INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

Recognizing Speech with Anthropomorphic Models For Voice Synthesis

Application to Humanoid Robotics

Cláudia Alexandra Magalhães Soares

Dissertação para o grau de Mestre em

Engenharia Electrotécnica e de Computadores

Júri

Presidente: Doutor Francisco Garcia

Orientador: Doutor Alexandre José Malheiro Bernardino

Vogais: Doutor Jorge Salvador Marques

Doutor José Santos-Victor

11/2007

Acknowledgements

Once again I want to thank my supervisor, Prof. Alexandre Bernardino, for his support and enthusiasm, for his scientific insight, for his ability to pursue new paths, and, last, but not least, for his taste for the genuine. I would also like to thank my co-supervisor, Prof. José Santos-Victor, for his constantly critical view of my work and results. His comments were absolutely fundamental for the theoretical strength and formal structure present in the arguments and claims in this thesis. I wouldn't forget Prof. Rodrigo Ventura's insights, comments and revisions on my ideas and papers. A big thanks to everybody at ISR, and specially at vislab, for their partership, and for putting up with the recordings.

Abstract

In order to emulate in robots the speech production and learning capabilities of human infants, exploratory strategies in articulatory synthesizers have been proposed for the creation of acoustic to motor associations. However, commonly used articulatory speech synthesis models are based on an unconstrained modeling of the physiology of the human vocal tract which contain many redundant parameters. In this thesis we show that vocalic speech requires, in fact, a very reduced number of parameters and, based on well-established linguistic knowledge, propose a two-dimensional articulatory space for VTCalcs articulatory synthesizer.The proposed space is generated through the convex combination of prototype vowels representing vocal tract extremal configurations. We also propose a speech learning architecture that can integrate unsupervised classification for vocalic speech sounds of a given language. This provides a low-dimensional and intuitive vowel production space, suited for automatic production, recognition and learning of speech in articulatory models.

Keywords

Speech Perception, Dimensionality Reduction, Hierarchical Clustering

Resumo

Para recriar em robots as capacidades de produção e aprendizagem de fala das crianças têem sido propostas estratégias esploratórias, com base em sintetizadores articulatórios com o objectivo de mapear a representação acústica na representação motora da fala. Contudo, os modelos de síntese de fala commumente utilizados não colocam restrições fisiológicas importantes às configurações motoras que efectuam, quando, na verdade, os parâmetros do tracto vocal contêm redundância. Nesta tese mostra-se que, de facto, as vogais orais são representáveis por um reduzido número de parâmetros no espaço articulatório, propondo-se ainda, com base em conhecimento linguístico bem estabelecido, um espaço articulatório bi-dimensional para o sintetizador articulatório *VTCalcs*. Este espaço articulatório é gerado pela combinação convexa de vogais prototípicas que representam configuraç ões extremas do tracto vocal humano. Propõe-se ainda uma arquitectura de aprendizagem de fala que integra classificação não supervisionada para sons de fala vocálicos de uma língua.

Palavras-Chave

Percepção de fala, Redução de dimensionalidade, agrupamento hierárquico.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis is inscribed as an epigenetic humanoid robotics perspective. Developmental robotics aims at studying how knowledge on human cognitive development can be exploited to allow robot to learn and adapt continuously to its morphology and environment [1]. The development of speech production involves the exploration of the vocal tract capabilities during the infant's early developmental stages. Also for speech perception development, the vocal tract's articulatory information may be of fundamental importance. So, one should understand the human cognitive functions associated with speech. In this thesis we exploit the vocal tract's articulatory structure and linguistic related knowledge to propose realistic constraints on speech production ""degrees-of-freedom"" what can facilitate the early stages of learning in a robotic system.

In order to better understand the human basic processes that underlie the speech production and perception functions, a short overview in speech related neuroscience and robotic implementations is presented. The approach of this thesis concludes this Chapter, along with the description of this thesis' structure and the brief statement of it's contributions.

1.1 Hypotheses from Neuroscience and Psychology

Speech is an intrinsically human characteristic. Indeed, it is not deliberately learned, in contrast to reading and writing, and only human beings demonstrate to possess it. It is the process of communicating with others using the voice, modulating and articulating the sound in order to convey meaning. Therefore, it is an interdisciplinary field of study for several areas of knowledge. In order to investigate new and more effective ways to recognize speech and construct a humanoid robot that is able to produce and identify speech sounds, this thesis takes a biologically motivated approach, trying to grasp the fruits of human evolutionary improvements and to better understand the human speech production and recognition systems. So, it calls for different branches of knowledge and its approach is motivated by their relevant results. In the present chapter some theories, experiments and results from the fields of neuroscience and psychology are briefly reviewed.

1.1.1 Neuroscience: Mirror Neurons and Broca's Area

Brief perspective through the human brain: actions and speech

Medical studies conducted in man and animals in the last two centuries have consolidated the assumption that the cerebral cortex can be divided in specialized functional areas. The delimitation of those areas is,

nevertheless, not established beyond doubt. Moreover, neuroscience knows that the division between motor and sensory cortical areas is erroneous, since motor responses can be obtained on stimulation of sensory areas. Having this in mind, this text refers to the terminology that associates functions to brain locations, because it is useful to illustrate functional proximities in the brain.

The observation of the cortical areas of interest to the present work, their functions and connections, can shed some light in the relationships between action and speech, and between production and perception. The areas are roughly illustrated in Figure 1.1, and a greater detail in description can be found in [2].

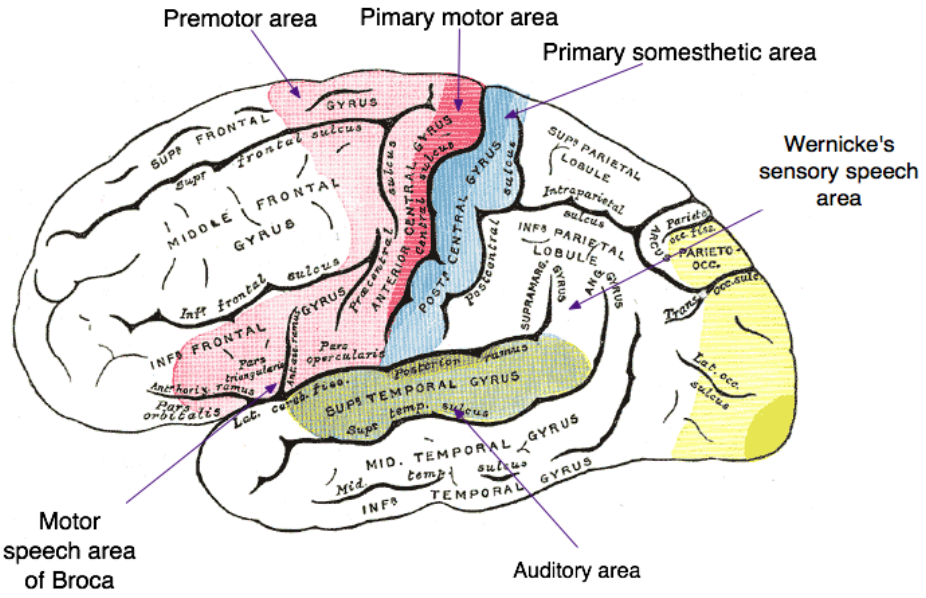


Figure 1.1: Functional localization of the cerebral cortex: lateral view of left hemisphere. (Brain image from [3], which is in the public domain. Text labels added, according to [2]).

The primary motor area carries out individual movements of different parts of the body, but it does not *design* the pattern of movement: it converts the afferent design into execution of movement.

The premotor area stores programs for motor execution that result from past experience and feeds them to the primary motor area. This area is also responsible for the control of coarse postural movements.

The motor speech area of Broca is responsible for the generation of speech sounds, by its neural connexions with the nearby primary motor area; as a result, the phonatory system is activated in order to produce speech. Broca's area is important only on the dominant brain hemisphere (the left one, for right-handed people).

The primary somesthetic area receives stimuli from body sensors, and the amount of cortex assigned to a part of the body is proportional to its functional importance and not to its size.

The auditory area has two parts; in the primary auditory area the sound frequencies are represented tonotopically[1]. The secondary auditory area is needed for interpreting general sounds.

The sensory speech area of Wernicke is connected to Broca's area by an aggregate of nerve fibers. It receives inputs from the visual cortex and from the auditory cortex. Wernicke's area allows us to understand the written and spoken language.

[1] Tonotopic: spatially organized by tone or frequency.

Mirror neurons: perceiving and performing

Experiments in macaque monkeys presented in [4] revealed a type of neurons, called *mirror neurons*, which discharged when the monkey performed active goal directed movements and also when the animal observed meaningful hand movements made by an experimenter.

All movements mimicking an action, but lacking the object of that action had a minimal effect, if any, in the discharge pattern of the mirror neurons. In contrast, the most prominent neural activity was in response to the experimenter's hand or mouth interaction with objects. The majority of the identified mirror neurons was selective to one single action. They discharge only when the monkey is presented (or performing) a specific action like grasping, manipulating or holding.

The authors of these experiments discuss possible functional roles for these neurons. They can code representations of actions and movements or even associate specific motor center codings with the *meaning* of the action, its goal. This facilitates, or even enables, prediction of consequences for one's own actions, as well as recognition of another subjects' movement, with the same representation for both cases.

These neural cells belong to the F5 brain area in monkeys, corresponding to Broca's area in humans. Evidence from medical imaging and clinical experiences in humans agree substantially with the existence of a *mirror system* in humans. The findings also suggest that Broca's area is not exclusively a speech area. It is suggested to have a hand movement representation, performing also action recognition.

The human action recognition system and Broca's Area

Several neuroscience studies were conducted on this subject since the mirror neuron system was first described, in order to better understand the physiology of action recognition and the connection between speech and manipulation. With the development of non invasive and non health-threatening functional brain imaging techniques like functional Magnetic Resonance Imaging (fMRI), it became possible to study brain activity during speech and language production and comprehension. In [5], using fMRI, the activation of both the language production system and the action recognition system showed an overlap, in accord with the hypothesis of a common functional architecture residing in Broca's area.

1.1.2 Psychology: Motor Theory of Speech Perception

The motor theory of speech perception, reviewed in [6], claims that speech perception occurs in a biologically specialized system or module that detects the intended articulatory gestures of the speaker. These intended gestures would be the basis for the phonetic categories that, in each language, define a set of speech sounds that can be classified as a phoneme. This perception module does a transparent conversion from acoustic features to articulatory gestures, delivering to the upper module the articulatory gestures and not an acoustic perception. Another claim of this theory is that speech perception and production are a single mode and its link is formed by motor invariants, gestural commands, that are elementary events of both production and perception. This module is believed to be in competition with other modules for the sound stimulus.

The theory argues that the objects of speech perception are the intended phonetic gestures, represented in the brain as invariant motor commands that call for movements of the articulators, targeting certain linguistically significant configurations. The authors point out that the traditional phonetic notions such as tongue backing, lip rounding and jaw raising have expression in the physical reality as gestural commands. These gestural commands are elementary events of speech production and groupings of them are phonetic segments. So, to perceive an utterance is to perceive a specific pattern of intended gestures. Those gestures

are not directly represented in the acoustic signal nor even in the observable articulatory movements due to different types of phenomena, such as noise, channel distortion or coarticulation, for example. The listener has to extract these segmented phonetic categories from a signal that is continuous and unsegmented by nature: the percept is segmented in a way that the signal is not, showing considerable overlap of phonetic information in the acoustic pattern.

Also, there is a great number of auditory cues for each speech percept, and each one of them is more or less sufficient to determine the underlying phonetic category. Nevertheless, none of those cues is truly necessary, because the absence of one can be compensated by others. To worsen the auditory theories' problem, these cues are subject to extensive variation due to context. So, we cannot define a phonetic category simply in acoustic terms; we should rather use the sound signal as a source of information about the gestures that properly define the category.

But the sound signal is a source of information not only for the speech mode. If the listener is not directed to perceive speech, he will apply other mode — or modes — in order to deliver a perception of some kind to upper cognitive levels. For instance, this acoustic signal can be perceived as a chirp and permit the spatial localization of the emitting subject.

Of course, the speech perception mode gathers information from other sensorial inputs, namely vision. The McGurk effect is a perceptual phenomenon that demonstrates that vision has its role in speech perception. It is experienced while watching a mouth articulation that does not correspond to the heard syllable. It is not an illusion, because it still has effect when the listener is aware of the discrepancy between the seen and the heard.

This theory proposes that the link between speech perception and production is not a learned association. Rather, it is innately specified, requiring epigenetic development. This claim arises from considering speech perception as a *module*, that is, a portion of neural architecture specialized in the necessary computations on the input, outputting representations of objects or events that belong to an ecologically significant class for the organism.

1.2 Computational and robotic implementations

In this Section we describe the main computational models related to speech and manipulatory gesture production and recognition invoking the use of motor information in the execution of predominantly perceptual tasks.

1.2.1 The DIVA Model

The Directions Into Velocities of Articulators (DIVA) model is a description of speech acquisition and production based on artificial neural networks, feedback and feedforward control. Different neural nets represent mappings which are related to regions of the cerebral cortex and cerebellum, as described in [7]. Its purpose is to facilitate the understanding of

Speech production phenomena like motor equivalence, contextual variability, coarticulation or speaking rate effects;

Child development processes. With the radical anatomical transformation of the vocal tract, articulators and the overall phonatory system, the child must continue the process of speech acquisition;

Auditory feedback role in speech production for normal hearing and hearing-impaired individuals.

This model focuses on the sensorimotor transformations that give rise to the control of motor articulatory movements. For this, an adaptive neural network learns to control a simulated vocal tract. The synthesizer used is the one described in [8]. The architecture of this network is related to brain functional areas, having specific artificial neural subnetworks as maps between different cognitive representations in the brain. The motor and premotor representations are coded in two maps:

The premotor cortex speech sound map corresponds to the premotor cortex and the posterior Broca's area. Its cells are functionally related to mirror neurons. Each of those cells should represent the motor command associated with a frequently used speech sound (phones and syllables).

The motor cortex velocity and position maps which are mental representations of the actual position and velocity of the articulators. These variables are determined from the feedforward and feedback control signals.

The feedback control subsystem acts after the model's production of the learned speech sound. It constructs the *auditory state map* with representations of the acoustic sound derived from the articulatory configuration learned. The *somatosensory state map* codes the proprioception, i.e, the internal state of the body, for the spoken sound. The DIVA model authors hypothesize the existence of *auditory and somatosensory targets* that would encode in an acoustic level a syllable or a phone. This proposed explanation leads to the *auditory and somatosensory maps* that code the difference between the target region and the current state. These errors are then fed back as motor velocities, correcting the position of the articulators.

The feedforward control subsystem is incrementally learned, tuning the projection functions of the premotor cortex and the cerebellum. As the speech sound map is learned, the error in the produced sound will diminish, and the feedback subsystem will be almost disconnected from the overall system. It will only come to stage when a change in the articulatory anatomy occurs, like the development of the infants' vocal tract, for instance.

The DIVA model is both a neuroscience and a computational approach and is primarily concerned in speech production at the phone and syllable level with good results for oral vowels.

1.2.2 Robotic Manipulation: Learning by Imitation

Motivated by the mirror neuron findings in neuroscience and the concept of *affordances*[2] in psychology, Lopes and Santos-Victor propose in [9] a general architecture for action and gesture recognition in a humanoid robotic platform taking advantage of the similarities between robot and human motion. Joining the knowledge of its own body's movements, the objects' affordances and the visual perception of the demonstrator's gesture or action, proved to outperform the traditional visual-only methods for action and gesture recognition.

1.3 An Articulatory Approach to Speech Recognition

The long-term goal pointed out by this work is to give rise to an architecture for speech acquisition, recognition, and production that does not need to be trained or programmed by specialists, but instead is prone to self-learning, with the stimuli available in the environment.

[2] Affordances are actions that an agent can potentially perform in the environment. More specifically, they are properties of an object or system having available *action possibilities*. The affordances suggest how an object may be used.

For now, this work's concern is to define a simpler architecture, dealing with stationary voiced signals — vowels, as produced in the chosen articulatory synthesizer. This architecture is meant to obtain acoustic cues only to generate the motor commands that represent phones, as they are perceived in some language, in this case, the portuguese language. The inclusion of visual cues, as suggested by the motor theory of speech perception discussed above, would improve the system's performance.

Unlike the DIVA approach, this work defines a two-dimensional subspace for oral voiced vowels in the broader six-dimensional space of the articulator coefficients. This dimensionality reduction, inspired by linguistic knowledge and human articulator constraints not present in the synthesizer's definitions, permits a complete mapping of the articulatory space. So, as DIVA computes its sensory motor map locally, by computing the tangent spaces to the synthesis function at some prototypical points, this approach can compute the whole map, in an exploratory fashion, as the system babbles and acquires some specific language. The synthesizer used in both approaches is the same.

A recognition performed in the motor articulatory space is expected to be easier, and less prone to noise and variation, than the traditional purely acoustic recognition, achieving more invariance and robustness, even with no other cues whatsoever.

1.4 Thesis Structure

The present thesis is organized as follows: in Chapter 2 speech mechanisms in humans are investigated, in its physiological and anatomical perspective as well as in the linguistic one. This Chapter's goals are to give an insight on the restrictions on human articulators, introduce the source-filter theory for vowels and, in another hand, to briefly introduce human speech acquisition and the linguistic approach to oral vowels. The Chapter concludes with a discussion on restrictions to the articulatory parameters that describe phonation in oral vowels.

Chapter 3 proposes a two-dimensional model for generating oral vowels in an articulatory synthesizer. Firstly, the principles underlying the synthesizer in use are introduced, The tube model for the vocal tract is briefly stated. Then, the new low-dimensional articulatory space is formally defined. This chapter aims at defining a formal basis for the following Chapters and future work.

In Chapter 4 the artificial system's speech acquisition architecture is presented. It consists in a three-stage process that starts with exploration of the system's own articulatory capabilities — babbling; follows the identification of a particular language's speech sounds and, finally, the recognition of groupings in speech sounds — the identification of specific phonological vowels. In the babbling phase the system learns the mapping between its acoustic productions and the articulatory gestures that originated them, and the model proposed in Chapter 3 is validated.

Classification of vocalic sounds as phonological vowels of a certain language is discussed in Chapter 5. Here a collected dataset of native speaker validated portuguese vowels is classified with unsupervised classification methods, and several issues regarding clustering techniques are discussed. Classification results are also discussed.

Finally, in Chapter 6, some conclusions are drawn and future work is proposed.

Chapter 2

Speech Mechanisms

In humans, speech sounds are produced thanks to the *phonatory system*. In order to understand the constraints that emerge from the architecture of this system and to introduce them in a model for the robot's vocal tract, it is necessary to investigate how and why people produce speech and to determine how far can we go when reducing degrees of freedom in this model.

2.1 Voice and Speech Physiology

Speech is one of the most important communication media for human beings. So, it should be an essential medium of communication between man and machine. Here are presented some important physiological and anatomic facts about speech and human voice production in order to understand the nature and complexity of the processes involved in human oral communication and to prospect ways to better replicate some of them. There are three subsystems involved in speech production, namely,

1. Pulmonary airflow,
2. Larynx and the vocal vibration,
3. Oral and nasal cavities.

There is no human organ uniquely dedicated to speech production, in contrast to other functions like feeding or supplying oxygen to the blood. The phonatory system is, in fact, a secondary use for life maintaining systems, with a little remake by evolution.

These topics will be briefly presented in the subsequent sections, mainly focused in the production of voiced speech sounds. For more information on this subject refer to [10]. The unvoiced and murmured sounds are produced without the vibration of the vocal folds. The source for these sounds can easily be modeled as noise.

2.1.1 Voice: The Sound Source

Voice is defined as the resulting sound wave from vocal fold vibration. It is the medium for spoken communication. Phonation is the set of physical and physiologic processes that lead to a sound vibration at the vocal fold level. There is a less strict definition for *voice*, that describes it as the support for spoken communication, for the expression of emotions and personality.

Pulmonary airflow

The phonation starts by the building up of a pressure gradient between the subglottis and the supraglottis, creating thus aerodynamic energy that will be transformed in acoustic energy by means of the approximating vocal folds. This sound is then filtered by the superior aerial cavities (buccopharyngeal resonators) producing speech sounds.

To produce phonation, the speaker needs to control and activate antagonic muscles in order to set the subglottic pressure at a constant level, which is a difficult control problem. The air pressure must be monitored constantly and it's value must be rapidly increased in milliseconds to about $1.015\ atm$. When the speaker needs to raise the sound's fundamental frequency the vocal folds stiffen and the air pressure must, again, be increased in order to maintain the vibration amplitude.

Larynx vibrating stucture

The larynx tops the tracheal tube and its rigidity keeps the airway open. It is also an effective resonance box for voiced sounds. It supports the vocal folds and supplies the muscular, nervous and cartilaginous support for their adduction and abduction movements. A view of the larynx and the vocal folds from above is depicted in Figure 2.1. They have a common static union point in the larynx and two mobile articulations permitting the

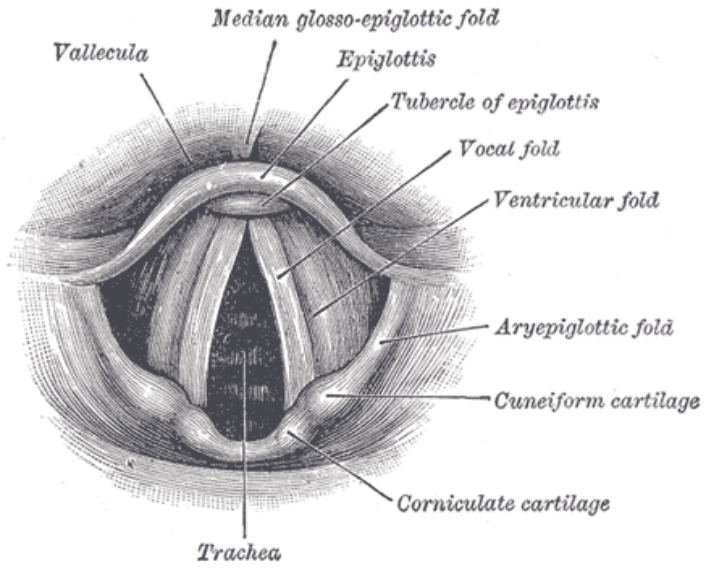


Figure 2.1: Laryngoscopic view of the interior of the larynx, depicting vocal folds and glottis (image from [3], which is in the public domain).

opening and closing of the airway.

Vocal Folds: A Dynamic System

Unlike the rest of the larynx, vocal folds specialized in voice production and their free vibrant edges are covered with a Malpighian epithelium, which is one of the most resistant tissues in the human body, better adapted to pressure and vibration forces. The vocal folds are equipped with shock absorbers to protect them when in contact with the non vibrating rigid structures of the larynx. Although this specialization equips the vocal folds in a way only found in humans, their main function is still the protection of the trachea and lungs from foreign objects.

The vocal fold vibration is different from the vibration of a tensioned string in a guitar. It's movement generates a periodic sequence of air puffs and not a sinusoidal sound vibration. It is also not plucked by fingers, but activated by a transglottic pressure gradient. The vocal fold vibration is a flow-induced oscillation. This oscillation is sustained and generates a sequence of glottal pulses, forming a source signal that will be passed through a filter.

The vocal folds are responsible for the voice's fundamental frequency and, in some languages in Asia and Africa, for tonal distinctions between speech sounds.

2.1.2 Articulators and Resonators: The Sound Filter

The resonators are supra-glottic cavities that filter the voiced stationary sound produced by the vocal folds. The main resonators are depicted in Figure 2.2. The nasal cavity is fixed. It is delimited by bone and cartilage.

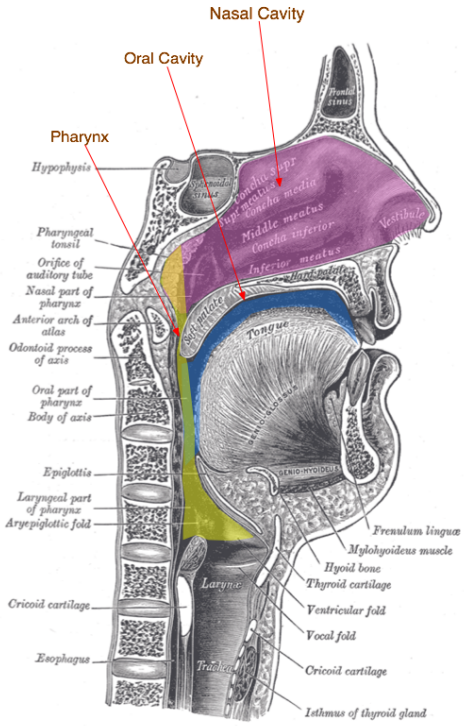


Figure 2.2: Sagittal view of the vocal tract, with resonator cavities delimited. Marked in red, there is the nasal cavity, in blue, the oral cavity and in yellow the pharynx (image from [3], which is in the public domain, labeled and colored by this thesis author).

The pharynx size and shape can be changed by the uvula, the back of the tongue and the pharingeal muscles. The oral cavity is the most changeable of the resonance spaces. It is delimited by the hard palate and the *active articulators* or, simply, *articulators*. The lower pharynx and oral cavity are commonly known as the vocal tract. The articulators can be moved in order to create constrictions and change the vocal tract's shape. The main articulators are:

Jaw: moves in coordination with the lips and tongue; this articulator can determine the openness of the vocal tract;

Lips: can protrude or round for certain positions of the jaw;

Tongue: it is the most mobile and flexible of the articulators. It is so flexible that it is considered as three different

articulators, although they constrain each other: *apex* and *lamina*, the most anterior of the three, the *dorsum* and, most posteriorly, the *radix*. The amplitude and nature of their movements depend also on the degree of openness of the jaw;

Soft palate: it is mainly important for the distinction between nasal and oral vowels. It is a valve that opens so that the flux of air can enter the nasal cavity.

All these articulators have a high degree of coupling and in an unequal way. For instance, it is possible, to some extent, to compensate a blocked jaw by tongue gestures, but the inverse is not possible.

The articulators can cause an obstruction in the airway to produce *consonants*. If this occlusion is complete, the resulting speech sounds are *plosives*. According to [11], in such an occlusion, there is a build up of pressure behind the articulator that is blocking the airway. In the sudden release of pressure that follows, the transitory turbulence creates an explosive sound. When the occlusion is not complete, the consonant is called *continuant*. In this group we can find, among others, the *fricatives* and *liquids*. Fricatives are generated by a sufficient velocity of the airflow through a constriction in the vocal tract as to create friction noise. Liquids have no real obstruction in the airway generating noise. They are defined by a narrowing of the vocal tract, permitting the free flux of air. In lateral consonants, a subset of liquids, an occlusion occurs along the tongue axis, leaving the lateral spaces in the mouth open for air to escape.

Brief notes on the source-filter theory for vowels

The acoustic properties of vowels have been traditionally studied in the context of the source-filter theory. According to [12], the theory considers the source as the quasi-periodic glottal airflow and the filter, the upper-glottic cavities, as mentioned earlier. So, the vocal tract reshapes the frequency envelope of the source, selecting some frequencies of the total spectrum of the source signal — the formants —, in order to radiate them from the mouth. The vocal tract is approximated by a cylindrical tube concatenation and, in most applications, assuming constant air density and sound velocity throughout the airway. This source-filter model is valid for vowel production and the synthesizers' implementations based exclusively on this theory have poor results on consonants.

2.2 Linguistic Knowledge

To understand speech mechanisms we must go beyond the simple consideration of morphology and physiology of structures that enable speech production or perception. We must also consider the linguistic aspects of speech. Linguistics is the scientific study of language. The relevancy of its contribution to this present work is considerable, since it permits access to speech acquisition, speech production and speech classification results that point out dependencies and possible simplifications in the degrees of freedom of the speech recognition problem.

2.2.1 Language vs. Speech

Language is a set of arbitrary symbols, associated with a meaning (the semantics), and a set of rules to combine and manipulate them (the syntax). Under this broad definition it is possible to find programming languages, formal languages or natural languages, among others. Natural languages evolved naturally from the need of general human communication and have native speakers, through whom it is possible to study the several aspects of that language.

Speech is the physical aspect of the oral representation of language. For linguistics, as explained in [13], it is the simultaneous representation of the production and perception systems.

The theories and proposals about language are constructed from natively produced speech and the understanding that each native speaker has of his own language.

Phonetics vs Phonology

The mainstream in linguistics defines two levels in the study of speech units. Those two disciplines are distinguished by the questions that each of them tries to answer:

Phonetics: What sounds are humans capable of producing?

What sounds are used in speech?

What operations are involved in the translation from the linguistic representations to the sound signal?

Phonology: What sounds are distinctive units in a particular language?

What is the specific organization of the sound system in a particular language?

So, phonetics is more concerned with the universal characteristics for all human spoken languages, while phonology is focused in the arbitrary and particular relations between symbols of a language and sounds. A phonological description is particular for each language. This difference is expressed also on the working unit of each level: for phonetics, the *phone* is the smallest discrete unit that a listener can perceive in a continuous sound sequence; the phone is the phonetic specification of the speech sound. For phonology, the *phoneme* is the abstract unit that is phonologically distinctive, i.e., that establishes a contrast in meaning in a minimal pair of sequences.

Linguistic Variability

There are multiple causes for linguistic variability, be it the large scale historical change that makes languages evolve, die or even derive themselves into new languages, or be it simply normal processes of coarticulation, used in order to fluently produce speech. The categories of linguistic variation whom this work is mostly concerned is, naturally, the phonetic and phonological ones. In these categories one can define particular groups:

Contextual variation: Since articulators are in constant motion during speech and the latencies for tongue muscle response to brain command are of approximately 30 ms (from [7]), certain articulatory gestures may not be completed in time or can co-occur with the segment that precedes or follows it. When the specification of a phonological segment is redundant, it can even be replaced by the combination of contiguous gestures. An example of this variation is allophony, the phenomenon that a phoneme has different (context-dependent) variants that occur in complementary distribution; those variants are called allophones. The phoneme identification is done by the *minimal pair test*[1]. In English, two allophones of the phoneme /p/ can be found in the word *paper* [pʰeɪpɚ]. [pʰ] and [p]. In Portuguese, a good example of allophony is the realization of the plural phoneme, /s/, that has 3 possible allophones, for three different contexts. It can be [z] as in *casas amarelas* [kazɐzɐmɐɾɛlɐʃ], [ʃ] as in *casas pequenas* [kazɐʃpɨkenɐʃ], or [ʒ] *casas bonitas* [kazɐʒbunitɐʃ].

[1] Tests if two phones are realizations of the same phoneme. Find two words that sound identical, except that one contains the phone P and the other the phone Q in the same position. If this difference leads to a difference in the meaning between those words, then phones P and Q represent different phonemes.

Regional variation: Certain phonemes have a characteristic realization in a region. The Scottish English rhotic accent contrasts with the traces of [r] in the precedent vowel in England's English. The Portuguese affricate phone [ʧ] for the /ʃ/ phoneme present in the surroundings of the city of Chaves, located in the northern region of Portugal, is another example of regional variation.

Ideossincratic variation: based on sociolects or even individual variants. A rapid speech accentuates coarticulation and creates variation within segments of speech belonging to the same speaker.

But the speech sound itself can vary in many different ways, such as with prosody, or with temporary conditions like obstruction of the nasal cavities, or inflamation of the vocal folds, for instance. The emotional state of the speaker is also a factor of variation.

2.2.2 Language Acquisition in Infants

In order to acquire speech, the infant must be in an environment where he is exposed to a spoken language, where he can interact verbally with his world, in addition to having the neural structures that enable the speech module in proper condition. The exterior stimuli are people speaking in the child's perceptive area, not necessarily with the child.

To the developing child, speech perception precedes speech production. This time delay is presumably due to the need of the child to be immersed in a specific language, before he can be ready to produce it.

In a first perceptive stage, the child associates systematically certain sound patterns with events, objects or situations that are familiar to him. The first meaningful speech feature is prosody. Afterwards, the child starts segmenting speech in words or blocks of words. For instance, the phrase "Let's take a bath" is frequently repeated, in the same way as "Let's take a nap". This can lead to an association between real-life situations (and later, concepts) and the sound sequence; afterwards, the situation can be related to the segmented sequences *bath* and *nap*.

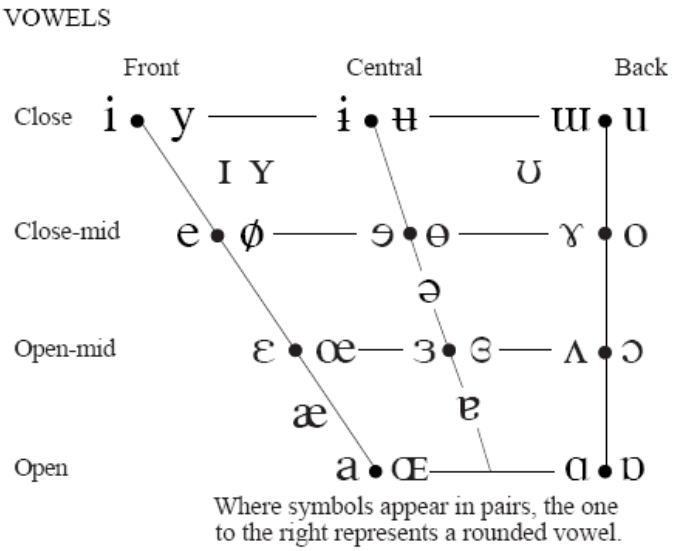
In a later stage, production starts to take place. Although the child babbles since birth, exploring his articulatory possibilities, only at the age of about six months, after accomplishing global comprehension of phrases and words, does he start to articulate the sounds of his own environment: repeated or isolated sounds or syllables and words-phrase. This is the inverse path of understanding: the infant recognizes the meaning of the phrase because of the presence of one single word ("bath") and produces the same single word to mean the whole phrase.

Language acquisition is complete when phrases and continuous speech are mastered. This usually happens when the child reaches the age of four or five years old.

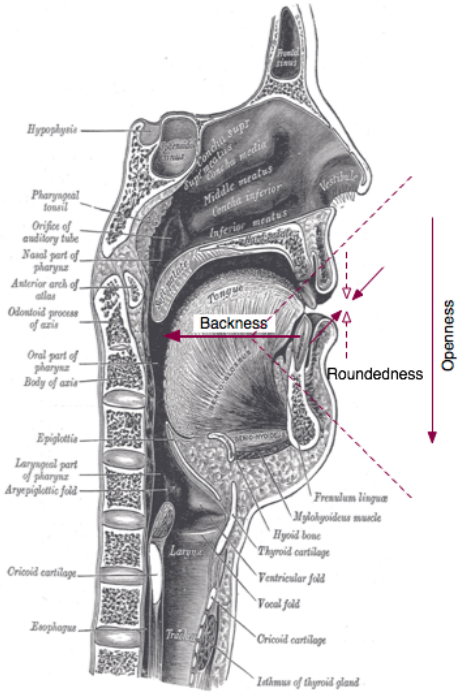
2.2.3 IPA, A System for Speech Classification: The Oral Vowel Chart

Since the beginning of Linguistics and Phonetics speech sounds are classified mainly by articulatory parameters. As the present thesis focuses only in the vowel space, we will restrict this overview to the vowel space.

One of the pioneer works in defining where are vowels located in the articulatory space was [14] in which the mathematician and phonetician Daniel Jones first proposed the Cardinal Vowel Diagram. This diagram was a subject of many discussions and contributions from the phonetics community and originated the generally accepted representation for oral vowels in use.



(a) International Phonetic Alphabet chart for oral vowels.



(b) Main degrees of freedom represented in the IPA chart. Figure from [3], with our labels.

Figure 2.3: Articulatory degrees of freedom in the IPA chart representation.

The diagram of the International Phonetic Alphabet (IPA) for oral vowels in Figure 2.3(a) shows the distribution of vocalic sounds in three dimensions relative to the human vocal tract: height (vertical axis), backness (horizontal axis) and roundedness (lip rounding) [15] as illustrated in Figure 2.3(b).

This choice of reference frame has roots in the physiology of the phonatory system. The vocal tract configuration for oral vowels depends on the tongue, jaw and lips. The jaw and the lips can have several degrees of openness, the tongue can assume the its articulatory position in anterior, middle or posterior of the oral cavity and the lips can also change the vocal tract by rounding. So, these three articulatory parameters are considered the main degrees of freedom of vowel production, the ones that better explain the inter-vowel

variation. Nevertheless, there are other static articulatory parameters that influence oral vowel quality, although they are not determinant in most spoken languages.

2.2.4 Degrees of Freedom in vocalic articulation: Corner Vowels

In most languages, rounded and not rounded vowels are not minimal pairs, i.e., for the same articulatory configuration, roundedness alone does not create two different phonological vowels. In addition to this, some studies support that roundedness is perceived mainly by vision in normal hearing-seeing subjects [16]. For these reasons, the main articulatory dimensions considered for oral vocalic sounds in the human vocal tract should be the height and backness, motivating the approximation proposed in this work — whatever the dimensionality of the articulatory space we consider, there is a two-dimensional subspace approximation that maps the vowel system of most languages. The phones [i], [a] and [u] define a set of axes in the 2D plane of the articulatory parameters of *height* and *backness*. These three vowels are called *corner vowels* because they represent extreme placements of the tongue, forming the corners of a triangle in articulatory space and also, importantly, in formant space (the second formant versus the first one).

In the following Chapter we will exploit this knowledge to propose constraints in the articulatory space that will prove useful in the simplification of speech learning and production.

Chapter 3

The Speech Production Model: A 2D Articulatory Space

In order to model restrictions due to articulator coupling and to reduce redundant complexity, it has been developed a two dimensional space of articulatory parameters for oral vowels. The parameter α models *frontness* (as oposed to backness), and β models *openness*. Roundedness is proven not to be a relevant degree of freedom in most languages, as suggested in Section 2.2.

To test and validate our proposal we use a well-known articulatory speech synthesizer. This will allow us to do systematic tests and quantify the errors arising from the proposed approximation. From realizations of the extremal phones [i], [a] and [u], we generate a dense representation of the feasible acoustic signals. Then, to evaluate the model, we compute the acoustic errors outside the feasible set.

3.1 Articulatory Synthesizer Platform

The synthesizer in use[1] is a Matlab™version of Shinji Maeda's Vocal Tract Calculator (*VTCalcs*) [8]. The seven articulatory parameters are *jaw, tongue, shape, apex, lip_ht* (lip height), *lip_pr*, (lip protrusion), and *larynx*. This last one is not relevant to oral vowel production, so it is kept constant. Thus we consider only six degrees of freedom and each one can assume any value in $[-3; 3]$. The articulator parameters are presumed to be independent, which is not the case in the human vocal tract, leading sometimes to improbable configurations of the articulators, producing a non human or even no sound. In fact, after a dense sampling of the six-dimensional hypercube and feeding the samples to the synthesizer, as explained later in this section, we realized that only 44.22% of the articulatory vectors generated sound, even if not a human-like one.

3.1.1 Source-Filter Model of Oral Vowel Production

We have already seen in Section 2.1 that vowel production can be understood as a cavity filter excited by a sound source, being usually modeled as a linear filter applied to the glottal pulse. This filter is the mathematical correspondent to the dynamics of sound propagation and resonance in concatenated tubes. When a sound wave travels in a sequence of tubes some boundary conditions have to be accounted for.

There are two important concepts when we consider the propagation and reflection of sound waves

[1] Available at the CNS Speech Lab webpage `http://speechlab.bu.edu/VTCalcs.php`

in a tube:

The acoustic impedance of a tube (Z) is defined in [12] as

$$Z = \frac{\rho\, c}{A} \qquad [kg^{-1}m^{-4}] \tag{3.1}$$

where ρ is the air density inside the tube, c is the sound velocity and A is the cross-sectional area of the tube.

Changes in the acoustic impedance across tubes will lead to changes in wave propagation. Examining equation 3.1 and noting that the sound velocity and air density are quite constant along an open tube, we see that a change in the cross-sectional area of the tube will determine a change in wave propagation.

Reflection coefficients between abutted tubes. The vocal tract is approximated by a series of cylindrical concatenated tubes with different diameters, conforming to the different cross-sectional areas of the vocal tract. Considering constant air density and sound velocity along the airway — which holds for vowels — the reflection coefficient between tubes i and j is

$$r = \frac{A_i - Aj}{A_i + A_j}. \tag{3.2}$$

Equation 3.2 presumes the convention that the wave travels from tube i to tube j. This is a very useful result because, again, the parameter depends only on the cross-sectional area of the tubes.

Waves in a tube can have multiple reflections and all of them can interfere in a constructive way, leading to *resonance*. A *formant* is a resonance in the vocal tract, and can be calculated using expression 3.3

$$F_n = (2n-1)\frac{c}{4L}. \tag{3.3}$$

Here L is the length of the tube and n the formant number.

3.1.2 A Tube Model for the Vocal Tract

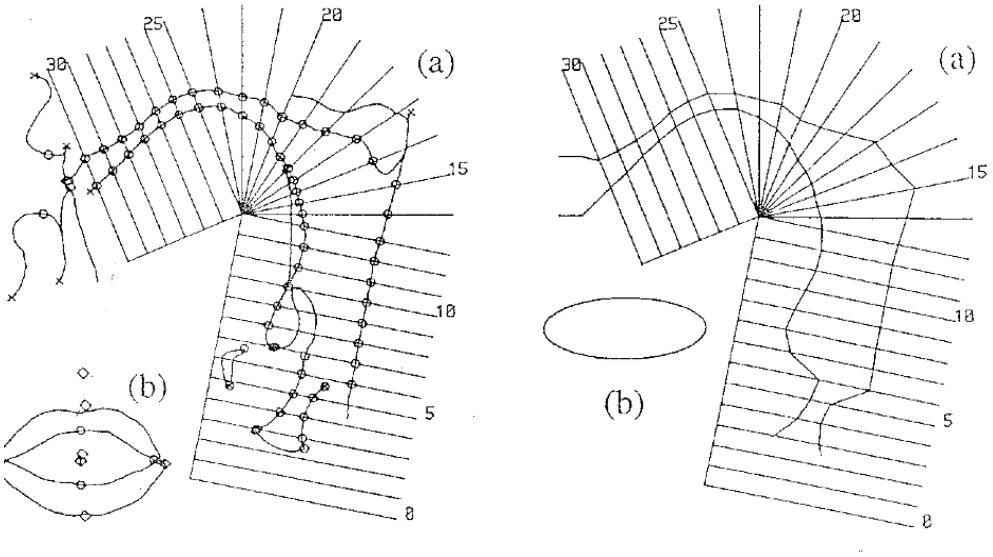
Shinji Maeda studied more than 1000 digitized tracings of vocal tract shapes from 10 French sentences uttered by two female speakers, along with the resulting utterances [8].

Measurements of specific points were made and, based on those landmarks, a vocal tract profile was generated. From this profile, based on the tube concatenation model, an area function was extracted as well as a transfer function from formants, as described before. Formants and area functions estimated from measurements were quite compatible with those obtained from Maeda's vocal tract model. This model is implemented in the referred Matlab™package *VTCalcs*.

3.2 An articulatory space for vowel production

The space of the articulators in *VTCalcs* is homographic to $\mathbb{R}^7$, but to produce vocalic voiced sounds only six parameters are distinctive, since larynx controls the voicing.

The synthesizer's output is a sound represented by its temporal amplitude. To analyze the sound waveform we use the Mel Frequency Cepstral Coefficients (MFCC)[17], using 12 coefficients.



(a) Vocal tract profile from measurements.

(b) Regenerated vocal tract profile

Figure 3.1: The Shinji Maeda vocal tract model extracted from real speech measurements. Images from [8].

Let vector $\mathbf{v} \in \mathcal{V} \subset \mathbb{R}^6$ represent a configuration of the six-dimensional synthesizer's articulatory space and $\mathbf{a} \in \mathcal{A} \subset \mathbb{R}^{12}$ be a vector of MFCC coefficients in the acoustic space. We define the synthesis function as:

$$\begin{aligned} f : \mathcal{V} &\mapsto \mathcal{A}, \\ \mathbf{a} &= f(\mathbf{v}) \end{aligned} \tag{3.4}$$

The function is not invertible — distinct articulatory configurations may lead to very similar sounds (in particular, many configurations generate no sound at all). Therefore, there is ambiguity in the identification of motor configurations corresponding to the listened acoustic signals, which may pose problems to motor-based learning and recognition algorithms. To deal with this we define a subspace of $\mathcal{V}$ where the restriction of f to this subspace is assumed to be invertible.

3.3 Dimensionality Reduction

We define a two-dimensional subspace of the full articulatory space, generated by a linear combination of vowels corresponding to extremal positions in the articulatory space. There are two major arguments that support this approach: a linguistic argument and an experimental one. As mentioned in Section 2.2, according to Linguistics and Phonetics knowledge, most of the vowel production capabilities of the human vocal tract can be explained by two parameters related to the height and frontness of the articulators. The experimental argument is that the Isomap, as discussed in Section **??**, shows that there is a good two dimensional approximation to the image of f.

Considering the extreme $\mathbb{R}^6$ prototypes for the phones [i],[a] and [u], it is possible to generate an affine space. Let a_0, u_0 and $i_0 \in \mathbb{R}^6$ be the chosen vowel prototypes for [i], [u] and [a] and a two-dimensional vector $\mathbf{p} \in \mathcal{V} : \mathbf{p} = (\alpha, \beta)$, with α and β real parameters. A linear combination of the given points forming a

two-dimensional polygon, can be defined by the function:

$$v : \mathcal{P} \subset \mathbb{R}^2 \mapsto \mathcal{M} \subset \mathcal{V}$$
$$v(\alpha, \beta) = \alpha\ i_0 + \beta\ a_0 + (1 - \alpha - \beta)\ u_0$$

where the input space $\mathcal{P}$ is defined as:

$$\mathcal{P} = \{(\alpha, \beta) : \alpha + \beta \leq 1 \wedge \alpha, \beta \geq 0\}$$

Let us denote $\mathcal{M}$ the image of v, and call it the *Motor Space*. We define the function f_2 as the restriction of the synthesizer's function f to the motor space, and call its image $\mathcal{A}_2$

$$f_2 : \mathcal{M} \mapsto \mathcal{A}_2 \subset \mathcal{A}. \tag{3.5}$$

We will denote f_2 as the *Motor-Acoustic Map*. The image of this function will produce a 2D manifold $\mathcal{A}_2$ in the MFCC acoustic space. Given the choice of the *Motor-Space*, the properties of the used synthesizer (assuming smoothness) and the dense sampling made on $\mathcal{M}$, there are strong reasons to believe that f_2 is invertible. Therefore, the inverse function of f_2, f_2^{-1}, is an acoustic to motor map. A schematic representation of the proposed vowel production model is shown in Figure 3.2.

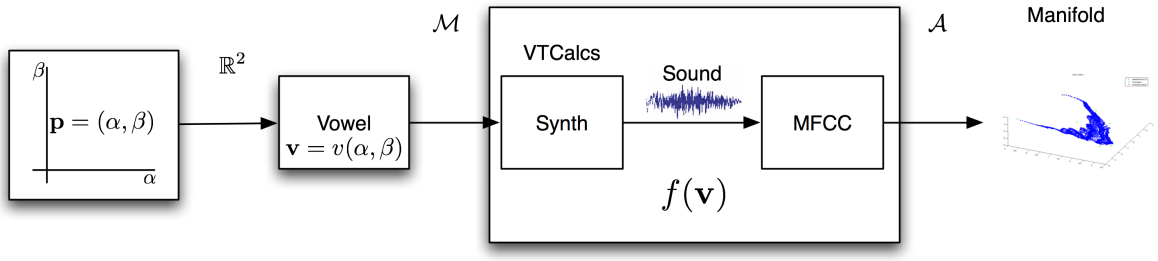


Figure 3.2: Vowel generation diagram.

The twelve-dimensional acoustic space was sampled twice; one using the motor map f_2, from the motor space $\mathcal{M}$, and another from $\mathcal{V}$.

In the following Chapter we will show that the model allows a good approximation to the full vocalic space, with obvious advantage of a low complexity representation.

Chapter 4

Speech Acquisition Architecture by motor babbling

One of the most prominent goals in humanoid robotics is to provide for an intuitive and non-specialized way for humans to teach and interact with robots. To achieve this, it should be possible for the robot to develop flexible skills by learning from a demonstrator, a caregiver or another robot with no need for reprogramming or external parameter tuning. The robot's capabilities must allow it to learn by replicating an observed action and must facilitate the interaction with humans, as in this work, by speech. The goal is, thus, to conceive artificial humanoids that can take part of everyday's life, meaning that a robot can be bought in a general store and taken home, being able to learn one's language, and serving one's specific needs, without the intervention of a specialized programmer.

Learning speech is of utmost importance for this goal, due to the relevance of spoken language in human communication. It is, along with vision, a main means of human-machine interaction. In this perspective, this work endeavors to assemble biologically plausible models, as well as to achieve performance gains in its field of study. The present Chapter explores the global learning architecture used, whose implementation and details are considered in the following Chapters.

Based on the constrained articulatory space exposed in Chapter 3, the architecture for physical production and sensory acquisition of speech sound data depicted in Figure 3.2 is the basis for a higher level in perception. On top of this, we establish a learning process that will enable the recognition of the vowels for a particular language.

The speech production and perception acquisition architecture adopted in this work has three main functions:

- Babbling, to develop a sensorimotor map,
- Listening and producing the speech sounds of a given language,
- Recognizing structure and groupings in the acquired sound set.

These functions and their relationships are illustrated in Figure 4.1. Although the main babbling phase is at the beginning of life, it will proceed, in a smaller scale, throughout the robot's early childhood. The acquisition of speech sounds should also continue, even after its mother tongue is completely known and all the unnecessary sounds are forgotten.

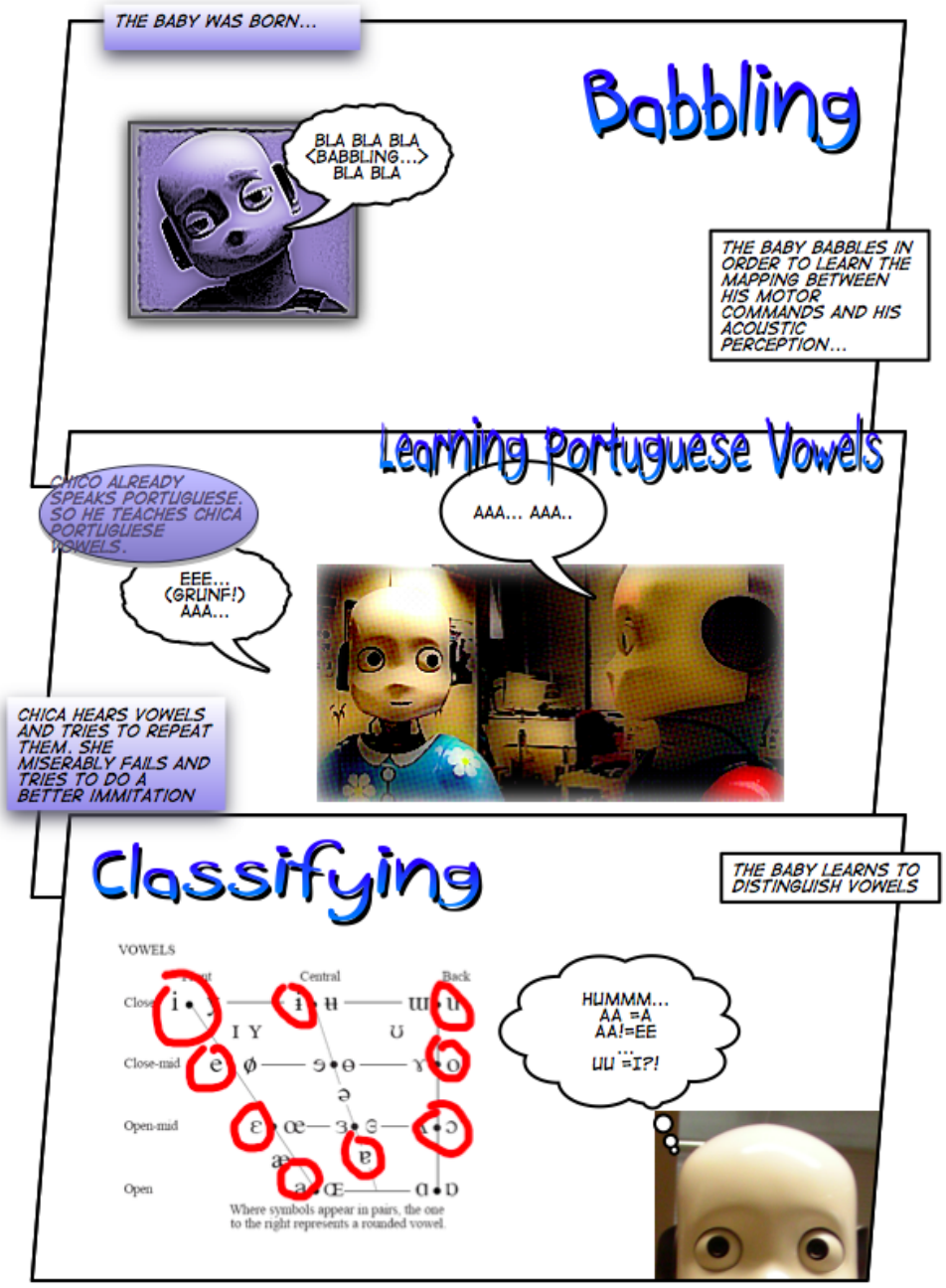


Figure 4.1: Developing robot.

4.1 Babbling: general sensorimotor map

The implemented sensorimotor map is estimated from a dataset generated by the robot.

To estimate the *acoustic manifold* $\mathcal{A}_2$ we have sampled the parameter space $\mathcal{P}$ applying steps of 0.01 to the α and β parameters, generating a discrete set of 5000 samples:

$$\mathcal{P}_d = \{\mathbf{p}_i,\ i = 1, \ldots, 5000\}$$

These samples were then used to generate a motor-space sample set, using function v:

$$\mathcal{M}_d = \{\mathbf{m}_i = v(\mathbf{m}_i),\ i = 1, \ldots, 5000\}$$

Thus, a discrete sampling of the acoustic manifold was created using the synthesizer's function:

$$\mathcal{A}_{2d} = \{\mathbf{a}_i = f_2(\mathbf{m}_i),\ i = 1, \ldots, 5000\} \tag{4.1}$$

The first three coordinates of the sampled acoustic manifold are plotted in Figure 4.2.

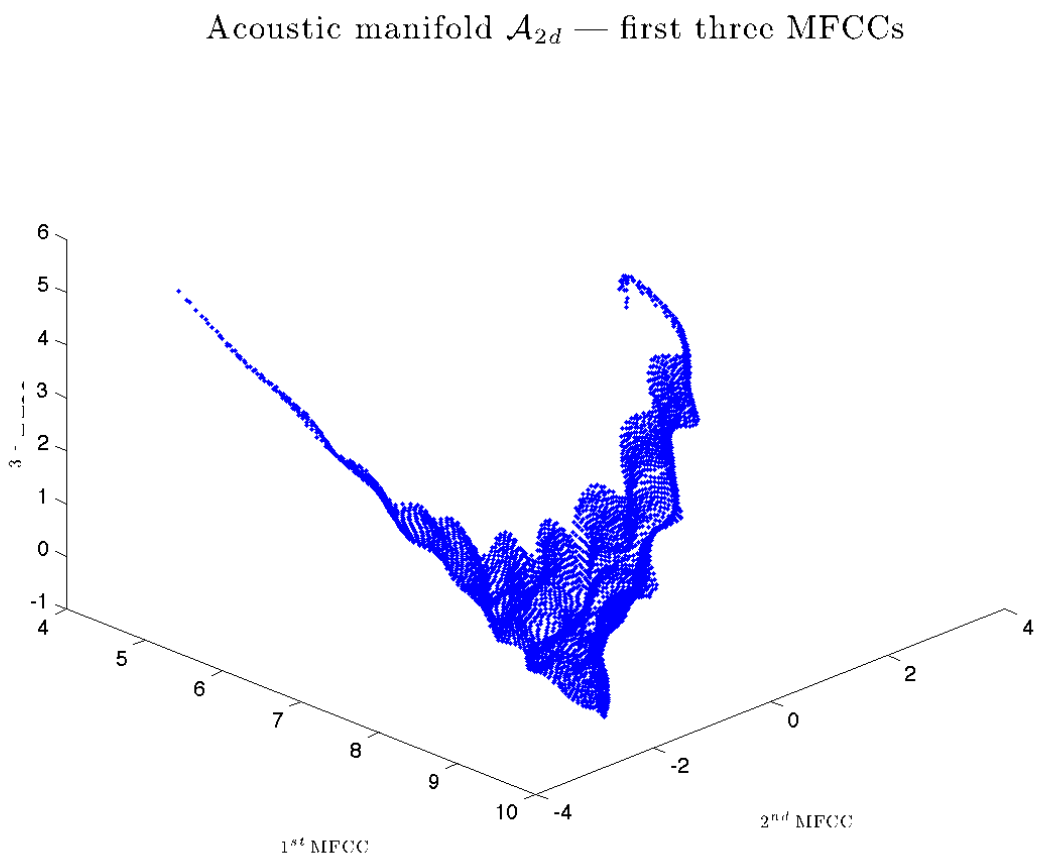


Figure 4.2: Representation of the first three Mel coefficients of the acoustic manifold.

The *VTCalcs* parameter's six-dimensional $\mathcal{V}$ space was also sampled in steps of 0.6 obtaining a grid with 10 samples per dimension. The point cloud has 10^6 samples:

$$\mathcal{V}_d = \left\{\mathbf{v}_i,\ i = 1, \ldots, 10^6\right\}$$

Again, the synthesizer's function was applied to the data:

$$\mathcal{A}_d = \left\{\mathbf{a}_i = f(\mathbf{v}_i),\ i = 1, \ldots, 10^6\right\} \tag{4.2}$$

From this data we removed the samples with zero sound amplitude, retaining 44.22% of the initial number.

4.2 Experimental Results

To validate the proposed model we generate a set of test vowels $\mathbf{a}^t$ and compute the error in the acoustic space (MFCC coefficients) between each one and its projection on the manifold $\mathcal{A}_{2d}$. We also consider the residual variance incurred in a two-dimensional approximation of $\mathcal{A}$.

Since we do not have an analytic expression for the $\mathcal{A}_2$ surface, we use its sampled version defined by equation (4.1). To compute the projection of each point we use the nearest neighbor operator:

$$nn(\mathbf{a}^t) = \left\{ \mathbf{a}_i \in \mathcal{A}_{2d} : i = \underset{i}{\operatorname{argmin}} \left\{ \|\mathbf{a}_i - \mathbf{a}^t\|_2 \right\} \right\} \tag{4.3}$$

The acoustic approximation error is then computed by:

$$E_a(\mathbf{a}^t) = \|\mathbf{a}^t - nn(\mathbf{a}^t)\|_2 \tag{4.4}$$

and, relatively to the size of the manifold, it is defined as

$$\delta_a(\mathbf{a}^t) = \frac{E_a(\mathbf{a}^t)}{\max(length(\mathcal{A}_{2d}))}\ 100\% \tag{4.5}$$

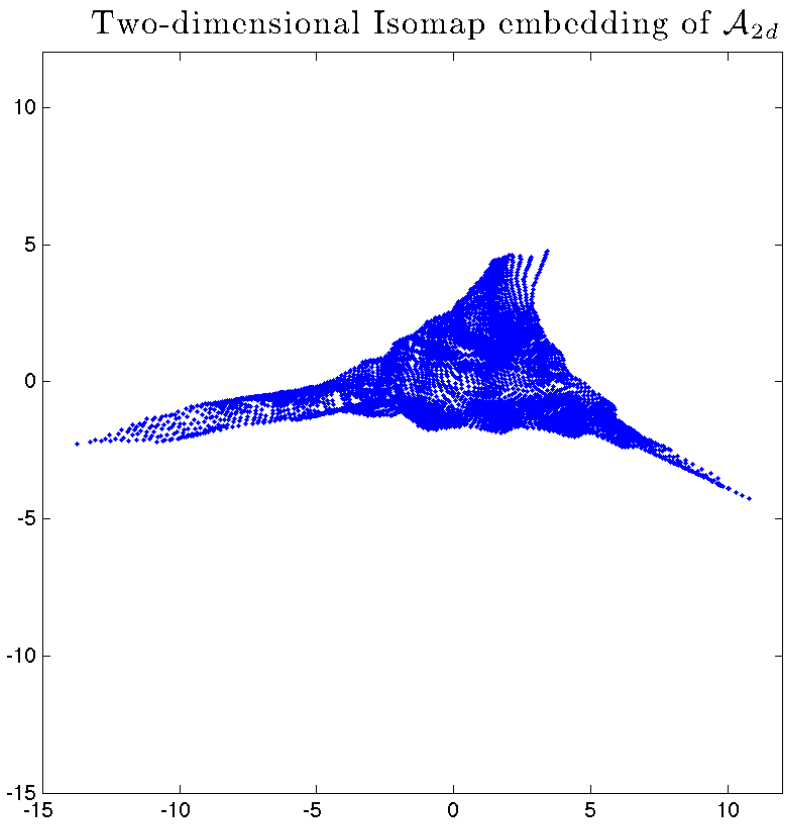


Figure 4.3: Isomap embedding for the two-dimensional manifold $\mathcal{A}_{2d}$.

This measure is dimensionless and gives an indication of how good is the approximation relative to the size of the approximating surface. We consider acceptable to use the maximum length of $\mathcal{A}_{2d}$ to normalize the error because the manifold's shape is not too discrepant, as it is possible to confirm in the Isomap embedding

shown in Figure 4.3. This embedding was determined with the Isomap algorithm as described in [18]. The *isometric feature mapping procedure* or Isomap recovers low-dimensional nonlinear structure in perceptual datasets. It finds a space embedding for the data, preserving its intrinsic metrics, by conserving distances measured through *geodesic paths* along the observation manifold. For $\mathcal{A}_{2d}$, Isomap output a reproduction in the two-dimensional space of the pairwise distances measured in the data twelve-dimensional space.

4.2.1 Dimensionality reduction: validation

To verify the approximation validity of a two-dimensional surface given the full space $\mathcal{A}$, the dimensionality of the sampled space $\mathcal{A}_d$ defined in equation 4.2 was investigated.

It was estimated through Isomap that the dimensionality of the image of f is 2, with a residual variance of 0.197, as illustrated in Figure 4.4.

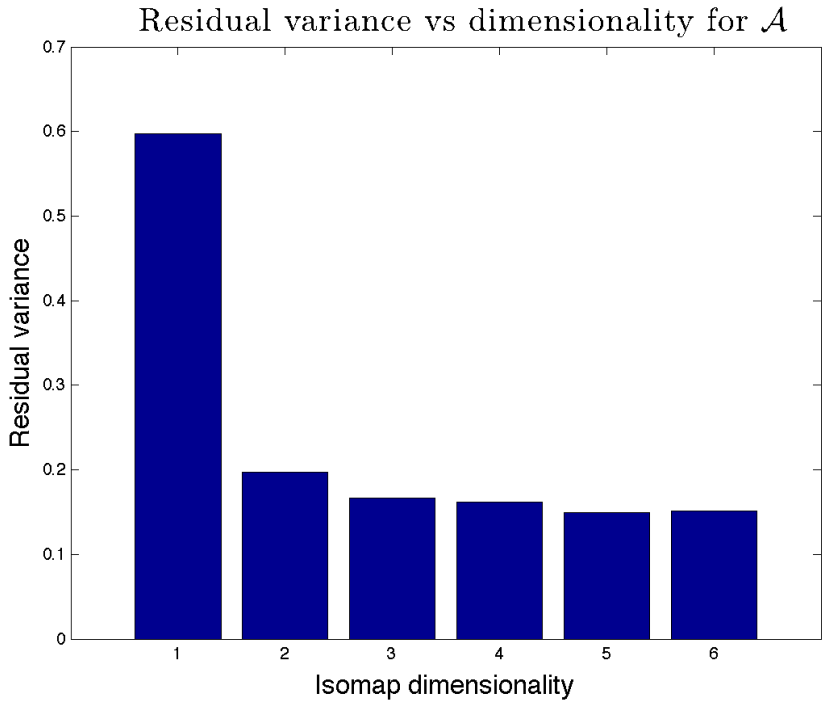


Figure 4.4: The Isomap algorithm provides the residual variance of the fit to the model's dimensionality. The greatest decrease in variance happens from one to two dimensions of the manifold representing the global acoustic space $\mathcal{A}$.

The global articulatory space $\mathcal{M}$ is six-dimensional; thus, due to the continuity of f, the maximum possible dimensionality for $\mathcal{A}$ is six. The residual variance of the data for six or more dimensions can be interpreted with regard to phenomena such as noise and numerical problems in the MFCCs calculation.

This experimental result confirms that there is a good two-dimensional approximation to the overall acoustic space $\mathcal{A}$. The residual variance present in the 2D approximation is partially due to the model simplification, but its slow decrease with increasing dimensionality of the model leads to the conclusion that it is caused mainly by non-informative phenomena.

4.2.2 Vowel prototypes: appropriateness

To investigate the performance of the approximating space with speech sounds of real languages, some experiments have been conducted with synthesized prototypes of several languages. Those prototypes naturally lie outside the motor space $\mathcal{M}$. Those that are integrated in the *VTCalcs* matlab package are preexistent to the experiment; the other sets were constructed by us and validated by naive native speakers. The speech sounds' intensity, fundamental frequency and duration were kept constant so as to strictly validate the model for vocal tract configuration.

In the *VTCalcs* package there are eleven prototypes for oral vowels which are found outside the two-dimensional polygon $\mathcal{M}$. They were used to evaluate the amount of error introduced in the two-dimensional approximation. The error was measured as described above, and the results are shown in Table 4.1. The oral vowels from two very distinct european languages were also used for the same purpose: vowels from Portuguese, an Indo-European, Romance language, and vowels from Finnish, a Finno-Ugric language. Nine Portuguese prototype vowels were used. The errors are shown in Table 4.2. From Finnish, the eight short vowels were investigated, with results that can be seen in Table 4.3.

Table 4.1: *Approximation error for the* VTCalcs *prototypes.*

vowel	symbol	$E_a(\mathbf{a}^t)$	$\delta_a(\mathbf{a}^t)$%
1	iy	0.40149	1.6295
2	ey	0.17829	0.72361
3	eh	0.1522	0.61771
4	ah	0.48633	1.9738
5	aa	0.24348	0.98818
6	ao	0.51035	2.0713
7	oh	0.58974	2.3935
8	uw	1.6111	6.5389
9	iw	1.4057	5.7053
10	ew	0.29547	1.1992
11	oe	0.18119	0.73536

Table 4.2: *Approximation error for the Portuguese prototypes.*

vowel	IPA symbol	$E_a(\mathbf{a}^t)$	$\delta_a(\mathbf{a}^t)$%
1	ɨ	0.13425	0.54487
2	ɐ	1.2335	5.0061
3	ɛ	0.37961	1.5406
4	ɔ	0.50396	2.0453
5	e	0.61689	2.5037
6	o	1.4141	5.739
7	a	0.24161	0.98057
8	u	1.6211	6.5792
9	i	0.39633	1.6085

The sample mean over the percent error $\delta_a(\mathbf{a}^t)$ is 2.95% in the Portuguese vowels set, 3.87% in the Finnish vowels, and 2.23% in the *VTCalcs* set. The standard deviation is 2.22%, 2.81% and 2.02% in the Portuguese, Finnish and *VTCalcs* sets, respectively. The maximum value for the percent error is 9.17% in the Finnish dataset.

So, in terms of the error, the two-dimensional convex space performs well with linguistically relevant synthesized speech sounds. Acoustically, the prototypes and the projections are hardly distinguishable. By inverting the projected points through f_2^{-1} back to the two dimensional motor space $\mathcal{M}$ and plotting the

Table 4.3: *Approximation error for the Finnish prototypes.*

vowel	IPA symbol	$E_a(\mathbf{a}^t)$	$\delta_a(\mathbf{a}^t)$%
1	i	0.28764	1.1674
2	ø̞	0.7918	3.2135
3	æ	0.99949	4.0564
4	ɒ̞	0.87593	3.555
5	ɑ	1.6373	6.645
6	u	0.5645	2.291
7	e̞	0.21044	0.85406
8	y	2.2605	9.1741

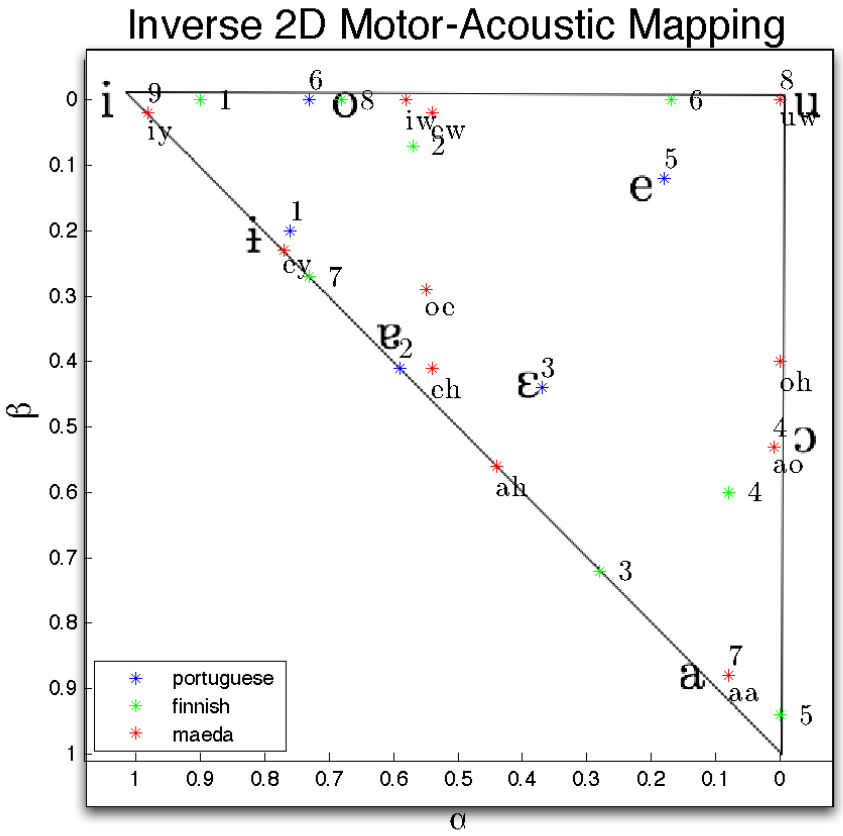


Figure 4.5: The inverse mapping of the vowel prototypes. The Portuguese vowels are numbered as in Table 4.2, and the Finnish as in Table 4.3. Some landmark IPA phonetic symbols are also represented.

result (Figure 4.5), we can recognize some similarities between the IPA openness and frontness and the motor space α and β parameters. The hypothesis that the restrictions in the construction of $\mathcal{M}$ are in fact modeling physiological constraints is corroborated by these experimental results.

4.3 Acquiring the speech sounds of a particular language: refining the sensorimotor map

After the pure babbling phase, the robot starts to hear and reproduce speech sounds from its environment. It hears a speech sound and uses its map to guess parameters in the articulatory space. It still babbles around the

guessed articulatory vowel in order to produce a smaller error in the acoustic space. When the error is small enough it stops introducing points in the map. This kind of directed learning aims at improving the robot's efforts to attain its goal, specializing to the symbols existing in the language it is embedded. The results for mimicking speech sounds of two different languages are presented in the previous Section.

4.4 Recognizing groupings in speech data from a particular language

When the robot imitates its caregiver, it collects data that brings it in a further step towards language. In fact, it is the speech sounds of a particular language that the robot has in its database. This is a big leap in terms of language layers, since we are in the border between phonetics and phonology. Now the robot starts to specialize in a certain set of vowels, say, the portuguese oral vowels, and tends to forget all the other tested but now useless vocalic possibilities.

To understand the collected data from caregivers, the robot must grasp the structure of its mother tongue's phonology. So, it must figure out how many groupings of sounds really exist and it must learn to distinguish between relevant differences and unimportant ones. This is something a human baby does by its own, with no explicit teaching about the rules to group speech sounds. In a similar fashion, unsupervised learning was used for the robot.

Starting from the babbling and speech sounds acquisition phases, the robot stores in memory the 1000 most recently heard speech sounds and tries to organize this data in a meaningful way. But it doesn't do this in the acoustic high-dimensional space; it rather uses the two-dimensional articulatory mappings of those sounds, even when the map is not good enough. While the developing robot perfects its skills, the mapping and the classification will grow better together.

So one of the first questions arising when we face a somewhat big amount of data is *how many meaningful groupings we have in this dataset?* In fact, this number is the input for many clustering algorithms and knowing or estimating it is also useful for validating the number of clusters suggested by other clustering algorithms. In any case, this number must conform with domain knowledge.

There is no rigorous definition for a data cluster: it is highly subjective and depends on scale and resolution. It is application dependent. This work's approach to this subject is presented in Chapter 5.

Chapter 5

Speech sounds classification

When one is presented with a large and heterogeneous set of objects it is common that one tries to find the natural groupings between them. The current approach is to map the objects to points in $\mathbb{R}^n$ and group them based on some similarity measure. This mapping in the present work is to process sound into MFCC, obtaining points in $\mathbb{R}^{12}$ and representing them in the two-dimensional articulatory space. The similarity measure is the euclidean distance in $\mathbb{R}^2$.

5.1 The ideal clustering algorithm

As already mentioned in Section 4.4, the goal of clustering is to organize data in groups such that similar objects are grouped together and dissimilar objects lie in different groups. The bottleneck in this problem is to define what is similar enough to group in the same cluster and what is dissimilar enough to belong to different clusters.

The ideal clustering algorithm must have three properties [19]:

Scale-invariance: it must be insensitive to changes in the units of distance measurements;

Richness: it must be able to generate any partition of the space;

Consistency: when we shrink distances between points inside a cluster and expand distances between points in different clusters, we get the same result.

But Kleinberg demonstrated Theorem 1 in [19], proving that it is impossible to have all three properties of the ideal clustering algorithm in a implementable form.

Theorem 1 (impossibility theorem for clustering) *For each $n \geq 2$, there is no clustering function f that satisfies scale-invariance, richness, and consistency.*

It is also proven that choosing between three different stopping conditions leads to a clustering function that satisfies two of the three properties of the ideal clustering algorithm.

Considering the above properties we choose to have scale invariance and consistency, because being able to generate all the possible partitions in this space is not necessary for this application. In the other hand, scale invariance is important and consistency is fundamental, since we can have different boundaries for vowels, depending on the language and the speaker.

5.2 Hierarchical clustering

Clustering is to group or segment a collection of objects into subsets — clusters — such that objects that share the same cluster are more closely related to those that do not [20].

Given a certain measure of dissimilarity, hierarchical clustering gives us an hierarchical representation where each cluster is formed by joining the two most similar clusters below it. We use an agglomerative approach, starting by having one cluster per data point and ending at the top having only one cluster for all the existing data.

For hierarchical clustering the stopping conditions referred in the previous section are

To specify k-clusters: to stop at k number of clusters leads to *scale invariance* and *consistency*;

To specify a distance-r: to stop at distances over r leads to *richness* and *consistency*;

To specify α-scale: Let $\rho^* = \max_{i,j} \{distance(i,j)\}$. Stop at distances larger than $\alpha\rho^*$. This leads to *richness* and *scale invariance*.

Since we have chosen to have scale invariance and consistency, our stopping condition will be the achieved number of clusters.

5.3 Investigating the natural groupings in data

For each level of the clustering process, we have different relationships between data groupings. So, the question is: what is the "natural" grouping for this dataset? We want an index that measures consistency and separability in the dataset. To meet such demands, Tibshirani et al. [21] proposed the Gap statistic in 2001.

The Gap statistic is a robust index for cluster compactness and cluster separation. Compactness is assessed by:

- measuring within cluster distances,
- measuring between cluster separation,
- comparing within cluster distances in the dataset with those obtained by performing the same clustering in a non-informative distribution of points.

To define the Gap statistic it is necessary to formalize the used distance metrics and the within cluster distance. Let $\{\mathbf{v_{ij}}\}, \quad i = 1, \ldots, n, \quad j = 1, 2$ be the dataset with n independent observations and $d_{ii'} = \sum_j (\mathbf{v_{ij}} - \mathbf{v_{i'j}})^2$ be the squared euclidean distance. If the data was clustered into k groups, then $C_r, \quad r = 1, \ldots, k$ is the set of observation indexes that were classified in cluster r and n_r is the total number of elements in that cluster. The sum of pairwise distances in cluster r is

$$D_r = \sum_{i,i' \in C_r} d_{ii'} \tag{5.1}$$

The within cluster dispersion (W_k) is an error measure that can be defined, for the squared euclidean distance, as

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r \tag{5.2}$$

By using only W_k, we can try to estimate the number of clusters by identifying the "elbow" on the graph of W_k. This heuristic needs formalization. So, the authors standardize the graph of $\log(W_k)$, by comparing it with a non-informative, null-reference distribution. The Gap statistic is, thus, defined as in expression 5.3

$$Gap_n(k) = E_n^*[\log(W_k)] - \log(W_k) \tag{5.3}$$

where $E_n^*[\ \ ]$ is the expected value of a sample set of size n from the null-reference distribution. The estimated number of clusters $\hat{k}$ is

$$\hat{k} = \underset{k}{\operatorname{argmax}}\ Gap_n(k) \tag{5.4}$$

The Gap statistic explained above with the algorithm described by Tibshirani et al. was implemented as a Matlab™function, using hierarchical agglomerative clustering.

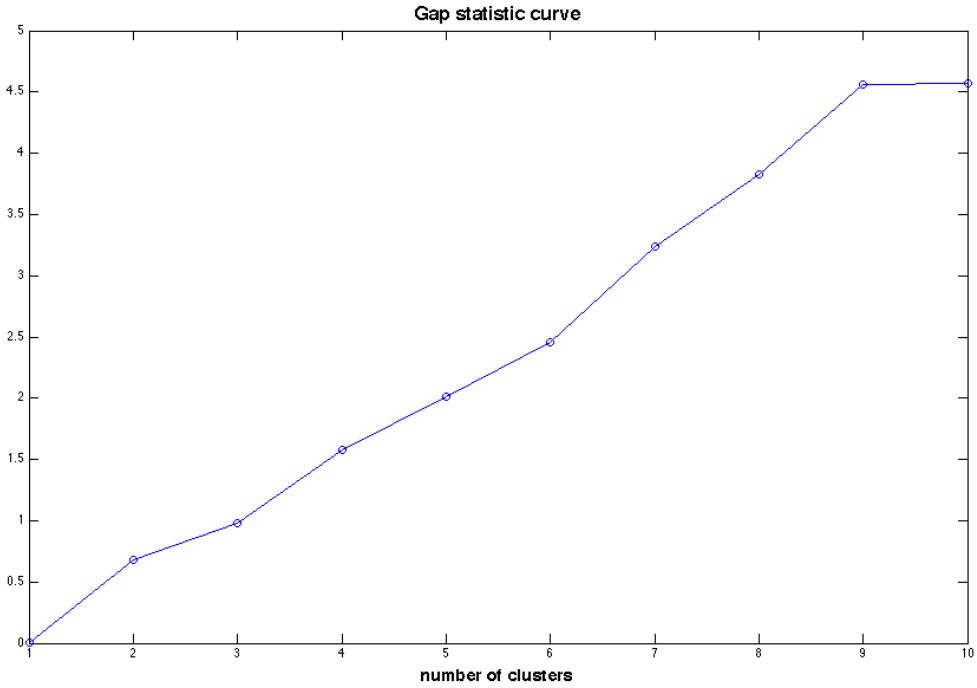


Figure 5.1: Gap statistic versus number of clusters. The growth of the curve stops at nine clusters.

This function compares the within-cluster dispersion of our data with that obtained by clustering a reference uniform distribution. This is to compare the gainings of raising the cluster number in a structured data with those that arise from adding another cluster to a non-informative and not structured set of points.

In ten performed trials, the Gap statistic consistently pointed to nine as the most natural number of clusters. One example of this result is presented in figure 5.1.

5.3.1 Dataset of portuguese synthesized vowels

To create a sufficient number of valid training vowels for the robot, we created a dataset with 900 vowels, and then submitted them to the evaluation of 16 native speakers, so that they rejected or approved each vowel as a valid portuguese vowel and — for those that were approved — agreed or not in their phonological classification. From these 900 vowels, 448 were considered appropriated.

The original dataset was generated from nine prototype vowels in the two-dimensional articulatory space, added with 10% of white noise.

Applying agglomerative hierarchical clustering to the present vowel dataset originated good results, as we can see in figure 5.2. The nine vowel groupings depicted in different colors are clearly visible.

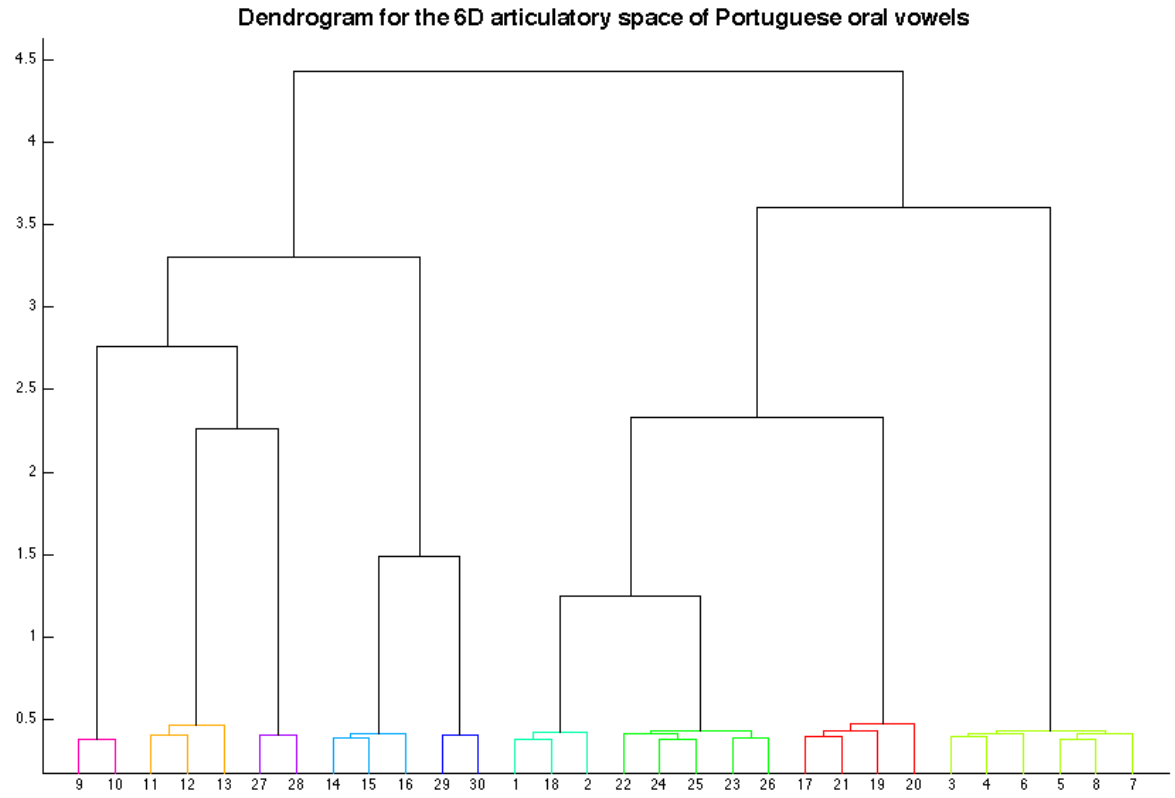


Figure 5.2: Dendrogram depicting the hierarchical clustering performed by the robot.

The dendrogram shown in Figure 5.2 summarizes the data structure that was detected by our simple dissimilarity measure: euclidean distance between vectors and average dissimilarity between groups.

5.3.2 Discussion

The coupling of dimensionality reduction with the clustering algorithm that joins Gap statistic and hierarchical clustering proves to give interesting results. Dimensionality reduction cleans out irrelevant variation in data, leaving an open field for a consistent and scale invariant cluster algorithm.

Although, there are some limitations, namely when clusters are not completely separated. But even with overlapping classes, the probability that the algorithm miscounts them grows linearly with the proportion of overlapping points.

Some comparisons should be made in the future with the Gap statistic and other well-known clustering algorithms, like k-means, for instance. The behavior of the Gap statistics should also be assessed in raw high dimensional data, in order to test the benefits of the present coupling.

Chapter 6

Conclusions

6.1 Consequences

In this thesis we have proposed a two-dimensional parameterization for the motor space of an available speech syntesizer, *VTCalcs*. The approach is able to generate acoustic signals that represent well all the vowels produced by the syntesizer. Namely, the euclidean error relative to the size of the two-dimensional approximating surface has its maximum of 9.17% in the used test sets, and the Isomap analysis of the residual variance versus the dimensionality of the approximating manifold confirms the validity of a two-dimensional model for the overall acoustic space.

The proposed model is important for two main reasons:

- The motor space is two-dimensional; thus, it can be densely sampled with low computational requirements. This simplifies creation and representation of the motor-acoustic map.
- The restriction of the synthesizer's function to the proposed motor space is invertible, allowing to map signals back from the acoustic to motor coordinates. This will facilitate the utilization of learning and classification algorithms.

The fact that this space is two dimensional facilitates its bootstrapping role in autonomously producing and recognizing speech. Once the system learns a good initial model of the motor-acoustic map using the low-dimensional manifold, it can expand the available degrees of freedom and refine its production capabilities. As in the ontogenesis of human infants, such a developmental strategy is more likely to succeed than learning from scratch with the whole system's complexity.

Based on this notion, the present work presents also a learning architecture and an unsupervised clustering technique coupled to it, revealing interesting and stable results, leading to an optimistic perspective about the future of man and machine interaction through speech.

6.2 Open Issues

Future work should quantify the approximation error to human spoken vowels and evaluate the model's performance in online learning and recognition. The problem of fundamental frequency and vocal tract length normalization must be addressed, as well as interconnection of the present structure with higher linguistic layers, like morphology, semantics or syntax that can be very helpful when ambiguities take place.

Visual cues should also be considered in the sensorimotor mapping. A replica of the McGurk effect should be implemented.

Based on radial basis functions, receptive fields that support linear local models, preliminary unsupervised incremental results were obtained, bur further exploration on the advantages of the differentiable approach should be considered, namely the benefits of having a good exploratory direction against the payload of the jacobian matrix computations.

The clustering should also be incremental, for the present work keeps a buffer for the last 1000 vocalic sounds, and does not implement an authentic online algorithm for class separation and classification.

Appendix A

Learning speech

This Chapter has some preliminary results concerning the implementation of an incremental online learning strategy for the sensorimotor map presented in Chapter 4. As stated in previous Chapters, this work has an articulatory approach to speech perception, and, in order to develop a motor-based recognition system, one must have a sensorimotor map that captures the coupling of the listener's articulatory parameters with the speech perception in its various cues. As mentioned earlier in Section 1.1, the motor theory of speech perception supports that there is no absolutely fundamental sensory cue to identify a speech sound. They all contribute to communication through sound. It is reasonable to believe that the more cues we have more we improve the quality of this identification. So, the acoustic cues could be complemented by visual cues, as suggested by the McGurk effect referred in Section 1.1.

So, the sensorimotor map that the system is trying to learn is the function f_2^{-1} defined in Section 3.3. This map is learned using Receptive Field Weighted Regression (RFWR) presented in [22] and discussed in Section A.1.

A.1 RFWR: Incremental online learning

Receptive Field Weighted Regression (RFWR) is an algorithm for function approximation, designed to work well in incremental online learning which is desirable for a system inspired in human speech acquisition.

The algorithm is based on nonparametric regression with locally linear models. It tries to determine the number of linear models K, the parameters β_k for each linear model k that represents the data and a region of validity, curiously named *receptive field*. In neuroscience, a receptive field of a sensory neuron is a region of space whose variations in the sensory specialization of the neuron alter its firing pattern. In the human auditory system, receptive fields relate or to areas of the physical space surrounding the listener or bins of sound frequencies. This locality of the mapping in humans motivates for the use of this algorithm in sensorimotor maps.

In RFWR the receptive field for the k^{th} linear model is a Gaussian kernel, centered in $\mathbf{c}_k$ with a positive definite distance metric $\mathbf{D}_k$, defined as in equation (A.1)

$$w_k = \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{c}_k)^T\ \mathbf{D}_k\ (\mathbf{x}-\mathbf{c}_k)\right) \tag{A.1}$$

with a fixed threshold of 0.001. The weight w_k is a measure of how much a data point falls into the region of validity of each linear model.

In the prediction mode, when a query point $\mathbf{x}$ is presented to RFWR, each one of the local models will determine their own prediction for that point, $\hat{\mathbf{y}}_k$. The overall prediction is

$$\hat{\mathbf{y}} = \frac{\sum_{k=1}^{K} w_k \hat{\mathbf{y}}_k}{\sum_{k=1}^{K} w_k} \tag{A.2}$$

Each $\hat{\mathbf{y}}_k$ is calculated as shown in the expression (A.3)

$$\hat{\mathbf{y}}_k = (\mathbf{x} - \mathbf{c}_k)^T \mathbf{b}_k + b_{0k} \tag{A.3}$$

or, assuming that

$$\tilde{\mathbf{x}} = \begin{pmatrix} (\mathbf{x} - \mathbf{c}_k)^T & 1 \end{pmatrix}^T$$

then

$$\hat{\mathbf{y}}_k = \tilde{\mathbf{x}}^T \beta_k \tag{A.4}$$

where β_k in equation A.4 are the linear model parameters, corresponding to the k^{th} receptive field and $\tilde{\mathbf{x}}$ is an augmented and centered form for the input vector, for simplicity sake.

The learning algorithm implemented in RFWR estimates $\mathbf{c}_k$, $\mathbf{D}_k$, and β_k for each receptive field independently. The number of receptive fields increases or decreases with the complexity of available data. So, for a given point, if there is no significant activation of any receptive field, a new one will be created. In a similar way, if one of the kernels explains a small quantity of data variance and there are others that overlap with it, then it will be removed.

In the present work two maps were created, one for the α and another for the β parameter. Here, each one of those parameters have their $\hat{\mathbf{y}}$ function, as shown in equation A.5:

$$\hat{\mathbf{v}} = \frac{\sum_{k=1}^{K} w_k \hat{\mathbf{v}}_k}{\sum_{k=1}^{K} w_k} \tag{A.5}$$

then, each linear prediction associated with a given receptive field k, can be written as in equation A.6

$$\hat{\mathbf{v}}_k = (\mathbf{a} - \mathbf{c}_k)^T \mathbf{b}_k + b_{0k} \tag{A.6}$$

or, similarly,

$$\hat{\mathbf{v}}_k = \tilde{\mathbf{a}}^T \beta_k \tag{A.7}$$

where $\tilde{\mathbf{a}} = \begin{pmatrix} (\mathbf{a} - \mathbf{c}_k)^T & 1 \end{pmatrix}^T$ from equation A.7 is the compact centered form for the local projection of the input point.

As it is suggested by the receptive field metaphor, this algorithm implements incremental learning, which means that knowledge acquired from old data is forgotten when integrating new data in the regression model. This is particularly interesting in this specific application, due to its evolving nature. In fact, the infant's first discoveries about his vocal tract and vocalization abilities will be rather unproductive latter on, when his anatomy is radically changed and most of the sound – articulation mappings are not useful for his mother-tongue.

In order to assemble a vocalic sound learning architecture for the present system, a two-step procedure was setup; the first step is *babbling* while the second step is *incremental vowel learning*. In the babbling stage, as presented in Section A.2, the system talks to itself. This means that it maps its articulations with the auditory stimuli of his own voice. The first map is produced, independently of the language that will be

learned later on. The incremental vowel learning stage will lead to the knowledge of the speech sounds of a particular language. All the unnecessary sounds will become dim when the robot is systematically exposed to organized language sound sets.

A.2 Babbling: general sensorimotor map

At this stage, the baby has not yet reached a linguistic competence that differentiates between languages. He is just exploring his physical possibilities. So, the system proposed in this thesis will just focus on learning the map between the produced sound and the intended one, possibly heard in the system's surroundings from a caregiver, for instance.

A Matlab™implementation of this behavior was elaborated. The structure of the `babble` function is depicted in Figure A.1.

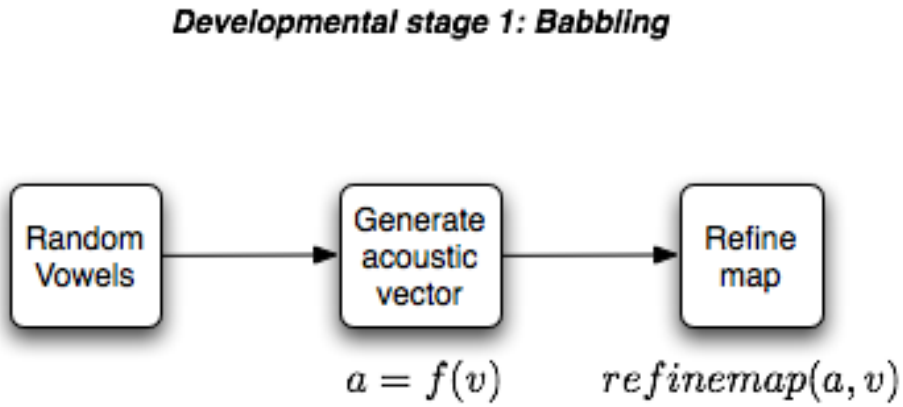


Figure A.1: Babbling stage function main components. `refinemap` is a function that introduces pairs of acoustic feature vectors and 2D articulatory parameters in the RFWR learning structure.

The function `f_p_a`, represented in Figure A.1 as the *Generate acoustic vector* block, receives as argument the two-dimensional vector $\mathbf{v}$ and outputs the twelveth-dimensional acoustic vowel representation $\mathbf{a}$. It is the synthesis function as defined in the following Section.

A.2.1 Random babbling

The implemented sensorimotor map described in Chapter A is trained with a dataset generated by the robot. It performs an uniform random sampling of the two-dimensional articulatory space, obtaining the MFCC output of the synthesis function. The map is defined as

$$\begin{aligned} g : \mathbb{R}^{12} \supset \mathcal{A}_2 &\mapsto \mathcal{P} \subset \mathbb{R}^2 \\ g(\mathbf{a}) &= \mathbf{v}. \end{aligned} \tag{A.8}$$

In the babble phase we used 780 vowels for training and 30 for testing. In Figure A.2 we show how the training works for a subset of only 80 vowels, so that the convergence can be made visible.

To assess the performance of the map we need to establish a measure of the distance between the desired output and the result the model provides. Due to the linear nature of the articulatory space, an

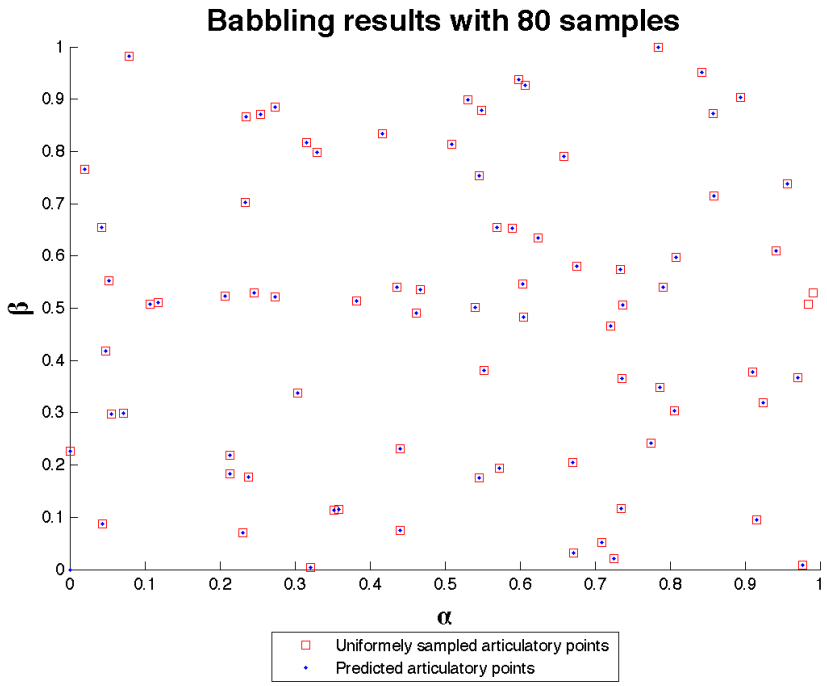


Figure A.2: Babbling stage results: the inverse map is learned by generating sound samples from articulatory positions.

euclidean distance measure is chosen. The error is normalized to the size of the dataset and the variance in the desired outputs.

The normalized mean square error is estimated as defined in equation A.9

$$E_v^2 = \frac{\|\mathbf{v}^t - \mathbf{v}\|_2^2}{n \ (var(v_1^t) + var(v_2^t))} \tag{A.9}$$

where $\mathbf{v}^t = [v_1^t \quad v_2^t]$ is the desired articulatory vowel, n is the number of vowels used in the trial and $\mathbf{v}$ is the predicted articulatory vowel.

Sampling randomly the articulatory space and training the inverse map leads to a low training error, and the test normalized square error is quite small, as can be seen in Figure A.3.

This result was found after an extensive search in the algorithm's parameters, in order to optimize the learning of the goal maps.

A.3 The implemented online learning architecture

To be able to learn with a teacher without knowing the articulatory parameters that generated a particular sound, the system invokes the `learnVowel` function that can be briefly described in the system's perspective as follows:

Predict: I have heard a sound. How should I configure my vocal tract in order to reproduce that sound? I will estimate a configuration based in my present coarse map.

Generate acoustic vector: Using the estimated configuration from the previous item, I will generate the acoustic representation.

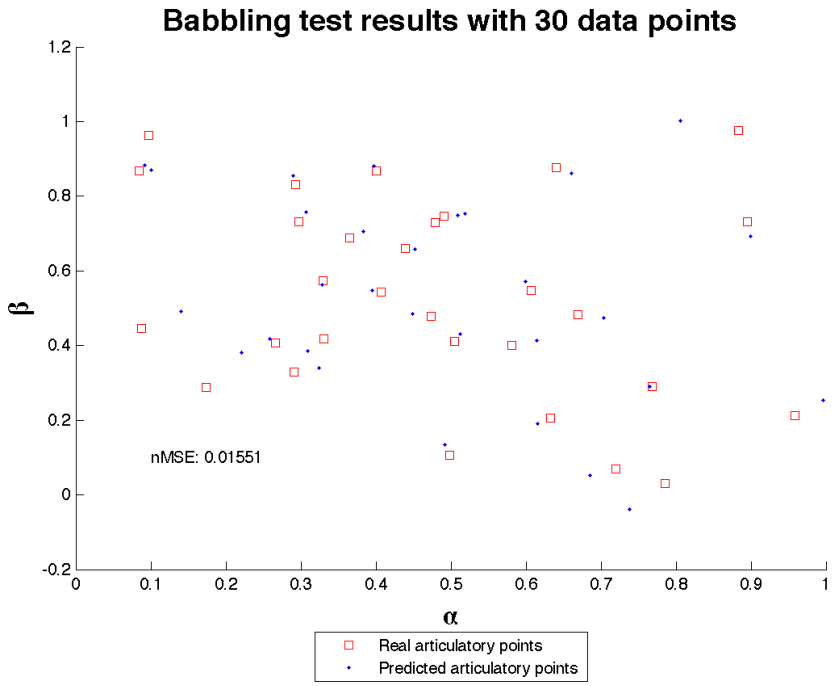


Figure A.3: Babbling stage results on the test set.

Get a closer prediction: I can see the difference between what my teacher told me and what I have produced. Let's try to get closer. I will change my vocal tract configuration in a good possible direction. This way I will refine my map in more interesting areas.

A good possible direction is, for instance, the negative of the gradient of the sensorimotor map, in order to shorten the distance between the teacher's acoustic vector and the system's attempt to reach it.

The differentiable structure of RFWR is very useful at this point. Unlike other methods, like neural networks, with RFWR one can use higher order information from the learning map and converge more easily to an acceptable result.

A diagram for this development stage is shown in Figure A.4.

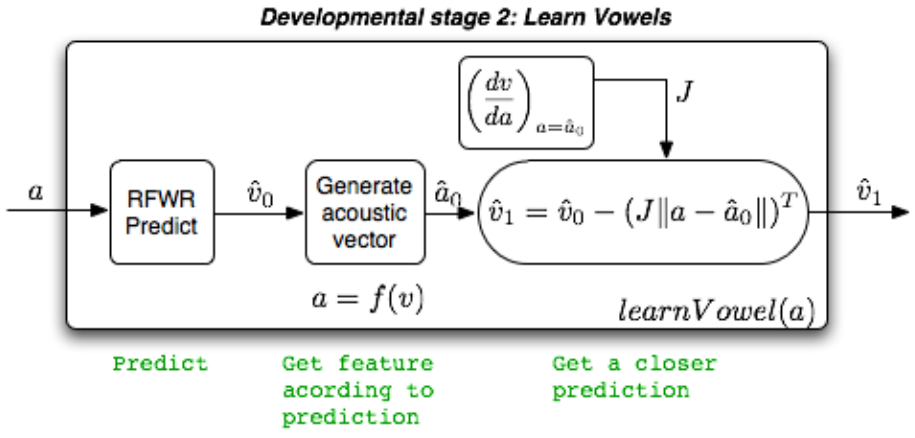


Figure A.4: Diagram representing the main steps involved in learning vowels of a specific language.

To calculate the derivative shown in the diagram of Figure A.4, one must recall the map defining equations A.5 A.1 and A.7, reproduced bellow, in equations A.10, A.11 and A.11:

$$\hat{\mathbf{v}} = \frac{\sum_{k=1}^{K} w_k \hat{\mathbf{v}}_k}{\sum_{k=1}^{K} w_k} \tag{A.10}$$

$$w_k = \exp\left(-\frac{1}{2}(\mathbf{a}-\mathbf{c}_k)^T \ \mathbf{D}_k \ (\mathbf{a}-\mathbf{c}_k)\right) \tag{A.11}$$

$$\hat{\mathbf{v}}_k = \tilde{\mathbf{a}}^T \beta_k \tag{A.12}$$

Let $f(a)$ stand for the numerator and $g(a)$ for the denominator in equation A.10. Then

$$\begin{aligned} \frac{dv}{da} &= \frac{df}{da} \cdot \frac{1}{g(a)} + f(a) \cdot \frac{d}{da}\left(\frac{1}{g(a)}\right) \\ \frac{df}{da} &= \sum_{k=1}^{K} \frac{dw_k}{da} \cdot \hat{v}_k + w_k \cdot \frac{d\hat{v}_k}{da} \\ \frac{d}{da}\left(\frac{1}{g(a)}\right) &= \frac{\sum_{k=1}^{K} w_k \cdot (a-c_k)^T \cdot D_k}{\left(\sum_{k=1}^{K} w_k\right)^2} \end{aligned}$$

and the resulting derivative is shown in equation A.13.

$$\frac{dv}{da} = \frac{\sum_{k=1}^{K} w_k \cdot \sum_{k=1}^{K} \left(w_k \cdot (a-c_k)^T \cdot D_k \cdot \hat{v_k} + w_k \cdot \beta_k^T\right) + \sum_{k=1}^{K} \left(w_k \hat{v}_k \cdot \sum_{k=1}^{K} w_k \cdot (a-c_k)^T \cdot D_k\right)}{\left(\sum_{k=1}^{K} w_k\right)^2} \tag{A.13}$$

The use of this differential structure is quite new. The author has no knowledge that such approach has ever been used in the literature.

Nevertheless, some deeper study must be made in order to compare the performance of this differential approach with a more straightforward one.

Bibliography

[1] M. Lungarella, G. Metta, R. Pfeifer, and G. Sandini, "Developmental robotics: a survey," *Connection Science*, vol. 15, no. 4, pp. 151 — 190, 2003.

[2] R. S. Snell, *Clinical Neuroanatomy for Medical Students*. Little, Brown and Company, 1992.

[3] H. Gray, *Anatomy of the Human Body*. Philadelphia: Lea & Febiger, 1918. [Online]. Available: http://www.bartleby.com/107/

[4] V. Gallese, L. Fadiga, L. Fogassi, and G. Rizzolatti, "Action recognition in the premotor cortex," *Brain*, vol. 119, no. 2, pp. 593–609, 1996. [Online]. Available: http://brain.oxfordjournals.org/cgi/content/abstract/119/2/593

[5] F. Hamzei, M. Rijntjes, C. Dettmers, V. Glauche, C. Weiller, and C. Buchel, "The human action recognition system and its relationship to broca's area: an fmri study," *NeuroImage*, vol. 19, no. 3, pp. 637–644, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/B6WNP-48J44GM-M/2/b15b65bb826b176a1d8a51dbec551cc7

[6] A. M. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, no. 1, pp. 1–36, October 1985. [Online]. Available: http://dx.doi.org/10.1016/0010-0277(85)90021-6

[7] F. H. Guenther, S. S. Ghosh, and J. A. Tourville, "Neural modeling and imaging of the cortical interactions underlying syllable production," *Brain and Language*, vol. 96, no. 3, pp. 280–301, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/B6WC0-4GP7P1X-2/2/2707dea48a65e9b46ce620018a79bdc3

[8] S. Maeda, *Speech production and speech modeling*, ser. NATO ASI Series. Dordrecht, Netherlands: Kluwer Academic Publisher, 1990, ch. Compensatory articulation during speech, evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model, pp. 131 — 149.

[9] M. Lopes and J. Santos-Victor, "Visual learning by imitation with motor representations," *Systems, Man and Cybernetics, Part B, IEEE Transactions on*, vol. 35, no. 3, pp. 438–449, 2005.

[10] A. Giovanni, M. Ouaknine, and R. Garrel, "Physiologie de la phonation," in *Encyclopédie Médico Chirurgicale*. Elsevier SAS, 2003, oto-rhino-laryngologie 20-632-A-10.

[11] A. Andrade and M. Viana, *Introdução à Linguística Geral e Portuguesa*, ser. Linguística. Lisbon, Portugal: Editorial Caminho, 1996, ch. Fonética, pp. 115 —167.

[12] I. R. Titze, *Principles of Voice Production*. Eaglewood Cliffs, New Jersey: Prentice Hall, 1994.

[13] M. R. Delgado-Martins, *Introdução à Linguística Geral e Portuguesaa Linguística Geral e Portuguesa*, ser. Linguística. Lisbon, Portugal: Editorial Caminho, 1996, ch. Representações da linguagem verbal, pp. 85 — 102.

[14] D. Jones, "An english pronouncing dictionary," in *Daniel Jones: Selected Works*. London: Routledge, 1917.

[15] I. P. Association, *Handbook of the International Phonetic Association*. Cambridge: CUP, 1999.

[16] H. Traunmüller, "Cross-modal interactions in visual as opposed to auditory perception of vowels," in *Proceedings of Fonetik 2006, the XIXth Swedish Phonetics Conference*. Department of Linguistics, Lund University, 2006, pp. 137—140.

[17] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech, and Signal Processing* [*see also IEEE Transactions on Signal Processing*], *IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, 1980.

[18] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction." *Science*, vol. 290, no. 5500, pp. 2319–2323, December 2000. [Online]. Available: http://dx.doi.org/10.1126/science.290.5500.2319

[19] J. Kleinberg, "An impossibility theorem for clustering," in *Proc. of the16th conference on Neural Information Processing Systems*, 2002.

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

[21] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, 2001. [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=4391834&site=ehost-live&scope=site

[22] S. Schaal and C. G. Atkeson, "Constructive incremental learning from only local information," *Neural Computation*, vol. 10, no. 8, pp. 2047–2084, 1998. [Online]. Available: http://neco.mitpress.org/cgi/content/abstract/10/8/2047