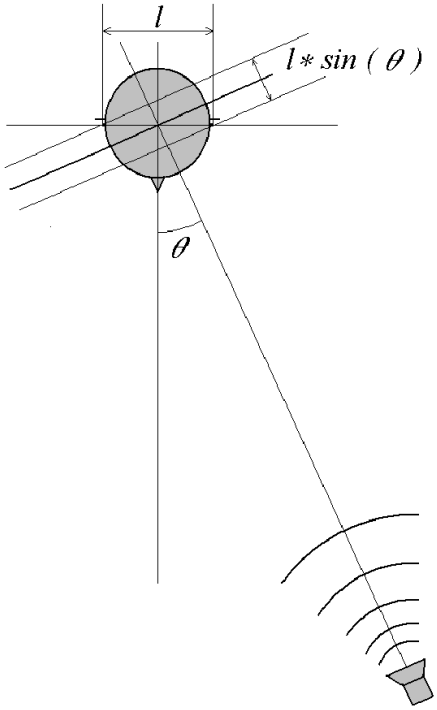

Figure 4.1: Interaural time difference

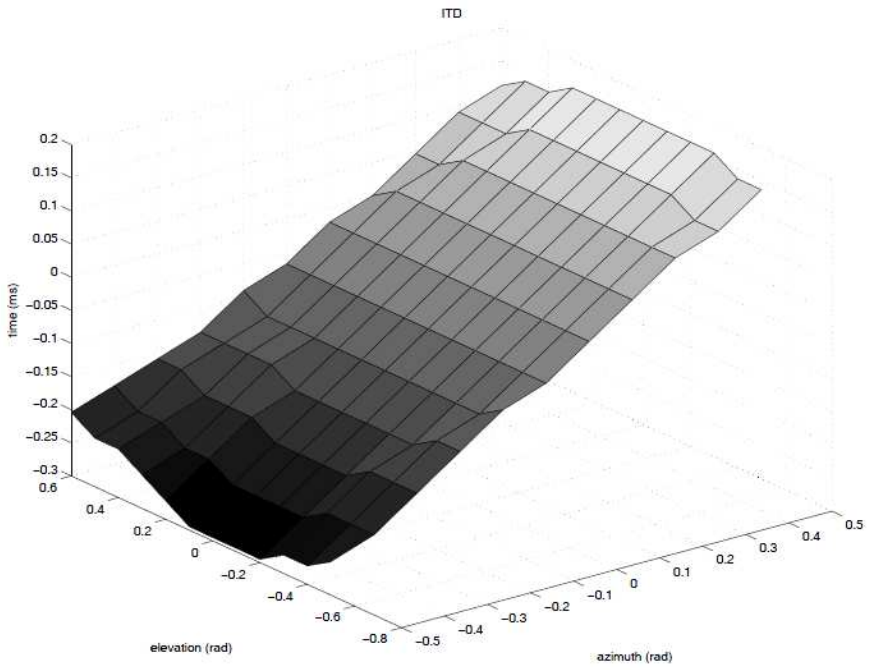

Figure 4.2: ITD for different positions of the sound source.

4.1.2 ILD

The interaural level difference ILD, is calculated as a function of the average power of the sound signals reaching the left and right ear.

$$ILD = 10 * log_{10}\left(\frac{\sum_k s_l^2(k)}{\sum_k s_r^2(k)}\right) \tag{4.4}$$

Sometimes the ILD is calculated from the frequency response rather than directly from the temporal signal. It is easy to go from the temporal signal to the frequency response by applying a fast Fourier transform FFT. The reason for working with the frequency response instead of the temporal signal is that it makes it easy to apply a high-pass, low-pass, or band-filter on the signal before calculating its average power. Different frequencies have different properties. Low frequencies typically pass more easily through the head and ears while higher frequencies tend to be reflected and their intensity more reduced. One type of filtering that is often used is dBA which corresponds to the type of filtering that goes on in human ears and which mainly takes into account the frequencies between 1000 Hz and 5000 Hz. In [90] a band-pass filter between 3-10 kHz have been used which gives them a better calculation of ILD. Different types of head and ears may benefit from enhancing different frequencies. Here we calculate the ILD directly from the temporal signal which is equivalent to considering all frequencies. The response for a sound source placed at different angles from the head is shown in Figure 4.3.

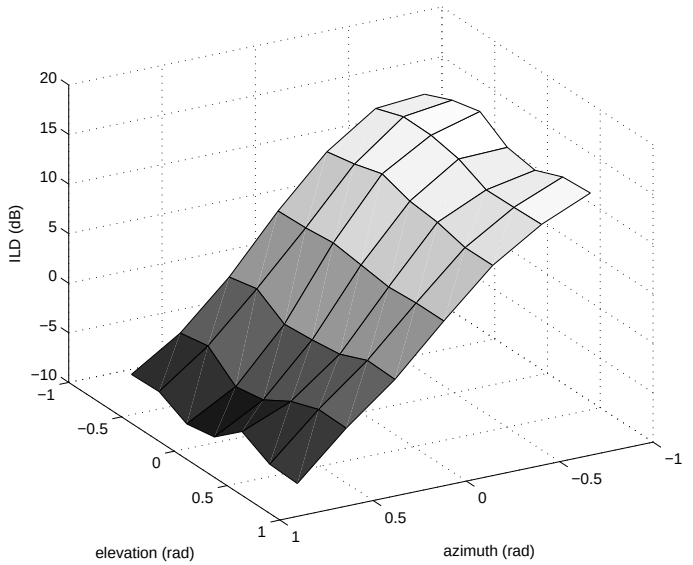

Figure 4.3: ILD for different positions of the sound source.

4.1.3 Spectral notches

Finding the spectral notches is a little more challenging. Existing methods for extracting spectral notches such as [98] [99] [100] focus on finding the notches in spectral diagrams obtained in anechoic chambers using white noise. For a humanoid robot that has to be able to turn towards any type of sound these methods are not suitable.

In [52] we therefore developed a novel method to extract the frequency notches is suggested that was reasonably fast and simple to implement, and gave better accuracy for calculating the elevation of the sound source than methods based on ILD. The method makes use of the fact that there is a slight asymmetry between the ears and has the following steps:

1. Calculate the power spectra density for each ear
2. Calculate the interaural spectral differences
3. Fit a polynomial to the resulting differences
4. Find minima for the fitted curve

To calculate the power spectra we use the Welch spectra [127]. Typical results for the power spectra density, $H_l(f)$ and $H_r(f)$, for the left and right ear respectively are shown in Figure 4.4. As seen, the notches disappear in the complex spectra of the sound, which makes it very hard to extract them directly from the power spectra. To get rid of the maxima and minima caused by the form of the free field sound, i.e. what is said, we calculate the interaural spectra difference as:

$$\Delta H(f) = 10 * log_{10} H_l(f) - 10 * log_{10} H_r(f) = 10 * log_{10} \left(\frac{H_l(f)}{H_r(f)} \right) \tag{4.5}$$

Finally we fit a polynomial to the interaural spectra difference. The best results seem to be obtained with a 12-14 degree polynomial. In this work we have used a polynomial of degree 12. As seen in Figure 4.4, the minimum more or less corresponds to the expected frequencies of the notches for the left ear. This is because we carefully designed the ears so that the notches from the two ears would not interfere with each other. In this case we could actually calculate the relationship between the frequency of the notch and the position of the sound source. However, in the general case it is better to let the robot learn the HRTF than trying to calculate it since the positions of the notches are critically affected by small changes in the shape of the pinna or the acoustic environment. Also, if we can learn the relationship rather than calculating it we do not have to worry about the fact that the minima that we find do not directly correspond to the theoretical notches as long as they change with the

elevation of the sound source. In Figure 4.5 we show the value of the notch feature with the sound source placed at a number of different positions in relation to the head.

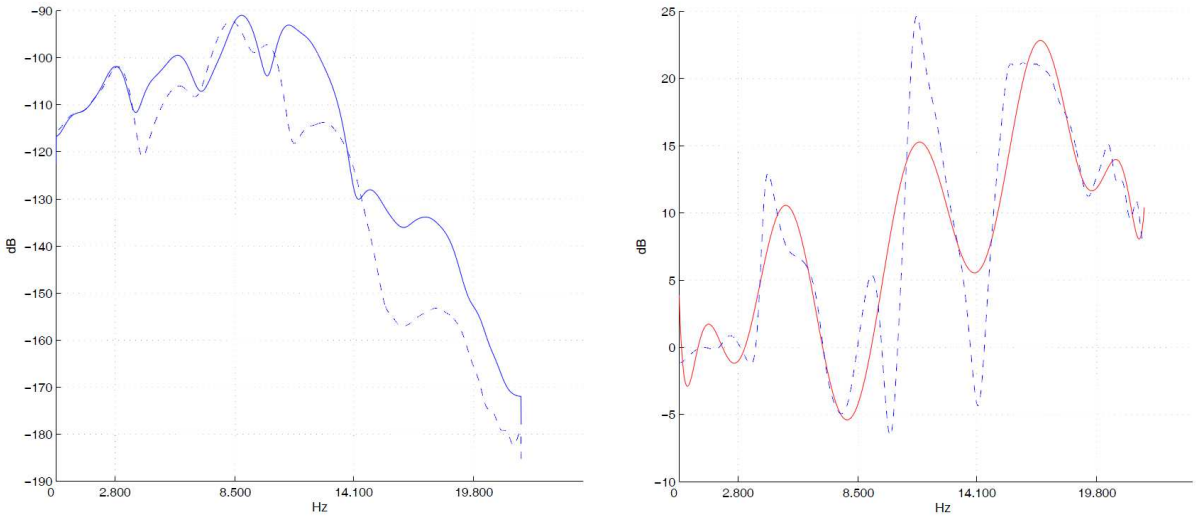

Figure 4.4: Left: The power spectra for left ear (solid line) and right ear (dotted line). Note that the spectra are shown in dB, i.e. $10 * log(H_x(f))$. The vertical lines represent the expected frequencies for the notches. Right: The interaural spectral difference (dotted line) and the fitted curve (solid line)

While the method described above works well in most situations and is fairly robust to different types of sound, it still has a few shortcomings. First of all, the steps involved are not very intuitive, and second, the method does not provide any measurement of the confidence of the estimated notch position. The latter is very important in a robot that receives stimuli from several different modalities and needs to evaluate how interesting, or salient, each of them are [102].

To address those problems we have developed an alternative way to estimate the notches, which is inspired by a method for computation of binocular stereo disparity with a single static camera [121]. The trick here is to divide the signal using several overlapping windows, with a single sample delay between each of them. For each window the frequency spectrum is calculated using FFT, and we then calculate the sum of each frequency component over all windows.

In Figure 4.6, we have used 300 windows, each with a length of 4000 samples (approximately 0.1 s sound). Instead of looking for the notch, i.e. the minimum caused by the reflecting wave, we have here chosen to look for the maximum situated between the notch at 8500 Hz and the notch at 14100 Hz in Figure 4.4. We estimate the maximum by calculating the first order moment over all frequency components between those notches, and then

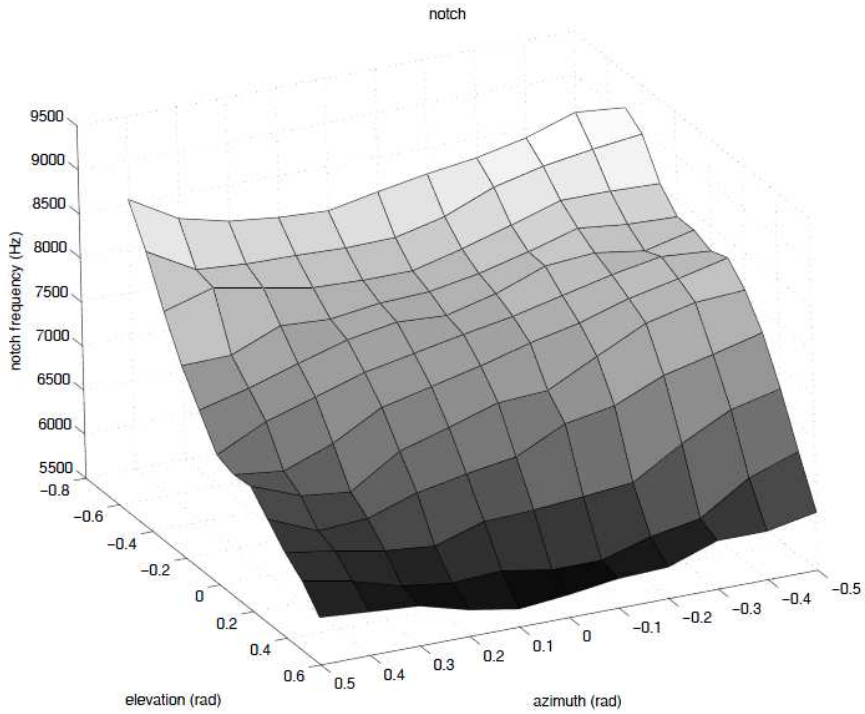

Figure 4.5: Notch frequencies for different positions of the sound source.

calculate the variance as the second order moment around this maximum. In Figure 4.6, the place of the maxima and their variance is illustrated with a Gaussian for each elevation angle.

It should be noticed that, in this new method, we have only used the signal from one of the ears. The use of multiple overlapping windows seems to overcome the problem of separating what is said from where it is said. Of course we can still subtract the signal from the other ear in order to further improve the estimate, see Figure 4.7.

4.2 Sound localization

To test the sound localization algorithm, we acquired a dataset in an anechoic chamber with a white-noise sound-source located 1.5m from the robot. We recorded 1 second of sound in 132 different head positions θ by moving the head with its own motors. A set of features was evaluated from this data consisting of ITD and the notch frequency evaluated on 0.1 second. This is used as the training dataset.

Using this training set we wanted to train a map m from the sound features C to its localization written in head spherical coordinates S_H, that can be used to direct the head toward a sound source. This map can be represented by $S_H = m(C)$. A simple way to move

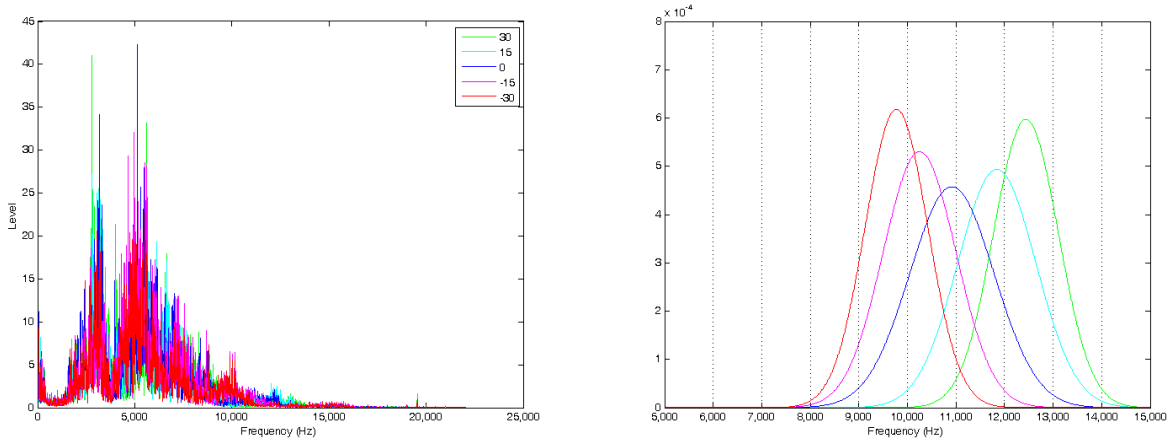

Figure 4.6: Estimated maxima for different elevation angles using one ear.

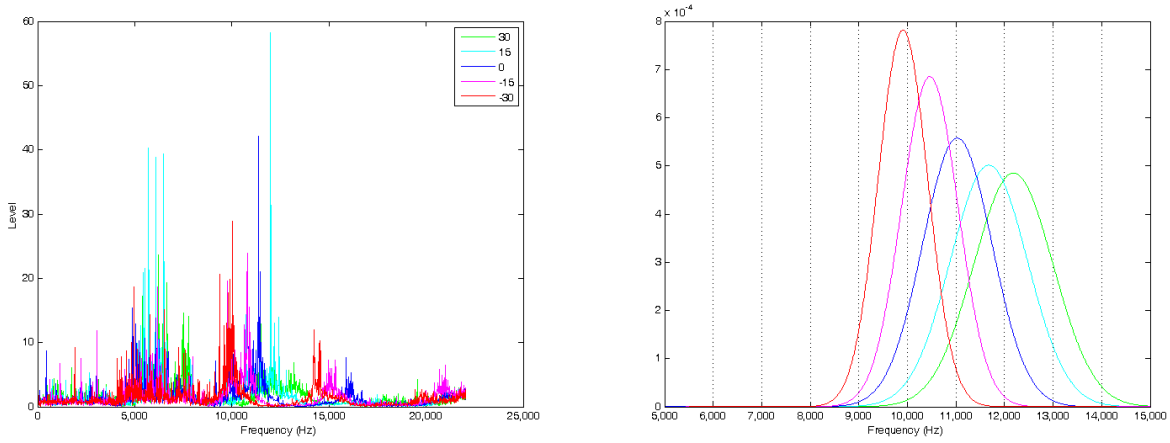

Figure 4.7: Estimated maxima for different elevation angles using two ears.

the head toward the target is to move the head pan and tilt by an increment $\Delta\Theta$ equal to the position of the sound source, i.e. $\Theta = S_H$. Although the function is not linear, if we restrict to the space of motion of the head in can be considered as such, we can observe this by noting that the features in Figure 4.2 and Figure 4.5 are almost planar. Because of this, the nonlinear function can be approximated by a linear function:

$$S_H = MC \tag{4.6}$$

This simpler model allows a faster and more robust learning. To estimate the value of M a linear regression with the standard error criteria was selected:

$$\hat{M} = argmin_M \sum_{i=1} \|\Delta\Theta_i - MC_i\| \tag{4.7}$$

After training the map, a second dataset was created to test the learning method. The procedure was similar to the previous one but the sound-source was replaced by a human voice sound. This was done because the system should operate in a human-robot interaction and also to evaluate the generalization properties of the method. Figure 4.8 presents the reconstruction error by showing for each head position the corresponding error in reconstruction. We can see that the worst case corresponds to the joint limits of the head but it is always less than 0.1 rad, which is very small. As a comparison we can note that 0.1 rad is the size of a adult human face when seen from 1.5 m of distance. The error increase near the limits is due to the nonlinearity of the features being approximated by a linear model, however with this small error the computational efficiency and robustness makes us choose this model.

Finally we want the robot to be able to compensate in the map for changes to the ears or the environment. Therefore the map has to be updated online. For this a Broyden update rule [33] was used. This rule is very robust and fast, with just one parameter and by only keeping the actual estimative of M in memory. Its structure is as follows:

$$\hat{M}(t+1) = \hat{M}(t) + \alpha \frac{(\Delta\Theta - \hat{M}(t)C)C^T}{C^T C} \tag{4.8}$$

where α is the learning rate. This method is typically used with supervised learning, where the positions corresponding to a certain sound are given. This is not the case for an autonomous system. A robot needs an automatic feedback mechanism to learn the audio-motor relation autonomously. Vision can provide information in a robust and non-intrusive way. Provided that the sound source is a human caregiver, a face detection algorithm can be used to provide the feedback, see Figure 4.9. Table 4.1 presents the algorithm for updating the audio-motor

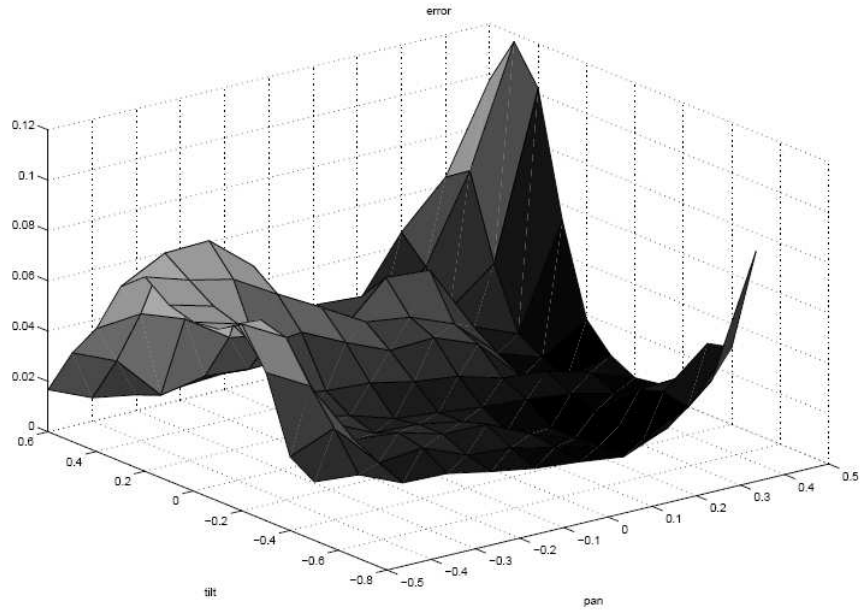

Figure 4.8: Audio-motor map reconstruction error for each head position (Unit: Radians)

map during interaction with a human.

Figure 4.9: Result of the face detection algorithm used as a feedback signal for the sound localization algorithm.

A second experiment was performed using direct interaction with the robot in our offices. After hearing a sound the head moves to the corresponding position. The previously learned

Table 4.1: Algorithm for autonomously learning an audio-motor map by interaction with a human

1) listen to sound
2) move head toward the sound using the map
3) locate human face in the image
4) if face not close enough to the center a) do a visual servoing loop to center the face b) update the map

function was used as a bootstrap. The map quality would always guarantee that the sound source was located in the camera images, even though it was not trained neither in the same environment nor considering the eye-neck coordinate transformation. In order to further improve the map, we followed the steps of the algorithm presented in Table 4.1. Figure 4.10 presents the evolution of the error during the experiment. The error represents the difference, in radians, from the position mapped by the model and the real localization of the person. We can see that this error decreased towards less than one degree.

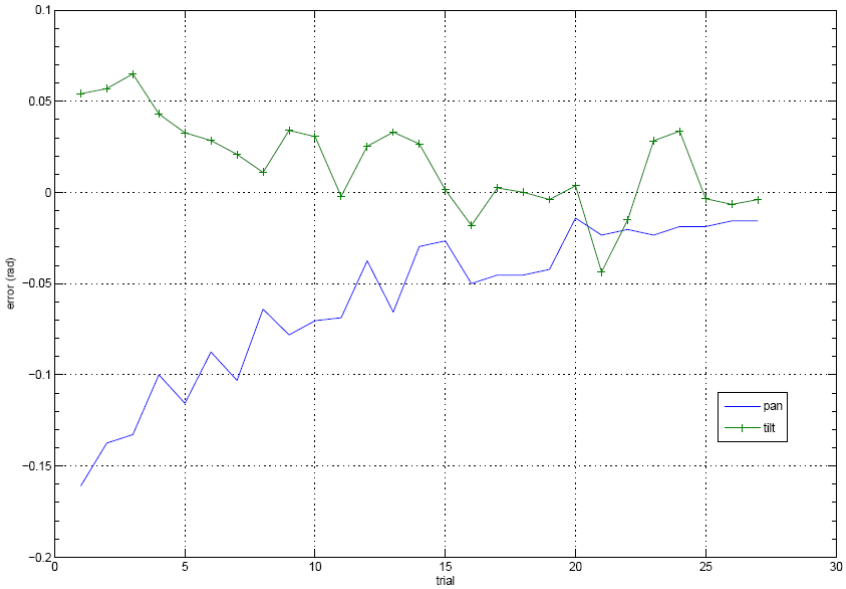

Figure 4.10: Convergence rate of the audio-motor map learning algorithm when running online with feedback given by a face detector.

4.3 Conclusions

For sound localization, ITD, ILD, and spectral notches are the most important features. A novel method for finding the notches is proposed, and it is shown that the robot is able to learn how to map these features to the location of the sound source, either by supervised learning or by using vision. The suggested method has good accuracy within the possible movements of the head used in the experiments. The error in the estimated azimuth and elevation is less than 0.1 radians for all angles, and less than 0.02 radians for the center position.

Chapter 5

Babbling and inversion mapping

In the previous chapters we have developed models of the human articulatory and auditory system, and demonstrated how the auditory models can be used for sound localization. Being able to produce and perceive speech sounds, as well as being able to turn towards a sound source are capabilities that can be seen as innate in infants, and should therefore be present also in an artificial system that aims at mimicking infant's language learning.

In the following chapters we will focus on how these models can be used to acquire language capabilities. We will follow the developmental path that was outlined in the introduction, and is shown in Figure 5.1. This consists of two parallel paths where the robot learns an initial word model in the upper path and initial speech units in the lower. These are learnt independently and are then combined in the final statistical model.

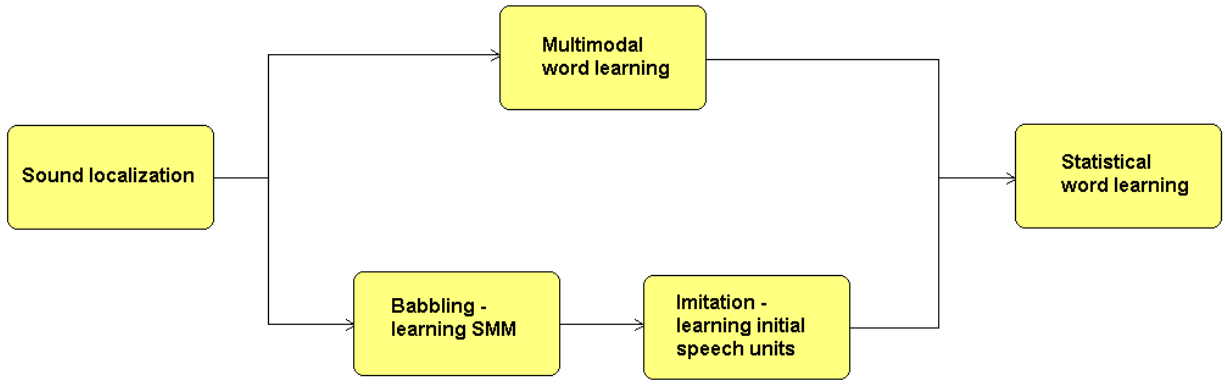

Figure 5.1: Developmental learning approach

Here we will first follow the lower path where the robot learns a set of initial speech units. This path uses motor-based learning and is closely related both to how infants learn to produce sounds and how they learn to imitate different arm gestures. In this chapter we

will look at how babbling can be used to learn the sensory-motor map, which is the first step towards being able to control the articulatory system and later on to be able to imitate speech sounds from other robots or humans.

5.1 Acoustic-articulatory inversion mapping

The acoustic-articulatory inversion mapping aims at finding the articulatory positions that can be used to produce a given speech sound. The articulatory positions are represented by the six parameters of jArticulator, and the speech sound is represented by 12 mel frequency cepstral coefficients (MFCC). Here we only look at the static map, and therefore only the actual values of the MFCC are used and not their derivatives. Hence, in the case of inversion mapping, we want to map the 12 MFCC back to the 6 articulatory positions used to create the sound.

This mapping is a well-studied, but difficult problem, since several articulatory positions can be used to produce the same or similar speech sounds. This many-to-one relationship has been demonstrated in several different studies where the subjects have had their articulatory space restricted using a bite plate. Despite this restriction, subjects have been able to pronounce both vowels and consonants correctly. It has also been shown that they do this mapping directly without the need to listen to the actual output.

One common solution to the inversion problem is to use a codebook that stores the resulting sound for different articulatory positions. If the articulatory space is densely sampled, this method can give accurate results. On the downside it requires lots of resources, both in terms of memory in order to store the codebook, and in processing power to search the codebook.

A more efficient approach in terms of memory and processing power, is to use an artificial neural network (ANN). Conventional neural networks can be used to represent any kind of function in order to map the input variables, i.e. the acoustic representation, to a set of output variable, i.e. the articulatory parameters. The usual approach to train the network involves minimizing the sum-of-squares error over a set of training data, by using back-propagation to update the network weights. Unfortunately, when there can be multiple target values for the same input, these networks will provide the average of the target values which is not necessarily a correct value. A better result would be if, for every input, the network would give us the conditional probability distribution of the target data. This can be achieved by using a Mixture Density Network (MDN) [12].

5.1.1 Mixture Density Networks

An MDN is a combination of an ANN and a mixture model that can represent arbitrary conditional probability distributions, as shown in Figure 5.2. In this case $\mathbf{x}$ represents the MFCC, and $\mathbf{t}$ is the target vector, i.e. the articulatory positions. Hence, we want the network output $\mathbf{z}$ to represent the conditional probability density

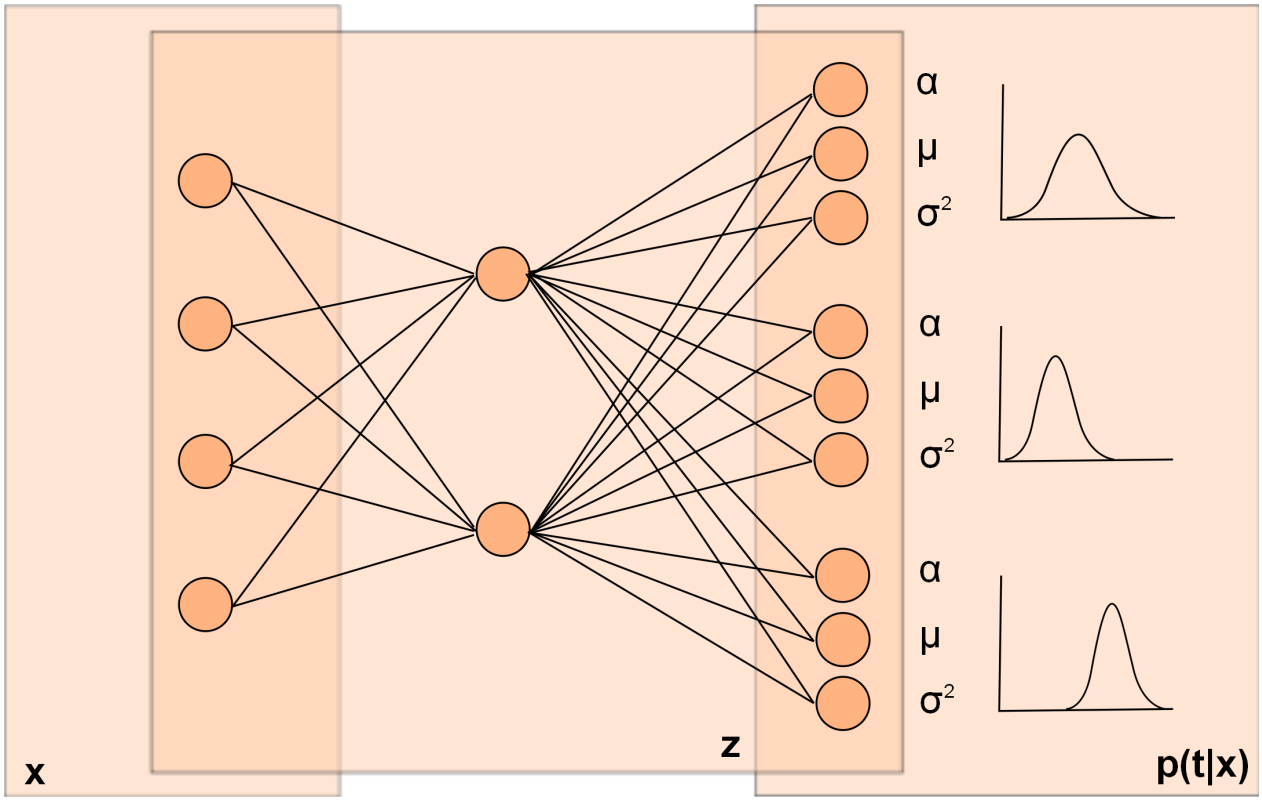

Figure 5.2: Mixture Density Network with three Gaussian mixture components

$$p(\mathbf{t}|\mathbf{x}) = \sum_{i=1}^{m} \alpha_i(\mathbf{x})\phi(\mathbf{t}|\mathbf{x}) \tag{5.1}$$

where m is the number of mixture components, $\alpha_i(\mathbf{x})$ are the mixing coefficients or prior probabilities that the target vector $\mathbf{t}$ having been generated from the ith component of the mixture, and $\phi_i(\mathbf{t}|\mathbf{x})$ are the conditional density. Typically the density is represented with a Gaussian:

$$\phi(\mathbf{t}|\mathbf{x}) = \frac{1}{(2\pi)^{c/2}\sigma_i(\mathbf{x})^c} exp\left\{-\frac{\|\mathbf{t} - \boldsymbol{\mu}_i(\mathbf{x})\|^2}{2\sigma_i(\mathbf{x})^2}\right\} \tag{5.2}$$

where the vector $\boldsymbol{\mu}_i(\mathbf{x})$ represent the centre of the ith mixture component, $\sigma_i(\mathbf{x})$ is the

variance, and c is the dimension of the target vector.

For the network part, a standard multilayer perceptron (MLP) feed-forward network is used. Here a single hidden layer 20 neurons is used. For the hidden layer sigmoidal activation is used, and for the output layer a linear function is used. The number of output neurons depend on the number of mixture components, M. In total $(c+2)*M$ output neurons are used. For each component that allows for c neurons that represents the center of the kernel, 1 neuron for the prior, and 1 neuron for modeling a spherical covariance.

To make sure that the network output $\mathbf{z}$ provides a valid probability density some additional constraints are needed. In order to constrain the mixing coefficients to lie within the range $0 \leq \alpha_i(\mathbf{x}) \leq 1$ and to sum to unity, the softmax function [14] is used to relate the output of the corresponding units in the neural network to the mixing coefficients.

$$\alpha_i = \frac{exp(z_i^\alpha)}{\sum_{j=1}^{M} exp(z_j^\alpha)} \tag{5.3}$$

where z_i^α is the output of the neural network corresponding to the mixture coefficient for the ith mixture component. The variances σ_i represent scale parameters. To avoid that the variance become less than or equal to zero it is convenient to represent them in terms of the exponentials of the corresponding network outputs

$$\sigma_i = exp(z_i^\sigma) \tag{5.4}$$

where z_i^α is the output of the neural network corresponding to the variance for the ith mixture component. Finally the means are represented directly by the corresponding outputs of the ANN:

$$\mu_{ij} = z_{ij}^\mu \tag{5.5}$$

The training of the network part of the MDN, i.e. the weights of the ANN, can be done using a standard back-propagation algorithm. To do this we need a suitable expression of the error at the network output.

We want the network output to correspond to the complete conditional probability density of the output variables. Hence we want to maximize the likelihood that the model gave rise to the particular set of data points. In practice this is done by minimizing the negative log likelihood of the observed target data points

$$E = -\sum_n ln \left\{ \sum_{i=1}^{M} \alpha_i(\mathbf{x}^n) \phi_i(\mathbf{t}^n | \mathbf{x}^n) \right\} \tag{5.6}$$

given the mixture model parameters. The derivatives, $\delta_k^q = \partial E^q / \partial z_k$ for a particular pattern can then be used as the errors which can be back-propagated through the network in order to update the weights. By introducing the following term for the posterior probabilities, which is obtained using Bayes theorem

$$\pi_i(\mathbf{x}, \mathbf{t}) = \frac{\alpha_i \phi_i}{\sum_{j=1}^{m} \alpha_j \phi_j} \tag{5.7}$$

the error values for the network outputs corresponding to α_i, μ_{ik}, and σ_i respectively, can be expressed as [12]:

$$E_\alpha = \alpha_i - \pi_i \tag{5.8}$$

$$E_\mu = \pi \left\{ \frac{(\mu_{ik} - t_k}{\sigma_i^2} \right\} \tag{5.9}$$

$$E_\sigma = -\pi \left\{ \frac{\|\mathbf{t} - \boldsymbol{\mu}_i\|}{\sigma_i^2} - c \right\} \tag{5.10}$$

After training, for each observation $\mathbf{x}$, the MDN will provide us with m possible target vectors and the a priori probability that each of them have been used to create the observed vector.

5.2 Babbling

As mentioned, the acoustic-articulatory inversion map must be trained in order to set the weights of the MDN. Just as an infant, the robot does this through babbling. During babbling the robot produces a speech sample using a given articulatory position and then uses this sample together with the known articulatory position to update the map. The complete path for generating a speech sample and mapping it back to the articulatory positions is illustrated in Figure 5.3.

In practice, this is not done one speech sample at a time. Instead a large number of speech samples are created and the articulatory positions together with the resulting MFCC are stored in a training set. This training set is then used to train MDN using batch training. After training, the resulting map is evaluated using a separate evaluation set.

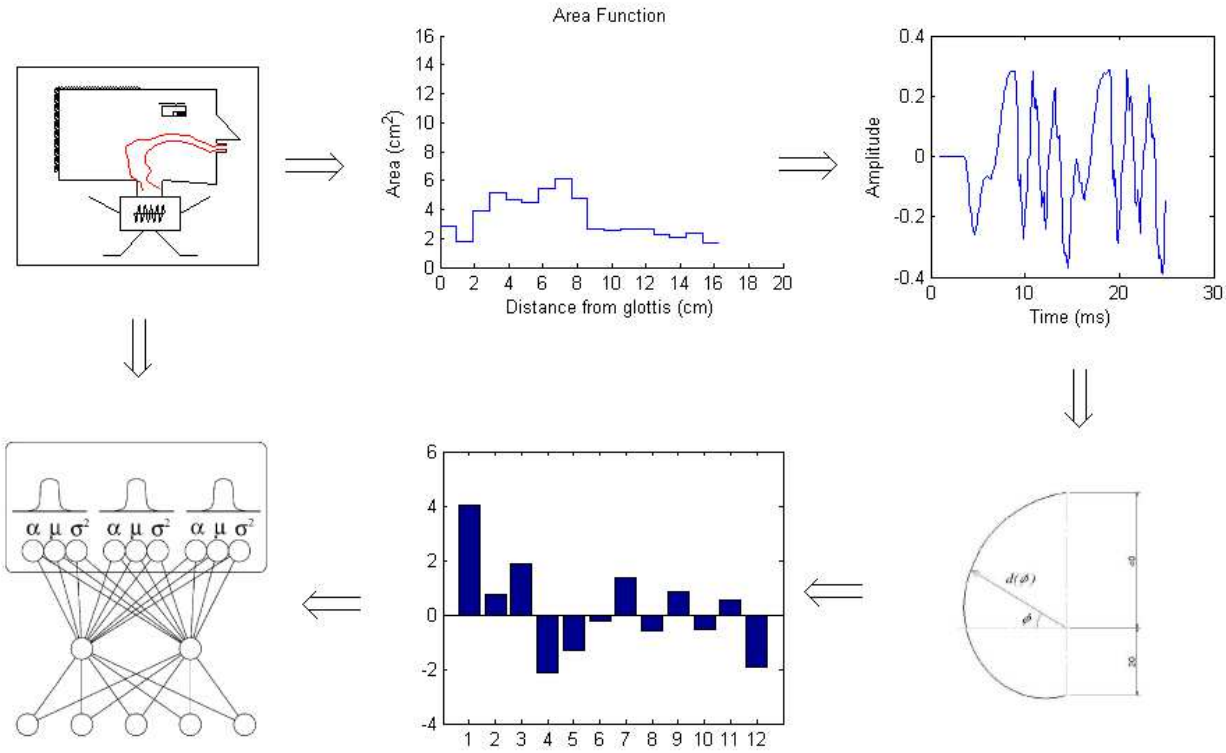

Figure 5.3: Babbling

5.2.1 Creating a training set

A simple way to create the training set is to randomly sample the complete articulatory space and synthesize speech from those positions through the articulatory model. While the articulatory model greatly reduces the search space and removes many unrealistic configurations of the underlying tube model, the resulting training set may still contain too much redundancy. This makes it difficult to learn the map without using very large datasets and a mixture with a large number of kernels.

Many of the redundant positions may never be used by human speaker because they are difficult to reach or perceived as uncomfortable. While our articulatory model does not contain any measure of how comfortable a certain position is, we have manually defined three corner vowels that specify the outer boundaries for what can be perceived as a comfortable zone. These positions are then used as a starting point for further exploration by adding Gaussian noise.

To compare the two approaches, two separate training sets were created, one using unrestricted random sampling containing 50,000 samples, and another set using restricted sampling around the corner vowels containing 10,000 samples.

5.2.2 Training the MDN

Before starting the actual training, the network weights must be initialized. To get a good starting point, the target vectors of the training data are first clustered into a defined number of groups using K-means. The number of groups is equal to the number of mixtures in the MDN, and the mean target values for each group are chosen as the bias for means of each mixture. The number of target vectors in each cluster is used to set the bias for the priors, and a fixed value is used to set the bias for the variance. The rest of the weights are initialized using random numbers.

During training, each example is presented to the network, which calculates the outputs given the current weights. The outputs are compared with the target values and the errors are calculated according to error expressions above. These errors are then back-propagated through the network and the weights are adjusted in order to decrease the error. Figure 5.4 shows how the training error decreases for each epoch, i.e. each time the complete training set has been presented to the network.

5.2.3 Evaluation

After training, the network was evaluated using a set of 12 vowels that are included in VT calcs, and also compatible with the model in jArticulator. The reason for using this

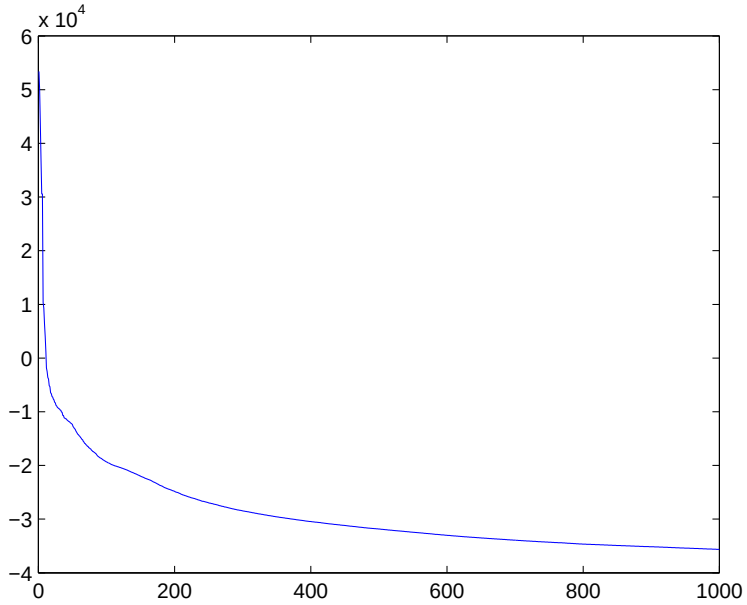

Figure 5.4: Training error of the MDN. The error decreases for each epoch, i.e. each time the training set has been presented to the network.

rather limited set for evaluation, instead of creating a large test set covering a wider range of possible articulatory positions, is that we are specifically interested in knowing how well the map works for typical speech sounds and not very interested in how well it might work for other types of sounds even if those can be produced by the articulatory model.

Both the error of the mapped positions and the error of the resynthesized speech sound was evaluated for different number of mixtures, see Figure 5.5 and Figure 5.6 respectively. For the acoustic error, we only show the result obtained when using the mixture with the highest prior. For the position error, we show three different results: 1) the result using only the mixture with the highest prior, 2) the best result obtained with any of the three mixtures with highest prior, and 3) the best result obtained using any of the M mixtures.

It can be noticed that the error of the resynthesized speech sound is less than half when using 10 mixtures compared to using a single Gaussian. On the other hand, there is no improvement in the estimated position when increasing the number of mixtures, as long as only the mixture with the highest prior is used. This should not come as a surprise though. If a speech sound can be produced using several different vocal tract positions there is no way to tell from which of these position it was actually produced. Using a single Gaussian will be equivalent to using a common least squares network, and the estimated position will be a weighted average of all possible solutions. Unfortunately, the average of two positions that create the same sound may result in a position that creates a completely different sound. If we

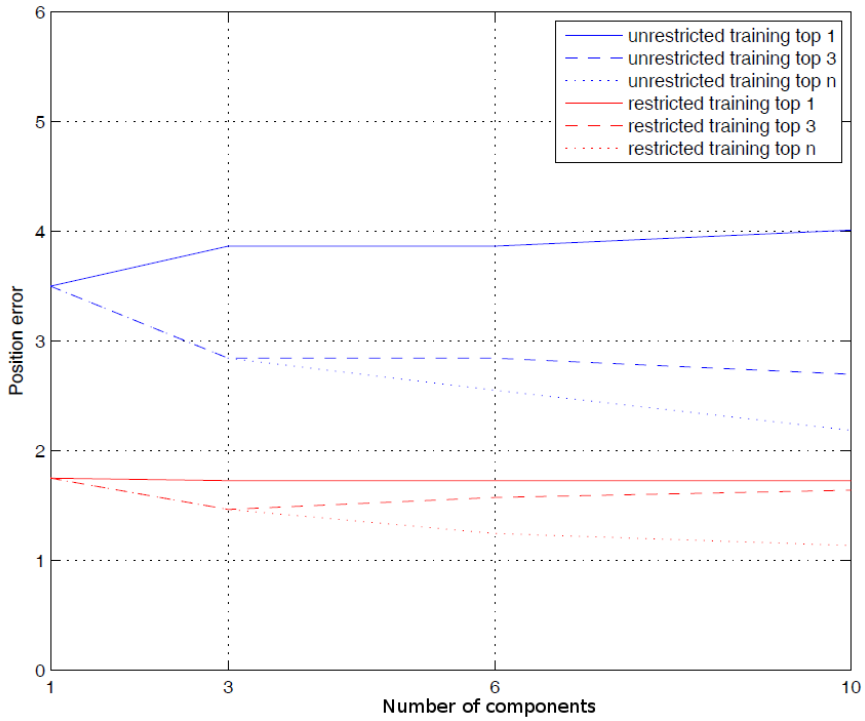

Figure 5.5: Sum of squared errors for the estimated vocal tract positions when training with different number of components

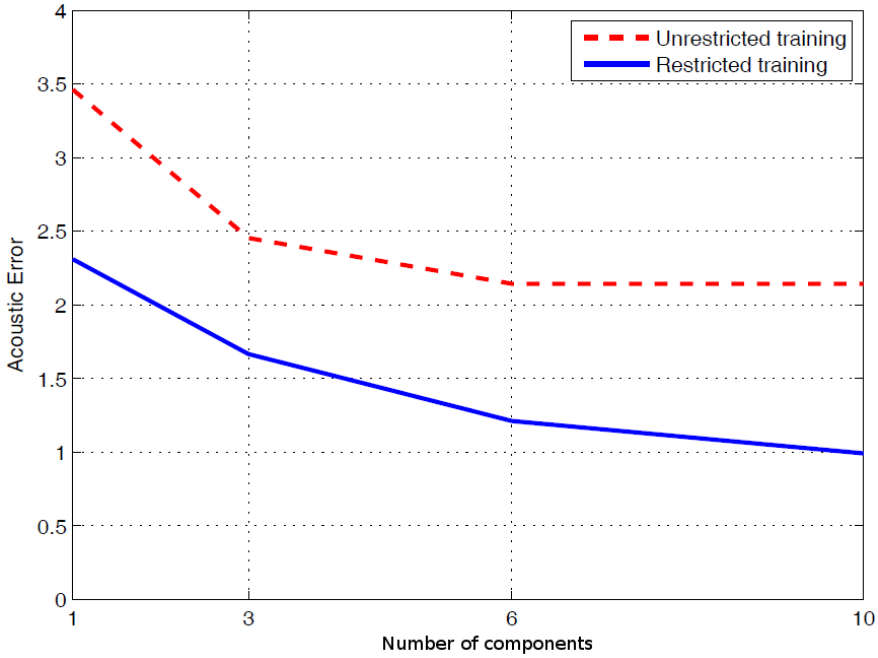

Figure 5.6: Sum of squared acoustic errors for the synthesized sound (using the vocal tract position with highest prior), after training with different number of mixture components

have several mixtures, but only select the position coming from the mixture with the highest prior, this position may be even further away from the position that actually produced the mapped sound, while still being able to reproduce the same sound. On the other hand, if we compare the original position with not the position with the highest prior, but several of the mixtures, the error should be reduced. This is also what can be seen in Figure 5.5.

Table 5.1: *The priors (i.e. mixing coefficients) for each mixture component when the network is presented with different vowel sounds.*

vowel	1	2	3	4	5	6	7	8	9	10
ie	0.0050	0.0001	**0.9511**	0.0004	0.0000	0.0000	0.0000	0.0000	0.0435	0.0000
ey	0.0445	0.0122	0.4854	0.0521	0.0006	0.0005	0.0002	0.0004	0.4014	0.0027
eh	0.0656	0.2279	0.0841	0.2414	0.0099	0.0104	0.0084	0.0845	0.0624	0.2054
ah	0.0112	0.2077	0.0031	0.1203	0.0216	0.0190	0.0195	0.2357	0.0058	0.3561
aa	0.0022	0.2370	0.0000	0.0807	0.0563	0.0710	0.0712	0.2443	0.0005	0.2369
ao	0.0020	0.0762	0.0000	0.0841	0.2415	0.1047	0.0734	0.2433	0.0005	0.1742
oh	0.0021	0.0815	0.0000	0.0709	0.3696	0.1108	0.0658	0.1740	0.0009	0.1243
uw	0.0246	0.0593	0.0000	0.0650	0.5746	0.1470	0.0220	0.0429	0.0158	0.0487
iw	0.4646	0.0780	0.0320	0.1123	0.0290	0.0089	0.0024	0.0246	0.2037	0.0445
ew	0.1757	0.2814	0.0052	0.1330	0.0363	0.0214	0.0066	0.0794	0.1703	0.0908
oe	0.1121	**0.2275**	0.0777	**0.1954**	0.0142	0.0131	0.0066	0.0782	0.1130	0.1623

Table 5.2: *Motor parameters for each mixture component for vowel oe*

mixture	jaw	tongue pos	tongue shape	apex	lip height	lip protrusion
1	-0.8249	-0.5310	0.5203	-3.7408	0.6561	-2.5013
2	**0.0686**	**0.2140**	**0.6580**	**-1.6138**	**0.5913**	**0.1685**
3	1.1385	1.5579	1.2080	-1.3217	-0.4465	-2.5630
4	**-0.9942**	**-0.5277**	**0.4814**	**-1.9700**	**0.0730**	**-1.4687**
5	0.3132	-0.1453	0.4649	-1.0988	0.1584	-1.4087
6	3.0514	0.6096	0.7904	0.0395	1.2104	1.1393
7	1.4757	0.5679	0.0509	-0.1324	0.9555	0.7901
8	0.7334	0.6216	0.2125	-0.3468	0.5694	0.0775
9	0.1459	0.3121	0.9534	-2.1644	0.7840	-1.4255
10	-0.3277	0.1779	0.3418	-0.9043	0.1715	-0.7665
original	-1.0	-0.5	0.5	-2.0	0.2	-0.5

Table 5.1 shows the priors for each component in the case of 10 mixture components. For some vowels, such as *ie*, there is a strong prior for a single component, while for other vowels, such as *oe*, there are several components with similar priors. Table 5.2 shows center positions for each of the mixtures for the vowel *oe*. Despite the relatively large differences in

articulatory positions, mixture component 2 and mixture component 4 produce very similar speech sounds. This can be seen by comparing the spectra of the two, Figure 5.7.

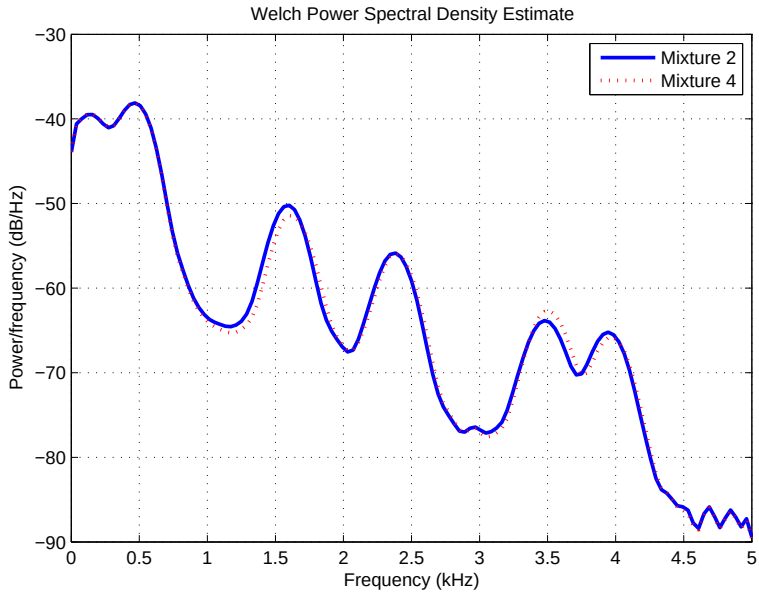

Figure 5.7: Frequency spectra of the synthesized speech signal, from two different positions mapped with the MDN.

5.3 Conclusions

This chapter has explained the inversion problem, i.e. how to retrieve the articulatory positions that have been used to create a speech sound. Due to the one-to-many relationship, an MDN has been used for the inversion map. It is shown that by increasing the number of mixture components we can both create a more accurate map and at the same time find several alternative articulatory positions that can be used to reproduce a given speech sound. It is also shown that restricting the babbling around what can be seen as comfortable positions make it easier to learn the map.

Chapter 6

Learning initial speech units

We have seen how to define and learn audio-motor maps doing babbling. This is the first step towards being able to use motor space, i.e. the articulatory space, for learning new speech sounds and recognizing those sounds based on motor primitives.

In the introduction of this thesis we identified two additional steps based on a similar work on grasping [80]: gaining source invariance, and imitations. We also discussed some of the similarities and differences between learning to grasp and learning to produce speech sounds. One of the main differences was the increased importance of interactions when learning to speak. There are two reasons for this. The first reason is that speech is specifically about communication and there is no natural reward for producing a speech sound, and interactions are therefore needed to provide that reward. For grasping, being able to grasp an object can be seen as rewarding in itself.

The other reason that interactions are more important when learning to speak is that gaining source invariance for interspeaker differences is a much harder problem than the related problem of overcoming view-point differences for grasping. Since a view-point transformation may be easier to do than to overcome acoustic inter-speaker differences, we will also look at how vision in the form of lip-reading can be used to facilitate the mapping. It has been shown that in order to babble naturally it is important for infants to establish visual contact with others in order to [88].

In this chapter we will discuss how interactions can be used both in order to gain speaker invariance and to learn an initial set of speech sounds that will later be used as building blocks.

6.1 Gaining speaker invariance

The robot's first interactions are simply extensions of the babbling where the robot articulates sound and then listen for a response from the caregiver. If the response from the caregiver is an imitation of the robot's utterance, the robot can use not only the sound produced by itself, but also that from the caregiver to update its sensorimotor maps. Repeating this with several caregivers allows the robot to create a speaker-independent map. However, if the caregiver is in fact not repeating the same utterance, doing this could potentially destroy the map.

As been explained in the previous section, in the case of infant-adult interactions, there is around 20% chance that an eventual response is actually an imitation. Even though this percentage can be made much higher in a human-robot interaction where the caregiver is aware of the learning strategy of the robot, we still want the learning to work under natural conditions. It is therefore necessary for the robot to be able to separate imitations from non-imitations. This can be done by comparing some simple prosodic features in the acoustic signals produced by the robot and the caregiver. In the study executed together with the linguistic department of Stockholms university [55], it was found that comparing the number of syllables and the duration of the last syllable was enough to get a good imitation classification for early infant-adult interactions. As the infant gets older more prosodic features need to be compared in order to get a good classification. A threshold is used to decide when the prosodic features are sufficiently similar to classify the caregiver's utterance as an imitation. This threshold is considered to be innate. Details on how we model speech imitations, including which features and threshold that were chosen will be given below. After that we will discuss how visual information can be used to facilitate the mapping.

6.1.1 Modeling speech imitations

While robots usually follow very strict imitation games as described above, adult-child interactions tend to be much more complex. For the robot to be able to learn its maps under such natural conditions it has to be able to separate imitations from non-imitations. Our hypothesis is that this can be done by comparing the prosody of the utterance from the infant or robot with that of the caregiver.

A dataset with natural adult-infant interactions was created at the Phonetics Laboratory, Stockholm University. Data was recorded during 15 half-hour sessions. Seven Swedish infants, with ages ranging from 185 to 628 days, participated in one, two or three sessions each. A lot of care was taken to allow for natural interactions during the experiments. The

recordings were made in a comfortable home-like environment. The infant and the adult were free to move around during the recordings and they were also provided with a number of toys. In total, these recordings generated an adult-infant interaction speech data base consisting of 4100 speech samples.

To get a ground-truth for which utterances that should be considered as imitations, a listening experiment was performed. A computer program created in LabView was used to select utterances from the database and present those for the user. The program randomly draws an utterance from the pool of adult utterances and then randomly selected the utterance that the infant produced within five seconds before or after the adult's utterance. For the imitation judgments 20 subjects were each presented with 150 pairs of utterances (50 from each age group). Of these the subjects evaluated 22% as imitations, 19% as uncertain, and 59% as non-imitations.

There was no significant difference in the number of perceived imitations between the three age groups. However, there was an obvious difference in the way subjects classified a pair of utterances based on the perceived age of the infants. The older the infant are the higher are the demands on matching parameters. This is nothing new and can be illustrated with the familiar example of [baba] that happily is rewarded as an imitation of both [mama] and [papa] when the infant is very young, but it is still worth to mention as it has implications on the way we chose to build our classifier. Further analysis of the results showed that in 52% of the cases that a pair of utterances was judged as an imitation it was the adult imitating the infant.

Next, we wanted to create a classifier that is able to find when there is a true imitation, based on the prosodic features of the utterances. We used the prosodic features explained in section 3.2.2.

For the number of syllables we simply classified every utterance-pair where the number of syllables of the infant and the adult utterance did not match as a non-imitation. Such a straightforward classification could not be done for the other features. Even for a true imitation we cannot expect the difference between the same features of the two utterances to be zero. We used a gamma-distribution to model the differences for each feature in the case of an imitation and a non-imitation. The parameters of the gamma-distribution were then estimated from the normalized distance between the feature values for each pair of utterances. This was done separately for utterances judged as imitations and non-imitations, as well as for each age group.

In order to mimic the way human listeners seemed to incrementally demand a more detailed match between the utterances as infants got older, we took a hierarchical approach by adding an extra classifier for each age group. For the youngest infants the length of

the last syllable showed to be the most efficient feature for separating imitations from non-imitations. We chose the crossing between the two distributions in 6.1 as a separation point for when an utterance should be classified as an imitation. Doing so 74% of the utterances classified as imitations had also been classified as imitations by the panel subject, while 26% were false positives. Using the same feature for the data in the second age group, we only get around 50% true classifications. However, when adding a second classifier based on the fourth feature, i.e. difference in length between the two last syllables, the combined classifier gave around 80% true positives for the second age group and around 65% for the third age group. Finally we added a third classifier based on the difference in pitch which was able to completely eliminate the false positives for the third test group.

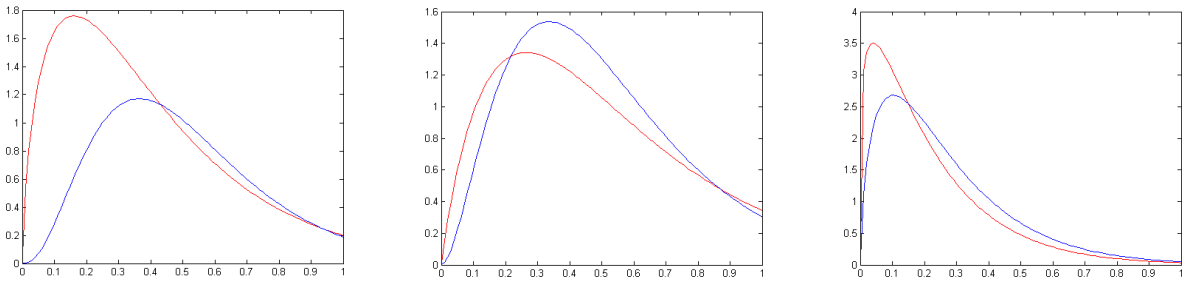

Figure 6.1: Expected feature values for imitations (red) and non-imitations (blue). From left to right, the plots show the features used for the first classifier (difference in length of the last syllable), second classifier (features of classifier one + difference in length between the two last syllables), and third classifier (features of classifier two + difference in pitch).

Finally the imitation classifier was tested during a simple human-robot interaction game, where the robot randomly selected a number of points in the vocal tract model and created trajectories between those. The utterance was then synthesized and played for the caregiver who tried to imitate the sound. When presenting the utterance from the robot and the caregiver to the classifier 66% were classified as imitations by the first classifier, 39% by the second, and only 3% by the third classifier.

We also tested how important it is for the robot to be able to detect false imitations when trying to learn the map between human speech sounds and its own vocal tract by doing a simulated imitation game. For this experiment we used a database of Portuguese vowel sounds from 14 different speakers and the setup described in [54]. Data from seven of the speakers were used for training the network and data from the other seven were used for testing. The robot was presented a mix of correct vowel imitations and false positives. First we tested how well the robot could learn the map without checking if it is a true imitation by using only 22% correct data. Then we simulated the use of the hierarchical classifier and first trained with 72% correct data, followed by 80%, and finally 100% correct data. We found

that the robot will not be able to learn the map from natural interactions without being able to separate imitations from non-imitations. However, when using the suggested hierarchical classifier it would be able to make use of imitations found during natural interactions in order to learn the map. We also notice that a residual number of false positives around 20% do not affect the quality of the map for other than the training data.

6.1.2 Using visual information

When available, the infant/robot can also make use of visual information in order to estimate the articulatory positions used to produce an utterance. As the robot cannot see its own lip movements, it is again dependent on the interaction with the caregiver. During face to face interaction with the caregiver the robot can estimate the openness of the mouth and given that the caregiver is imitating the utterance from the robot, it is possible for the robot to map this to its own articulatory positions. This is done simultaneously as the robot updates its audio-motor map and hence it is again the comparison of the prosodic features of the utterances produced by the robot and the caregiver that tells the robot if it should use the response to update its maps or not.

Lip tracking

The purpose of the visual sensor is to provide visual information that can help estimating the parameter values of the vocal tract parameters. While there are methods to find the exact contour of the lips, like the usage of snakes or active contour methods [67], these methods are typically too complex to use in speech recognition. With no a priori assumption of the shape of the lips the estimation becomes slow and more error prone. Furthermore, the complexity of the final description makes further data processing costly. For practical applications where we need to track the movements of the lips in real-time, and are interested in some simple feature like the area of the mouth opening rather than the exact contour, we need a compact representation of the lips. In this work we have chosen to represent the lips by an ellipse, which is fitted to the pixels that belong to the lips. The pixels that belong to the lips are found by using color segmentation. The color segmentation can be done in several different ways. It is usual to extract the color from the first frame using the initial position of the lips. In [118] the whole color distribution of the lip region is calculated and modeled as a Gaussian mixture and the EM method is used to estimate both the mixture weights and the underlying Gaussian parameters. Here we use a much simpler method and simply model a lip by its redness, where we define the redness as:

$$Redness = R^2/(R^2 + G^2 + B^2) \tag{6.1}$$

where R, G, and B are the red, green, and blue value of an RGB-image. If the redness of a pixel is above some threshold we define the pixel as a lip. The threshold can be calculated from the initial frame, but we have chosen a fixed threshold of 0.9. This threshold seems to work well even for different persons. Of course there are other pixels apart from the lip pixels that are classified as red so we need to know the approximate position of the lips and only use those pixels to fit the ellipse. Here we use a face detection algorithm, based on [124] and [78]. The face detection algorithm not only gives us an initial estimate for the position of the lips, but also gives us the size of the face which is later used to normalize the area of the mouth opening. However, the face detection algorithm is rather slow so the position and size of the head is therefore only calculated once in the beginning of every experiment and the subject with which the robot interacts is assumed to maintain approximately the same distance to the robot during each experiment.

To fit the ellipse to the lip pixels we use a least square method described in [34]. The result is shown in Figure 6.2. We then use the ellipse to calculate the area of the mouth opening. The ratio between the area of the mouth opening, given by the lip tracker, and the area of the face given by the face tracker, is used as a visual feature and is sent to the vision-motor map.

As said before, the face detection is too slow to be useful for tracking the movements of the lips between two frames in the video stream. We therefore use the method suggested by Lien et. al [77]. They use Lucas-Kanade tracking algorithm [82] to track the movements of the lips between adjacent frames. One problem with the tracking algorithms is that it is sensitive to the initial feature point selection as most points on the lips have ambiguities around the lip edges. Here we solve this by looking for Harris features [45] around the lips and use these as initial points that will be tracked. The result gives us a sufficiently good estimate to maintain an initial estimate of the lip position over the video sequences used in our experiments.

6.2 Learning speech units by imitation

In order to find useful target positions for creating speech sounds, the robot tries to imitate the speech sounds pronounced by its caregiver. Given that the robot is able to correctly map the caregiver's utterances to its own articulatory space, it should be able to recreate the utterance. However, due to the difficulties explained above, the robot is not able to fully learn the map before it has found the target positions. Still the initial map should be

Figure 6.2: Lip tracking

accurate enough in many cases. By trying to recreate the caregiver's utterance using the mapped positions, the robot is able to verify if the mapping was correct.

This scenario is equivalent to having the infant trying to imitate the caregiver. As explained in the previous chapter it is common that the caregiver needs to repeat the utterances several times and adapting its voice in order to help the infant to find the right positions. The same technique is used in the human-robot interaction. If the caregiver finds the robot's utterance is an acceptable imitation it gives a positive reinforcement that causes the robot to store the positions used to create the utterance. If the caregiver does not consider the utterance an acceptable imitation the caregiver can try to adapt the voice in order to overcome shortcomings in the map.

6.2.1 Clustering of target positions

Initial speech units can be found through an hierarchical clustering algorithm [46]. The algorithm starts by creating one cluster for each target position. It then iteratively joins the two clusters that have the smallest average distance between their items until only one cluster remains.

The resulting hierarchical tree is then 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 [119]. 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 speech unit in the motor vocabulary.

6.2.2 Imitation experiment

The objective of this experiment is to show how the robot can learn speech units, i.e. articulatory target positions, corresponding to 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 neighborhood 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 jArticulator. 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 behavior 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 6.3.

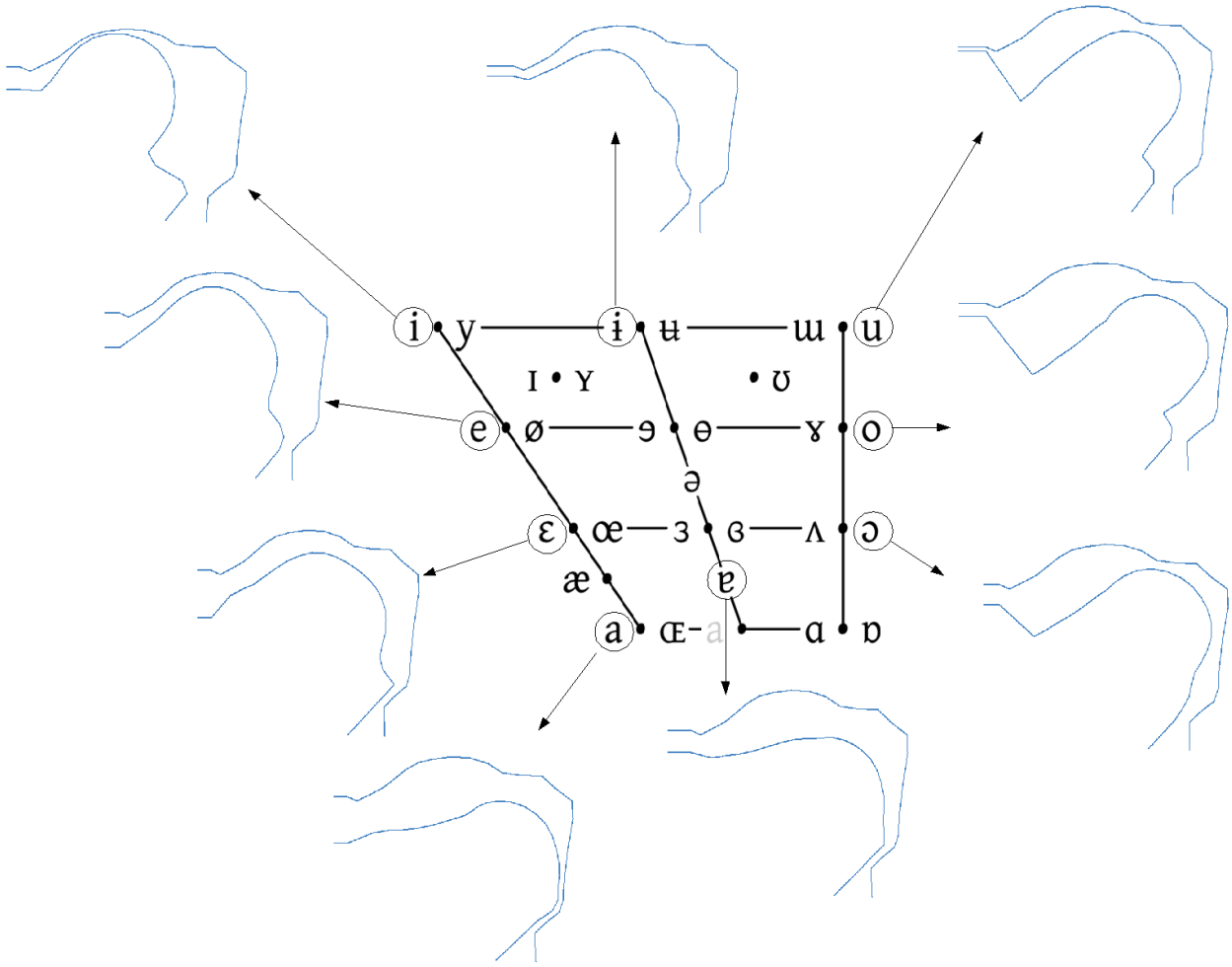

Figure 6.3: Articulator positions used by the robot for the Portuguese vowels. In the center we show the positions of the vowels in the International Phonetic Alphabet (IPA). The vertical axis in the IPA corresponds to the vertical position of the tongue and the horizontal axis to the front-back position 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 speech unit. In [54] 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 speech units. This showed that the robot automatically

would group the positions into nine speech units 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 reads the words several times, and the vowels were hand labeled 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 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 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 interaction with humans, 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.

Table 6.1: 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%

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 6.1.

6.3 Conclusions

In this chapter we have shown how the robot can learn speech units by imitating the caregiver. However it is also found that imitation is not only about the robot or the infant imitating its caregiver, having the caregiver imitating can be just as important for language learning. When the caregiver imitates the robot it gives the robot a possibility to extend its audio-motor map so that it maps utterances from the caregiver to its own articulatory positions. By making use of this during interactions with several different people, the map can continuously become more speaker invariant. This type of interactions, where it is the caregiver that imitates, has been found common also in adult-infant interactions. It is shown that prosodic features can be used in order to detect when there is an imitation.

Chapter 7

Initial word learning

In the previous chapters we have shown how the robot can learn a sensory-motor map and use this to acquire an initial set of articulatory speech units. We believe that this sensory-motor map is important not only for being able to produce speech sounds, but also to recognize those, and that articulatory speech units can be important building blocks for creating word models.

However, infants start to segment words and recognize those even before they are able to produce them. The hypothesis here is that the initial set of words can be acquired without the need for any linguistic structure such as speech units. This follows the ecological and emergent approach to language learning.

In order to create this initial word model the robot looks for recurring acoustic events and associates those to visual objects in its environment. A short term memory (10-20 s length) is used to restrict the search space and increase the possibility that the recurring acoustic patterns that are found refer to the same object. Recurring patterns in the short-term memory are paired with the visual object and send to a long-term memory where they are organized in hierarchical trees. Finally, the mutual information criterion [18] is used to find which words that are consistently used for a certain object.

This approach was first described in CELL [101], Cross-channel Early Lexical Learning. There, an architecture 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 [117], no external database is used. Instead 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 makes 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 are instead seen as emergent consequences imposed by increasing representation needs [92] [74], and will be discussed in the next chapter.

7.1 Detecting visual objects

Starting with the object detector, the robot takes a snapshot of the camera's view and segments 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 transformation of an image to the object representation is illustrated in Figure 7.1.

7.2 Finding recurring events

In order to find recurring patterns, the auditory sound stream is first 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 utterances in the memory in order to find recurring 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 7.2. By averaging over 200 ms we exclude local matches that are too short and can find word candidates by simply looking for minima in the distance matrix. Starting from a minimum 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'

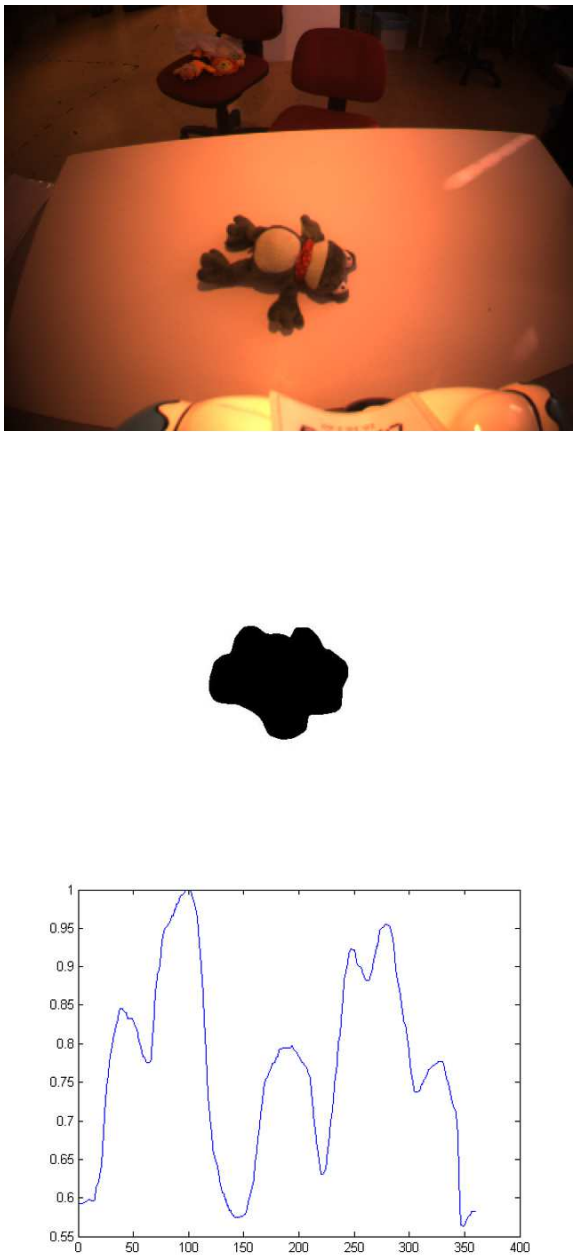

Figure 7.1: Original image (top), silhouette image after background subtraction and morphologic operations (center), and the silhouette perimeter in polar coordinates (bottom).

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, i.e. the pitch. 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 [32]. 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.

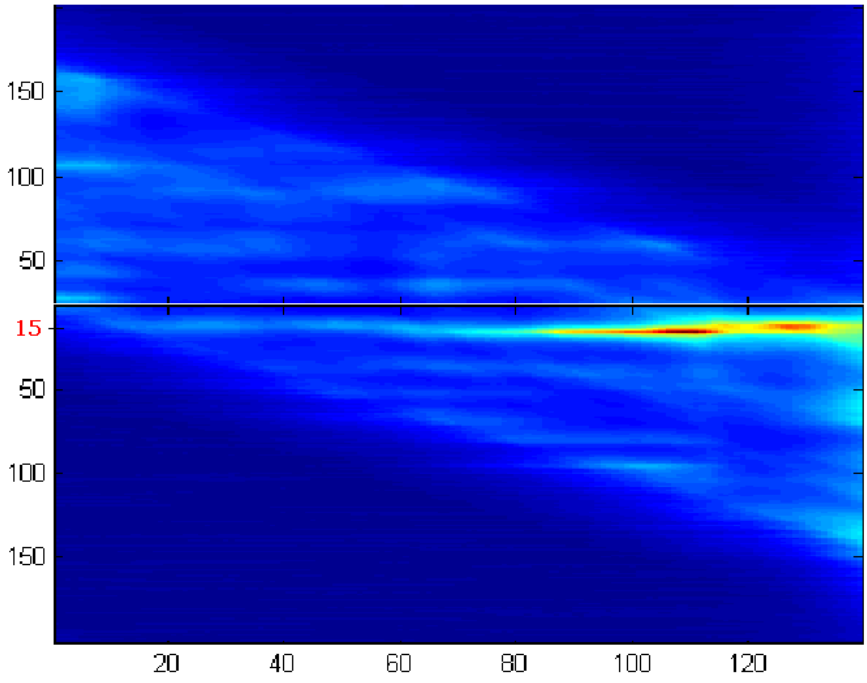

Figure 7.2: Finding word candidates from the utterances "Titta här är den söta Dappan" and "Se på lilla Dappan". The unit for the axis is time (in number of windows). The horizontal axis represents the first utterance, and the vertical axis the number of windows that the second utterance has been slided to the left or right. The color shows how well each window of the two utterances matches each other. The best match is found in the last part of the first utterance, when the second utterance is shifted 15 windows to the right. This match corresponds to the word "Dappan"

The same pattern matching technique 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 matching, 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.

7.3 Hierarchical clustering

When both a word candidate and a visual object are found, their representations are paired and stored in a long term memory. To organize the information we use a hierarchical clustering algorithm [46]. Word candidates and visual objects are organized independently into two different tree clusters. The algorithm starts by creating one cluster for each item. It then iteratively joins the two clusters that have the smallest average distance between their items until only one cluster remains.

While the algorithm is the same for both 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 Warping (DTW) [105] 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.

7.4 Multimodal integration

When we have interconnected multimodal representations, which is the case for the word candidates and visual objects that assumingly refer 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 [18]. In the general form this can be written as

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log \left(\frac{p(x,y)}{p_1(x)p_2(y)} \right) \tag{7.1}$$

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.

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 7.3.

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 was always referred to using the Swedish word "bollen".

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

Figure 7.3: Experimental setup for robot test

acoustically close. The result from the hierarchical clustering of the word candidates and the visual objects can be seen in Figure 7.4. 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 simpler and it is tempting to select the five clusters in the bottom as our objects. However, in this case it is actually the clusters one level up that represents our visual objects.

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 highest 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 highest 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.

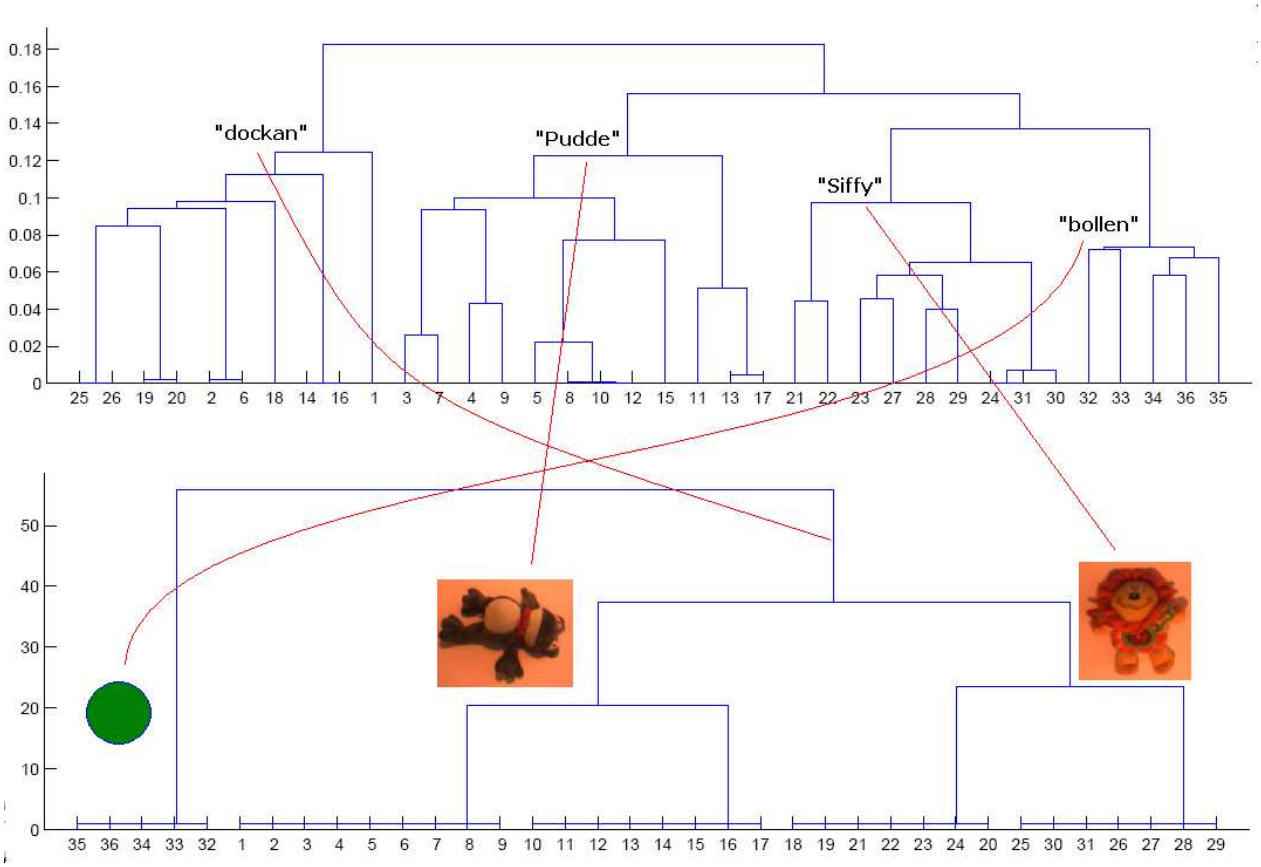

Figure 7.4: Above: Clusters of the extracted word candidates during the robot experiment. Word candidates 1-17 are paired with object Pudde, nr 18-29 with object Siffy, and 32-36 with object bollen. Below: Clusters of the extracted visual objects during the robot experiment. Objects 1-17 corresponds to object Pudde, nr 18-29 to object Siffy, and 32-36 to object bollen.

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 was taken to create 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 cases that an object was present. In 33 of 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.

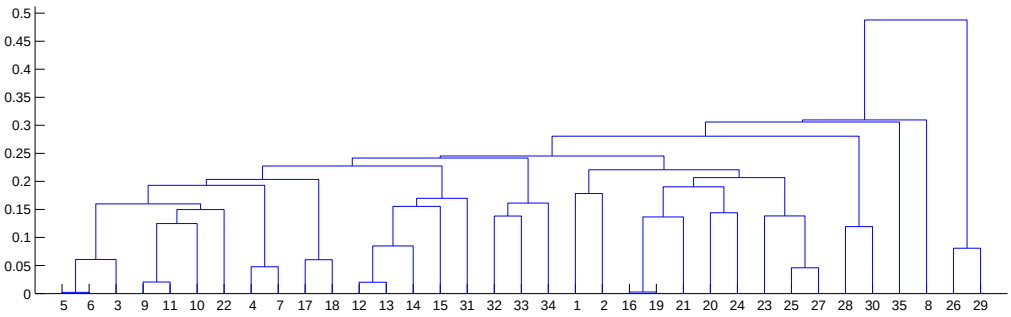

Figure 7.5: Cluster formations from word candidates taken from infant directed speech. Word candidates between 1 and 30 are paired with object Kuckan and word candidates between 31 and 35 are paired with Siffy. Using the mutual information criterion, cluster (32 33 34) gets associated with Siffy and cluster (5 6 3 9 11 10 22 4 7 17 18) gets associated with Kucka.

After searching all utterances for word candidates we cluster all the candidates in the long-term memory. The result can be found in Figure 7.5. 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 resemble 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.

7.5 Conclusions

By making use of some characteristics, typically found in infant directed speech (IDS), the robot is able to learn a set of initial words by looking for recurring patterns in the speech stream and associating those with objects in the visual field.

The proposed method does not make use of any pre-programmed linguistic knowledge such as phoneme-models, and is able to learn words both from recorded adult-infant interactions and during direct interactions in an online experiment.

Chapter 8

Learning statistical models of words and speech units

In the previous chapter we showed how a robot can learn an initial set of words using pattern matching and multimodal learning. The described method works well for creating a small vocabulary from a limited number of demonstrations, but unfortunately it does not scale very well as the vocabulary grows. One of the issues is that the hierarchical trees continuously grow as the number of demonstrations increase, which increases the time it takes to search the trees. Another issue is that the acoustic distance between each word tends to get smaller as the vocabulary grows, making it increasingly difficult to separate between the words. Similar limitations can be found in other related works on pattern based word learning such as [1] [116] [117], which are all limited to vocabularies below 50 words.

Infants, on the other hand, do not show any difficulties in learning larger vocabularies. Instead there is an increase in the word learning rate, often referred to as a vocabulary spurt, which seems to occur between the age of 12-18 months when the infant's vocabulary has typically grown to around 50 words [13] [25]. While it is still disputed whether there exists a specific point in the developmental path where this spurt takes place or if it is more of a gradual change [39], there are some evidence that the infants change the way that they interpret the speech signal at this stage. Using a switch task experiment it has been shown that infants of 14 months of age can learn to associate two dissimilar sounding words, such as "lif" and "neem", to two different objects, but fail on this task when the words are phonetically similar, such as "bih" versus "dih" [110]. Using the same experiment with infants of age 17-20 months showed that they were able correctly associate even similar sounding words with different objects [128]. This increased sensitivity to phonetic details indicates that the infants have started to learn the underlying structure of words, i.e. phonemes or speech units, and make use of this for word recognition.

These observations fit well with the ecological theory of language acquisition, where the linguistic structure is seen as something that emerge from the need to handle the growing vocabulary. By finding a number of speech units that represents the sounds used in the particular language, words can then be expressed as a sequence these speech sounds. In systems for Automatic Speech Recognition (ASR), a predefined phone model is typically used as the speech units, and a large number of examples of each phone is used to create a statistical model that captures the differences in pronunciation between different speakers. A words is then created by using a Hidden Markov Model (HMM) that describes the phone sequence and the transition probabilities between each phone. In this work however, we specifically want to avoid having to provide linguistic structure in the form of a given phone model and instead learn the structure.

In Chapter 6 we showed how different types of interaction games can be used to learn an initial set of speech units. However, this set was mainly limited to vowels and a few stop vowels. While this was partly due to limitations in the vocal tract model, it is still unlikely that a complete set of speech units will emerge from these imitation games. We therefore need an additional method for acquiring speech units that will allow us to create statistical word models. A number of different methods have been proposed in related work on phone acquisition. In [106] speech units are created with a bottom-up approach where similar speech sounds are clustered in an hierarchical tree, and Bayes Information Criterion (BIC) is used to select the proper number of speech units. One of the difficulties with this method is the lack of a good measure for how useful the resulting speech units are in terms of word learning. In [50] [51] the speech units are chosen specifically from the word recognition rate obtained with HMMs based on those speech units. However, the use of a predefined number of speech units and a given training vocabulary make this method less ecological. A more ecological approach is taken in [62] where the number of speech units is increased iteratively and evaluated using a multimodal word learning approach similar to that described in the previous chapter.

Inspired by those methods, we propose a slightly adapted approach of that described in [62], where the speech units are evaluated on the vocabulary learnt using our initial word learning technique [58]. The main drawback with the original approach is the need to store the complete utterances and visual stimuli received up until the point where the speech units are created. While this may work well in a small and controlled experiment it becomes less likely to work in more natural settings, especially considering that infants may develop these models somewhere between the age of 12-18 months. In our approach the complete utterances only need to be stored in a short-term memory for a few seconds until the pattern matching has been done, and only recurring patterns are stored in a long-term memory.

Other differences with the method proposed in this chapter is that we take advantage of the speech units found during the imitation experiments, and that our speech units are defined in both articulatory and auditory space.

In summary, the aim of this chapter is to acquire a set of speech units that can be combined into words, and to express those in a statistical model that is able to capture differences in rhythm and pronunciation. The first section is used to outline the statistical model while the remaining sections focus on how each part of the model is learned.

8.1 Defining the statistical model

The statistical model used in this work is similar to those used in a standard ASR-system, but with essential differences in how the model is learned.

In an ASR-system, the words are typically modeled using Hidden Markov Models (HMM). An HMM consists of a fixed number of hidden states $x^{1..n}$, where each state represents a specific phoneme or speech unit. The states are called hidden since an observer cannot see which of the states that is currently active. Instead, a vector of output parameters is observed y. Each state has different probability distributions over the output parameters $p(y|x)$. In ASR-systems, the output vector typically consists of MFCC, and the conditional probability distribution over the output vector is modeled with a Gaussian or mixture of Gaussians. The speech unit model is illustrated in Figure 8.1.

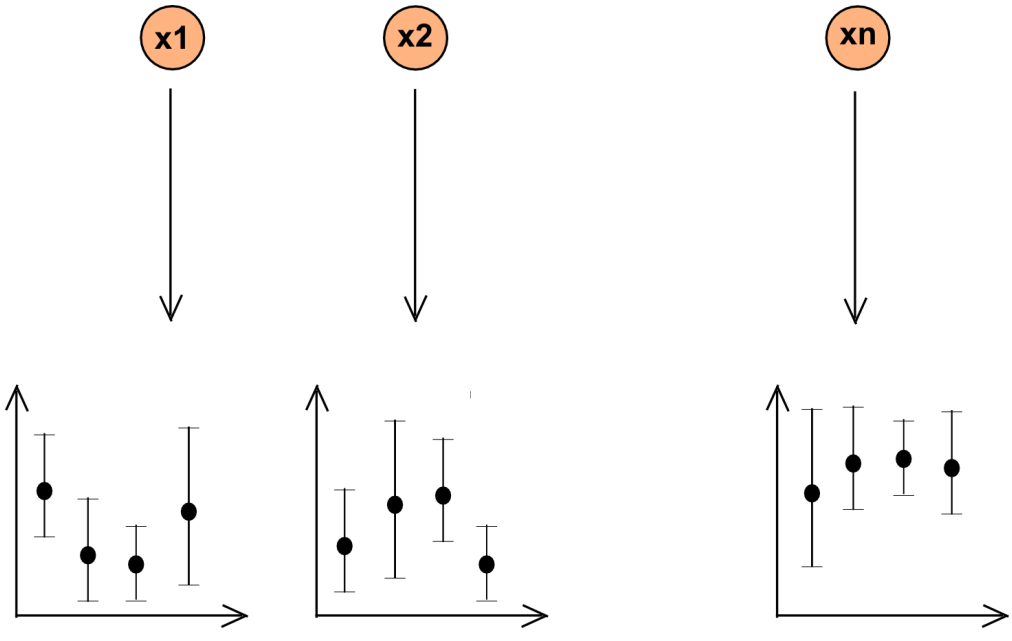

Figure 8.1: Statistical model of the speech units

Compared with the traditional methods, one of the differences in this work is the addition of articulatory features in the output vector. Another difference is the way we estimate the probability distribution. In an ASR-system those are estimated using a labeled phone database, specifically defined for the language that should be learnt. As explained in the introduction we want to avoid using a predefined phone model and instead use a combination of speech units learned through the interaction with the caregiver and speech units that are learned in an unsupervised manner using clustering techniques.

For a given language, each speech unit has a specific probability to occur during natural speech, and the probability of occurring at a specific time depend on which speech unit that was used before. In an HMM it is assumed that the probability of being in state i at time t only depend on the state at time $t-1$, and a state transition matrix is used to model the probability of each transition. By connecting each speech unit with all other it is possible to create a general model of the language, also called a phonotactic model, see Figure 8.2.

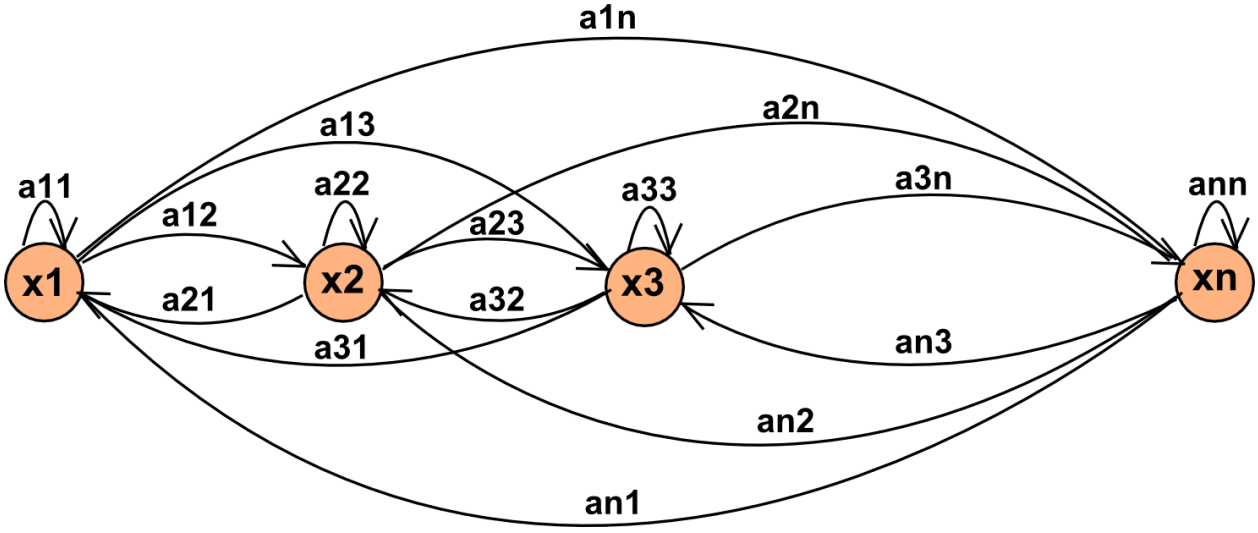

Figure 8.2: Phonotactic HMM based on the speech units

The phonotactic model makes it is possible to calculate the most likely state sequence from a given speech input. We will use this to transform our initial word models into statistical models. By calculating the most likely state sequence for a given word, we can then create a specific HMM for that word. The word models differ from the phonotactic model in that each state is only connected to itself and to the next state in the sequence, see Figure 8.3.

By evaluating the resulting word models we get a measure on how well the chosen number of speech units works. Finding the optimal number of speech units is therefore an iterative process, starting with a low number of speech units and increasing the number as long as the word models improve.

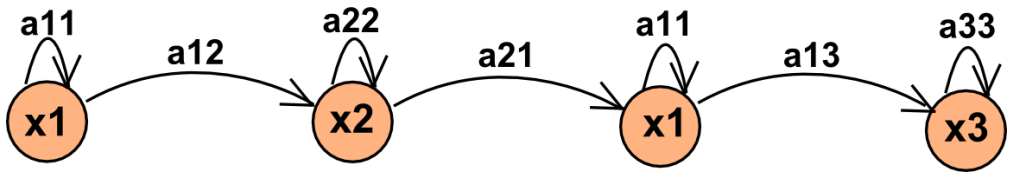

Figure 8.3: Example of statistical word model

The steps involved in creating the statistical models of words and speech units is summarized below. In the following sections we will then explain the learning methods used for each of the steps.

1. Defining the speech units
2. Creating a phonotactic model.
3. Finding the parameters of the phonotactic model.
4. Create word models based on the speech units.
5. Evaluate the word models.

8.2 Defining the speech units

No predefined phone model or database is used to create the statistical model. Instead an arbitrary number of speech units are chosen and the models are trained from unlabeled natural speech in a completely unsupervised way.

For the first iteration a very small number of speech units are chosen. K-means is used to cluster the speech data into the specified number of speech units. In the general case, random speech samples are used to initialize the K-means algorithm. Here we can alternatively make use of our initial speech units, found through the interaction with the caregiver, and use those as starting points for the K-means algorithm. The K-means algorithm then has the following steps:

1. The starting points are used as mean values for each of the clusters.
2. Each speech sample is associated with the closest cluster.
3. The centroid of each of the k clusters becomes the new means.
4. Steps 2 and 3 are repeated until convergence has been reached

The clusters from the K-means algorithm are then used to initialize the statistical model. The probability distribution for the output vector is estimated by calculated the means and covariance for all the samples within each cluster.

Next we want to estimate the transition probabilities for each speech unit. This is done by creating a single HMM where each state, i.e. speech unit, is connected to all other states. This can be seen as a phonotactic language model that shows which transitions are possible, and how common they are.

8.3 Finding the parameters of the phonotactic model

While the initialization provides an initial guess of the parameters for the phonotactic model, the HMM must be trained in order to optimize the parameter values. The phonotactic model is trained using Baum-Welch EM algorithm, and the same speech data as was used by the k-means algorithm for the initialization. The Baum-Welch algorithm not only calculates the transition probabilities, but also reestimates the probability distributions for the output vector. It is an iterative process with two steps.

In the first step, the state probabilities are calculated given the observations and the current emission probabilities for each state. This is done using a forward-backward algorithm. The forward calculation gives the total probability of all paths to a state given the observations, and the backward calculation gives the total probability of all paths from the state to the stop given the observations. The forward and backward probabilities are then multiplied and divided with the total probability of the observations.

Given the probability of being in a state at each time instant, and knowing the observed output vector at that time, it is possible to reestimate the emission probabilities for each state. In the same way the transition probabilities can be reestimated. This gives a new estimate of the parameters of the HMM. However, the state probabilities were calculated using the previous parameters and may now have changed. It is therefore necessary to repeat those steps until the algorithm converges.

8.4 Modeling words

From the phonotactic HMM and a given speech utterance, it is possible to calculate the most likely sequence of speech units that resulted in the given utterance. This is typically done using the Viterbi algorithm. The Viterbi algorithm is related to the forward calculation in Baum-Welch, but takes the maximum probability of a single path going to a state instead of total probability of all paths.

To create statistical models of the words acquired through the initial word learning, we select the word candidate in the center of each cluster and present that to the phonotactic HMM. Using the Viterbi algorithm we get the most likely sequence of speech units for the given utterance. From this sequence we create a new single path HMM, where repeated speech units within the sequence are collapsed into a single state. Transition probabilities are taken from the phonotactic HMM, and are normalized so that they sum to unit.

8.5 Evaluating the model

To evaluate the statistical word models, we make use of the remaining word candidates within each cluster of the initial word models. These are presented to each of the statistical word models and for each word model we calculate the likelihood that the utterance was created with the current model. The utterance is then associated with the statistical word model given the highest likelihood. If this statistical model is indeed the one that represents the cluster from which the utterance was drawn, this is counted as a correct classification. Doing this for all word candidates in the initial word model we get the recognition rate for the current statistical model.

The whole process is then iterated by increasing the number of states and creating new statistical models for as long as the recognition-rate improves.

8.6 Experimental results

The multimodal word learning has been implemented in a humanoid robot. In a previous experiment the robot was able to learn the names of a number of toys that the caregiver placed in front of the robot Figure 7.4. In this experiment we were mainly interested in testing the statistical model. The robot was therefore taught a number of additional words. Like in the previous experiments only full sentences and no single words were given to the robot. However, this time no images were used. Instead the utterances were labeled with a number representing the object. The pattern matching resulted in 88 word candidates that were divided into 8 different clusters by using hierarchical clustering and the mutual information criterion.

For each cluster we then created a statistical model using the method described above. This was done both with and without bootstrapping. With bootstrapping, the vowels learnt by imitation were used as initial guesses for the positions of the speech units. Without bootstrapping, random samples from the speech data was used for initializing the K-means algorithm. We started with only 5 speech units and iteratively increased the number until 12

speech units when there was no longer any improvement in the recognition rate. The results are shown in Figure 8.4.

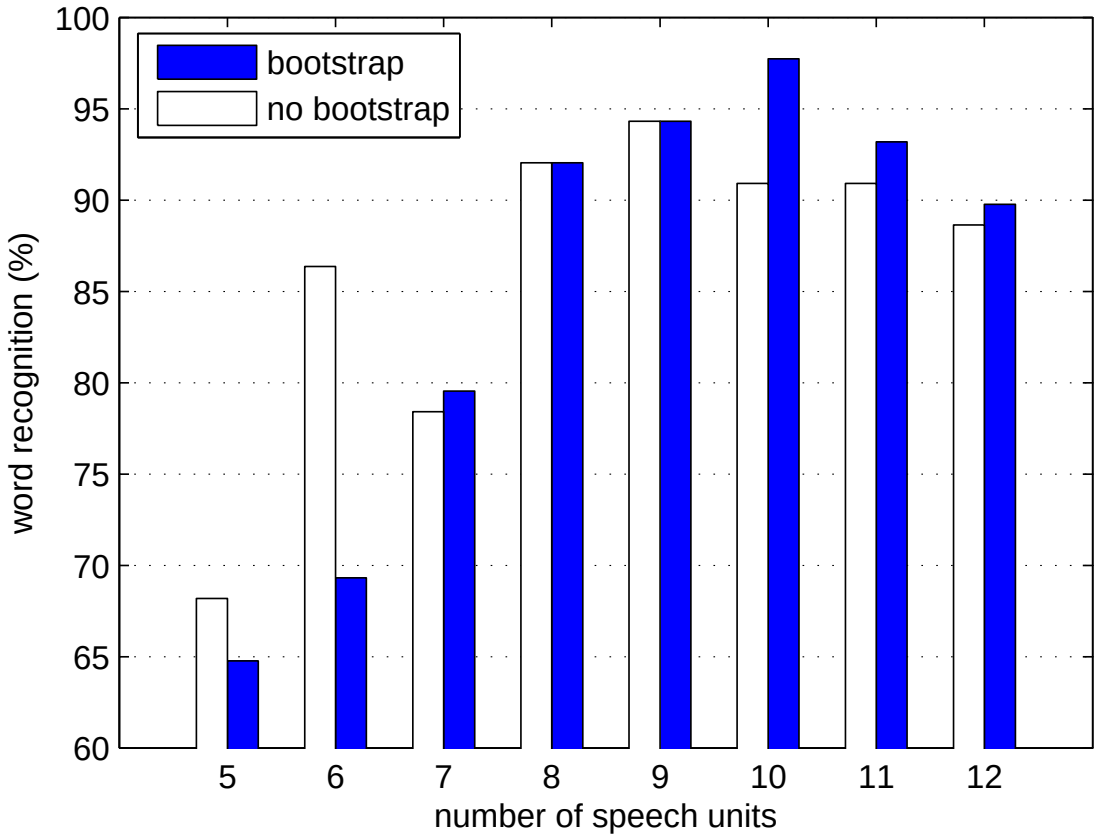

Figure 8.4: Word recognition rates for different number of speech units with 8 target words.

The best result, 98% recognition rate, was obtained when using 10 speech units and bootstrapping. The resulting word models are shown in Table 8.1. Note that some of the speech units have a close to one-to-one relation with real phonemes, such as 1=a and 3=m.

A second, larger experiment was also performed where the robot listened to sentences containing a total of 50 target words. In this case, only 38 clusters with at least five word examples were found. A statistical model was created for each of these clusters. Again this was done both with and without bootstrapping. For this case, the best results was obtained for 17 speech units, Figure 8.5.

8.7 Conclusions

In the previous chapter we showed that it is possible to learn an initial word model without the need for predefined linguistic knowledge, and it was argued that structures such as phonemes

Table 8.1: *Statistical word models for 10 speech units with bootstrapping*

word	representation
siffy	5 7 5 7
pudde	9 6 10 5 8
docka	6 10 6 10 5 8
pappa	6 1 6 10 6 1
mamma	3 1 3 1
lampa	7 9 1 2 3 6 10 6 1
pippi	7 5 10 6 7
vovve	4 2 6 2 6 5 2 9

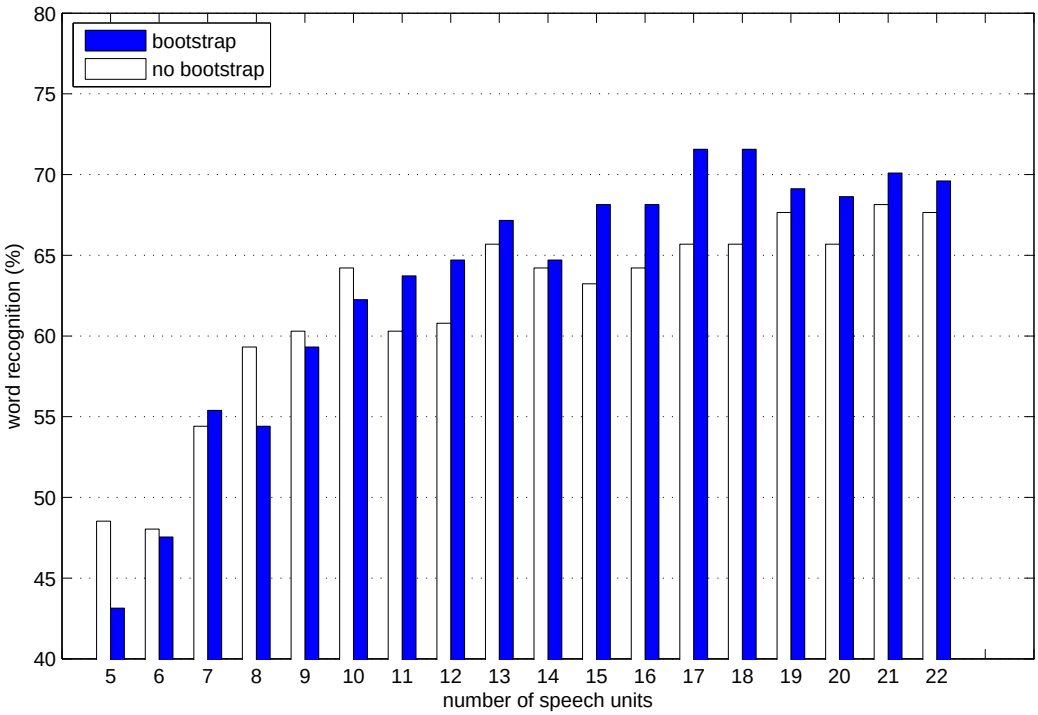

Figure 8.5: Word recognition rates for different number of speech units with 38 target words.

instead emerge when needed in order to handle the growing vocabulary. In this chapter we studied how this structure can emerge and developed a method for creating statistical models of words and speech units.

While related methods often start by trying to cluster speech sounds into suitable units before doing anything else, we take advantage of the fact that these structures actually may emerge later in the development. By delaying the creation of speech units to a stage where we have already been able to acquire an initial vocabulary we can use this vocabulary to

evaluate how useful a set of speech units is for word learning. Another advantage by delaying the creation of speech units is that it allows us to create a useful auditory-motor map, which makes it possible to define the speech units in both auditory and articulatory space. An additional benefit is that the speech units learned during the early imitation games can be used to bootstrap the learning of the statistical models.

The experimental results show that this bootstrap has a positive effect on the word recognition rates obtained with the statistical models.

Chapter 9

Discussion and future work

This thesis proposes a developmental and ecological approach to language acquisition in humanoid robots, where embodiment and human interactions can relax the need for preprogrammed linguistic knowledge and labeled training data.

The work provides models for simulating the human ears, eyes, vocal tract, and memory functions. On top of that, a cognitive model has been developed that provides the necessary learning processes for acquiring words and underlying speech units. The cognitive model makes use of motor learning and simultaneously learns how to produce and recognize speech. The models have been implemented and tested in a humanoid robot. It has been shown that the robot is able to acquire words and speech units directly through the interaction with a caregiver, without the need for preprogrammed linguistic knowledge.

The speech units in this work are defined as target positions in motor space rather than as auditory goals. While this approach is inspired by the Motor Theory of speech perception, unlike the latter we do not consider speech to be special in that it needs any specific circuitry in the brain to do this mapping. With the discovery of mirror neurons, similar mappings has also been found between vision and motor space when recognizing different types of grasps, and for grasping it has been shown that the map can be learnt through the use of motor babbling and that different grasps can be learnt through imitation. In this work, the same approach has therefore been used to learn the sound-motor map. However, it is found that babbling alone is not sufficient to learn the audio-motor map. The main difficulty is to overcome the large differences in the pronunciation between different speakers and in particular differences between the sound produced by the robot and that of the caregiver. In the case of grasping this difference is mainly restricted to a viewpoint transformation, but in the case of speech researchers have not been able to find a direct transformation that can resolve interspeaker differences. To compensate for these differences an additional learning step is necessary in which the caregiver repeats utterances produced by the robot and

allows the robot to generalize the map. Studies of the interactions between infants and their caregivers have shown that this behavior is very common. One difficulty with this approach is that the robot needs to be able to distinguish between imitations and non-imitations so that only actual imitations are used to update the map. It has been shown in this work that comparing prosodic features between the two utterances is sufficient for initial infant-adult interactions. After learning the map, the robot has been able to learn target positions for vowels and stop consonants by imitating the caregiver. Due to limitations in the vocal tract model other speech sounds such as fricatives cannot be reproduced and can therefore not be learnt using the current approach.

In parallel with the speech units, the robot is also able to acquire initial word models. This is done by looking for recurring speech patterns and ground these in visual objects. While the same multimodal approach can be found in earlier works on word learning, the approach taken here is different in that it does not rely on any predefined phonemes. This is an important difference since, according to the ecological approach, initial word learning precedes the stage where infant starts to group speech sounds into language specific speech units. Finding recurring pattern directly from the acoustic signal is very difficult, especially in fluent speech, and the very reason why predefined phoneme models are typically used. However, as the review of some of the typical characteristics found in IDS indicates, target words are typically highlighted using utterance final position and focal stress. Infants are very sensitive to these kinds of cues, and by taking advantage of the same cues the robot can also facilitate the word acquisition task. The experiments with the humanoid robot clearly show that it is possible to acquire a small vocabulary using the proposed method. While no larger tests have been performed, the limitations of the suggested pattern matching approach are well known in ASR, and the robot will eventually have to abandon this approach in favor for statistical learning based on a limited number of speech units.

By creating an initial lexicon based on pattern matching, the robot can then use this lexicon for training and evaluation of the statistical models. An iterative process is used to find the optimal number of speech units for separating between the words in the initial lexicon. Using the initial speech units, found during imitation, to initialize the statistical model can improve the model by decreasing the risk of getting stuck at a local minimum.

To conclude, the presented work demonstrates the feasibility of the ecological and emergent approach to language acquisition. It has been showed that the robot is able to acquire both words and a set of underlying speech units without any ”innate” linguistic knowledge. Further it provides a new and flexible way of language learning in humanoid robots. While the final statistical representation of words and speech units are very similar to that of traditional ASR-systems, no hand labeled data is needed in order to train the model. In addition

to the statistical model, words referring to objects are also represented by the object's visual appearance. Finally the mapping to motor space and use of interactions make it possible for the robot to adapt to different speakers and continuously gain speaker invariance by updating the audio-motor map.

9.1 Future work

There are several possible direction that can be followed in order to improve or extend the current model.

One possible extension would be to try learning not only names of the objects, but also words that describe events, object properties and relationships between objects, by repeating the experiment done in [69]. This would demand a more advanced object detector that include object size, color, direction, and velocity, and that is also able to detect more than one object at a time.

The auditory sensor also needs to be improved with respect to noise filtering. Especially it would be desired to model internal noises caused by the fan and the motors in the robot's head, and to be able to focus the attention to sound coming from a single direction. This can be implemented by template subtraction and beamforming.

Finally, the speech production needs more attention. The current implementation simply create a linear trajectory between target positions but may be made more natural by creating a smoother path (i.e. coarticulate), and by adapting the prosody. It would also be desirable to extend the current synthesizer with noise sources in order to produce fricative sounds.

Bibliography

[1] Aimetti, G., "Modelling early language acquisition skills: towards a general statistical learning mechanism", Proceedings of the EACL 2009 Student Research Workshop, pp 1-9, Athens, Greece, 2 April, 2009

[2] 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

[3] Algazi, V., Duda, R., Morrison, R., Thompson, D., "Structural composition and decomposition of hrtfs", in Proc. IEEE WASPAA01, New Paltz, NY, pp. 103-106., 2001

[4] Algazi, V., Avendano, C., Duda, R., "Estimation of a spherical-head model from anthropometry", J. Aud. Eng. Soc., 2001

[5] Algazi, V., Duda, R., Thompson, D., "The cipic hrtf database," in IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, NY, USA, 2001.

[6] 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

[7] Avendano, C., Duda, R., Algazi, V., "Modeling the contralateral hrtf", in Proc. AES 16th International Conference on Spatial Sound Reproduction, Rovaniemi, Finland, 1999, pp. 313-318.

[8] Batliner, A., Möbius, B., Möhler, G., Schweitzer, A., Nöth, E., "Prosodic models, automatic speech undeerstanding, and speech synthesis: towards the common ground", Eurospeech 2001

[9] 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

[10] Beira, R., Lopes, M., Praga, M., Santos-Victor, J., Bernardino, A., Metta, G., Becchi, F., Saltaren, R., "Design of the robot-cub (icub) head", in IEEE ICRA, 2006

[11] Birkholz, P., Jackèl, D., Kröger, B. J., "Construction and control of a three-dimensional vocal tract model", In Proc. of the International Conference on Acoustics, Speech, and Signal Processing, pp. 873-876, Toulouse, France, 2006

[12] Bishop, C. M., "Mixture Density Networks", Technical Report NCRG/94/004, Department of Computer Science and Applied Mathematics, Aston University, UK, 1994

[13] Bloom, L., "One word at a time: The use of single-word utterances before syntax", The Hague: Mouton, 1973

[14] Bridle, J.S., "'Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition", In: Neurocomputing: Algorithms, Architectures and Applications. Springer, Berlin, pp. 227236, 1990

[15] Burnham, D., "What's new pussycat? On talking to babies and animnals", Science, 296, p 1435, 2002

[16] Chomsky, N., "Rules and representations", Oxford: Basil Blackwell, 1980

[17] CONTACT, Learning and Development of Contextual Action, NEST project 5010, http://eris.liralab.it/contact

[18] Cover, T. M., Thomas, J. A., "Elements of information theory", Wiley, July 2006

[19] Crystal, D. "Non-segmental phonology in language acquisition: A review of the issues", Lingua, 32, 1-45, 1973

[20] Davis, S. B., "Acoustic Characteristics of Normal and Pathological Voices", Haskins Laboratories: Status Report on Speech Research, 54, 133-164, 1978

[21] 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

[22] de Boer, B. (2005)., "Infant directed speech and evolution of language", In Evolutionary Prerequisites for Language, Oxford: Oxford University Press, 2005, pp. 100121

[23] Deutsch, S., Deutsch, A., "Understanding the Nervous System - an engineering perspective", IEEE Press, ISBN 0-87942-296-3, 1993

[24] Duda, R., Avendano, C., Algazi, V., "An adaptable ellipsoidal head model for the interaural time difference,", in Proc. ICASSP, 1999

[25] Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., and Plunkett, K., "Rethinking innateness: A connectionist perspective on development.", Cambridge, MA: MIT Press., 1996

[26] Engwall, O., "Modeling of the vocal tract in three dimensions", In EUROSPEECH'99, pp. 113-116, Budapest, Hungary, 1999

[27] 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

[28] Fant, G., Liljencrants, J., Lin, Q., "A four-parameter model of glottal flow", STL-QPSR 4, pp. 1-13, 1985

[29] Ferguson, C. A., "Baby talk in six languages", American Anthropologist, 66, pp 103-114, 1964

[30] Fernald, Al, "The perceptual and affective salience of mothers' speech to infants", In The origins and growth of communication, Norwood, N.J, Ablex., 1984

[31] Fernald, A., "Four-month-old infants prefer to listen to Motherese", Infant Behavior and Development, 8, pp 181-195, 1985

[32] Fernald, A., and Mazzie, C., "Prosody and focus in speech to infants and adults", Developmental Psychology, 27, pp. 209-221, 1991

[33] Fletcher, R., Practical Methods of Optimization, 2nd ed. Chichester, 1987

[34] Fitzgibbon, A., Pilu, M., and Risher, R. B., "Direct least square fitting of ellipses", Tern Analysis and Machine Intelligence, 21., 1999

[35] Fitzpatrick, P., Varchavskaia, P., Breazeal, C., "Characterizing and processing robotdirected speech", In Proceedings of the International IEEE/RSJ Conference on Humanoid Robotics, 2001

[36] 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

[37] Galantucci, B., Fowler, C. A., Turvey, M. T., "The motor theory of speech perception reviewed", Psychon Bull Rev. 2006 June; 13(3): 361-377.

[38] Gallese, V. and Fadiga, L. and Fogassi, L. and Rizzolatti, G. "Action Recognition in the Premotor Cortex", Brain, 199:593-609, 1996

[39] Ganger, J., and Brent, M., R., "Reexamining the vocabulary spurt", Developmental Psychology, Vol 40., No. 4, pp 621-632, 2004

[40] Gardner, B., Martin, K., "Hrtf measurements of a kemar dummy-head microphone", MIT Media Lab Perceptual Computing, Tech. Rep. 280, May 1994.

[41] Grimaldi, M., Gili Fivela B., Sigona F., Tavella M., Fitzpatrick P., Metta G., Craighero L., Fadiga L., Sandini G., "New Technologies for Simultaneous Acquisition of Speech Articulatory Data: Ultrasound, 3D Articulograph and Electroglottograph", in Cristina Delogu, Mauro Falcone (eds.), LangTech 2008, Fondazione Ugo Bordoni, Roma, pp 81-85, 2008

[42] Guentchev, K., Weng, J., "Learning based three dimensional sound localization using a compact non-coplanar array of microphones", in AAAI Spring Symp. on Int. Env., Stanford CA, March 1998

[43] 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

[44] 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

[45] Harris, C. and Stephens, M., "A Combined Corner and Edge Detector", Proceedings of the 4th Alvey Vision Conference, pp 147-151, 1988

[46] Hastie, T., "The elements of statistical learning data mining inference and prediction", Springer, 2001

[47] 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

[48] Hill, D., Manzara, L., Schock, C., "Real-time articulatory speech-synthesis-by-rules", in the Proceedings of AVIOS '95, the 14th Annual International Voice Technologies Applications Conference of the American Voice I/O Society, pp 27-44, San Jose, 1995

[49] Hirsh-Pasek, K., "Doggerel: motherese in a new context", Journal of Child Language, 9, pp. 229-237, 1982

[50] Holter, T., Svendsen, T., "A comparison of lexicon-building methods for subword-based speech recognisers", TENCON '96, Proceedings 1996 IEEE TENCON. Digital Signal Processing Applications, Vol 1, pp. 102-106, 1996

[51] Holter, T., Svendsen, T., "Combined optimasation of baseforms and model parameters in speech recognition based on acoustic subword units", IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 199-206, 1997

[52] 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

[53] 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

[54] 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

[55] 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.

[56] Hörnstein, J., Gustavsson, L., Lacerda, F., Santos-Victor, J., Multimodal Word Learning from Infant Directed Speech, IEEE/RSJ International Conference on Intelligent Robotic Systems, St. Louis, USA, October 2009.

[57] Hörnstein, J., Gustavsson, L., Santos-Victor, J., Lacerda, F., "Multimodal Language Acquisition based on Motor Learning and Interaction", book chapter in From Motor Learning to Interaction Learning in Robots, ed. Sigaud, O., Peters, J., Springer, January 2010

[58] Hörnstein, J., and Santos-Victor, J., "Learning words and speech units through natural interactions", Interspeech 2010, Tokyo, Japan, September 2010

[59] Hörnstein, J., jArticulator - an open source articulatory speech synthesizer, Available from: http://sourceforge.net/projects/jarticulator/

[60] Huopaniemi, J., Savioja, L., Takala, T., "Diva virtual audio reality system", in Proc. Int. Conf. Auditory Display (ICAD-96), Palo Alto, California, pp. 111-116., Nov 1996

[61] Hwang, S., Park, Y., Park, Y. S., "Sound source localization using hrtf database", in ICCAS2005, Gyenggi-Do, Korea, June 2005.

[62] Iwahashi, N., "Language acquisition through a human-robot interface by combining speech, visual, and behavioral information", Information Sciences 153, pp 109-121, 2003

[63] Jin, C., Corderoy, A., Carlile, S., van Schaik, A., "Contrasting monaural and inteaural spectral cues for human sound localization", J. Acoust. Soc. Am., vol. 6, no. 115, pp. 3124-3141, June 2004

[64] 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

[65] Juaczyk, P., Aslin, R. N., "Infants' Detection of the Sound Patterns of Words in Fluent Speech", Cognitive psychology 29, pp 1-23, 1995

[66] 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. 18461851

[67] Kass,M., Witkin, A., and Terzopoulus, D., "Snakes: Active contour models", International Journal of Computer Vision., 1987

[68] Krstulovic, S., "LPC modeling with speech production constraints", in proc. 5th speech production seminar, 2000

[69] Krunic, V., Salvi, G., Bernardino, A., Montesano, L., Santos-Victor, J., "Associating word descriptions to learned manipulation task models", IROS-2008 WORKSHOP on Grasp and Task Learning by Imitation, Nice, France, September 2008

[70] 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

[71] 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

[72] 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

[73] Lacerda, F., Klintfors, E., Gustavsson, L., Lagerkvist, L., Marklund, E., Sundberg, U., "Ecological Theory of Language Acquisition", In Genova: Epirob 2004, 2004

[74] Lacerda, F., "Phonology: An emergent consequence of memory constraints and sonsory input", Reading and Writing: An Interdisciplinary Journal, 16, pp 41-59, 2003

[75] Lenneberg, E., Biological Foundations of Language, New York: Wiley., 1967

[76] Liberman, A. and Mattingly, I., "The motor theory of speech perception revisited", Cognition, 21:1-36, 1985

[77] 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

[78] Lienhart, R. and Maydt, J. , "An extended set of haar-like features for rapid object detection", IEEE ICIP, 2002, pp. 900903

[79] Liljencrants, J. and Fant, G., "Computer program for VT-resonance frequency calculations", STL-QPSR, pp. 15-20, 1975

[80] Lopes, M. C., Santos-Victor, J., "A Developmental Roadmap for Learning by Imitation in Robots", IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics, Vol.37, No.2, April 2007.

[81] Lopez-Poveda, E. A., Meddis, R., "Sound diffraction and reflections in the human concha", J. Acoust. Soc. Amer., vol. 100, pp. 3248-3259, 1996.

[82] Lucas, B. D. and Kanade, T., "An Iterative Image Registration Technique with an Application to Stereo Vision", Proceedings of Imaging Understanding Workshop, pp 121-130, 1981

[83] 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

[84] Mattys, S. L., Jusczyk, P. W., "Phonotactic and Prosodic Effects on Word Segmentation in Infants", Cognitive Psychology 38, pp 465-494, 1999

[85] Mehler, J., Jusczyk, P.W., Lambertz, G., Halsted, N., Bertoncini, J., Amiel-Tison, C., "A percursor of language acquisition in young infants", Cognition 29, pp 144-178, 1988

[86] Moore, B. C. J., "Interference effects and phase sensitivity in hearing", Philosophical Transactions: Mathematical, Physical and Engineering Sciences, vol. 360, no. 1794, pp. 833-858, May 2002

[87] Moore, R. K., "PRESENCE: A Human-Inspired Architecture for Speech-Based Human-Machine Interaction", IEEE Transactions on Computers, vol. 56, no. 9, September 2007

[88] 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

[89] Nakamura, M. and Sawada, H., "Talking Robot and the Analysis of Autonomous Voice Acquisition" in proc. International Conference on Intelligent Robots and Systems, Beijing, China, pp 4684-4689, October 2006

[90] Natale, S. G. L., Metta, G., "Development of auditory-evoked reflexes: Visuo-acoustic cues integration in a binocular head", Robotica and Autonomous Systems, vol. 39, pp. 87-106, 2002

[91] Norberg, R. A., "Skull asymmetry, ear structure and function, and auditory localization in tengmalms owl, aegolius funereus (linne)", Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, vol. 282, no. 991, pp. 325-410, Mar 1978

[92] Nowak, M. A., Plotkin, J. B., Jansen, V. A. A., "The evolution of syntactic communication", Nature, 404, pp 495-498, 2000

[93] Okuno, H., "Human-robot interaction through real-time auditory and visual multipletalker tracking", in IROS, 2001.

[94] Okuno, H., Nakadai, Kl, Kitano, H., "Social interaction of humanoid robot based on audio-visual tracking", in Proc. of 18th Intern. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE-2002), June 2002.

[95] Park, A. S., "Unsupervised Pattern Discovery in Speech", IEEE Transactions on audio, speech, and language processing, vol 16, no 1, January 2008

[96] Ploner-Bernard, H., "Speech Synthesis by Articulatory Models", Tech. rep., Graz University of Technology, 2003

[97] Pitz, M., Ney, H., "Vocal tract normalization equals linear transformation in cepstral space", IEEE Trans. Speech and Audio Processing, 13, 5, pp 930944, 2005

[98] Ramirez, S. G. R. M. A., "Extracting and modeling approximated pinnarelated transfer functions from hrtf data", in Proceedings of ICAD 05- Eleventh Meeting of the International Conference on Auditory Display, Limerick, Ireland, July 2005.

[99] Raykar, V. C., Duraiswami, R., Davis, L., Yegnanarayana, B., "Extracting significant features from the hrtf", in Proceedings of the 2003 International Conference on Auditory Display, Boston, MA, USA, July 2003

[100] Raykar, V. C., Duraiswami, R., "Extracting the frequencies of the pinna spectral notches in measured head related impulse responses", J. Acoust. Soc. Am., vol. 118, no. 1, pp. 364-374, July 2005

[101] Roy, D. and Pentland, A., "Learning words from sights and sounds: A computational model", Cognitive Science, 2002, vol 26, pp 113-146

[102] Ruesch, J., Lopes, M., Bernardino, A., Hörnstein, J., Santos-Victor, J., Pfeifer, R., "Multimodal Saliency-Based Bottom-Up Attention A Framework for the Humanoid Robot iCub", VisLab-TR 2/2008, 2008 IEEE International Conference on Robotics and Automation Pasadena, CA, USA, May 19-23, 2008.

[103] Saffran, J. R., Aslin, R. N., Newport, E. L., "Statistical Learning by 8-Months-Old Infants", Science, vol 274, December 1996

[104] 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

[105] 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

[106] Salvi, G., "Ecological language acquisition via incremental model-based clustering", In INTERSPEECH-2005, 1181-1184.

[107] Savioja, L., Houpaniemi, J., Huotilainen, T., Takala, T., "Real-time virtual audio reality", in Proceedings of ICMC, pp. 107-110., 1986

[108] Schmid, P., Cole, R., Fanty, M., "Automatically Generated Word Pronunciations from Phoneme Classifier Output", Proceedings of ICASSP, Minneapolis, MN, April 1993.

[109] Slud, E., Stone, M., Smith, P., Goldstein, M., "Principal Component Representation of the Two-Dimensional Coronal Tongue Surface", Phonetica, 59, pp 108-133, 2002

[110] Stager, C. L., and Werker, J. F., "Infants listen for more phonetic detail in speech perception than in word-learning tasks", Nature, 388, pp. 381382, 1997

[111] 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.

[112] Sturim, D., Silverman, H., "Tracking multiple talkers using microphone-array measurements", Acoustics, Speech, and Signal Processing, ICASSP-97, 1997

[113] Sundberg, U, and Lacerda, F., "Voice onset time in speech to infants and adults", Phonetica, 56, pp 186-199, 1999

[114] Sundberg, U., "Mother tongue Phonetic aspects of infant-directed speech", Department of Linguistics, Stockholm University, 1998

[115] Talkin, D., "A Robust Algorithm for Pitch Tracking (RAPT)", in Speech Coding & Synthesis, W B Kleijn, K K Paliwal eds, Elsevier ISBN 0444821694, 1995

[116] ten Bosch, L., and Cranen., B., "A Computational Model for Unsupervised Word Discovery", INTERSPEECH 2007, pp. 1481-1484., 2007

[117] ten Bosch, L., Van hamme, H., Boves, L., "A computational model of language acquisition: focus on word discovery", In Interspeech 2008, Brisbane, 2008

[118] Tian, Y. L., Kanade, K., Cohn, J. F., "Multi-state based facial feature tracking and detection", technical report, Robotics Institute, Carnegie Mellon University, 1999

[119] 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

[120] Valin, J., Michaud, F., Rouat, J., Létourneau, D., "Robust sound source localization using a microphone array on a mobile robot", in Proceedings International Conference on Intelligent Robots and Systems, 2003.

[121] Vernon, D., "An Optical Device for Computation of Binocular Stereo Disparity with a Single Static Camera", Proceedings OPTO-Ireland 2002, SPIE Vol. 4877, pp. 38-46, 2002

[122] Vihman, M and McCune, L., "When is a word a word?", Journal of Child Language, 21, 1994, pp. 517-42

[123] Vihman, M. M., Phonological development, Blackwell:Oxford., 1996

[124] Viola, P. and Jones, M. J., "Rapid object detection using a boosted cascade of simple features", IEEE CVPR., 2001

[125] Vogt, F., Guenther, O., Hannam, A., Doel, K., Lloyd, J., Vilhan, L., Chander, R., Lam, J., Wilson, C., Tait, K., Derrick, D., Wilson, I., Jaeger, C., Gick, B., Vatikiotis-Bateson, E., and Fels, S., "ArtiSynth Designing a modular 3d articulatory speech Synthesizer", Journal of the Acoustical Society of America, 117(4):2542, May 2005.

[126] Webster, A. G., "Acoustical impedance, and the theory of horns and of the phonograph", Proc. Natl. Acad. Sci. USA, 5, pp. 275282, 1919; reprinted in J. Audio Engineering Soc., 25, pp. 2428, 1977

[127] Welch, P., "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms", IEEE Trans. Audio Electroacoust., vol. AU-15, pp. 70-73, June 1967

[128] Werker, J. F., Fennell, C. E. T., Corcoran, K. M., and Stager, C. L., "Infants ability to learn phonetically similar words: Effects of age and vocabulary size", Infancy, 3, pp. 130, 2002

[129] 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

[130] Zwiers, M., Opstal, A. V., Cruysberg, J., "A spatial hearing deficit in early-blind humans", JNeurosci, vol. 21, 2001.