**UNIVERSIDAD AUTONOMA DE MADRID**

**ESCUELA POLITECNICA SUPERIOR**



# PROYECTO FIN DE CARRERA

## Pronunciation Training of Swedish Vowels Using Speech Technology, Embodied Conversational Agents and an Interactive Game

David Lucas Escribano

Junio 2008

**UNIVERSIDAD AUTONOMA DE MADRID**

**ESCUELA POLITECNICA SUPERIOR**



# PROYECTO FIN DE CARRERA

## Pronunciation Training of Swedish Vowels Using Speech Technology, Embodied Conversational Agents and an Interactive Game

David Lucas Escribano

Junio 2008

# Pronunciation Training of Swedish Vowels Using Speech Technology, Embodied Conversational Agents and an Interactive Game

AUTOR: David Lucas Escribano
TUTOR: Preben Wik
PONENTE: Doroteo Torre Toledano

Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid

Junio 2008

# PROYECTO FIN DE CARRERA

**Título:** *Pronunciation Training of Swedish Vowels Using Speech Technology, Embodied Conversational Agents and an Interactive Game*

**Autor:** David Lucas Escribano

**Tutor:** Preben Wik

**Ponente:** Doroteo Torre Toledano

**Tribunal:**

**Presidente:** Joaquín González Rodríguez

**Vocal:** Pablo Haya Coll

**Vocal secretario:** Doroteo Torre Toledano

**Presidente suplente:** Javier Ortega García

**Vocal suplente:** Michael O'Donnell

**Fecha de lectura:**

**Calificación:**

# MASTER THESIS

**Title:** *Pronunciation Training of Swedish Vowels Using Speech Technology, Embodied Conversational Agents and an Interactive Game*

**Author:** David Lucas Escribano

**Tutor:** Preben Wik

**Ponente:** Doroteo Torre Toledano

**Examiner:**

**President:** Joaquín González Rodríguez

**Vocal:** Pablo Haya Coll

**Vocal secretary:** Doroteo Torre Toledano

**Secondary president:** Javier Ortega García

**Secondary vocal:** Michael O'Donnell

**Date of exposition:**

**Mark:**

**Abstract:**

*Learning pronunciation in a foreign language is difficult. The native pronunciation of the new language is always a hard task for non native speakers. This thesis intends to help the pronunciation training by giving an immediate visual feedback to the speaker. A moving 3D ball in a 3D environment moves around a canvas in response to the pronunciation of the user. An Embodied Conversational Agent will guide the user to calibrate the voice parameters of the user. A normalization method is developed which makes the software work independent of the size of the users vocal tract. Included is also an interactive game to encourage the user to practice the pronunciation.*

*Several tests were performed by users and an analysis from the results obtained is included. A special non Swedish tester used the pronunciation trainer for longer time. Diagrams with the results of the tests are shown in the study. Improvements in the pronunciation and a discussion about the efficiency of the method are also presented.*

**Resumen:**

*Aprender la pronunciación de un nuevo lenguaje extranjero es difícil, La pronunciación nativa del nuevo lenguaje es siempre una tarea dura para los hablantes no nativos de la lengua. Este proyecto de fin de carrera pretende ayudar a la mejora de la pronunciación de la lengua extranjera dando un feedback inmediato al hablante en su pronunciación. Una pelota en 3D es capaz de moverse en un entorno 3D a través de un plano con tan solo la pronunciación del hablante. Un agente embebido conversacional será el encargado de guiar al usuario en la etapa de calibración de los parámetros de la voz. Se ha desarrollado también un método de normalización que hace el software independiente del usuario. Además se incluye un juego interactivo para motivar al usuario a seguir entrenando su pronunciación.*

*Se incluyen varios test realizados por usuarios así como análisis del os resultados obtenido como consecuencia de éstos. Un usuario no nativo de la lengua usó el software durante más tiempo. Se incluyen diagramas con todos los test realizados por los usuarios, además de comparaciones entre éstos y el usuario que ha realizado un mayor número de test. Se prestan como objeto de estudio las mejoras en la pronunciación y la eficiencia del método propuesto.*

Palabras clave:

Pronunciación, formantes, espectrograma, normalización, calibración, ECA, 3D, feedback.

Keywords:

Pronunciation, formants, spectrogram, normalization, calibration, ECA, 3D, feedback.

# ACKNOWLEDGEMENTS

This master's thesis could not be written without the great help of my family. My parents and my lovely sister, who supported me during all this year, and stayed there for everything that I needed during my first year abroad. But also my friends, my very close friends played an important role, because all of them made the person which is writing this. Thank you.

I do not forget my corridor mates, because later all of you became my big family during this year (Aida, from Spain, thanks for being so sweet always with me; Bea, also from Spain, keep on having fun always; Derrick, from Singapore, thanks for smiling every day; Fernando, the most loudly Spanish speaker; Florian, the German funniest person; Francesca, from Italy, my closest person this year; Hassan and Aquib, where are the Pakistanis guys if I switch off the light?; Jerôme, my French brother, I will visit the Paris suburbs; Maral, from Canada, I love your banana cake; Nico, the French mountain climber; Paco, of course, the best Spanish chef; Paola, from Italy; and Viktoria, my personal Swedish teacher, from Ukraine).

In the academic field I have to thank Preben for all the support, not only with the thesis itself, where the help was constant and very productive, but also for the chance I got to spent one of my best years in my life. Thanks for everything. Last but not least, thanks to Jenny and Teodor, partners during most of my work in the Speech, Music and Hearing department in KTH, whom shared with me my software and report problems, and gave me a great help.

David.

# INDICE - OUTLINE

# LIST OF FIGURES

# LIST OF TERMS

**Spectrogram.** The result of calculating the frequency spectrum of windowed frames of a compound signal. It is a three-dimensional plot of the energy of the frequency content of a signal as it changes over time.

**Khz.** Hertz is the International System of Units base unit of frequency. Its is cycle/s or $s^{-1}$. Kilohertz = $10^3$ Hz.

**dB.** The decibel is a logarithmic unit of measurement that expresses the magnitude of a physical quantity relative to a specified or implied reference level.

**Sampling.** Is the act of taking a portion, or sample, of one sound recording.

**Resonance.** Is the potentially deadly tendency of a system to oscillate at maximum amplitude at certain frequencies.

**API.** An application programming interface (API) is a source code interface that an operating system, library or service provides to support requests made by computer programs.

**Widget.** Window in a graphical user interface that has a particular appearance and behaviour. Includes buttons, scrollbars, menus and text windows.

**Cardinal vowels.** Set of reference vowels used by phoneticians in describing the sounds of languages. A cardinal vowel is a vowel sound produced when the tongue or the jaw are in an extreme position, either front or back, high or low.

**Calibration.** Process of establishing the relationship between a measuring device and the units of measure.

**Broker.** Intermediate between two points in a connection which takes care of administrate the requests that they produce.

**Bark scale.** Is a psychoacoustical scale proposed by Eberhard Zwicker in 1961.

# 1. INTRODUCCIÓN

La comunicación es una de las necesidades básicas en la raza humana, pero desde que nace más de un lenguaje, ésta se convierte en un delicado problema. Cada lenguaje no es sólo una mera colección de palabras, entonaciones, símbolos o distintas reglas que le describen. Cada idioma usa incluso distintos músculos en la boca, con la obvia repercusión de una pronunciación diferente.

Normalmente es muy difícil aprender nuevos idiomas, debido al enorme número de palabras, expresiones y frases hechas, pero si además nos centramos en el hecho de querer adquirir una pronunciación nativa en la nueva lengua, esta tarea se convierte en casi imposible.

Las personas que crecen con una determinada lengua nativa especifica aprenderán y desarrollaran el uso apropiado de los músculos de la boca de una forma natural, lo que les conducirá por tanto a una correcta pronunciación nativa de la lengua. De alguna forma, esta temprana adquisición del uso de los músculos en la boca permanecerá así para el resto de nuestra vida. Esto nos proporcionará la denominada pronunciación nativa, es decir, la aprendida desde la niñez. Los problemas aparecen cuando personas que previamente han asentado una pronunciación nativa en su lengua, intentan adaptar su pronunciación también a una pronunciación nativa en una nueva lengua diferente.

Hoy en día es muy común que la gente viaje alrededor del mundo, trabaje en diferentes países o estudie fuera de su propio país durante determinado tiempo. Como se ha mencionado anteriormente, cada lenguaje posee sus propias cualidades, por lo tanto alguien que no esté acostumbrado a ellas y las quiera asentar como segunda o sucesivas lenguas, encontrará muchos problemas intentando alcanzar la considerada pronunciación nativa, incluso si prolonga los estudios del nuevo lenguaje durante varios años.

La percepción del nuevo lenguaje es otra parte critica para estudiantes de nuevas lenguas. En algunos idiomas, los estudiantes pueden tener problemas incluso al reconocer diferencias entre distintos sonidos vocálicos. Este hecho se convierte en una barrera para el aprendizaje de la nueva pronunciación, debido a que no serán capaces de hacer las correcciones adecuadas a la pronunciación si no son conscientes de las diferencias claves entre los distintos sonidos.

El objetivo de este proyecto de fin de carrera por tanto, es ayudar al estudiante de nuevas lenguas a adquirir una mejor pronunciación en el nuevo lenguaje, al menos en el punto de vista de los sonidos vocálicos. El campo de estudio en nuestro caso será la pronunciación de los diferentes sonidos vocálicos suecos, aunque demostraremos más adelante que se puede aplicar fácilmente a sonidos vocálicos en la mayoría de los lenguajes.

Existen muchas aplicaciones que se preocupan de las necesidades en la comunicación para personas discapacitadas, tales como el uso de síntesis de diálogo en máquinas lectoras, lectores de pantallas para personas con problemas de visión, o diferentes prótesis para personas mudas. Aún así, existen todavía numerosas soluciones diferentes que pueden minimizar el problema y proporcionar una gran ayuda para este tipo de personas con incapacidad. Por lo tanto, este proyecto pretende proporcionar esa ayuda para ciertos grupos de gente con problemas de audición. Este proyecto pretende establecer un punto de referencia en el campo de la pronunciación vocálica, no sólo para hablantes no nativos, sino también para gente con un determinado problema en alcanzar la pronunciación correcta en ciertos sonidos vocálicos, debido por ejemplo a una incapacidad.

Para desarrollar nuestro estudio nos basaremos en el método de las formantes. De hecho este método es el usado más comúnmente en la discriminación entre vocales, con multitud de estudios y publicaciones sobre teoría de formantes. A menudo estos métodos se centran en el primer y segundo formante, pero en el caso de las vocales suecas, estas dos dimensiones no serán suficientes para localizar todos los sonidos vocálicos suecos.

El problema empieza en nuestro caso en la diferencia entre los alófonos [i] e [y] en el idioma sueco, donde tan sólo el primer y segundo formantes (se explicará con detalle el término formante en el cuerpo del proyecto) no nos da una clara distinción entre ambos. Por esta razón se convierte en necesario extraer también el tercer formante en el análisis del sonido, para hacer una discriminación correcta entre los distintos sonidos vocálicos. Esta razón nos lleva a desarrollar un sistema capaz de distinguir entre esos dos fonemas en particular, El tercer formante será extraído y procesado para establecer esta diferenciación. Como podremos ver en el cuerpo de este proyecto, en el lenguaje sueco, el solapamiento solamente ocurre entre estos dos fonemas [i] e [y], por tanto esta extracción del tercer formante solo será usada satisfactoriamente para diferenciar ambos. Sin embargo, en el resto del mapa vocálico para las vocales suecas, como podremos ver en la sección

2.4, los diferentes sonidos vocálicos son fácilmente diferenciables sólo con la extracción del primer y segundo formante.

Por otra parte, cada persona tiene diferentes parámetros en su voz, por lo tanto, también tendrá diferentes valores en los valores de los formantes y en la frecuencia fundamental de su propia voz. Por esta razón es importante normalizar estos parámetros. Un método capaz de ajustar los diferentes parámetros de la voz de distintos hablantes, para su posterior análisis, es objeto de estudio e investigación [1]. La normalización es un punto principal en nuestro estudio, y secciones posteriores mostrarán que ésta se convertirá en una parte critica cuando se desarrolla el software propuesto para el entrenamiento de la pronunciación.

Antes de desarrollar la normalización, es necesaria previamente una etapa de calibración, donde se recogen los parámetros de la voz del hablante. La etapa de calibración así como la normalización también ocupará un papel importante en nuestro proyecto, debido a que es el primer paso para hacer el software independiente de cada usuario. Por lo tanto necesitamos un método de calibración lo más robusto posible, para así obtener los datos mas fiables posibles por parte del usuario.

Inicialmente, este proyecto estaba planeado como una útil extensión de VILLE, un software desarrollado para aprendizaje de idiomas en el departamento de Speech, Music and Hearing en KTH, Estocolmo. VILLE es un agente embebido conversacional con el que puedes hablar y obtener respuestas por su parte. Este agente guía y da una respuesta a cualquiera que desee mejorar sus habilidades en la nueva lengua. En cambio, el tiempo ha hecho de este proyecto un software independiente, de modo que podemos usarlo enteramente para entrenar la pronunciación de las distintas vocales suecas.

## *1.1. Objetivos de estudio*

El principal objetivo de este proyecto ha sido desarrollar un software capaz de dar una respuesta inmediata al hablante cuando éste pronuncia los diferentes fonemas de las vocales suecas. El reto ha sido crear una región en 3D donde una pelota es capaz de moverse por un plano a través de la pronunciación del hablante. Más allá, se pretende que ese movimiento sea con la suficiente suavidad como para no contener errores aislados en la extracción de las formantes, y ser aún capaz de proporcionar la respuesta instantánea deseada.

Se demandaba crear un método eficaz de normalización de modo que se pudiera representar diferentes voces en el mismo plano mencionado anteriormente, independientemente de la edad o el sexo del hablante.

Otra tarea era también añadir un agente conversacional que proporcione interacción entre el ordenador y el usuario, especialmente útil en la etapa de calibración.

Además otra labor era crear un juego interactivo similar a [2], donde se desarrolla también un juego interactivo para la mejora en la pronunciación de las vocales finlandesas. Este juego debe ser capaz de medir la mejora en la pronunciación del usuario, y después posibilitar la comparación y análisis de los diferentes resultados en el tiempo usado en completar el juego.

La última demanda era recoger los datos obtenidos de los tests realizados a los usuarios, para averiguar que realmente se ha producido alguna mejora palpable en la pronunciación de los fonemas suecos.

## *1.2. Resumen de la memoria*

Esta memoria se divide en 7 capítulos diferentes:

Este primer capítulo contiene esta breve introducción al proyecto.

El segundo capitulo da una profunda presentación del estado del arte relacionado con nuestro estudio. Se describe una breve introducción a la fonética y al método de las formantes. Explicaremos el por qué del uso de este método para clasificar los fonemas vocálicos en el mapa y discutiremos las limitaciones que este método posee. Se darán además métodos alternativos a este método de las formantes usados a su vez para el mismo objetivo. Al final de este capítulo se introducirá al agente conversacional, explicando las principales cualidades que éste posee.

En el tercer capítulo se presenta el objetivo de este estudio, así como una descripción del paso a paso del método utilizado en él. Se presentarán con detalle los problemas que se han encontrado además de nuevas partes que se convertirán en críticas en el posterior desarrollo del proyecto. También se introduce Open GL, usado para la parte de gráfico sen 3D, así como TCL, usado como lenguaje de programación en nuestro software.

El capítulo cuarto describe el software como si de un manual al usuario de tratara, exponiendo las múltiples características que éste posee, así como los distintos modos de operación que se incluyen en él.

El quinto capítulo contiene la etapa de recogida de datos, además de las posibilidades de análisis que éstos ofrecen al investigador.

Finalmente los capítulos seis y siete contendrán las conclusiones y el trabajo futuro de este estudio. Se expondrán y analizarán los datos obtenidos a partir de los tests realizados, así como de las pequeñas encuestas realizadas a los usuarios que pretenden analizar el grado de satisfacción del usuario en la interacción con el ordenador.

# 1. INTRODUCTION

Communication is one of the basics of the human nature, but since more than one language exists, it becomes a problem. Each different language is not just a huge collection of new words, intonations, symbols or even different rules which describe the behaviour of the language, each language even uses different muscles in the mouth, with the obvious relevance of different pronunciation.

It is commonly very hard to learn diverse languages, because of huge vocabularies, expressions and common phrases, but also we know that trying to achieve a perfect native pronunciation different from your mother tongue becomes an almost impossible task.

People who grow up with a specific mother language will learn and develop the particular muscles of the mouth in a natural way, which obviously will lead them to pronounce the mentioned pile of words, symbols and rules correctly. Somehow, this early developed 'configuration' of the muscles remains in our mouth for the rest of the life. This is the pronunciation that we denominate native for each language; the one which we have been learning naturally since our childhood. Problems occur when people who have previously settled their native pronunciation, try to change their mouth structure in order to achieve a new native pronunciation in a new language.

Nowadays, it is very common that people travel around the world, work in different countries or study abroad during some time. As mentioned earlier, each language has different pronunciations, so one who is not used to speak that language as his/her mother tongue, will face a lot of problems to reach the considered 'native pronunciation', even if he or she has been studying or speaking the language for several years.

Another critical issue that emerge for students while trying to learn a new language is the perception. In some languages students may have difficulties in recognizing differences between different vowel sounds. This fact is also a barrier for the pronunciation learning, because if they cannot perceive the key differences, they will not be able to make the proper corrections to their own pronunciation.

The aim of this thesis is to help people to achieve a better pronunciation in a new language different from their mother tongue, at least in view of the vowels pronunciation. The field of the study will be the Swedish vowels, but we will demonstrate that this study can be extended to include all the different vowel sounds in almost any language.

There are already many applications which take care of the communicative needs for disabled persons, such as the use of speech synthesis in reading machines and screen readers for visually impaired persons, and a speech prosthesis for non-vocal persons. But still there are a lot of different solutions that can minimize the problem and provide a great help. Therefore, this thesis also intends to provide an useful help for certain groups of impaired people. This master's thesis aims to become a reference in the field of vowel pronunciation training, not only for non-native speakers, but also for some people with a defined problem in reaching the correct pronunciation, or for those that due to a handicap, cannot achieve the native pronunciation of certain vowel sounds.

To perform this task, the formants method will play a fundamental role in this master's thesis. A lot of papers and studies have been conducted about formants theory, as a matter of fact this theory has become the major method to discriminate the vowels in speech technology. All these studies are often focused on the first and second formant, but in case of Swedish vowels, a two dimensional vowel chart is not enough to find out which vowel has been pronounced by the speaker.

The main issue in this case is the difference between the close front allophones [i] and [y], where the first and second formants do not give us a clear distinction between the phonemes mentioned above. For this reason we need to extract also the third formant in the sound analysis, in order to make a proper discrimination of the different vowel sounds. This reason lead us to develop a system able to distinguish these two particular phonemes. The third formant will be extracted and processed in order to make the proper discrimination. As we can see in the body of this thesis, in the Swedish language, the overlapping in the vowel discrimination occurs just between those particular phonemes [i] and [y], therefore this extraction will be useful used just in the difference between them. However, in the rest of the vowel chart for the Swedish vowels, as we can see in the section 2.4., the vowels phonemes are easy to classify with only the first and second formant.

On the other hand, each speaker has different voice parameters, thus, will have different frequency pitch and different values in the formant

frequencies. For this reason it is important to normalize this parameters of the speakers. A method able to fit the voice parameters independent from the speaker, and later on being able to process the data obtained is subject of research and study [1]. Normalization is a main point in our study, and the sections below will show that this becomes a critical part when the software is developed.

To develop the normalization, the software needs a stage of calibration, where it collects the voice parameters from the user. The calibration stage will also play an important role in the project, because it is the first step to make the software independent from the user. Thus, we need a robust calibration method, in order to get reliable data from the user.

Initially, this thesis was planned to be an extension of VILLE, a software developed for language learning at the department of Speech, Music and Hearing, at KTH. VILLE is an embodied conversational agent (ECA) that you can talk to, and that talks back to you. An agent that guides, encourages and gives feedback to anyone who wish to develop or improve his language skills. However, time has shaped this thesis into a software by itself, so we can use it entirely for the training of pronunciation in the domain of the Swedish vowels.

## *1.1. Goal of the research*

The aim of the project has been to develop a software able to give an immediate feedback to the speaker when he pronounces the different vowel phonemes in Swedish language. The challenge has been to create a 3D region where a three dimensional ball is able to move just with the students pronunciation. Moreover, that movement should be soft enough to not contain spare errors in the formant extraction, and still being able to give to the user instant feedback.

The demand was to create a normalization method able to plot different voices in the same canvas, independent of the age or gender of the speaker.

A talking head providing interaction between human and computer was an extra task which would give an useful help in the calibration guidance.

Another goal was create an interactive game similar to [2], where an interactive game is also developed for training the Finnish vowels pronunciation. This game has to be able to measure the improvement of

the user and later on to make it easy to compare and analyse the different results of the time spent in completing the game.

The last demand was to collect the data obtained from the user's tests in order to find out if any improvement of the pronunciation in the Swedish vowels had occurred.

## *1.2. Outline of the report*

This report is divided into 7 different chapters:

This first chapter including this brief introduction to the project.

The second chapter gives a thorough presentation of the theoretical background related to the study. A brief introduction to phonetics and the method of the formants is described. It is explained why it is chosen this method to classify the vowel phonemes in the vowel chart and discuss the limitations that it has. It gives a small collection of alternative methods used for the same aim. In the end of this part the Embodied Conversational Agent (ECA), is also introduced explaining the main features that facial animation process has.

In the third chapter the purpose of the study, and a step by step description of the work are described. The problems that were encountered and later issues that became critical in the development of the work are presented in detail. There is also an introduction to Open GL, used to perform the 3D drawings, and to TCL, used to program all the software.

The forth chapter describes the software developed as a kind of manual for the user, explaining the main features and the operation of the several modes that were built in it.

The fifth chapter will contain the data collection process, such as the possibilities of analysis that are open to the researcher.

Finally chapters sixth and seven will give us the conclusions and the future work of this study. The data obtained in the tests will be displayed and analysed, and also the data obtained from the research questions in a small questionnaire that intends to analyse the grade of satisfaction of the user in the Human Computer Interaction (HCI).

# 2. THEORETICAL BACKGROUND

## *2. 1. Phonetics*

Phonetics is the systematic study of human speech-sounds. For any person who works in language learning is good to have a basic knowledge of phonetics. Even a language teacher has to be able to diagnose pronunciation errors made by students in order to make the proper corrections, and although there exist a lot of language teachers without this phonetical background, this would be an important and useful qualification [3][4].

Phonetics is also useful in learning various aspects of the mother tongue, such as having a better understanding of orthographic problems and the relationship between the spelling and the spoken language. And also in order to study the numerous dialects and foreign accents.

It is common to classify the phonetics theory into three main groups:

- **Articulatory phonetics** studies the speaker, the production and the articulation of sounds. This is the older branch of the phonetics studies and has reached a considerable level of development. Describe with high precision the articulatory organs for each sound, and also the energy spent to expulse the air from the lungs, the muscle tension in that expulsion, etc.

- **Acoustic phonetics** studies the physical features of the sound and its transmission. Studies the intensity, length, pitch, etc. of the language sounds.

- **Auditory phonetics** studies the sound from the point of view of the listener, and the mechanisms related to the human audition and hearing [5].

### 2.1.1. History of phonetics

The first studies known in phonetics were developed more than 2000 years ago. At that time (about $4^{th}$ century BC) the Indian grammarian Panini researched in phonetic articulation in order to analyze the pronunciation in the songs from the ceremonies and rites. The first modern phonetician was J. Matthias, from Denmark, who wrote *De Litteris* (1586). The mathematician John Wallis classified the vowels

according to their articulation point (1653). The german C. F. Hellwag invented the vocalic triangle (1781), and just ten years later, the Austrian Physic Wolfang Von Kempelen invented the first machine able to produce sounds. The famous german doctor Hermann Helmholtz inaugurated the studies in acoustics phonetics with his book *Sensations of Tone* (1863); Jean Pierre Rousselot was the first researcher in experimental phonetics. He also wrote *Principes de phonétique experimentale* published between 1897-1908.

In the nineteenth century is developed the phoneme theory, by Ferdinand de Saussure. In United States, Leonard Bloomfeld and Edward Sapir made a decisive contribution to the phonetic theory, while Roman Jakobson developed the theory about the universal characteristics of all the phonemic systems. Also Gunnar Fant, professor emeritus of KTH, has made important contributions to phonetics. Gunnar is the inventor of OVE, a cascade formant synthesizer for vocals, and also his more than 200 publications have become a useful library in speech technology [6].

### 2.1.2. Phases of speech

In order to communicate something, such as a word or a sentence, the speaker has to conceptualize neurologically the idea of what to say, and encode it in a form laid down by the grammar of his language. Then the organs of the speaker move in a determined way to produce the message described. They compress or dilate the air, and set it moving in various ways –in rapid puffs, in sudden bursts, in a smooth flow, etc. Then a flow of air from the lungs, joined with a dilation and compression in the vocal tract, produce the sound, which propagates in sound-waves from the speakers vocal tract to the hearer of the listener.

The listener, and anyone within the hearing distance, including the speaker himself, receives the sound-waves and is able to decode the signal and extract the information transmitted [3].

This is a very wide explanation of how the communication occurs. But it is quite enough to understand the process of pronunciation, which is going to be the matter of this thesis.

## 2.1.3. The vocal tract

All sounds of speech are produced in the vocal tract, when considering the vocal tract as the entire respiratory tract from the lungs to the nose, plus the mouth. The vocal tract has to create the conditions to set the air in motion and control the flow of air in ways that generate sounds.

The vocal tract can be considered as a kind of pneumatic device consisting of a bellows, various tubes, valves and chambers whose function is to set the air in motion and control its flow.

The bellows, in our case the lungs, can expand to give a flow of air. There are two tubes leading from the lungs (the bronchi) which unite in a larger tube (trachea).

Near the end of the trachea there is a piston (larynx) that can slide up and down, specially noticeable when you swallow, and usually more prominent in men than in women. Within the larynx there is a valve (glottis) that can be tightly closed or opened in the course of the speech.

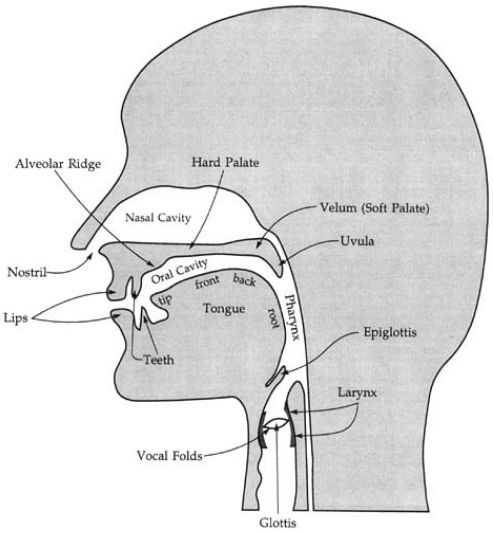



*Figure 1. Scheme of the vocal tract.*

Above the larynx there are three chambers, the pharynx, the oral cavity and the nasal cavity, which can be put into communication with each other by the valves velum, and tongue. The tongue is highly mobile and can control the flow in the mouth in a number of different ways and at different places. Finally the outer end of the mouth has a double valve, named lower and upper lip.
Depending on the movement and position of the several organs in the vocal tract we can classify the phonemes in different ways. For example we can classify the phonemes as fricative if there exists a narrow of passage of air between the tongue and the teeth ([f], [s] or [z] phonemes among others), plosives if there is a silence followed by a suddenly plosion of the air flow, usually generated by the closing of the lips and the sudden opening of them, which creates an explosive burst of air ([b], [k], [t] and [p] phonemes for example), or nasal if the flow of air is also expulsed by the nasal cavity (such as [m] and [n] phonemes). Apart from these examples there exists a whole classification of the phonemes, depending on the position of the tongue in the mouth, or even

if there is a thrill in the vocal folds, such as in the consonant [r]. Hence it is also presented the international phonetic alphabet chart from 2005 in Appendix 9.1 at the end of this thesis, where most of classification of the phonemes are shown.

## *2.2. Formants*

One of several methods to classify the vowels is by the named formants. The formants are maximums in density of energy of the spectrogram in a sound, therefore, they are maximums in energy in a determined frequency, and they corresponds to the resonances in the vocal tract [7].

Normally, the vowel space is reduced to a plane showing the position of the first two formants, but this information is not sufficient to express vowel identity. The higher formants have considerable influence on the identity, especially for the close front vowels [8]. This will be explained with more details below in section 2.4.1.

### 2.2.1. Formant spectrum section

When a formant extraction is done, it is possible to see in a graph the several maximums of energy that exist on each frequency. As can be shown, there is not only one peak of energy. At any given moment of time there exists several peaks of energy or relative maximums. Vowels will, in most cases, have four or more distinguishable formants; sometimes there are even more than six. It can be observed this fact in Figure 2.

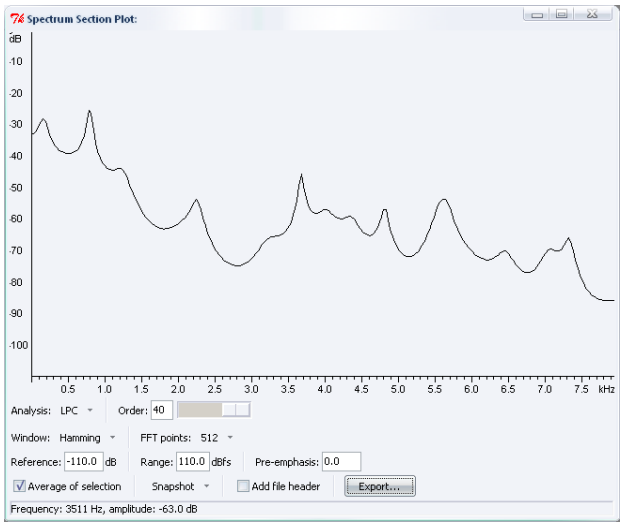



*Figure 2. Formants belonging to the vowel 'a'.*

Figure 2 represents an instant of time when pronouncing the vowel 'a' as in 'mamma'. The horizontal axis represents the frequency values of the sound wave, while the vertical axis represents the values in dB, that can be translated to the amount of energy that a determined frequency has. For example the first relative maximum that appears in the figure corresponds to a frequency of approximately 200 Hz and a value of around −28 dB. This maximum will represent the first resonance frequency, called $f_0$.

### 2.2.2. Spectrogram

When the formant spectrum sections are calculated through time, they create a spectrogram. A spectrogram is a compilation of spectrum sections, where the horizontal axis is the time and the vertical axis is the frequency domain. In a spectrogram the intensity of the graph also has a very important role, because it will determine the energy that a frequency has in an exact moment of time. If we cut the spectrogram at any given moment, we will obtain the formant spectrum section at that time, both looking at the frequency axis and the level of intensity that the graph has for each frequency. Therefore, formants can be seen very clearly in a wideband spectrogram, where they are displayed as dark bands. The darker the colour of the formant reproduced in the spectrogram, the stronger it is (the more energy is there, or the more audible it is).

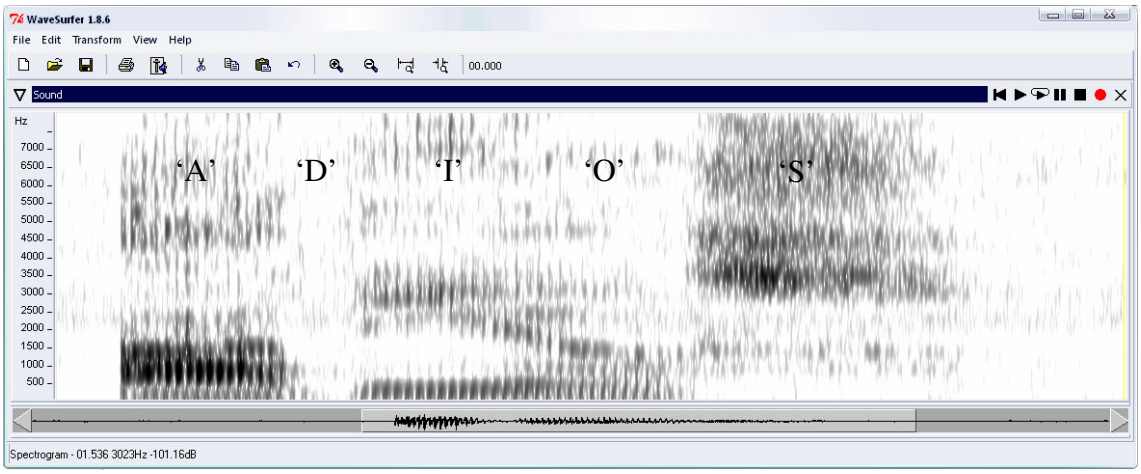



*Figure 3. Example of spectrogram of the Spanish word 'adios'.*

Formants are seen on spectrograms around frequencies that correspond to the resonances of the vocal tract. But there is a difference between oral vowels on the one hand, and consonants and nasal vowels on the

other. For consonants, there are also antiresonances in the vocal tract at one or more frequencies due to oral constrictions. An antiresonace is the opposite of a resonance, such that the impedance is relatively high rather than low. Consequently, formants are attenuated or eliminated at or near these frequencies, so that they appear weakened or are missing altogether when someone look at spectrograms. That is why, for example, it is difficult to see formants below 3000 Hz for the 's' phoneme in the spectrogram shown in the Figure 3 [7].

In addition, for nasal consonants and nasal vowels, the vocal tract divides into a nasal branch and an oral branch, and interference between these branches produces more antiresonances. Furthermore, nasal consonants and nasal vowels can exhibit additional formants, nasal formants, arising from resonance within the nasal branch. Consequently, nasal vowels may show one or more additional formants due to nasal resonance, while even one or more oral formants may be weakened or missing due to nasal antiresonance.

### 2.2.3. Formant extraction is not easy

Looking at Figure 2 one can find more than six peaks of energy, even more than ten. The problem is that not all the maximums in the figure are formants, just some of them. So, how do we know which peak belongs to a formant and which one does not? Formants are relative maximums in energy, not absolute maximums [7]. Therefore in Figure 2, not more than six clear formants are extracted. For example we can see a peak of energy at around 1 KHz and another maximum at approximately 1.2 KHz, but as they are very close to each other we will just consider the first one as a formant, due to a higher energy, and just omit the second one. If we follow this way of analysis we can find formants at 1, 2.3, 3.7, 4.7, 5.6 KHz, and maybe also at 6.5 and 7.4 KHz, which is a total number of seven formants, instead of more than ten if we just look at the maximums in the figure.

This analysis can vary depending on the researcher, or the software used, so maybe some of them will not consider more than five formants in that figure. On the other hand, different software or researchers might even consider another different peak of energy as formant. We can see that this is a subjective area looking at Figure 4.

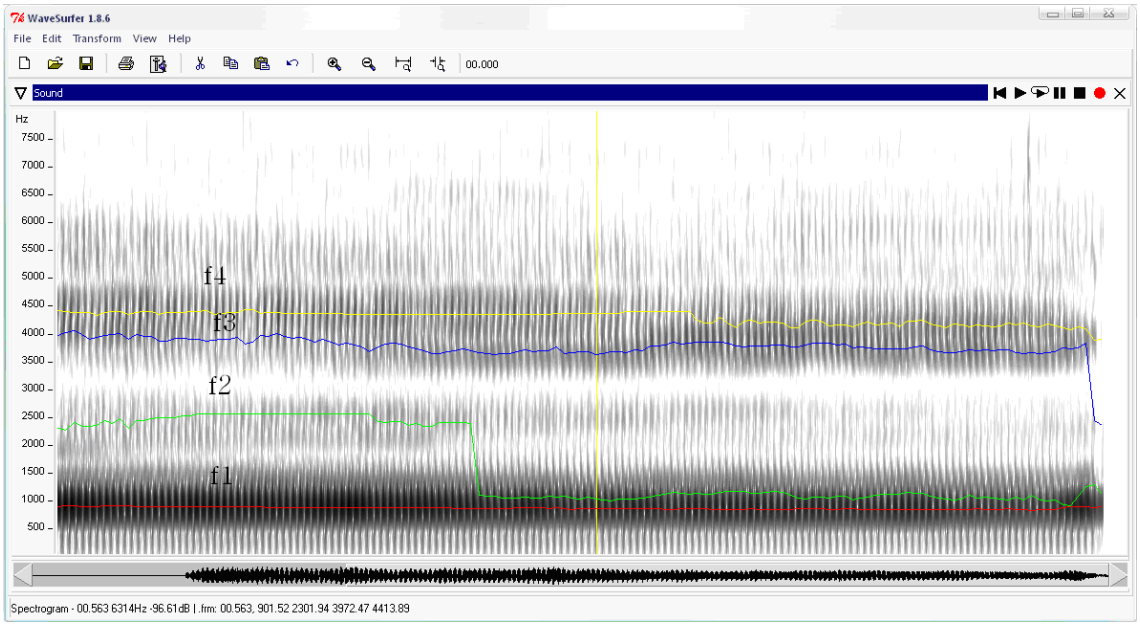



*Figure 4. Change in the second formant for the phoneme [a:].*

Figure 4 is a screenshot of the software WaveSurfer 1.8.6, a program used in audio and speech technology research at the department of speech, music and hearing in KTH. In the figure we can distinguish the four first formants of the vowel phoneme [a:], but at some point we see that the second formant suddenly drops down from around 2300 Hz to 1000 Hz. Why? Because the software detects at that point the relative maximum at 1000 Hz due to a decrease of energy in 2300 Hz, if the energy there now is not big enough, the software does not recognize as a maximum the one at 2300 Hz. Therefore, this fact will condition the analysis in our formant extraction. Hence, we have to rely in Wavesurfer, or Snack (which is the real engine of the interface) in order to trust in the formant extraction that is used in our software.

### 2.2.4. Why formants?

Analysis based on the formants is specially useful in the vowel domain, because the sound is being generated completely in the vocal tract, and does not have oral constrictions that can distort or modify the formants. Some examples of consonant characteristics are:

- Nasals usually have an additional formant around 2500 Hz. The liquid [l] usually has an extra formant at 1500 Hz, while the English "r" sound ([ɹ]) is distinguished by virtue of a very low third formant (well below 2000 Hz).

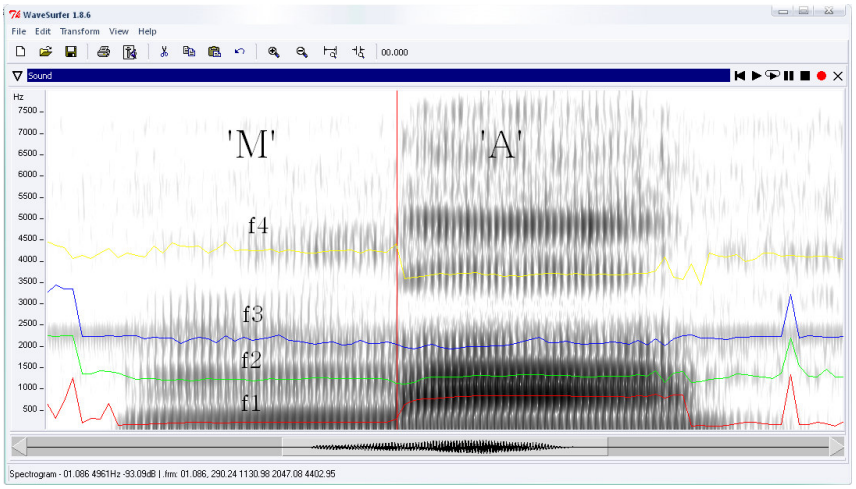



*Figure 5. The additional formant in the nasal phoneme 'm'.*

- Plosives (and, to some degree, fricatives) modify the placement of formants in the surrounding vowels. Bilabial sounds (such as 'b' and 'p' as in "ball" or "sap") cause a lowering of the formants; velar sounds ('k' and 'g' in English) almost always show $f_2$ and $f_3$ coming together in a 'velar pinch' before the velar and separating from the same 'pinch' as the velar is released; alveolar sounds (English 't' and 'd') cause less systematic changes in neighbouring vowel formants, depending partially on which vowel is present. The time-course of these changes in vowel formant frequencies are referred to as 'formant transitions'.

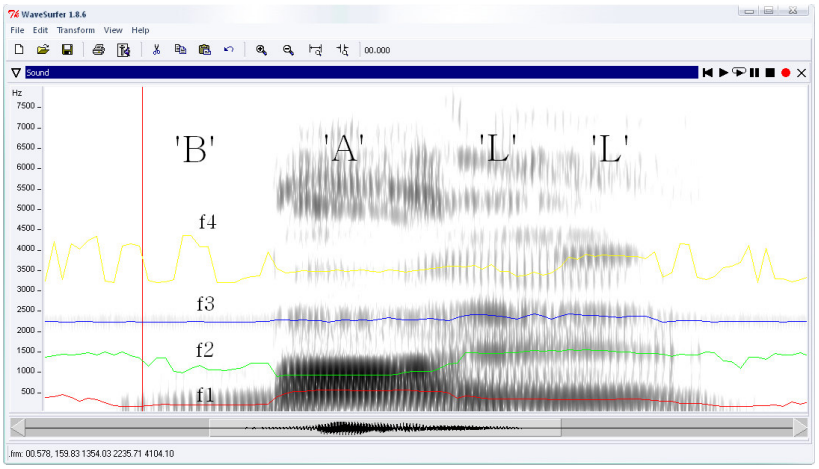



*Figure 6. Example of formant analysis of the word 'ball'.*

## *2.3. Alternative methods to formants*

There are also other methods apart from formants to analyze vowel sounds. Some of them are based in the formant method but with some additional features, others use different tools that lets them to get better, or simply different results in the vowel analysis. It is included a brief description below, with their own characteristics:

- In [9] Dowd et al. describe a method of "*Learning to pronounce vowel sounds in a foreign language using acoustic measurements of the vocal tract as feedback in real time*". An acoustic impedance spectrometer is used to measure the resonances directly in the vocal tract. Also the first and second resonances in this vocal tract are measured. The measurement was made just outside the mouth, in parallel with the free field, using a technique that provides precise information about the acoustic response of the vocal tract in real time. The values previously measured for native speakers for a particular vowel were used as target parameters for subjects who used a visual display as real-time feedback to realise the vocal tract configuration required to pronounce the target vowel. They report the values (R1,R2) for eleven non-nasalised vowels of French. These values are similar to the formant frequencies measured for these vowels, and also their relative positions in the (R2,R1) plane are similar to those of the same vowels in the ($f_2$, $f_1$) formant plane.

  The results of the attempts to imitate six French vowels by monolingual anglophone subjects are reported. One group used a traditional method of learning pronunciation: they heard the vowel sounds and then attempted to imitate them. Another group also heard the sounds, but were assisted by the vocal tract feedback described above when imitating the target sounds. The acoustic properties and recognizability of the vowels were significantly improved when the subjects used vocal-tract feedback.

- In [10] Van Der Stelt et al. describe "*Exploring the acoustic vowel space in two-year-old children*". A frequency domain band filtering analysis method that minimizes the dependency of the results on $f_0$ was developed to measure the spectral envelopes in children's utterances automatically, and was applied to existing utterance data sets of Dutch and Hungarian. One further advantage of the method is that it selects a maximum of 10 measurement points along the length of the utterance. A reference plane is created. Perceptually judged as being correctly pronounced corner vowels of Dutch- and Hungarian-speaking two-year-old boys

were mapped onto this common Dutch-Hungarian reference plane. The band filtering method has shown to be robust with regard to signal-to-noise ratios and to the differences in numbers of measurements.

- "*Second derivative analysis of consonant-vowel transition waveforms*" is described in [11] by Norian. The second derivative of the first cycle of the transition region results in a modulated sinusoidal waveform with a spectral peak that is a function of the site of articulation of the consonant in the consonant-vowel syllable. The velars and alveolars give the lowest and highest spectral peaks, respectively, while the bilabial peaks occupy the mid-frequency range.

- Also in [12] Lee and Soh describe "*Nonlinear Dynamical Analysis of the Vowels in Korean Traditional Folk Songs*". It is used two kinds of vocalization techniques, strongly pressing vibration and frequently pressing vibration, for Korean traditional folk songs were analyzed for four Korean vowels by using nonlinear dynamical analysis methods. Since all the sounds analysed lasted for a long time, more data points are used for calculating dynamical invariants, such as the correlation dimensions. The power spectrum, the phase portrait, the correlation dimension, and the second-order Kolmogorov entropy are investigated. Also the nonlinearity in the dynamics of the time series of the folk song is confirmed by using surrogate data.

- In [13] Carlson et al. also describe "*Some studies concerning perception of isolated vowels*": The objective is to study the feasibility of two parameter model of vowel perception by involving the matching of two-formants synthetic vowels with four formant values, identification tests, and various means of signal transformation and reduction. $f_2$ is substituted by $f_2'$, a mathematical creation calculated from $f_1$, $f_2$ and $f_3$, in order to improve the separation between rounded and unrounded front vowels.

## *2.4. Vowel charts*

With the theory of formants, each vowel sound can be classified by the different values of its formant values, and so we can make a vowel chart with these values, and later use this chart for the aim of our project. We can take as a first reference the vowel chart made by Gunnar Fant in [14] for the three first formants obtained from the mean over 24 male speakers.

| | F1 | F2 | F3 |
|---|---|---|---|
| iː | 255 | 2190 | 3150 |
| eː | 345 | 2250 | 2850 |
| ɛː | 505 | 1935 | 2540 |
| æː | 625 | 1720 | 2500 |
| yː | 260 | 2060 | 2675 |
| øː | 380 | 1730 | 2290 |
| ɑː | 691 | 1228 | 2540 |
| oː | 390 | 690 | 2415 |
| uː | 290 | 595 | 2330 |
| ʉː | 285 | 1640 | 2250 |

| | |
|---|---|
| iː | rit, riv, is |
| eː | ret-, rev, ek |
| ɛː | rät, räv, även |
| æː | här, ära |
| yː | ryt, fy |
| øː | röt, ösa |
| ɑː | rat-, hat, mat |
| oː | råt, hål, kål |
| uː | rot, not, mos |
| ʉː | Rut, ute |

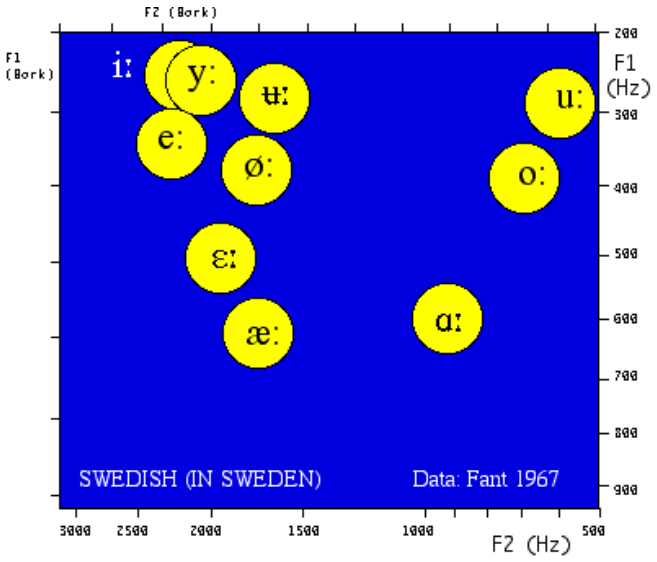



*Figure 7. Vowel chart by Gunnar Fant (1967).*

### 2.4.1. Swedish vowel system

Swedish has a quite rich vowel system, the orthographic base of which is three back vowels (/O/, /Å/ /A/), three front vowels (/I/, /E/, /Ä/), and three rounded front vowels (/Y/, /U/, /Ö/). These occur in pairs of long and short vowels, thus in all 18 phonemes.

Within a pair there usually exists a quality difference, which might be small or absent as in the allophones [ae:] and [ae] of the phoneme /Ä/ and in the allophones [oe:] and [oe] of the phoneme /Ö/. On the other hand, like in the distinction between the allophones [ɑ] and [a] for the long and short /A/ phoneme, have formant patterns significantly different.

This specially bigger difference between the long and short allophones in /A/ will lead us later to take it into consideration in the development of

our software, while the differences between long and short allophones in the rest of the vowel phonemes are not going to be included in our study.

The Figure below shows the typical relations between with the exception of [y:] [15].

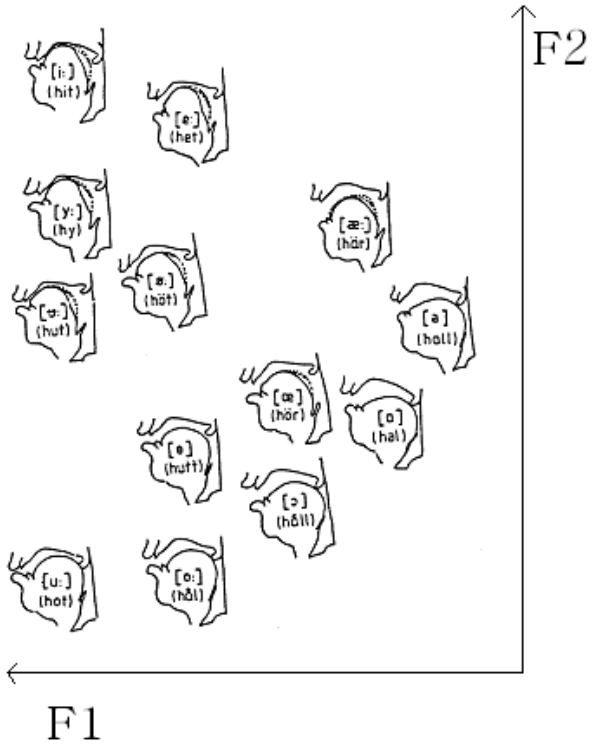



*Figure 8. Lateral view of sustained Swedish vowels.*

## *2.5. Talking heads*

Several studies have demonstrated that in many situations visual signals may be more important than verbal signals [16][17]. Speakers rely on gestures to supply their own speech production. These have a compensatory function in production, often substituting for an unknown word or phrase. Listeners may also be helped by gestures to aid the conversational flow between the speakers and them, and also to facilitate the actual learning experience. It has therefore been explored to use verbal and visual cues to signal prominence, emotion, encouragement, affirmation, confirmation and turntaking [18].

Talking heads provide this visual feedback. They are 3D deformable wireframe objects, controlled by rules [19]. The surfaces of the face can be made (semi)transparent to display the internal parts of the model.

This capability of the model is especially useful in explaining non-visible articulations in the language learning situation.

The internal part includes meshes of the tongue, palate, jaw and the vocal tract walls based on the analysis of three-dimensional MRI data of a reference subject [20].

It is crucial for this talking head that articulations and articulatory movements are natural and also that the timing between the facial and tongue movements are correct. Simultaneous measurements of the face (with optical motion tracking of reflective markers) and tongue movements (with electromagnetic articulography) were used to train the two models in a coherent way.

*Figure 9. Example of Talking head*

This work requires expertise in several areas including man-machine interaction, speech therapy, pedagogy, and computer science. The development is hence made using participatory design that includes all expert areas as well as the students.

The flexibility of the talking heads is a great advantage. The articulatory feedback can be shown using a midsagittal profile with a 2D tongue contour or in 3D, showing the tongue in different reference frames (by changing the visibility or transparency of surrounding articulators), at different scales and from different viewpoints [21][22]. However, this features are not going to be included in our study, but can be part of future work.

# 3. PURPOSE AND METHOD

## *3.1. Purpose*

The main purpose in this thesis is to create a software for training Swedish vowels pronunciation. This software will include a 3D region where a 3D canvas and 3D ball are placed inside. With the pronunciation of the user, the 3D ball will move around the 3D canvas. Different parts of the canvas will represent different vowels, as described in section 2.4. The need of having immediate feedback is also an important goal in this work. When the user talks, the software will show the consequences of his speech. This immediate feedback provides an useful help when errors in pronunciation occur, and will guide the user to correct them easily.

The 3D canvas will represent all the different vowel sounds that the speaker can pronounce. On one hand, with a single vocalic sound, the ball will be placed in a determined place in the canvas. On the other hand, there are some determined places in the same canvas where lie the Swedish vowel phonemes. The aim then is to place the ball in those Swedish places, in order to pronounce the Swedish vowel phonemes.

We can imagine the mouth in a human being as a box able to produce many sounds. In the case of the vowels, we can generate different sounds depending on where the tongue is placed, how much the mouth is open or how round the lips are. If we can draw that box in a 2D or 3D plane, where each dimension is a feature, we can get a map of the sounds that are produced in the mouth. When we have the map of sounds, we can also identify the native sound for a determined vowel, placing it inside of this box as a 'target sound'. This 'box' is the previously mentioned vowel chart.

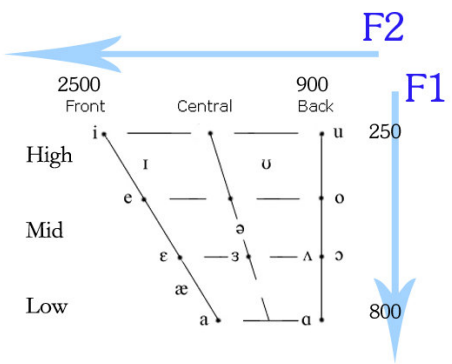



*Figure 10. Example of a vowel chart. The values are expressed in Hz and can vary depending on the speaker.*

Hence, practising for some time, users can train in placing the ball in the 3D canvas, and thus, placing the ball inside of the region where the pronunciation of the Swedish vowels are.

As mentioned in the introduction, each different language has different vowel sounds, even if they are written with the same letter. For example the differences between [i] and [y] differ depending on if we are Swedish native speakers or if we are Viennese German. In the Swedish language the difference remains mainly in the rounding of the lips, and is very slight if we just look at the mouth opening or the tongue placement. In Viennese German these differences are more evident in where the tongue is placed. If it can be identified these differences and those are able to plot it in the 'box of sounds' that is the mouth, it is just needed to locate the new 'target sounds' proper for the new language to have a new vowel chart for training. This is one of the powerful features that this master thesis has to offer [23].

There is a relationship between the formants named in the theoretical background and positions of the tongue and jaw. The first formant has a higher frequency when the jaw is more opened, while the second formant increases when the tongue is placed near the teeth, therefore the back vowels have a lower second formant than front vowels[24].

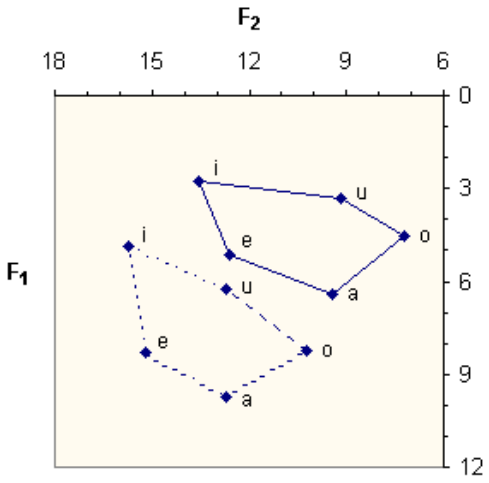
These formant frequencies are not exactly the same for all the people. Men, women, and children, can have different values in Hz for each vowel phoneme. These values can vary with the height, since these resonance frequencies depend very much on how large the vocal tract is [24].

Therefore the 'box of sounds' will vary and be bigger or smaller depending on these formant frequencies. A correct normalization of this 'box of sounds' in our voice parameters is needed. If we are able to normalize these variables we can get a design more or less 'universal', and we could specify easily which sound is produced regardless of the person talking. As Gunnar Fant mentioned in [25], a normalization is needed due to a nonuniform scaling of the female vocal tract with respect to the male vocal tract. But also between children and adults. I will explain more about the normalization method used in this master thesis in the section 3.3.4.

Apart from the variations proper from the same speaker in different moods, there are also variations in speech between different people, (or the same speaker at different ages). The main cause for the differences is the length of the vocal tract, related to the age and the gender of the

speaker. This length increases with the age, and this decreases the value of the formant frequencies [24].

Also the length of the glottis increases with the age, decreasing the value of $f_0$. Extra differences are also observed between children and adults. In the figure below it can be observed a diagram with the average values of the formants for five different Japanese vowels, in children of 4–5 years old (– – –) and adults (———). The diagram has Bark scale and the values of the axis are expressed in hundreds of Hz. One can be surprised when it is noticed that the region for children and adults is so different that even there is not a overlap between them [24].




Average values for F1 and F2 in Hz

| Vocal | i | e | a | o | u |
|---|---|---|---|---|---|
| F1, adult | 274 | 524 | 686 | 458 | 328 |
| F2, adult | 2172 | 1822 | 1158 | 828 | 1108 |
| F1, child | 494 | 958 | 1216 | 954 | 664 |
| F2, child | 3030 | 2788 | 1922 | 1318 | 1918 |

*Figure 11. Depending on the vocal tract, the 'box' of sounds can be bigger or smaller, and can be placed in different formant frequencies, but will keep the relative positions of the vowel sounds between them. This size is settled by the formant values of the three cardinal vowels, /A/, /I/ and /O/. Diagram and data of the table obtained from [24].*

## *3.2. Equipment*

### 3.2.1. Hardware

For the development of this project it has been used a computer with these features:

- HP® Pavilion dv2000.
- Intel® Core Duo™ processor T5500 @ 1.66 GHz.
- 1 Gb RAM memory.
- Hard disk 120 Gb.
- Graphic card Intel® 945GM Express Chipset Family 128 Mb.
- Operating System: Microsoft Windows XP® Professional Media Center Edition (Service Pack 2).

Programming language Tcl/Tk (v.8.4.15) was used for developing the software tool. The Open GL API (v.1.4.0) was used to take care of the 3D graphics, enabling the drawing of the 3D region and moving the ball in 3D.

### 3.2.2. Open GL

Open GL is a specification which describes several methods and tools for developing portable, 2D and 3D graphics applications. Since 1992, Open GL is a consistent and widely available API which lets the programmer hide hardware complexities. Open GL is used in thousands of applications in a wide variety of computer platforms. Open GL accepts each primitive like points, lines or polygons and convert them into pixels. The state machine of Open GL is the one which takes cares of the conversion and from version 2.0 it is fully programmable [26].


*Figure 12.*

### 3.2.3. Tcl and Tk


*Figure 13.*

Tcl is a string-based command language with few fundamental constructs, which makes it easy to learn. The program is interpreted when the application runs, and the interpreter makes it easy to build your application in an interactive manner.

The Tk toolkit is the graphical user-interface. Tk provides a set of Tcl commands that create and manipulate widgets. These widgets are organized as a hierarchy, this means that there is a primary window, and inside that window there can be a number of child windows, just like a folder system [27].

## *3.3. Method*

### 3.3.1. Drawing the 3D region and the ball

As introduced in section 3.1, the first thing that has to be done is to create a 3D canvas where it will be possible to plot the 3D ball, and move it later on.

This part of the project is programmed with the Open GL API, introduced in section 3.2.2. This API lets us program all the graphical part in an easy way, and makes it very intuitive when some changes have to be introduced.

The first for drawing the 3D canvas is to set up the initial values, such as background colour, light features, deep testing or load textures. This 3D canvas will consist on a square panel able to move in 3D motion in order to get some other perspectives.

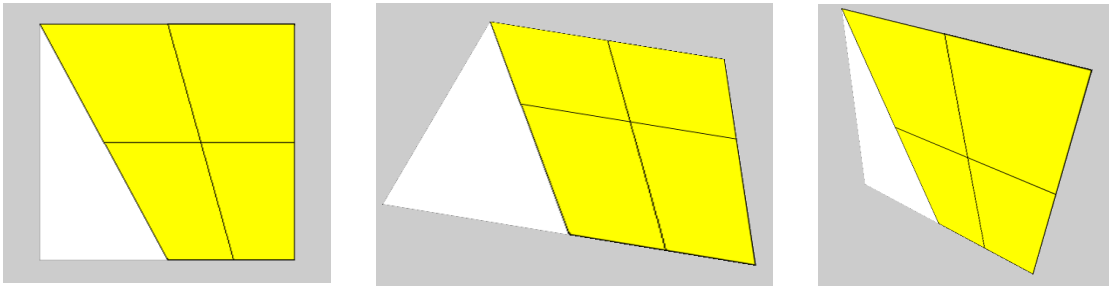


*Figure 14. Some perspectives of the 3D canvas.*

Textures are introduced in the 3D canvas instead of solid colours because in this way one can easily change different combinations of colours and shapes in the canvas, to make it easier to see the movement of the ball introduced in section 3.1. inside of the 3D region. It is possible to rotate the 3D canvas with the mouse in order to move to a different perspective. Due to this, some properties of the faces of the canvas had to be modified. It is selected one face as solid and other transparent. So when one wants to see the default point of view can see the canvas, but when we see the canvas from behind, with the ball being also behind the 3D canvas, this canvas becomes transparent, and the 3D ball is still visible. A part of the drawing code showing more deeply the features for the 3D region is in Appendix 9.2.1.

The ball on the other hand has both faces solid, and of course a different texture to make it distinguishable from the background and the pane. A fun texture is chosen to make the software more attractive and closer to the users. Sample code for this design is written in Appendix 9.2.2.

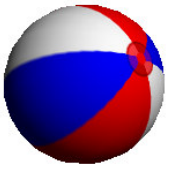


*Figure 15.*

### 3.3.2. Moving the ball

Moving the ball becomes the most critical part of our software. There are a lot of different factors that have to be balanced in order to achieve a natural and soft movement. These factors are based on the length of the portions of sound, extracted directly from the student pronunciation, that are given to the software for being processed.

If this portion is very long, the feedback from the ball is not going to be immediate, because we have to wait to process all the samples from the sound, and extract the formant values for every sample. That will create a slow response, losing the immediate feedback. This fact might confuse the student, when he/she tries to correct the mistakes it will be difficult to appreciate which part of the speech is wrong, due to the delayed response. For this reason, it is important to have an immediate feedback. Immediate feedback will help the student to easier correct the mistakes in the pronunciation, as he/she will perceive on the screen the result of the sound produced instantly.

On the other hand, if the portion of sound selected is not long enough, the movement of the ball is not going to be smooth. Due to the properties of the formants, very slight variations of the sound produced, have a critical role in the formant extraction, and the movement of the 3D ball will become a chain of jumps around the canvas, instead of having the

softness that we want. The refresh rate of the drawings was about 20 frames per second. This gave a balance between the softness and the immediate feedback wanted. To prevent from outliers in the formant extraction, a buffer is added in where the median of the samples obtained from the student is calculated before.

More factors such as choosing the relevant part of the sound that should be processed are important. When the user is asked to produce a sound during the calibration part (one of the three corner vowels), the recording will contain both silence and speech. Due to this, a pitch analysis is developed in order to separate voiced from unvoiced speech, and so eliminate the silence from the later formant extraction. Once the silence is deleted, the middle part of the voiced speech is chosen in order to avoid the beginning and the end of the sound, where transitory changes can lead us to wrong values.

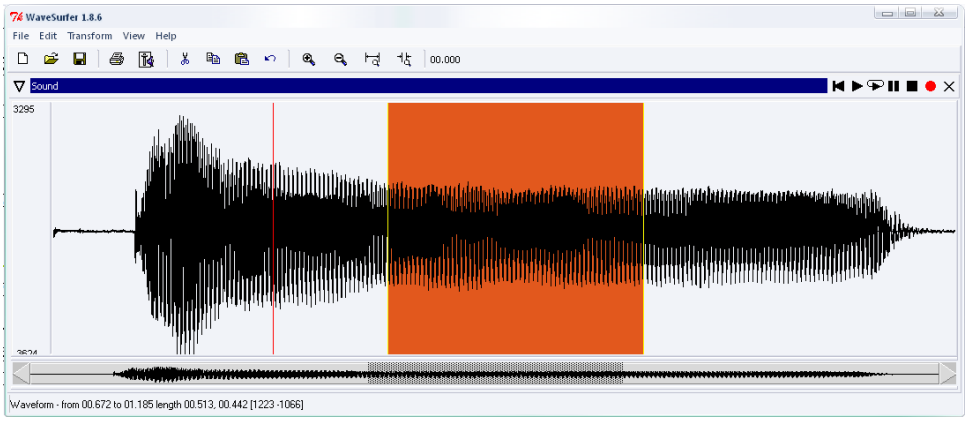



*Figure 16. We can see the piece of the recording that is chosen for the formant extraction highlighted.*

### 3.3.3. Target spheres

This software is being created to learn and train the pronunciation of the Swedish vowels, and this training has to have some target to reach. The student can move the ball around in the 3D canvas, but we also need to establish some places in this region that are going to be the goals. For this reason we also draw some fixed spheres in the 3D canvas, as aim for the students. These spheres are carefully placed in the 3D canvas where the sound of the target vowel is supposed to lie. If the student places the moving ball inside of the fixed sphere, it means that the student is pronouncing the correct sound for the vowel specified. In order to find the correct areas in the 3D canvas, several Swedish native speakers were used. They were asked to pronounce the different vowel

phonemes, and data was analysed and extracted from their recordings, creating a spatial region for each vowel phoneme.
To make it more intuitive for the student, these target spheres are not drawn with a solid colour, but are wired view, so the student can see if the moving ball is inside of the target or not.

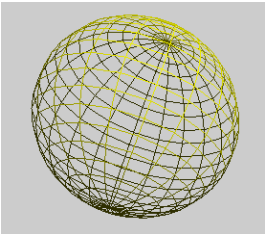


*Figure 17.*

As mentioned in section 3.1, each place in the 3D canvas will represent a single vocalic sound, and all the vocalic sounds can be placed inside of the canvas. Thus, it can be saved the coordinates in the 3D canvas for all the places where the sound pronounced belongs to any vowel phoneme. Also as mentioned before, those places will also correspond to the place where the target spheres are drawn. It is going to be easy to plot the target spheres wherever in the canvas, independent from the language to train, because they will attend to coordinates in relative positions. Therefore, if there exist previous studies with native speakers, and it can be located where the vowel phonemes are placed in the 3D canvas for that language, is easy to make the proper changes and get a pronunciation trainer for another different language (different to Swedish language in this case). This gives the software an interesting flexibility for learning not only Swedish sounds, but any vowel sounds in many different languages, just loading these new coordinates in the 3D region.

The part of the code in Tcl written for this aim is written in Appendix 9.2.3.

### 3.3.4. Performing the normalization

We have talked very broad about normalization, and about the need of having a rule which makes it possible for different people to use the software, regardless of age, height, or gender. Therefore, we intend to have a robust normalization method that allows the software to analyse the data obtained efficiently despite who is using it.

This normalization method not only covers the way to process the data obtained from the user in order to create a normalized map where the 3D ball is going to move around, it also takes care of guide the user for giving the correct data asked. This normalization is so strong part in the software, that later will become essential and will have strong consequences in the future use of the software. Hence, it is crucial to know exactly the parameters of the user's voice in advance, in order to plot correctly the movements of the 3D ball later on. Errors in the

acquisition data from the speaker's voice would ruin the normalized vowel chart and, therefore, would have wrong results when the speaker try to use the software afterwards.

During the calibration process the talking head will ask for some basic data to the speaker, as it is going to be explained in the section 3.3.4.1. The user is expected to answer in a way as close as possible with the question asked before by the agent, and not a random answer or a wrong one made on purpose. Otherwise the data will be wrong and the normalized vowel chart will have also wrong values. Let's explain more carefully this last fact.

### 3.3.4.1 Calibration

The parameters of the voice from the user are needed to be obtained before using the software, in order to process the data acquired properly. The user needs to give to the software some key parameters that later will help in the normalization process. These parameters, first proposed by Daniel Jones in [28], organize the vowel space between the two most extreme tongue body positions: high front [i] and low back [ɑ]. In order to facilitate the calibration, it is used two different phonemes extra, due to the difficulty for some speakers in pronouncing the low back [ɑ] phoneme. Thus, the final vowels selected were the [a], [i] and [u] phonemes, which give us the full range of the mouth. The importance of choosing those three corner vowels is that those phonemes exist in almost every language, so there will be a wide range of people able to pronounce them correctly.

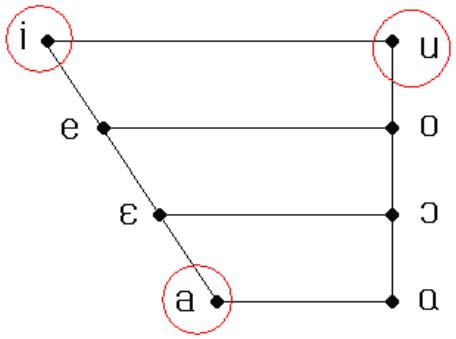



*Figure 18. [a], [i] and [u] can be seen as the corner vowels of our mouth, while the rest of the vowel sounds will be inside of those limit values.*

In the corner vowels there are exceptions, for example for southern British English pronunciation the [u] phoneme is not placed in the top right corner in the vowel chart. In this case, the [u] phoneme is placed between the Swedish [u] and [o]. In this case the right limit for the vowel chart is the phoneme [o:] [29].

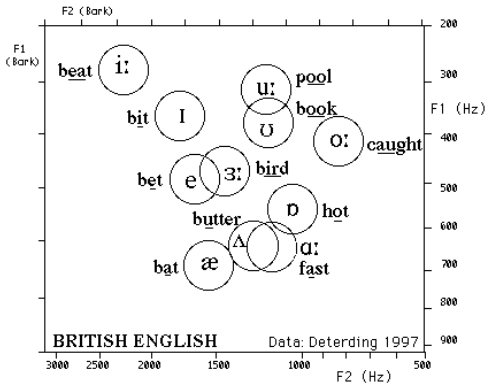



Figure 19. Vowel chart for British English speakers.

The frequency values for the three corner vowels proposed are going to become the edge of the 'box' of sounds, because, as it can be seen in the figure number 10, the formant extraction from the [i] phoneme will enclose the lowest value of the $f_1$ formant and the highest value for $f_2$ formant, the [u] phoneme will be the lowest value for $f_2$, and finally the [a] phoneme will contain the highest value for the $f_1$ formant. Thus, we will extract from these three vowels the maximum and the minimum values for $f_1$ and $f_2$ that our mouth is able to pronounce.

Hence, the user is asked to pronounce those three vowels separately. Sound data is filtered and analysed in several stages. First of all the signal has a threshold to avoid wrong formant extraction due to noise. If the threshold is not reached during the recording, we would obtain the mean between correct formant values, from the speech, and wrong ones, from the silence. Thus, the end value would be erroneous. The result after this stage is to adjust the canvas scale. It will contain the maximum and minimum values of the formants, so we get the dimension of the mouth.

The parameters acquired from the calibration stage will be used in the next normalization process of the data. Therefore, once we have completed the calibration part successfully, we can start to use the software and train the pronunciation in our Swedish vowels.

### 3.3.4.2. Normalization

The normalization itself is based on the calibration due to the personalized vowel chart obtained in the calibration stage. We use the formants frequency limits as reference and as relative maximums of our 'box of sounds'. It has to be pointed out that several methods have been tried in order to improve the throughput of the software, such as Bark

scale [30], or log scale, but the results were of the same kind, and the computational cost was bigger than a linear scale. Hence, it is created a linear normalized vowel chart where the limits are from (-0.5, 0.5) both for the horizontal and vertical axis, representing the first and second formant. This top limits are going to be the maximum or minimum values in the calibration stage for the first and second formants, and the rest of the vowel phonemes are going to be inside of this chart, placed in a relative position.

The idea of the normalization method is that with the calibration we know the relative places where the sounds are for the correct pronunciation of the vowels. It is easy to guess a third vowel phoneme when other phonemes in the vowel chart are known. There exist distance relations in the chart, from other phonemes, or even from the neutral vowel phoneme schwa [ə].

So for example if we know that the [e] phoneme is between the [i] and the [a] in the vowel chart, using the phonemes of [i] and [a] of the user, as reference obtained from the calibration of the corner vowels, we can also place in the 3D canvas the correct sound of the [e] phoneme in a relative form. We can use this method, for placing all the vowel sounds in the 3D canvas taking as reference the position in the 3D canvas of our three corner vowels.

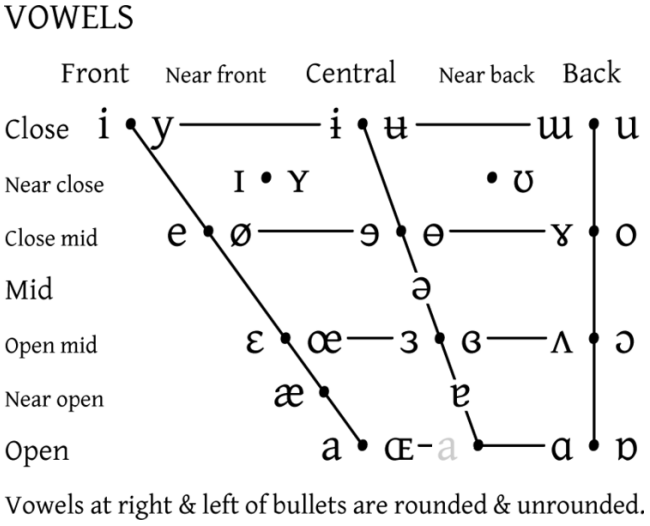



*Figure 20. Complete vowel chart.*

If this software is supposed to be used for training my pronunciation in a new language, or correct spare errors in some vowel phonemes in my mother tongue, one can make these questions:

- What happens if I make mistakes while recording the three corner vowels?

- If for example I am southern British. Why am I asked to pronounce a sound that I might not know how to pronounce correctly?

The answer to these questions is easier than it seems. Firstly, it is compulsory for the software to have a reference from the user, even if this is not completely correct, but it would be impossible to make any analysis without any previous references. Also this fact is applicable to language learning in general, because one teacher cannot give a lesson to students if he or she has not any idea of where is the level of knowledge of the students in the class. For this reason, the software is supposed to be calibrated more then once, as the user improves slowly his pronunciation. So the three corner vowels will be corrected every time the calibration is done.

Answering the second question, the software contains a feature that gives a very useful help, the talking head. The user is supposed to be guided by this talking head, so he can use as reference the example sound in the cardinal vowels that the talking head gives to him, in order to have a starting point in our learning process.

### 3.3.5. Adding the talking head

The talking head supplies an useful help to the user in the calibration stage. It will guide the user in several steps to make a proper recording to be used in the calibration of his voice, and also prevent from unexpected errors during this stage.

Basically the head is an external application which is run in parallel from the main .tcl file. As the main use for the head is going to be in the calibration part, the communications between the main part of the software and the head itself is pretty easy. The main software will give the control to the talking head during the calibration stage. The software will wait until the head write the acquisition data in one file, and when this is finished, the agent return the control again to the main software. Then the software uses that file to create the canvas and save the values in the normalization stage.

The head is opened in an external window in order to preserve the original software and also to give the opportunity to make research on the help that is provided with or without the agent. The agent and the software are connected by a *broker* connectivity solution. The agent acts as server while the main software will act as client. When the button 'calibrate' is pressed in the software, the agent starts to guide and extract the data from the speaker recordings.

When the agent finish collecting data, it gives an answer to the main software, which uses the files written by the agent to start to extract and normalize the formant values previously obtained.


*Figure 21. Screenshot of the talking head.*

# 4. THE SOFTWARE

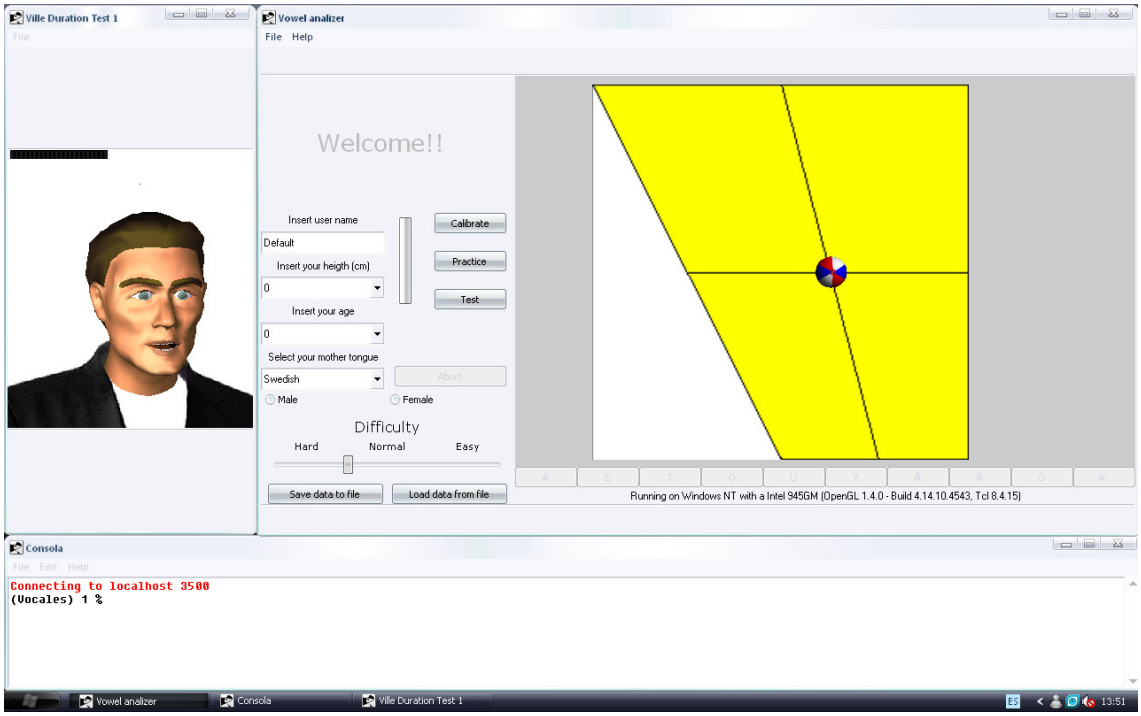



*Figure 22. Screenshot of the software.*

This section explains the software itself, how it basically works, and the most of the possibilities that are offered to the user. The software is being built with the always present idea to make it intuitive for the user, and all the features that it has are programmed thinking about this fact. In language learning it is important to give facilities to the user, in order to ease and make it faster, thus, the software has to be very intuitive. It is also important to give some motivation to the student so he can feel comfortable to continue using the software and training the pronunciation, instead of giving up the training and stop his progress.

This last fact is very important in language learning, because it is common that students feel very receptive to use some new tool for learning or improving a secondary language, but later they become lazy and lose the motivation to continue training with the tool. Thus, the learning is suddenly stopped, and the software becomes now useless.

The main reason for performing an interactive game is because it is a tool that the student can play with and motivate himself having a look at the highest scores, and the chances to beat his own top scores, or even the top scores of partners and other users of the same software,

providing this as an extra stimulus to continue using the software, and consequently continue training the pronunciation skills.

The main features are presented now, as a guide for the user.

## *4.1. Difficulty level*

The difficulty level can be changed in a very intuitive way during the Practice Mode or the Test Mode. A slide bar is implemented which makes the target spheres to grow or diminish and thus is easier or more difficult to fit the 3D moving ball inside of it, depending on the wishes of the student.

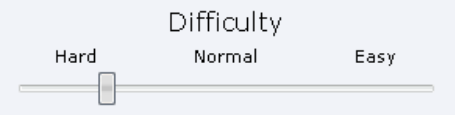



*Figure 23. Screenshot of the difficulty slider.*

## *4.2. User profile*

The student can also make a profile, adding his name/nickname, age, height, mother tongue and gender. Then try the software and save the data, and later on use his user account to continue playing or training the pronunciation loading his user name again with the calibration values previously saved.

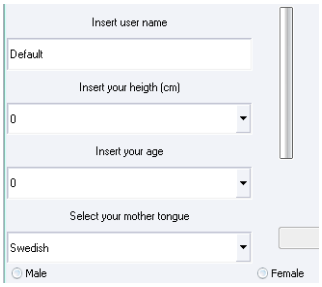



*Figure 24. Screenshot of the user profile area.*

## *4.3. Statistics saved*

Also for the goals of research, the software will save all the data related to the user, scores, time to reach target spheres, games finished, games aborted, etc. This data allows for comparisons and analysis about the

average and total spent time in the games, independent vowel timing which can be useful to discover problems in some vowels phonemes and follow the correct or incorrect evolution of these or other problems through the development of the training.

There are three main modes in the softare:

- Calibration mode.
- Practice mode.
- Test (interactive game).

## *4.4. Calibration Mode*

As discussed in previous pages, this part is the most critical in the development of the software. A big amount of the time spent in programming and developing the software, has been spent on this part.

This stage begins when the button of the 'calibration' is pressed. This starts the communication between the software and the talking head. When the agent receive the petition, it starts to ask the user for the three corner vowel phonemes, in order to get the formant frequency limits of the voice of the speaker.

The agent follows a path where he asks the user for the three corner vowel phonemes [a], [i], [u] one by one, and saves the three different successful recordings into files for its processing by the main software. If one of the recordings from the speaker is not successful during the interaction with the talking head, then the user is asked again to repeat the vowel phoneme.

When all this process is finished, the agent will return the control to the main software, so it can process the audio files recorded during the calibration stage as mentioned in the section 3.3.4.1.

During this stage, the movement of the 3D ball is disabled, because the software has not yet any references about the formant frequency limits from the user.

## *4.5. Practice Mode*

One of the basics of the pronunciation training is the opportunity to practise as much as you want in order to improve your skills in the pronunciation of the Swedish vowels. The more that you practise, the better results you will have later in your pronunciation skills. For this reason a Practice Mode has been developed. This mode lets the user train the pronunciation of the vowels without any time limit.

When this mode is selected the speaker is able to move the ball freely around the 3D canvas, and so he/she can recognize and get little by little the key parameters of the mouth that are able to move the ball where the speaker wants, such as open or close mouth, or a tongue placed in the back or in the front. Once the user get this, and he/she is able to move the ball more or less where he/she wish, can play with the software trying to place the ball in a chosen position in the 3D canvas. These positions would be the places in the canvas where the sounds of the Swedish vowel phonemes are. There are several buttons corresponding to all the different vowel phonemes in Swedish just below the 3D region. When the button that corresponds to any vowel phoneme is pressed, a semi-transparent sphere, placed in the part of the map corresponding to the sound of the phoneme, will be drawn in the canvas. Therefore, it can be drawn in the map any Swedish vowel and try to put the ball inside of it in order to get its pronunciation. Notice that the IPA-symbols are not used in the software, instead the written characters of standard Swedish of the corresponding phonemes are used.

The facts of draw the sphere that the speaker wants, and the unlimited time to practice, are specially useful if someone has special problems to reach certain sounds or vowels, because the user can train as much as he/she want without time limit, and focus the learning on the vowels that may result more difficult to get.

The facility that the software gives is that it can draw any vowel sphere, and it means that it can be trained any vowel. It can be loaded coordinates from the vowel charts of different languages, and not only from the Swedish language.

## *4.6. Interactive Game (Test Mode)*

As said before, it is considered important to motivate the student to continue training the pronunciation of the vowels, because to achieve a native pronunciation requires a considerable amount of time. For this reason we have developed a mini-game, where the student can make his own challenges, and get a good motivation to continue using the pronunciation trainer.

Briefly, this game consist of a 'try to catch the sphere' task, so several target spheres drawn in red are going to show up in the 3D canvas, but just one at a time. The student has to place the moving ball inside of the target sphere and hold it there during a short frame of time, approximately 500 milliseconds, and game progression will be without any kind of external help. When the moving ball is placed inside of the target sphere, this target sphere will become green. The student must hold the ball in that place for a while in order to complete it and try to catch a new sphere that will be drawn at a different location. If the user spend more than 10-12 seconds trying to catch the same sphere, the software will jump automatically to the next target and show the previous sphere as not completed. The maximum time for catching all the proposed spheres is thus two minutes (10 spheres x 12 seconds each sphere).

There are several alternatives to measure improvements in the pronunciation of the student. We have considered two of them, the first is to make a long list of target spheres (remember that each target sphere is a vowel phoneme sound) so long that it is impossible to finish the whole list in the two minutes of time proposed, and measure how many spheres the student managed to catch. The second alternative is construct a short list and give the chance to the student to be able to complete all the target spheres (one for each vowel), and measure the time that the speaker took to complete the task.

Both have different points of view. A long list will measure the improvement in number of spheres done, so presumably the student will get a higher number of spheres each time that he/she uses the software, and consequently it can be measured also the speed to get the spheres (more spheres done, less time spent on each sphere) that is also an important fact in pronunciation training. Anyway, the main goal of this method is to get a correct pronunciation despite of the time spent, because the list of targets is not supposed to be finished and the time issue plays here a secondary role.

On the other hand, having a list that is easier to complete, gives us the chance to measure the improvement in time spent, so the user can get better timing each time he/she uses the software. The user will have the change to try all the vowel phonemes and the main issue is complete them in the less time possible. Thus, this method gives us an idea of how fast the user is in pronouncing the vowel phonemes.

The discussion here is how many chances do we give the student to repeat and reinforce the learning in the pronunciation training. Because of this, the second method has being adopted, so there is a short list, able to be completed in two minutes. In this way it can be measured first the number of spheres completed within two minutes, so the first part of the learning is focused on the quality of the sound perceived. Secondly, when the student improves, he/she will be able to complete all the list successfully (so he/she will be able to say all the Swedish vowel phonemes), and the analysis can be focused on the time spent for completing this list, also important in pronunciation learning.

# 5. DATA COLLECTION

During the test stage, we recorded the time spent for each vowel phoneme. Each time that a sphere is completed, a time mark is written into a file, for later processing. Hence every user will have a complete statistic of his timings when using the software, such as an extra indication showing if the test is either aborted or finished successfully.

These values will be recorded into files, providing data for later analysis. It can show the improvement of the student in the global timing of the test, but also it can point out concrete improvements, such as specific vowel phonemes. The time spent on each sphere will decrease when the student starts to control his pronunciation, and he/she starts to learn the key parameters which lead him/her to reach the target spheres. Values as average total time, average time per sphere or comparison between average times of other users are easy to calculate with this data previously obtained.

It is also possible to make comparison between users, showing statistics for all the Swedish vowels. Researches on the most problematic phonemes for them (those which take longer to reach) and on efficient learning methods to improve them can be developed. This analysis of the most problematic phonemes is going to be discussed in section 6.1.

It should be pointed out that depending on the mother tongue, some phonemes can be hard for some people, and those do not have to be same for all. For example for German people phonemes like [ö] or [ä] should be familiar, and thus German people should not have special problems in pronouncing them, while Spanish, French or Italian people would have never seen those phonemes before, and those might be harder for them to pronounce. In the other hand, other example is the close back phoneme [u], which is hard to pronounce by southern British English speakers, as mentioned in 3.3.4.1.

In our study 18 different speakers from different countries, with and without any knowledge of Swedish language, tested our software. For each speaker there were implemented two different test sessions. Each session consisted on a first stage of calibration, followed by approximately five minutes of free practice, where the speaker could play with the 3D ball and the features of the software. Afterwards the test mode was run three times in a row, saving the timing as said in the beginning of this section.

Few days later the same process was repeated. The same speakers were asked to perform the same test again, after a new calibration and practice between five and ten minutes. The results of all the different tests of the speakers are shown in the Appendix 9.4.

One female non Swedish speaker was also asked to train her pronunciation with a higher number of sessions developed. During one week, several identical sessions of were tested on her. Before each new session, she was informed of which vowels were problematic in the previous one, due to reinforce the practice in those problematic phonemes. In total there were performed four different sessions, and these results can be shown in the Appendix 9.4.3.

# 6. CONCLUSIONES

## *6.1. Conclusiones de los tests*

Las figures 25 y 26 comparan el tiempo medio empleado en cada fonema vocálico propuesto en los tests realizados. Como se puede ver, la tendencia en los hablantes extranjeros es tener unos tiempos mas altos en los fonemas [y], [ä], [ö], [å] y [a] larga, mientras en los hablantes suecos los valores en aquellos fonemas no son especialmente destacables comparados con el resto.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 3,595 | 5,305 | 4,271 |
| E | 4,818 | 3,987 | 4,981 |
| I | 6,493 | 6,606 | 6,381 |
| O | 5,500 | 6,218 | 5,121 |
| U | 3,893 | 4,820 | 4,640 |
| Y | 5,232 | 4,742 | 4,656 |
| Ä | 2,039 | 2,864 | 2,191 |
| Å | 5,253 | 6,744 | 5,344 |
| Ö | 2,929 | 2,059 | 2,239 |
| long A | 4,689 | 4,712 | 3,556 |
| Total time spent | 44,439 | 48,057 | 43,380 |
| Average time / sphere | 4,444 | 4,806 | 4,338 |

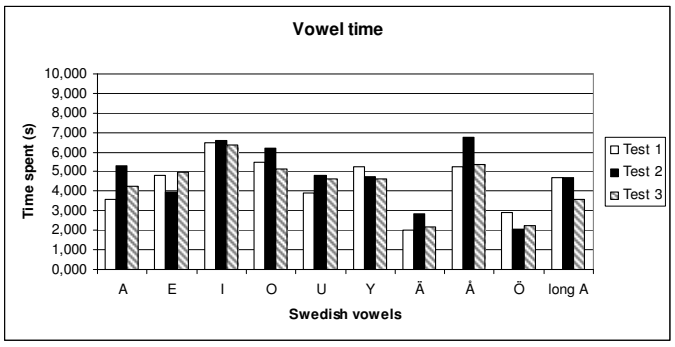



*Figura 25. Gráfica correspondiente al tiempo medio utilizado por los usuarios suecos en la primera sesión de tests.*

Las diferencias entre /Ä/ u /Ö/ y el resto de las vocales suecas puede explicarse con los diferentes dialectos de sueco que existen. Las esferas objetivo están emplazadas en un lugar medio en el mapa, de acuerdo a los tests previos nombrados en la sección 3.3.3. Pero como hemos dicho, este lugar puede tener ligeras variaciones dependiendo del distinto dialecto. Este hecho nos lleva a entender que algunas veces los usuarios suecos, incluso siendo hablantes nativos, tienen dificultades o no son capaces de alcanzar el lugar correcto en el mapa de las vocales cuando intentan pronunciar determinadas vocales suecas.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 5,068 | 4,381 | 2,875 |
| E | 4,830 | 7,353 | 6,684 |
| I | 4,428 | 3,820 | 4,303 |
| O | 4,854 | 2,870 | 3,155 |
| U | 5,096 | 6,141 | 3,616 |
| Y | 7,734 | 6,304 | 6,250 |
| Ä | 8,685 | 7,948 | 5,983 |
| Å | 6,178 | 6,374 | 5,594 |
| Ö | 8,264 | 6,808 | 7,249 |
| long A | 5,911 | 7,891 | 6,431 |
| Total time spent | 61,047 | 59,890 | 52,140 |
| Average time / sphere | 6,105 | 5,989 | 5,214 |

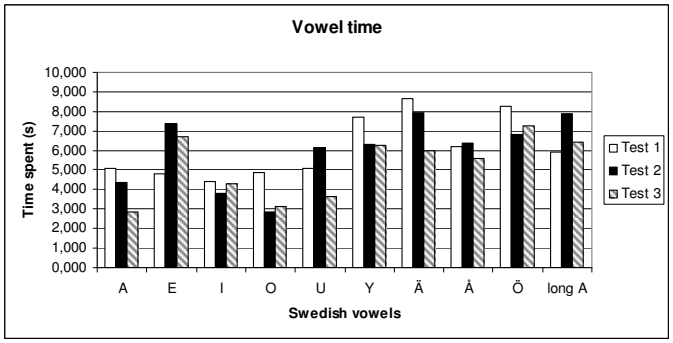



*Figura 26. Gráfica correspondiente al tiempo medio utilizado por los usuarios extranjeros en la primera sesión de tests.*

Como muestra la figura 26, la vocal /E/ es una de las mas complicadas, incluso existiendo en italiano, francés o español. La principal razón para esto es que el sonido de la vocal /E/ en estos idiomas se pronuncia con una mayor apertura de la boca. Este hecho hace que no sea tan fácil alcanzar el sonido vocálico /E/ respecto de otros como /A/, /I/ o /O/.

Se puede observar que el tiempo empleado en la primera sesión es ligeramente menor en los hablantes suecos, debido a una mejor pronunciación en los fonemas [y], [ä], [ö], [å] y [a:]. Pero aun así las diferencias entre suecos y extranjeros no están del todo claras. Una posible explicación para esto es que los hablantes suecos son más capaces de acertar con el lugar correcto directamente. Sin embargo, tanto suecos como extranjeros no son del todo capaces de corregir la posición de la pelota 3D cuando ésta se encuentra fuera de la esfera objetivo.

Como muestran las figuras 25 y 26, el tiempo empleado en los distintos tests de cada sesión usualmente decrece desde el primer hasta el tercer test realizado, obteniendo los resultados que se esperan. Gráficas más específicas sobre la primera sesión de test se pueden encontrar en el Apéndice 11.4.1.

Lamentablemente no todos los usuarios que realizaron el primer test pudieron atender a la segunda sesión de test. Sin embargo, las figuras 27 y 28 muestran que existe una mejora entre la primera y la segunda

sesión de test en aquellos que si completaron ambas sesiones. El tiempo empleado en cada vocal, tiempo total y por lo tanto el tiempo medio empleado en cada vocal durante los tres tests, fueron mejorados durante la segunda sesión.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 1,826 | 1,184 | 2,851 |
| E | 1,161 | 1,440 | 2,850 |
| I | 6,671 | 3,236 | 2,268 |
| O | 2,471 | 1,670 | 1,724 |
| U | 1,158 | 1,045 | 0,767 |
| Y | 2,246 | 1,482 | 1,116 |
| Ä | 2,001 | 2,434 | 2,013 |
| Å | 0,832 | 3,810 | 1,306 |
| Ö | 4,335 | 3,968 | 1,870 |
| long A | 3,493 | 4,411 | 1,438 |
| Total time | 25.246 | 29.825 | 23.284 |
| Average time | 2.525 | 2.983 | 2.328 |

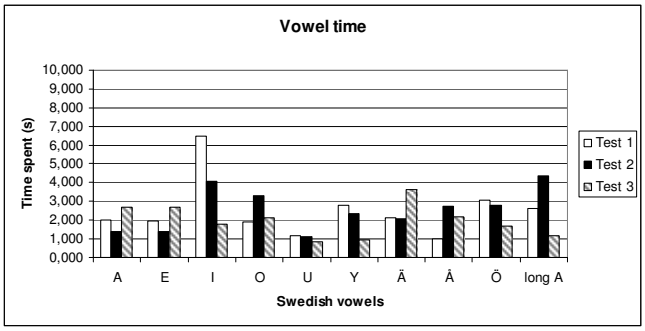



*Figura 27. Gráfica correspondiente al tiempo medio utilizado por los usuarios suecos en la segunda sesión de tests.*

.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 1,005 | 2,997 | 1,601 |
| E | 2,232 | 3,066 | 3,697 |
| I | 4,209 | 1,225 | 3,173 |
| O | 5,068 | 3,555 | 2,014 |
| U | 5,581 | 4,386 | 3,565 |
| Y | 2,253 | 2,335 | 3,227 |
| Ä | 6,240 | 6,015 | 3,570 |
| Å | 5,974 | 3,484 | 3,073 |
| Ö | 4,373 | 4,678 | 2,756 |
| long A | 5,242 | 3,650 | 5,338 |
| Total time | 42,176 | 35,391 | 32,014 |
| Average time | 4,218 | 3,539 | 3,201 |

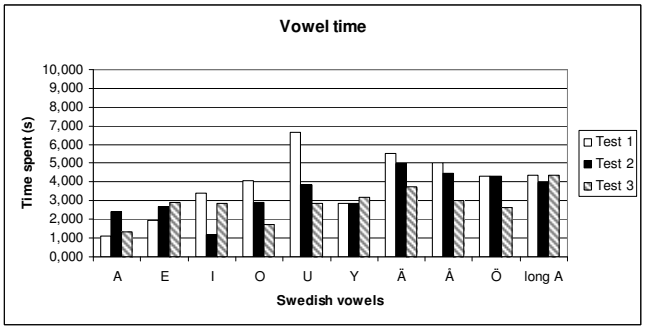



*Figura 28. Gráfica correspondiente al tiempo medio utilizado por los usuarios extranjeros en la segunda sesión de tests.*

Los hablantes no nativos fueron capaces de mejorar considerablemente sus tiempos empleados en cada esfera, además de corregir la pronunciación en los fonemas específicamente suecos. Como muestra la figura 28, los fonemas específicos suecos no aparecen con un tiempo especialmente mayor que el resto de los fonemas. Además, los hablantes suecos fueron capaces de mejorar el tiempo empleado debido a un mejor control en el movimiento de la pelota 3D durante la segunda sesión de tests.

De una forma más particular, algunos de los usuarios tenían una experiencia previa con el software antes del a primera sesión de tests. Se ha percibido en ellos que durante la primera sesión ya poseían unos tiempos aceptables. Por lo tanto, la mejora con respecto a la segunda tanda de tests ha sido ligeramente menor, y los resultados han sido del a misma índole. La mayoría de estos usuarios tenían un tiempo medio por esfera inferior a dos segundos.

El diagrama con el tiempo medio empleado en la primera y segunda sesión de tests se presenta a continuación. Como se puede ver, este tiempo es considerablemente menor en la segunda sesión respecto de la primera.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 4,332 | 4,843 | 3,573 |
| E | 4,824 | 5,670 | 5,832 |
| I | 5,460 | 5,213 | 5,342 |
| O | 5,177 | 4,544 | 4,138 |
| U | 4,494 | 5,480 | 4,128 |
| Y | 6,483 | 5,523 | 5,453 |
| Ä | 5,362 | 5,406 | 4,087 |
| Å | 5,715 | 6,559 | 5,469 |
| Ö | 5,596 | 4,434 | 4,744 |
| long A | 5,300 | 6,302 | 4,994 |
| Total time | 52,743 | 53,974 | 47,760 |
| Average time | 5,274 | 5,397 | 4,776 |

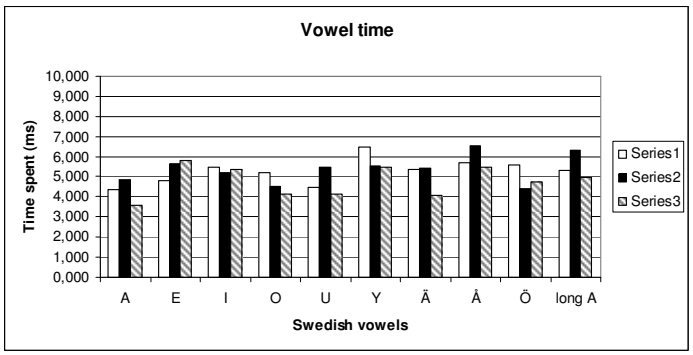



*Figura 29. Gráfica correspondiente al tiempo medio utilizado por todos los usuarios en la primera sesión de tests.*

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 1,415 | 2,090 | 2,226 |
| E | 1,697 | 2,253 | 3,274 |
| I | 5,440 | 2,231 | 2,721 |
| O | 3,770 | 2,613 | 1,869 |
| U | 3,370 | 2,715 | 2,166 |
| Y | 2,249 | 1,909 | 2,171 |
| Ä | 4,121 | 4,225 | 2,792 |
| Å | 3,403 | 3,647 | 2,190 |
| Ö | 4,354 | 4,323 | 2,313 |
| long A | 4,368 | 4,030 | 3,388 |
| | | | |
| Total time | 34,185 | 30,036 | 25,109 |
| Average time | 3,419 | 3,004 | 2,511 |

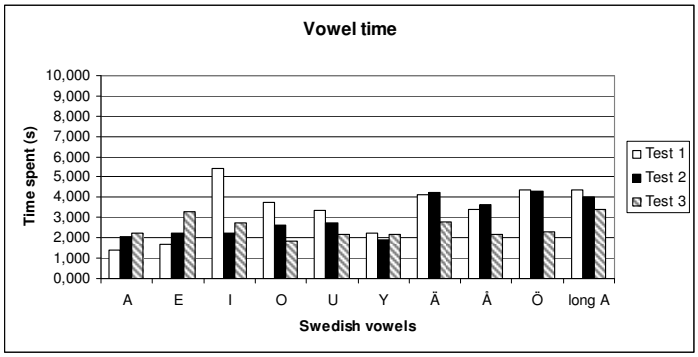



*Figura 30. Gráfica correspondiente al tiempo medio utilizado por todos los usuarios en la segunda sesión de tests.*

Además se aprecia una mayor mejora en el usuario especial que se utilizó, debido a un número mayor de sesiones de práctica. Los fonemas problemáticos fueron resolviéndose mediante los tests y las prácticas, alcanzando una velocidad considerable en el ultimo test desarrollado. De hecho, los resultados de sus últimos tests se podrían comparar directamente con un usuario sueco. Estas gráficas se han añadido al Apéndice 11.4.3. Como ejemplo podemos observar la evolución de un fonema en particular durante todos los tests completados por este usuario.

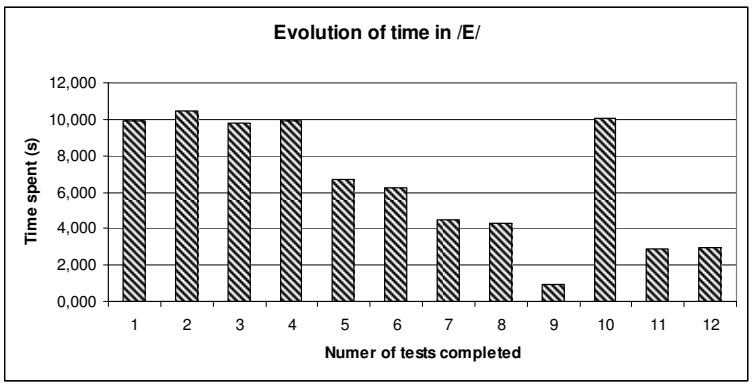



*Figura 31. Evolución del fonema /E/ en los tests completados.*

Para terminar podemos hacernos esta pregunta, ¿Cuál es el nivel de calidad y velocidad en la cual deberíamos parar nuestro aprendizaje?

Probablemente el aprendizaje de un nuevo idioma no se debe interrumpir nunca incluso después de varios años de estudio. Ni siquiera puede que sea posible alcanzar la pronunciación nativa completamente, pero el objetivo de este proyecto es apuntar a las miras mas altas posibles, es decir, el mayor nivel posible de calidad en la pronunciación de los sonidos vocálicos suecos. Por tanto, basándonos en los resultados de los tests anteriores, cuando un usuario practica con el software desarrollado durante algún tiempo, adquiere una gran velocidad en alcanzar correctamente las esferas objetivo, utilizando muy poco tiempo para ello. Podemos afirmar que un tiempo menor que dos segundos en alcanzar correctamente los sonidos vocálicos, se puede considerar lo suficientemente bueno como para creer que la pronunciación ha adquirido la velocidad y la calidad suficiente. Es de mencionar que en estos dos segundos de tiempo se incluyen tanto el tiempo que se debe mantener la pelota en el interior de la esfera objetivo, como el tiempo que se tarda en reaccionar al leer la nueva vocal a pronunciar.

## *6.2. Análisis de los objetivos previos de estudio*

Los objetivos de este proyecto fueron:

- Desarrollar un sistema capaz de plasmar en tiempo real el resultado de la pronunciación de vocales suecas en los usuarios.

- Crear un sistema capaz de normalizar los parámetros de la voz de diferentes hablantes, independientemente de su edad o sexo.

- Introducir la ayuda del Agente Embebido Conversacional estableciendo conectividad con él.

- Crear un juego capaz de medir la mejora en los usuarios en el entrenamiento de su pronunciación.

El primer requerimiento fue completado dibujando la región 3D y siendo capaz de introducir un apelota móvil en ella. Se utilizaron alrededor de 20 frames por segundo en el movimiento de la pelota, y este resultado fue lo suficientemente bueno como para dar un movimiento suave y continuo en el plano 3D. El resultado fue lo suficientemente natural como para que

ningún usuario se quejara sobre la calidad de la región 3D. Además, para ayudar a esta naturalidad se añadieron efectos de luces y sombras en esta región 3D.

El sistema de normalización también se implementó satisfactoriamente, y la etapa de calibración fue lo suficientemente robusta como par atener resultados positivos en la inmensa mayoría de las calibraciones. El método de normalizar a una región entre [-0.5, 0.5] tanto en el primer como en el segundo formante, y ajustar el mapa vocálico a estas medidas, ha demostrado ser fiable, y las posiciones relativas de las esferas objetivo en la región dieron, perceptualmente, un sonido nativo correcto para la vocal correspondiente en la mayoría de los casos. El movimiento de la pelota fue satisfactorio para los usuarios, debido a que podían percibir cambio sen la posición de la pelota cuando modificaban su pronunciación.

Incluso aunque la integración del agente no fue posible, su introducción en el software fue satisfactoria. La conexión entre el software y el agente fue muy sencilla, y la solución *broker* fue desarrollada sin ningún tipo de problemas y dando siempre resultados positivos.

El juego para mejorar la pronunciación también se completó con éxito usando la solución de tener suficiente tiempo para completar todas las esferas. Esto facilita el análisis y las comparaciones entre usuarios, además del posterior análisis de la dificultad para el usuario de cada vocal por separado.

Las encuestas entregadas a los usuarios (Apéndices 11.3.1. y 11.3.2.) nos revelan que la interacción entre el agente 3D y los usuarios fue satisfactoria, ya que ningún usuario se quejó o hizo ningún comentario negativo sobre el mismo. En cambio se obtuvieron sugerencias sobre un mayor grado de interacción entre el agente y el usuario durante el test o el modo práctica, y por tanto, éstas se incluirán en el trabajo futuro a desarrollar.

# 6. CONCLUSIONS

## *6.1. Tests conclusions*

Figures 25 and 26 compare the average time spent on each Swedish vowel phoneme proposed for the tests. As can be shown, the tendency in non Swedish speakers is to stay longer time to get the specific vowel phonemes [y], [ä], [ö], [å] and long [a], while in Swedish speakers those phonemes are not specially high compared with the rest of the vowel phonemes.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 3,595 | 5,305 | 4,271 |
| E | 4,818 | 3,987 | 4,981 |
| I | 6,493 | 6,606 | 6,381 |
| O | 5,500 | 6,218 | 5,121 |
| U | 3,893 | 4,820 | 4,640 |
| Y | 5,232 | 4,742 | 4,656 |
| Ä | 2,039 | 2,864 | 2,191 |
| Å | 5,253 | 6,744 | 5,344 |
| Ö | 2,929 | 2,059 | 2,239 |
| long A | 4,689 | 4,712 | 3,556 |
| Total time spent | 44,439 | 48,057 | 43,380 |
| Average time / sphere | 4,444 | 4,806 | 4,338 |

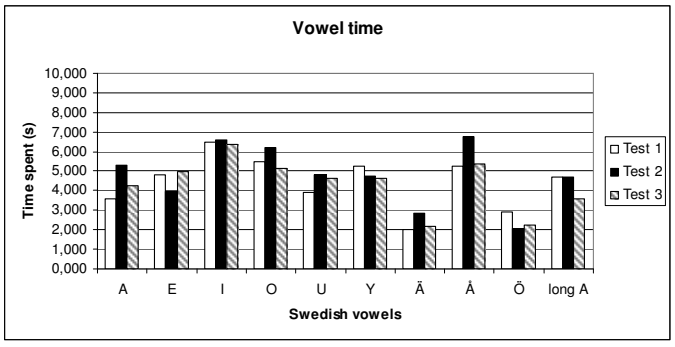



*Figure 32. Diagram corresponding to the average of all the Swedish speakers tested in the first session.*

Differences between /Ä/ or /Ö/ and the rest of the Swedish vowels among Swedish can be explained due to the different dialects that in Swedish language exist. The target spheres are placed in an average place in the canvas, according to previous tests among Swedes as explained in section 3.3.3., but this place can have slight variations depending on the dialect. This fact lead us to understand that sometimes some Swedish speakers, even being native speakers, are not able to reach the correct place in the canvas when pronouncing some Swedish vowels.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 5,068 | 4,381 | 2,875 |
| E | 4,830 | 7,353 | 6,684 |
| I | 4,428 | 3,820 | 4,303 |
| O | 4,854 | 2,870 | 3,155 |
| U | 5,096 | 6,141 | 3,616 |
| Y | 7,734 | 6,304 | 6,250 |
| Ä | 8,685 | 7,948 | 5,983 |
| Å | 6,178 | 6,374 | 5,594 |
| Ö | 8,264 | 6,808 | 7,249 |
| long A | 5,911 | 7,891 | 6,431 |
| Total time spent | 61,047 | 59,890 | 52,140 |
| Average time / sphere | 6,105 | 5,989 | 5,214 |

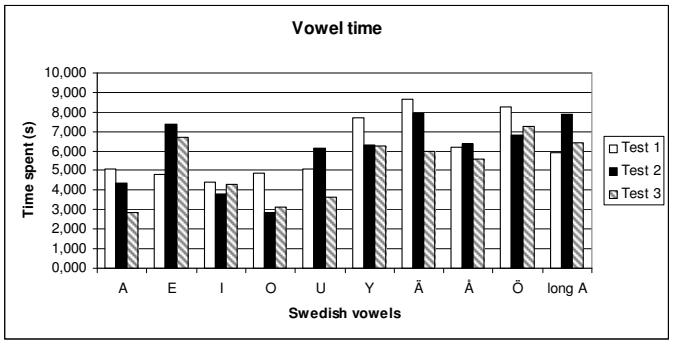



*Figure 33. Diagram corresponding to the average of all the non Swedish speakers tested in the first session.*

As Figure 26 can show, the /E/ is one of the hardest vowels, even if it exists in Italian, French or Spanish. The main reason for this is that in all those languages the /E/ sound is produced with a bigger opening of the mouth. This fact makes that users do not reach that vowel sound as fast as /A/, /I/ or /O/.

It can be observed that the timing during the first session of test is slightly lower in Swedish speakers, due to the best performance in the phonemes [y], [ä], [ö], [å] and long [a]. But still the differences between them are not clear enough. An explanation of this is that Swedish speakers are more able to hit the right place directly. However, still exists a small difference between them because in the first session, both Swedes and foreigners are not able to correct the position of the 3D ball when this is out of the target sphere.

As shown in the Figures 25 and 26, the time spent in the several tests usually decreases from the first to the third test performed, having the results that are expected. More specific diagrams of the first session can be found in the Appendix 9.4.1.

Not all the speakers attended to the second session of tests. However, Figures 27 and 28 reveal that it exists an improvement between the first and the second session of test in those who completed both sessions. The time spent on each vowel, total time and thus average time per sphere in the three test were improved during the second test.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 1,826 | 1,184 | 2,851 |
| E | 1,161 | 1,440 | 2,850 |
| I | 6,671 | 3,236 | 2,268 |
| O | 2,471 | 1,670 | 1,724 |
| U | 1,158 | 1,045 | 0,767 |
| Y | 2,246 | 1,482 | 1,116 |
| Ä | 2,001 | 2,434 | 2,013 |
| Å | 0,832 | 3,810 | 1,306 |
| Ö | 4,335 | 3,968 | 1,870 |
| long A | 3,493 | 4,411 | 1,438 |
| | | | |
| Total time | 25.246 | 29.825 | 23.284 |
| Average time | 2.525 | 2.983 | 2.328 |

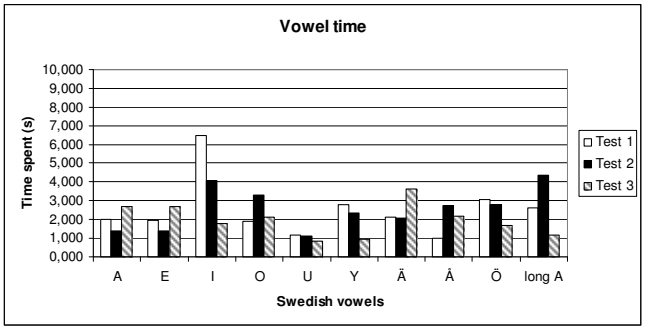



*Figure 34. Diagram corresponding to the average of all the Swedish speakers tested in the second session.*

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 1,005 | 2,997 | 1,601 |
| E | 2,232 | 3,066 | 3,697 |
| I | 4,209 | 1,225 | 3,173 |
| O | 5,068 | 3,555 | 2,014 |
| U | 5,581 | 4,386 | 3,565 |
| Y | 2,253 | 2,335 | 3,227 |
| Ä | 6,240 | 6,015 | 3,570 |
| Å | 5,974 | 3,484 | 3,073 |
| Ö | 4,373 | 4,678 | 2,756 |
| long A | 5,242 | 3,650 | 5,338 |
| | | | |
| Total time | 42,176 | 35,391 | 32,014 |
| Average time | 4,218 | 3,539 | 3,201 |

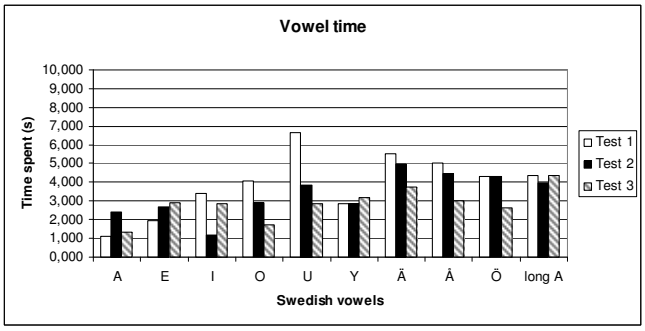



*Figure 35. Diagram corresponding to the average of all the non Swedish speakers tested in the second session.*

Non native Swedish speakers were able to improve considerably their times per sphere, and also were able to correct the pronunciation in the special Swedish phonemes. As shown in the Figure 28, the chart does not reveal specially higher timing in those special Swedish phonemes compared to the rest of the vowel sounds. Also Swedish speakers were able to improve the timing due to a better control in the movement of the 3D ball during the second session of tests.

Particularly, some of the testers had a previous training with the software before performing the first session of tests. It has been noticed that they already had a good timing in the first session of tests, thus, the second session does not reveal great improvement, however the results obtained are in the same kind for them. Most of them reached an average time per sphere of less than two seconds.

The chart for the average time spent in the game in the first and second session of tests is presented below. As can be noticed, the time is considerably lower in the second session than in the first one.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 4,332 | 4,843 | 3,573 |
| E | 4,824 | 5,670 | 5,832 |
| I | 5,460 | 5,213 | 5,342 |
| O | 5,177 | 4,544 | 4,138 |
| U | 4,494 | 5,480 | 4,128 |
| Y | 6,483 | 5,523 | 5,453 |
| Ä | 5,362 | 5,406 | 4,087 |
| Å | 5,715 | 6,559 | 5,469 |
| Ö | 5,596 | 4,434 | 4,744 |
| long A | 5,300 | 6,302 | 4,994 |
| Total time | 52,743 | 53,974 | 47,760 |
| Average time | 5,274 | 5,397 | 4,776 |

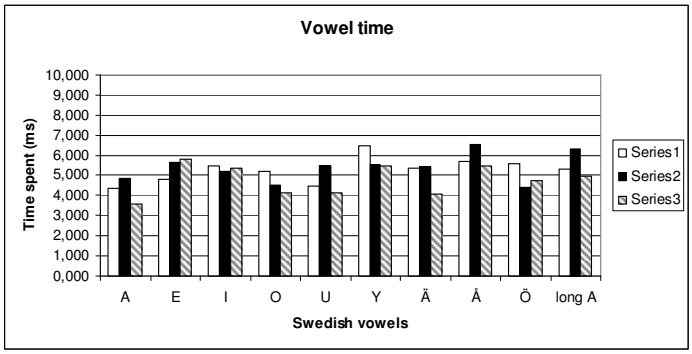



*Figure 36. Diagram corresponding to the average of all the speakers tested in the first session.*

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 1,415 | 2,090 | 2,226 |
| E | 1,697 | 2,253 | 3,274 |
| I | 5,440 | 2,231 | 2,721 |
| O | 3,770 | 2,613 | 1,869 |
| U | 3,370 | 2,715 | 2,166 |
| Y | 2,249 | 1,909 | 2,171 |
| Ä | 4,121 | 4,225 | 2,792 |
| Å | 3,403 | 3,647 | 2,190 |
| Ö | 4,354 | 4,323 | 2,313 |
| long A | 4,368 | 4,030 | 3,388 |
| Total time | 34,185 | 30,036 | 25,109 |
| Average time | 3,419 | 3,004 | 2,511 |

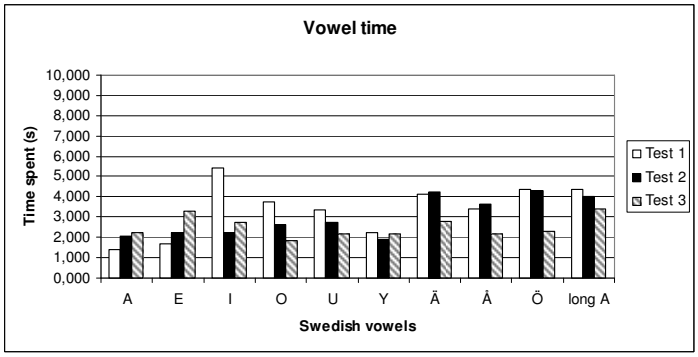



*Figure 37. Diagram corresponding to the average of all the speakers tested in the second session.*

A bigger improvement of the pronunciation is noticed in the special speaker due to a higher number of practice sessions. Problematic vowels were solved through the tests and practice, reaching a considerable speed in the last test performed. In fact, results from her last tests can be directly compared to a Swedish speaker. This charts are added in the Appendix 9.4.3. As an example we can observe the evolution of a particular phoneme during all the tests completed by the special tester.

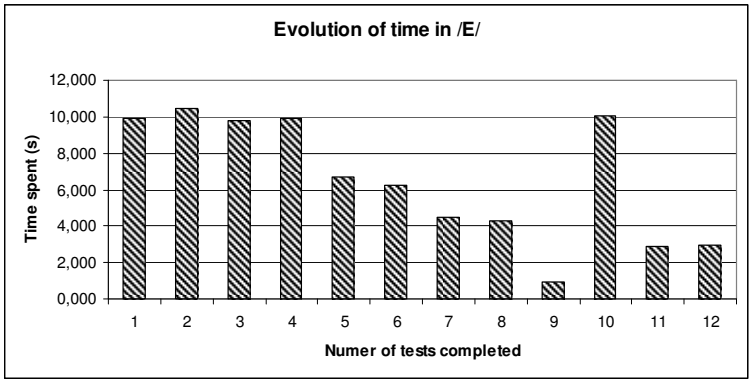



*Figure 38. Evolution of the phoneme /E/ with the tests completed.*

In conclusion, where is the level of quality in the pronunciation in which we should finish our learning?

Probably the learning of a new language will not stop even if one is studying it for long time. Also the native pronunciation may not be reached completely, but the aim of this project is to point to the highest quality possible of the pronunciation of the Swedish vowels. Therefore, during the tests explained before, when a speaker tries the software for long time, and get the vowel phonemes very fast, he/she will spend very few time on each sphere. One can believe that spend two seconds or less in get the sphere will be considered good enough to say that my vowel phonemes are good in quality and speed. In this two seconds are included the 500 milliseconds spent inside of the target sphere and the time spent in identify and read the label in the main menu which ask to pronounce the target vowel phoneme.

## *6.2. Analysis of the previous goals of the study*

The goals for this project were to:

- Develop a system capable to plot in real time the pronunciation of vowels.

- Create a system able to normalize the voice parameters of several speakers, so it is possible to compare speakers despite of their age, or gender.

- Introduce the help of an Embodied Conversational Agent establishing connectivity with it.

- Create a game able to measure the improvement of the user in the pronunciation learning.

The first requirement was fulfilled by drawing the 3D region and being able to move the ball around it. The refresh rate of the drawings was about 20 frames per second. This result was good enough to give us a continued movement around the 3D canvas. The result was natural enough and no users complained about the quality of the 3D region. Also as an external feature were added some shadows and lights effects to make that region more natural.

The normalization system also was implemented successfully, and the calibration stage was robust enough to have positive results in almost every attempt of calibration. The method to normalize into a region from -0.5 to 0.5 both in the $f_1$ and $f_2$ formants scales and then resize the canvas for the vowel chart has demonstrated to be reliable enough, and the positions of the relative normalized spheres in that region gave a successful native sound required for the vowel in most of cases. The movement of the 3D ball was satisfactory enough for the users, as they could perceive changes in the position of the ball when modifying their own pronunciation.

Even though the integration was not possible, the introduction of the talking head was also performed successfully. The connection between the software and the talking head was pretty easy. The broker connectivity was performed without any problems and gave always positive results.

The game for improving the pronunciation learning was also fulfilled using the solution of having enough time to complete all the spheres. This facilitate the analysis and comparisons between speakers, and the analysis of the difficulty of each vowel phoneme separately.

The surveys given to the testers (Appendix 9.3.1 and 9.3.2) reveal that interaction between the talking head and user has been fulfilled successfully, and no one complained about the features of the agent. Some comments about more grade of interaction between agent and user in the test or practice mode were asked in the surveys, and thus, this becomes part of the future work.

# 7. TRABAJO FUTURO

Sería interesante realizar más investigaciones en la dirección de la etapa de calibración. Se puede añadir una segunda comprobación al proceso de calibración, para hacer más robusta la recogida de datos. En la calibración actual se ajusta el mapa para cada usuario. El usuario puede pronunciar más tarde diferentes palabras o sílabas que contengan las vocales deseadas. Se puede extraer el fonema vocálico de la palabra obtenida, y comprobar en el mapa creado anteriormente, si la extracción de formantes de esta vocal esta en el mismo lugar que durante la etapa para el ajuste del mapa.

Métodos alternativos de normalización son otro campo de estudio. En este estudio se probaron con antelación varios métodos de normalización con distintos resultados. Se utilizó una normalización dinámica, donde los límites del mapa se ajustaban automáticamente cuando se recibía el sonido, perol os resultados obtenidos no fueron tan buenos comparados con el método finalmente implementado.

Se confía en que una investigación más profunda del tercer formante sería de gran ayuda. No sólo una clasificación automática entre los fonemas [i] e [y], sino también una investigación profunda en los parámetros clave que provocan cambios significativos en los valores del tercer formante. Otra posibilidad es encontrar métodos alternativos que puedan hacer una distinción favorable entre esos dos fonemas.

Se puede añadir algún tipo de ayuda en el test o más interacción con el agente 3D, como por ejemplo tener grabaciones de los fonemas y hacer ala gente pronunciarlos cuando el usuario selecciona la esfera a practicar o cuando la esfera aparece en el modo test. Instrucciones extra sobre como mover la pelota también pueden implementarse con la ayuda del agente 3D.

Como se mencionó en la parte de la introducción, este software pretendía ser parte de VILLE. Por tanto, la completa integración de éste en VILLE sería otro interesante campo de estudio.

Posibles mejoras en la interfaz del usuario, tales como añadir top scores, mejorar la interfaz para salvar y cargar los datos de los usuarios, o grabar el sonido de los usuarios al menos cuando realizan el test, para utilizarlo en posteriores análisis, sería parte también del trabajo futuro. El uso de este software para el análisis de las vocales dentro de palabras quizás sea el siguiente paso a seguir.

# 7. FUTURE WORK

Further researches can be done in the direction of the calibration stage. A secondary checking can be added when the process of calibration is done, in order to have more robust data to process. In the actual calibration stage it is conformed the normalized canvas for each speaker. The user can be asked this time to pronounce different words or syllables which contain the vowel sounds required. Later the vowel sound from the word or syllable will be extracted, and it can be checked if the formant extraction for that vowel sound extracted from the word, fits in the same place than the first formant extraction.

Alternative methods for normalization are other area of study. In this study were previously tested several normalization methods with distinct results. A dynamic normalization where the limits of the canvas were updated automatically with the pronunciation was developed, but the results were not as good as compared with the method implemented.

We also believe that a thorough research in the third formant classification would be helpful. Not only an automatic and normalized distinction between the phonemes [i] and [y], but also a deep research on the key parameters that make significant changes in the third formant values. Another possibility is find out alternative methods to make distinctions between those two phonemes.

Some guiding in the test or more interaction with the talking head can be implemented, such as having recordings of the vowel sounds that the talking head can say to the user when he is testing or practising. Extra instructions on how to move the 3D ball can be also implemented with the help of the agent.

As mentioned in the introduction part, this software intended to be part of VILLE. A integration of the software in VILLE, and the use of the talking head that VILLE provides is another interesting field of study.

Possible future improvements in the software interface, such as adding top scores, improve the interface for saving and loading data, or recording the speaker pronunciation for later analysis could be interesting future work to do. The use of this software for the vowel analysis in words would be also an interesting field of study.

# 8. REFERENCES


1. L. J. Gerstman, "*Classification of self-normalized vowels*". IEEE transactions on audio and electroacoustics. Vol. AU-16 NO. 1, (1968).

2. A. Paganus, V. P. Mikkonen, T. Mäntylä, S. Nuuttilla, J. Isoaho, O. Aaltonen and T. Salakoski, "*The vowel game: Continous Real-time visualization for pronunciation learning with vowel charts*", University of Turku, Department of information technology and department of phonetics. FI-20014 Turku, Finland. (2006).

3. J.C. Catford, "*A practical introduction to phonetics*" Second Edition, (2001).

4. K. Johnson, "*Acoustic & Auditory Phonetics*", Second Edition, (1997).

5. http://www.memo.com.co/fenonino/aprenda/castellano/castellano36.html. 2008.

6. http://www.geocities.com/sergiozamorab/fonetica.html. 2008.

7. S. Wood. "*Praat for beginners*". (2005).

8. R. Carlson, B. Granström, & G. Fant. "*Some studies concerning perception of isolated vowels*". STL-QPSR, 11(2-3), 019-035. (1970).

9. A. Dowd, J. R. Smith, J. Wolfe. "*Learning to pronounce vowel sounds in a foreign language using acoustic measurements of the vocal tract as feedback in real time*". Language and Speech, 41 1-20. (1998)

10. J. M . Van Der Stelt, K. Zajdo, T. G. Wempe. "*Exploring the acoustic vowel space in two-year-old children: Results for Dutch and Hungarian*". Speech communication (Speech commun.) ISSN 0167- 6393 CODEN SCOMDH.

11. K. H. Norian, "*Second derivative analysis of consonant-vowel transition waveforms*". (2003).

12. M. Lee and K. Soh, "*Nonlinear Dynamical Analysis of the Vowels in a Korean Traditional Folk Songs*". Department of Physics Education, Seoul National University, Seoul 151-742. (1999).

13. R. Carlson, B. Granström, & G. Fant. "*Some studies concerning perception of isolated vowels*". STL-QPSR, 11(2-3), 019-035. (1970).

14. G. Fant. "*Speech Sounds and Features*". Cambridge, Mass./London: M.I.T. Press. (1973)

15. G. Fant, "*Swedish vowels and new three-parameter model*", TMH-QPSR, 42 (1), 043-049 (2001).

16. D. Burnham and S. Lau. "*The integration of auditory and visual speech information with foreign speakers: The role of expectancy*". in Proc of AVSP, pp. 80.85. (1999).

17. B. Granström, D. House, and M. Lundeberg. "*Prosodic cues in multimodal speech perception*". in Proc of ICPhS, pp. 655.658. (1999).

18. J. Beskow, B. Granström, D. House, and M. Lundeberg. *"Experiments with verbal and visual conversational signals for an automatic language tutor".* in Proc of InSTIL , pp. 138.142. (2000).

19. J. Beskow. "*Talking Heads - Models and Applications for Multimodal Speech Síntesis*". KTH, Department of Speech, Music and Hearing. (2003).

20. O. Engwall. "*Combining MRI, EMA & EPG measurements in a three-dimensional tongue model*". Speech Comm, 41, 303-329. (2003).

21. O. Engwall, P. Wik, J. Beskow, B. Granström. "*Design strategies for a virtual language tutor*". ICSLP 2004, vol. 3, 1693-1696 (invited contribution to special session on "Second language learning and spoken language processing") (2004).

22. B. Granström. "*Towards a virtual language tutor*". Proc InSTIL/ICALL2004 - NLP and Speech Technologies in Advanced Language Learning Systems, 1-8 (Invited paper) (2004).

23. A. Livonen. Vowel charts. University of Helsinki. Department of Speech Sciences. (2003) http://www.helsinki.fi/speechsciences/projects/vowelcharts/

24. H. Traunmüller. "*En tur i fonetikens marker*". Stockholms Universitet. (1996).

25. G. Fant, "*Sound, features, and perception*". STL-QPSR, 8 (2-3), 001-014 (1967).

26. www.opengl.org (2007).

27. Brent B. Welch, "*Practical programming in Tcl and Tk*". Third edition (2000).

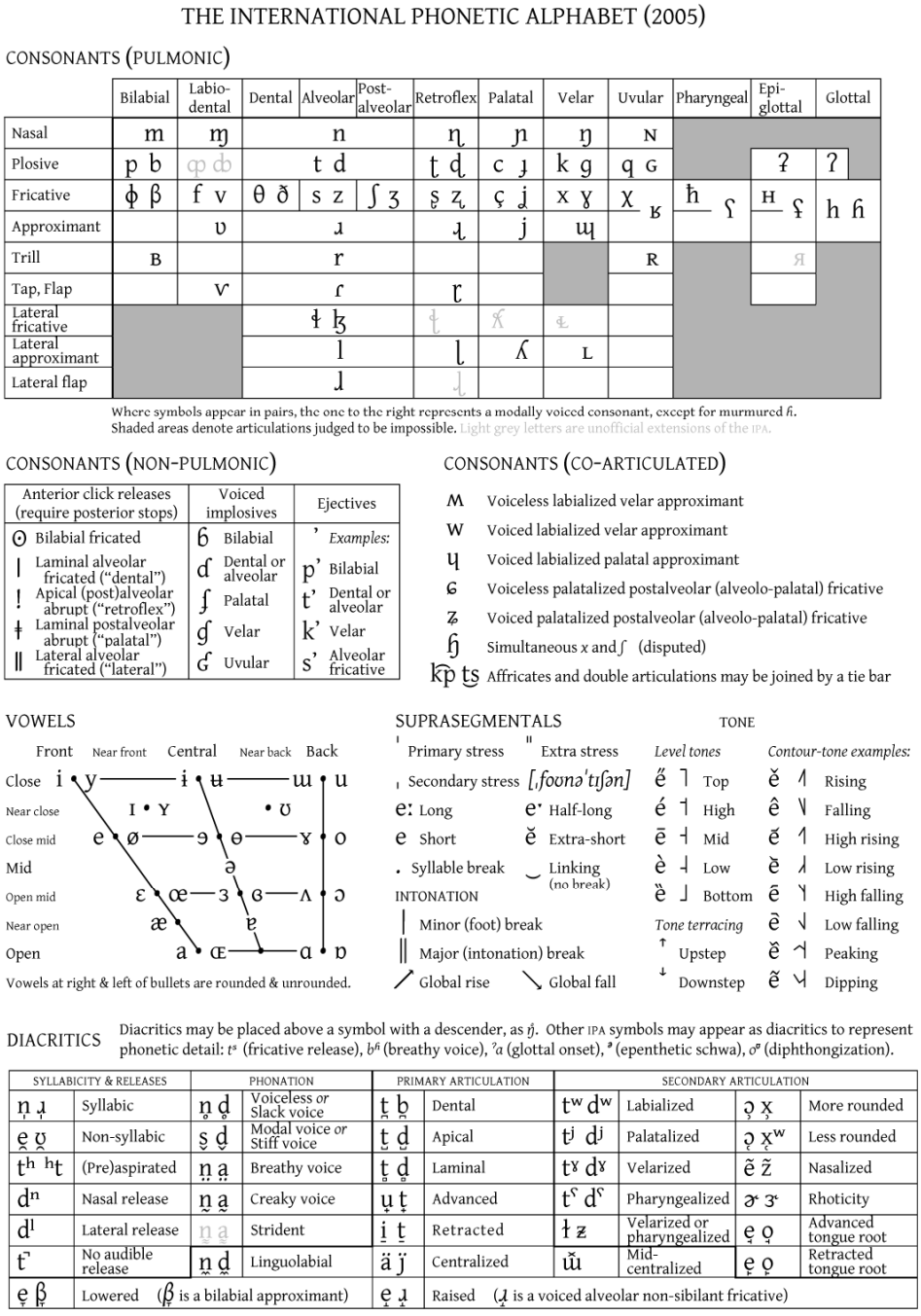
28. D. Jones (1918/1967). "*An Outline of English Phonetics*". Cambridge: Heffer. Ninth edition. (1967).

29. D. Deterding. "*The formants monophthong vowels in Standard Southern British English pronunciation*". Journal of the International Phonetic Association 27, 47-55. (1997).

30. J. O. Smith III, J. S. Abel. "*The Bark and ERB Bilinear Transforms*". IEEE Transactions on Speech and Audio Processing, (1999).

# 9. APPENDIX

## *9.1. International phonetic alphabet*



## *9.2. Code examples*

### 9.2.1. 3D region main features

```
glDisable GL_LIGHTING                              ; # Enable Lighting
glPolygonMode GL_BACK GL_FILL                      ; # Back Face Is Solid
glPolygonMode GL_FRONT GL_LINE                     ; # Front Face Is Made Of Lines
glBindTexture GL_TEXTURE_2D [$::texture get 1]
glBegin GL_POLYGON                                ; # Drawing
   # Back Face
    glTexCoord2f 0.0 0.0 ; glVertex3f -0.5 -0.5 -0.5 ;
    glTexCoord2f 0.0 1.0 ; glVertex3f -0.5  0.5 -0.5 ;
    glTexCoord2f 1.0 1.0 ; glVertex3f  0.5  0.5 -0.5 ;
    glTexCoord2f 1.0 0.0 ; glVertex3f  0.5 -0.5 -0.5 ;
glEnd                                              ; # Finished Drawing
```

### 9.2.2. Drawing of the moving ball

```
proc DrawObject {} {
    glColor3f 1.0 1.0 1.0                                ; # Set Color To White
    glEnable GL_LIGHT0                                   ; # Enable Light 0
    glEnable GL_LIGHTING                                 ; # Enable Lighting
    glPolygonMode GL_FRONT_AND_BACK GL_FILL
    glBindTexture GL_TEXTURE_2D [$::texture get 2]       ; # Select Texture 2 (1)
    gluSphere $::quadric 0.05 20 20                      ; # Draw First Sphere
}
```

### 9.2.3. Target Spheres code

```
proc Vowel {vow color} {
  glTranslatef [lindex $vow 0][lindex $vow 1][lindex $vow 2] ; # Position The Object
  set ::testVowel [list [lindex $vow 0] [lindex $vow 1] [lindex $vow 2]]
  glColor3f 1.0 1.0 1.0                                      ; # Set Color To White
  glEnable GL_LIGHT0                                         ; # Enable Light 0
  glEnable GL_LIGHTING                                       ; # Enable Lighting
  glPolygonMode GL_FRONT_AND_BACK GL_LINE
  glBindTexture GL_TEXTURE_2D [$::texture get $color]
  gluSphere $::quadrica $::ratio 20 20                      ; # Draw First Sphere
}
```

### 9.2.4. Calibration code

```
proc villeCalibrate {} {

  set answer [$::in2 callFunc Calibration "Initializing calibration..."]

#Initialize values
  set ::test 0
  set ::calibrate 1

  Initialize $::frameR
  DisableButtons $::frameL

  set ::info(test,say) "Calibrate"

  sound info(soundObj,a)
  info(soundObj,a) read A.wav

#limits for the maximum in f2
  set f [FormantExtraction info(soundObj,a)]
```

```
  if       {[lindex $f 0] > 550} {
      set ::maxY [lindex $f 0]
      puts "maxY $::maxY"
  }

  sound info(soundObj,i)
  info(soundObj,i) read I.wav

#limits for the maximum in f2 and minimum in f1
  set f [FormantExtraction info(soundObj,i)]
  if       {[lindex $f 1] > 1700} {
     set ::maxX [lindex $f 1]
     puts "maxX $::maxX"
  }
  if {[lindex $f 0] < 400} {
      set ::minY [lindex $f 0]
      puts "minY $::minY"
  }
  set ::f3Normalized [lindex $f 2]

  sound info(soundObj,u)
  info(soundObj,u) read O.wav

#limits for the minimum in f2 and minimum of f1
  set f [FormantExtraction info(soundObj,u)]
  if       {[lindex $f 0] < 800} {
      set ::minX [lindex $f 1]
      puts "minX $::minX"
  }
  if {[lindex $f 1] < $::minY} {
    set ::minY [lindex $f 0]
    puts "minY $::minY"
  }

  set ::info(test,say) "End of calibration"
  set ::calibrate 0

  #write into a file the results
  set user $::info(user,name)
  append user .usr
  set f [open ./users/$user w]
    puts $f $::info(user,name)
    puts $f $::info(user,height)
    puts $f $::info(user,age)
    puts $f $::info(user,language)
    puts $f $::info(user,sex)
    puts $f "### Values of normalization ###################################"
    puts $f $::minX
    puts $f $::maxX
    puts $f $::minY
    puts $f $::maxY
  close $f

  EnableButtons $::frameL
}
```

## 9.2.5. Normalization code

```
proc Normalize {val min max} {

#  Update the buffer
    for {set i 0} {$i<[llength $::bufferAux]} {incr i} {
           lset ::bufferAux $i [lindex $::bufferAux [expr $i+1]]
    }
    lset ::bufferAux [expr [llength $::bufferAux]-1] [lindex $val]

    set mean [::math::statistics::mean $::bufferAux]
    set val $mean

#if there is a calibration running
  if {$::calibrate == 1} {
    set auxmin [expr $$min+0]
```

```
    set auxmax [expr $$max+0]

    if {$auxmax < $val} {
           set $max $val
    }
    if {$auxmin > $val} {
           set $min $val
    }
  }

    set interval [expr $$max-$$min]
    set fnormal [expr ($val-$$min)/($interval*1.0)-0.5]
    return $fnormal
}
```

## *9.3. Surveys*

### 9.3.1. Profile of the testers

1. Which is your nationality?

2. Which is your mother tongue?

3. How many languages are you able to speak?

4. When did you arrive in Sweden?

5. Do you have relatives living in Sweden, Swedish husband/wife?

6. Do you have Swedish friends?

7. Do you try to speak Swedish with friends or in the city?

8. Do you have lectures in Swedish?

9. Why do you want to learn Swedish, what are your goals with the new language?

10. Is it important for you to achieve a good Swedish pronunciation?

11. Do you think that your Swedish level is good, on a scale from 1 (very poor) – 5 (native speaker) how good is it?

12. How tall are you?

13. Male or female?

14. Do you have any hearing problem?

### 9.3.2. The software

1. Do you think that any previous knowledge about computers is necessary to work with the system?

2. How did the system meet with your expectations?

3. Was the system intuitive for you?

4. Did the talking head help you to use the software?

5. How natural did you feel that the interaction with the talking head was in the calibration stage?

6. Did you want more instructions about the program before starting with it?

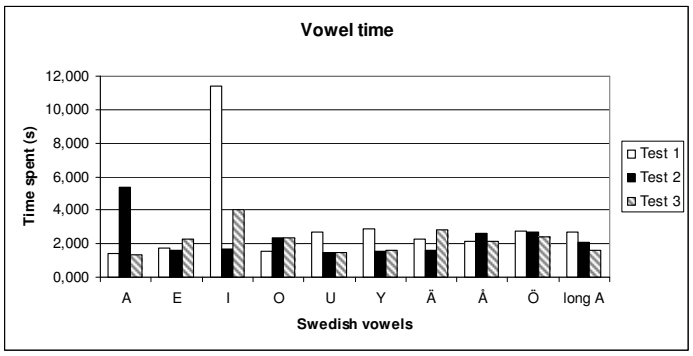
Any comments you would like to add:

## *9.4. Test results*

### 9.4.1. Session 1

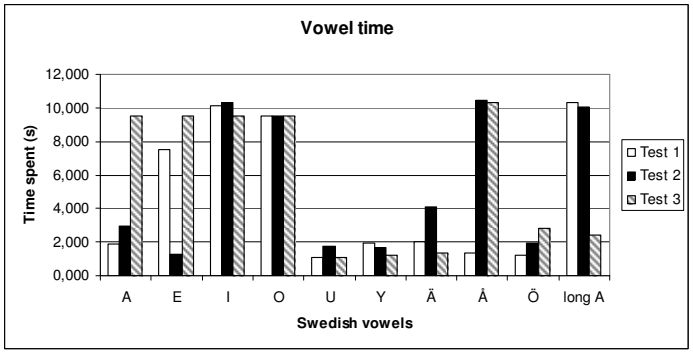
Tester 1. Swedish male.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 1,424 | 5,337 | 1,348 |
| E | 1,758 | 1,597 | 2,281 |
| I | 11,375 | 1,690 | 4,013 |
| O | 1,549 | 2,336 | 2,352 |
| U | 2,656 | 1,507 | 1,492 |
| Y | 2,870 | 1,517 | 1,590 |
| Ä | 2,310 | 1,606 | 2,831 |
| Å | 2,119 | 2,614 | 2,140 |
| Ö | 2,750 | 2,680 | 2,440 |
| long A | 2,684 | 2,048 | 1,589 |
| | | | |
| Total time | 31.495 | 22.932 | 22.076 |
| Average time | 3.150 | 2.293 | 2.208 |



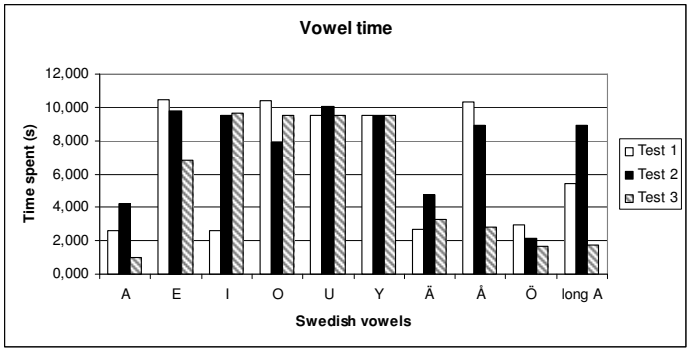

Tester 2. Swedish female.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 1,878 | 2,954 | 9,536 |
| E | 7,537 | 1,261 | 9,514 |
| I | 10,134 | 10,319 | 9,529 |
| O | 9,533 | 9,508 | 9,532 |
| U | 1,043 | 1,762 | 1,093 |
| Y | 1,971 | 1,677 | 1,201 |
| Ä | 2,011 | 4,103 | 1,320 |
| Å | 1,356 | 10,453 | 10,296 |
| Ö | 1,215 | 1,914 | 2,827 |
| long A | 10,354 | 10,063 | 2,441 |
| | | | |
| Total time | 47.032 | 54.014 | 57.289 |
| Average time | 4.703 | 5.401 | 5.729 |



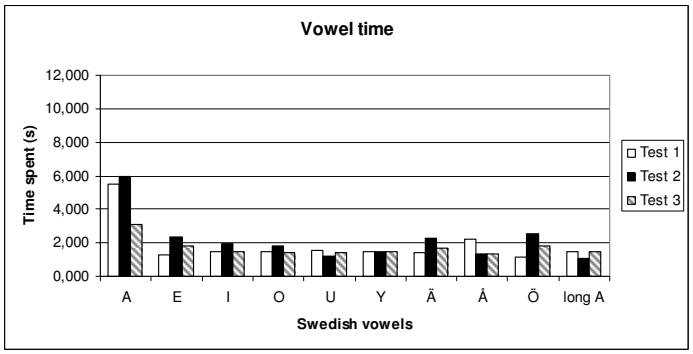
## Tester 3. Swedish male.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 2,612 | 4,223 | 1,038 |
| E | 10,447 | 9,791 | 6,841 |
| I | 2,593 | 9,510 | 9,661 |
| O | 10,367 | 7,909 | 9,511 |
| U | 9,509 | 10,079 | 9,509 |
| Y | 9,515 | 9,509 | 9,531 |
| Ä | 2,670 | 4,783 | 3,265 |
| Å | 10,301 | 8,904 | 2,810 |
| Ö | 2,983 | 2,147 | 1,652 |
| long A | 5,420 | 8,932 | 1,735 |
| Total time | 66.417 | 75.787 | 55.553 |
| Average time | 6.642 | 7.579 | 5.555 |



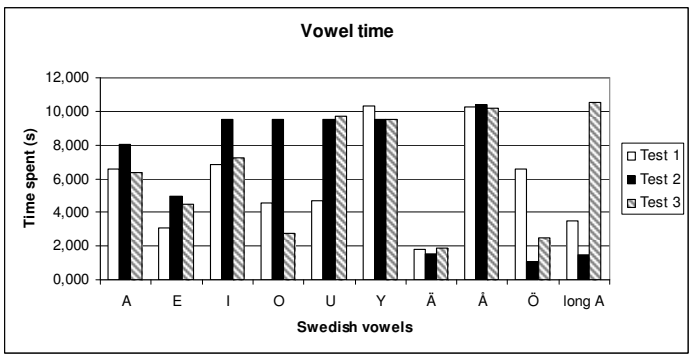

## Tester 4. Swedish female.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 5,510 | 5,970 | 3,073 |
| E | 1,287 | 2,354 | 1,790 |
| I | 1,497 | 1,964 | 1,496 |
| O | 1,503 | 1,788 | 1,432 |
| U | 1,548 | 1,221 | 1,418 |
| Y | 1,479 | 1,494 | 1,443 |
| Ä | 1,422 | 2,265 | 1,682 |
| Å | 2,239 | 1,341 | 1,310 |
| Ö | 1,146 | 2,514 | 1,796 |
| long A | 1,491 | 1,040 | 1,460 |
| Total time | 19.122 | 21.951 | 16.900 |
| Average time | 1.912 | 2.195 | 1.690 |



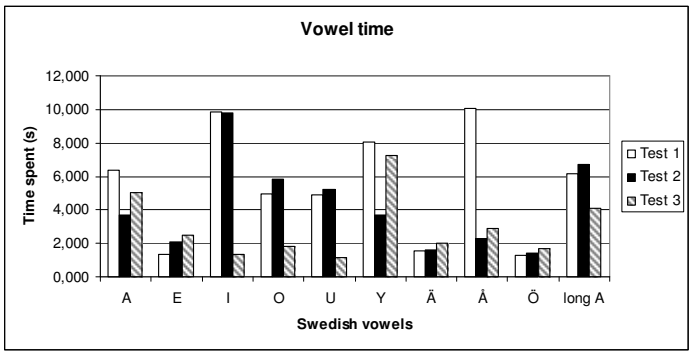

## Tester 5. Swedish male.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 6,551 | 8,043 | 6,362 |
| E | 3,061 | 4,930 | 4,478 |
| I | 6,864 | 9,547 | 7,207 |
| O | 4,546 | 9,549 | 2,780 |
| U | 4,707 | 9,530 | 9,688 |
| Y | 10,323 | 9,512 | 9,513 |
| Ä | 1,781 | 1,564 | 1,855 |
| Å | 10,248 | 10,408 | 10,164 |
| Ö | 6,553 | 1,040 | 2,481 |
| long A | 3,496 | 1,477 | 10,556 |
| Total time | 58.130 | 65.600 | 65.084 |
| Average time | 5.813 | 6.560 | 6.508 |



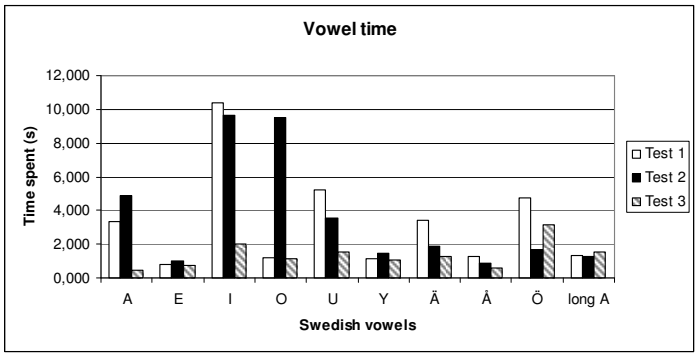
## Tester 6. Swedish male.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 6,392 | 3,718 | 5,011 |
| E | 1,352 | 2,055 | 2,472 |
| I | 9,860 | 9,755 | 1,356 |
| O | 4,969 | 5,827 | 1,840 |
| U | 4,865 | 5,231 | 1,116 |
| Y | 8,052 | 3,709 | 7,244 |
| Ä | 1,510 | 1,600 | 2,023 |
| Å | 10,060 | 2,284 | 2,855 |
| Ö | 1,264 | 1,394 | 1,678 |
| long A | 6,193 | 6,685 | 4,099 |
| | | | |
| Total time | 54.517 | 42.258 | 29.694 |
| Average time | 5.452 | 4.226 | 2.969 |



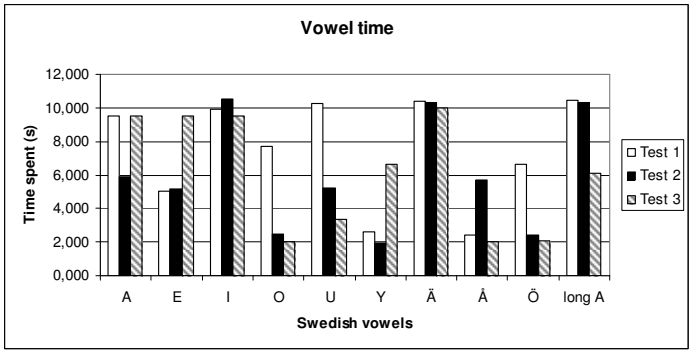

## Tester 7. Swedish female.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 3,356 | 4,867 | 0,485 |
| E | 0,807 | 1,020 | 0,709 |
| I | 10,414 | 9,640 | 2,039 |
| O | 1,209 | 9,540 | 1,146 |
| U | 5,196 | 3,571 | 1,525 |
| Y | 1,163 | 1,465 | 1,082 |
| Ä | 3,447 | 1,893 | 1,270 |
| Å | 1,295 | 0,871 | 0,588 |
| Ö | 4,754 | 1,700 | 3,133 |
| long A | 1,352 | 1,266 | 1,513 |
| | | | |
| Total time | 32.993 | 35.833 | 13.490 |
| Average time | 3.299 | 3.583 | 1.349 |



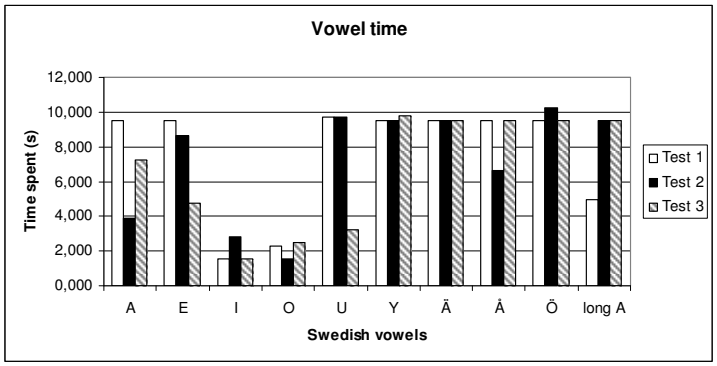

## Tester 8. Spanish female.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 9,508 | 5,889 | 9,530 |
| E | 5,050 | 5,191 | 9,516 |
| I | 9,950 | 10,514 | 9,515 |
| O | 7,726 | 2,457 | 2,031 |
| U | 10,274 | 5,259 | 3,336 |
| Y | 2,647 | 1,966 | 6,645 |
| Ä | 10,400 | 10,318 | 9,957 |
| Å | 2,402 | 5,691 | 2,014 |
| Ö | 6,654 | 2,406 | 2,083 |
| long A | 10,429 | 10,325 | 6,088 |
| | | | |
| Total time | 75.040 | 60.016 | 60.715 |
| Average time | 7.504 | 6.002 | 6.072 |



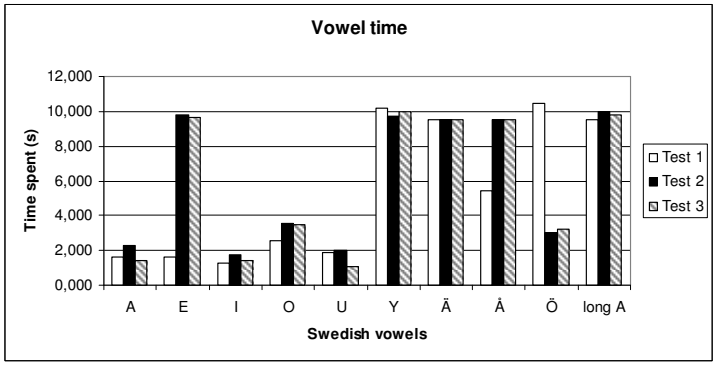
Tester 9. Italian female.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 9,512 | 3,867 | 7,257 |
| E | 9,515 | 8,641 | 4,791 |
| I | 1,532 | 2,799 | 1,530 |
| O | 2,311 | 1,537 | 2,481 |
| U | 9,736 | 9,699 | 3,186 |
| Y | 9,546 | 9,547 | 9,782 |
| Ä | 9,516 | 9,547 | 9,516 |
| Å | 9,516 | 6,666 | 9,516 |
| Ö | 9,515 | 10,240 | 9,547 |
| long A | 4,960 | 9,516 | 9,515 |
| Total time | 75.659 | 72.059 | 67.121 |
| Average time | 7.566 | 7.206 | 6.712 |



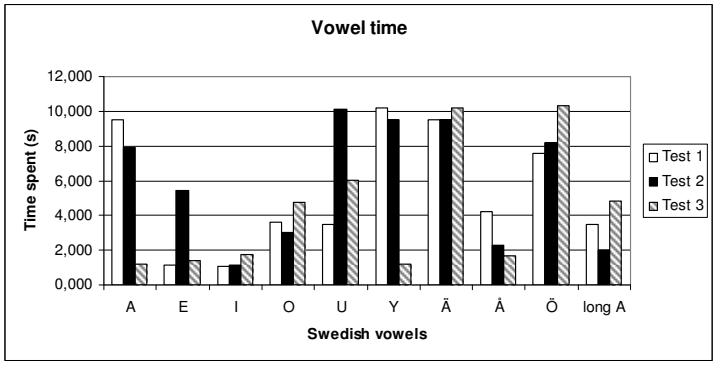

Tester 10. Spanish male.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 1,577 | 2,310 | 1,404 |
| E | 1,592 | 9,768 | 9,626 |
| I | 1,300 | 1,744 | 1,441 |
| O | 2,549 | 3,548 | 3,506 |
| U | 1,906 | 1,981 | 1,074 |
| Y | 10,164 | 9,696 | 9,978 |
| Ä | 9,516 | 9,515 | 9,516 |
| Å | 5,435 | 9,516 | 9,516 |
| Ö | 10,471 | 3,037 | 3,185 |
| long A | 9,516 | 9,979 | 9,783 |
| Total time | 54.026 | 61.094 | 59.029 |
| Average time | 5.403 | 6.109 | 5.903 |



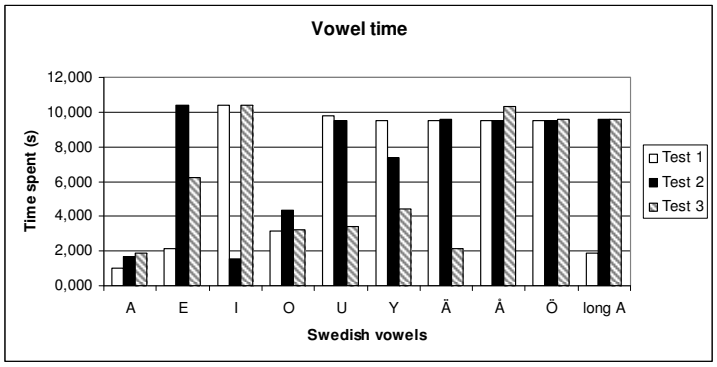

Tester 11. Spanish male.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 9,515 | 7,902 | 1,235 |
| E | 1,122 | 5,400 | 1,427 |
| I | 1,075 | 1,109 | 1,741 |
| O | 3,642 | 3,002 | 4,749 |
| U | 3,500 | 10,142 | 6,029 |
| Y | 10,178 | 9,506 | 1,222 |
| Ä | 9,552 | 9,534 | 10,188 |
| Å | 4,231 | 2,250 | 1,658 |
| Ö | 7,582 | 8,211 | 10,298 |
| long A | 3,458 | 2,028 | 4,817 |
| Total time | 53.855 | 59.084 | 43.364 |
| Average time | 5.386 | 5.908 | 4.336 |



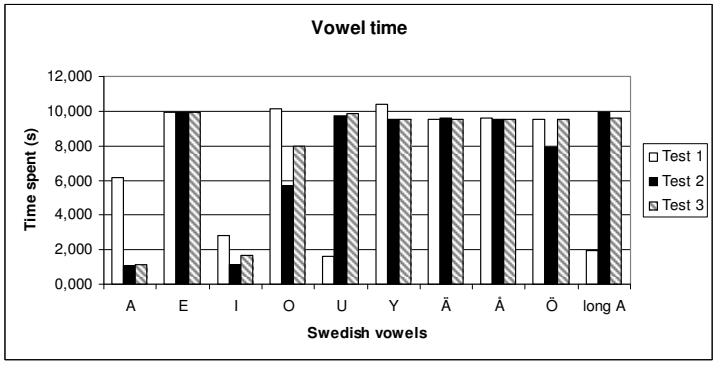
## Tester 12. Italian female.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 1,026 | 1,681 | 1,851 |
| E | 2,134 | 10,396 | 6,263 |
| I | 10,418 | 1,572 | 10,417 |
| O | 3,128 | 4,338 | 3,249 |
| U | 9,809 | 9,543 | 3,402 |
| Y | 9,516 | 7,407 | 4,412 |
| Ä | 9,547 | 9,578 | 2,122 |
| Å | 9,531 | 9,515 | 10,331 |
| Ö | 9,515 | 9,516 | 9,562 |
| long A | 1,861 | 9,563 | 9,578 |
| Total time | 66.485 | 73.109 | 61.187 |
| Average time | 6.649 | 7.311 | 6.119 |



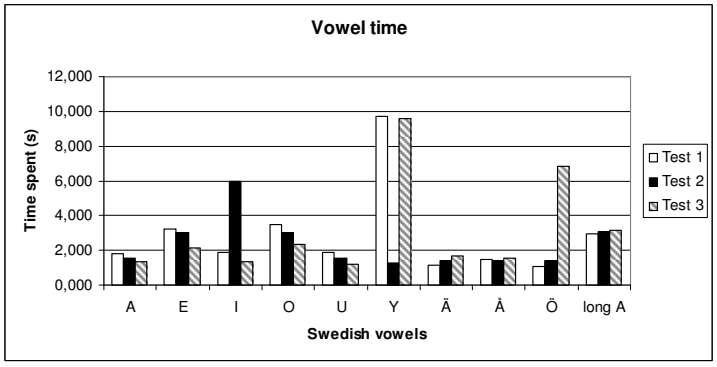

## Tester 13. Italian female.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 6,166 | 1,090 | 1,113 |
| E | 9,938 | 9,938 | 9,930 |
| I | 2,833 | 1,142 | 1,653 |
| O | 10,136 | 5,720 | 7,987 |
| U | 1,584 | 9,716 | 9,876 |
| Y | 10,400 | 9,516 | 9,515 |
| Ä | 9,516 | 9,578 | 9,532 |
| Å | 9,562 | 9,516 | 9,531 |
| Ö | 9,516 | 7,908 | 9,547 |
| long A | 1,944 | 9,935 | 9,562 |
| Total time | 71.595 | 74.059 | 78.246 |
| Average time | 7.160 | 7.406 | 7.825 |



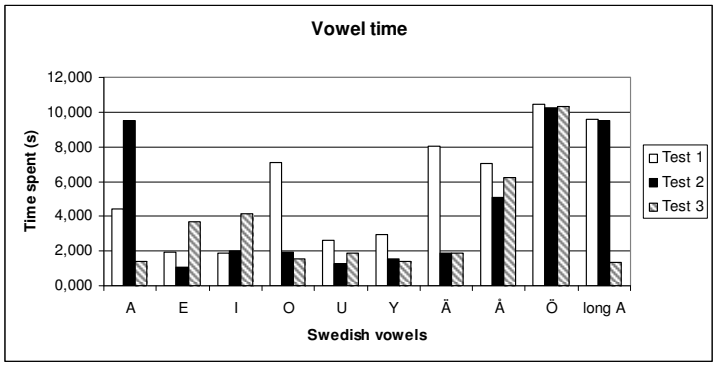

## Tester 14. Spanish male.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 1,821 | 1,528 | 1,335 |
| E | 3,198 | 3,009 | 2,172 |
| I | 1,855 | 5,954 | 1,357 |
| O | 3,517 | 2,986 | 2,359 |
| U | 1,890 | 1,548 | 1,217 |
| Y | 9,753 | 1,287 | 9,594 |
| Ä | 1,114 | 1,421 | 1,644 |
| Å | 1,481 | 1,398 | 1,553 |
| Ö | 1,078 | 1,410 | 6,870 |
| long A | 2,941 | 3,054 | 3,157 |
| Total time | 28.648 | 23.595 | 31.258 |
| Average time | 2.865 | 2.360 | 3.126 |



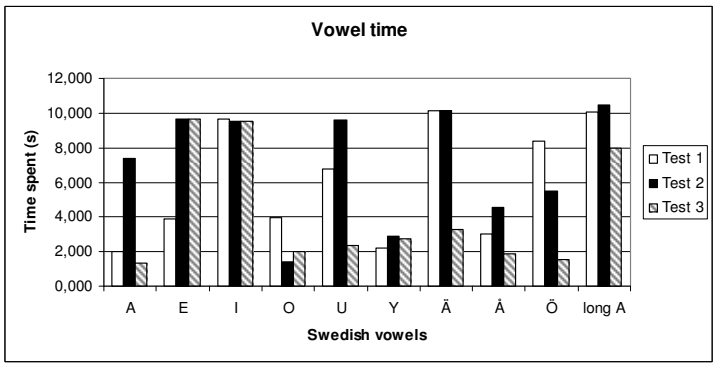
## Tester 15. French female.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 4,440 | 9,546 | 1,440 |
| E | 1,964 | 1,082 | 3,663 |
| I | 1,862 | 1,985 | 4,128 |
| O | 7,135 | 1,959 | 1,554 |
| U | 2,648 | 1,261 | 1,896 |
| Y | 2,977 | 1,557 | 1,383 |
| Ä | 8,075 | 1,853 | 1,877 |
| Å | 7,030 | 5,063 | 6,263 |
| Ö | 10,427 | 10,255 | 10,342 |
| long A | 9,578 | 9,516 | 1,331 |
| | | | |
| Total time | 56.136 | 44.077 | 33.877 |
| Average time | 5.614 | 4.408 | 3.388 |



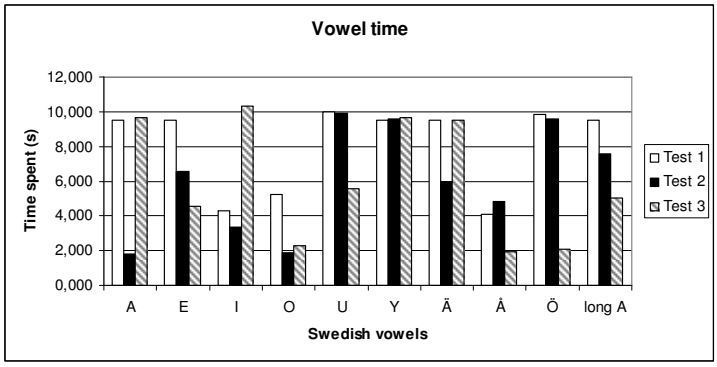

## Tester 16. Syrian male.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 1,981 | 7,373 | 1,342 |
| E | 3,870 | 9,670 | 9,667 |
| I | 9,676 | 9,518 | 9,510 |
| O | 3,956 | 1,398 | 2,018 |
| U | 6,798 | 9,583 | 2,320 |
| Y | 2,190 | 2,889 | 2,736 |
| Ä | 10,104 | 10,093 | 3,283 |
| Å | 3,035 | 4,527 | 1,877 |
| Ö | 8,360 | 5,487 | 1,517 |
| long A | 10,070 | 10,466 | 8,004 |
| | | | |
| Total time | 60.040 | 71.004 | 42.274 |
| Average time | 6.004 | 7.100 | 4.227 |




## Tester 17. German male.

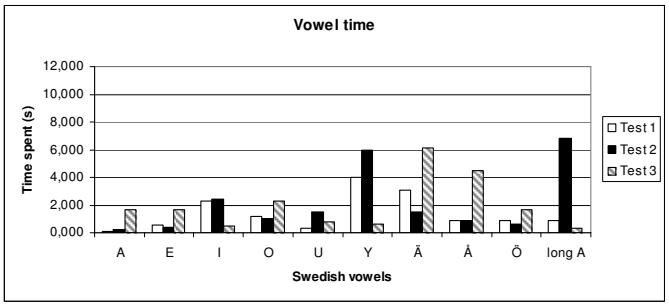
| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 9,500 | 1,834 | 9,670 |
| E | 9,505 | 6,590 | 4,532 |
| I | 4,263 | 3,367 | 10,350 |
| O | 5,240 | 1,855 | 2,262 |
| U | 10,003 | 9,905 | 5,558 |
| Y | 9,515 | 9,555 | 9,675 |
| Ä | 9,546 | 5,996 | 9,544 |
| Å | 4,096 | 4,816 | 1,971 |
| Ö | 9,869 | 9,585 | 2,095 |
| long A | 9,535 | 7,593 | 4,999 |
| | | | |
| Total time | 81.072 | 61.096 | 60.656 |
| Average time | 8.107 | 6.110 | 6.066 |



## 9.4.2. Session 2

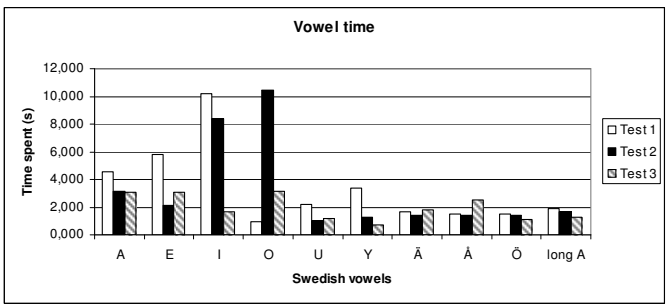
Tester 1. Swedish male.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 0,041 | 0,202 | 1,686 |
| E | 0,530 | 0,398 | 1,686 |
| I | 2,236 | 2,424 | 0,505 |
| O | 1,175 | 1,006 | 2,289 |
| U | 0,313 | 1,526 | 0,778 |
| Y | 3,983 | 5,960 | 0,607 |
| Ä | 3,068 | 1,508 | 6,150 |
| Å | 0,895 | 0,844 | 4,497 |
| Ö | 0,836 | 0,658 | 1,681 |
| long A | 0,885 | 6,817 | 0,341 |
| | | | |
| Total time | 13.962 | 21.343 | 21.513 |
| Average time | 1.396 | 2.134 | 2.151 |



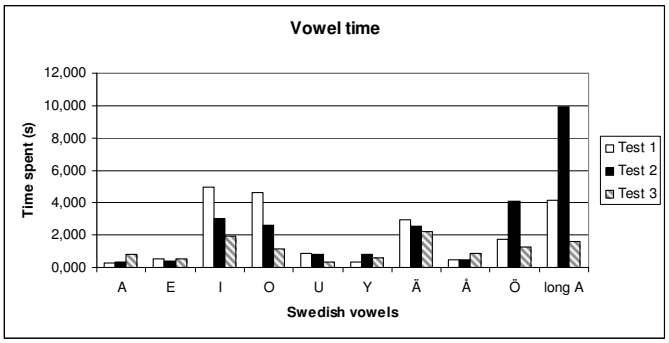

Tester 3. Swedish male.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 4,538 | 3,099 | 3,055 |
| E | 5,789 | 2,125 | 3,055 |
| I | 10,189 | 8,358 | 1,636 |
| O | 0,977 | 10,450 | 3,132 |
| U | 2,180 | 1,045 | 1,168 |
| Y | 3,381 | 1,277 | 0,682 |
| Ä | 1,655 | 1,439 | 1,843 |
| Å | 1,498 | 1,439 | 2,477 |
| Ö | 1,498 | 1,381 | 1,090 |
| long A | 1,847 | 1,628 | 1,271 |
| | | | |
| Total time | 33.552 | 32.241 | 18.753 |
| Average time | 3.355 | 3.224 | 1.875 |



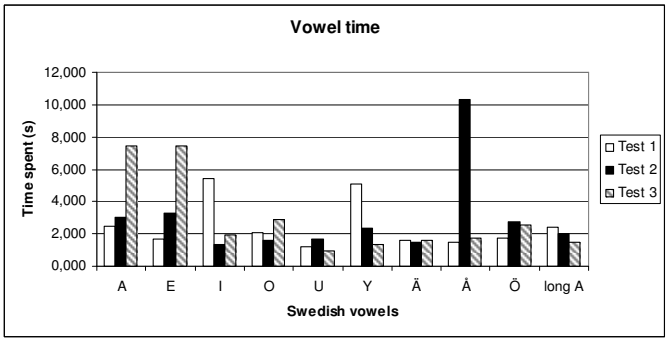
Tester 4. Swedish female.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 0,284 | 0,302 | 0,796 |
| E | 0,547 | 0,424 | 0,539 |
| I | 4,972 | 3,027 | 1,946 |
| O | 4,603 | 2,647 | 1,166 |
| U | 0,894 | 0,817 | 0,365 |
| Y | 0,314 | 0,807 | 0,610 |
| Ä | 2,945 | 2,516 | 2,194 |
| Å | 0,451 | 0,467 | 0,890 |
| Ö | 1,718 | 4,083 | 1,300 |
| long A | 4,147 | 9,944 | 1,597 |
| | | | |
| Total time | 20.875 | 25.034 | 11.403 |
| Average time | 2.088 | 2.503 | 1.140 |



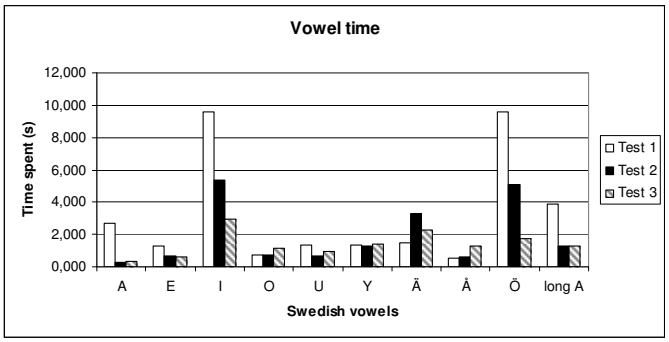

Tester 5. Swedish male.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 2,512 | 3,005 | 7,432 |
| E | 1,680 | 3,256 | 7,432 |
| I | 5,455 | 1,331 | 1,933 |
| O | 2,064 | 1,610 | 2,869 |
| U | 1,230 | 1,676 | 0,970 |
| Y | 5,084 | 2,352 | 1,347 |
| Ä | 1,590 | 1,474 | 1,580 |
| Å | 1,501 | 10,351 | 1,763 |
| Ö | 1,715 | 2,744 | 2,563 |
| long A | 2,415 | 2,026 | 1,460 |
| | | | |
| Total time | 25.246 | 29.825 | 23.284 |
| Average time | 2.525 | 2.983 | 2.328 |



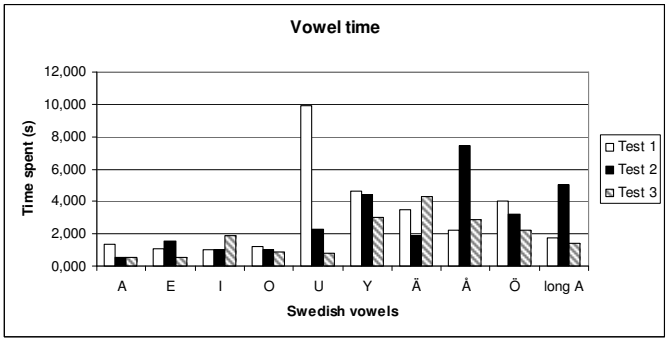

Tester 7. Swedish female.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 2,681 | 0,244 | 0,326 |
| E | 1,257 | 0,641 | 0,579 |
| I | 9,585 | 5,350 | 2,925 |
| O | 0,746 | 0,753 | 1,138 |
| U | 1,350 | 0,642 | 0,965 |
| Y | 1,340 | 1,287 | 1,391 |
| Ä | 1,468 | 3,313 | 2,266 |
| Å | 0,544 | 0,612 | 1,266 |
| Ö | 9,573 | 5,076 | 1,747 |
| long A | 3,918 | 1,263 | 1,257 |
| | | | |
| Total time | 32.462 | 19.181 | 13.860 |
| Average time | 3.246 | 1.918 | 1.386 |



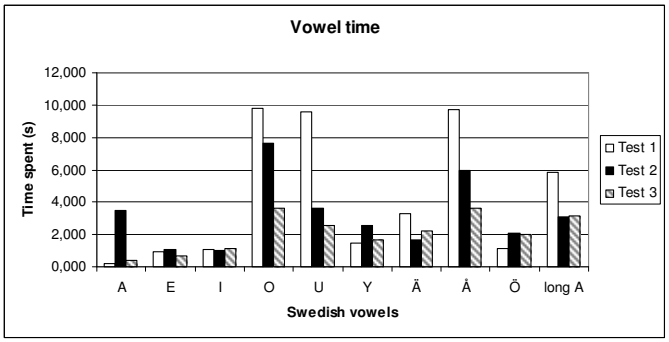
Tester 9. Italian female.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 1,372 | 0,511 | 0,544 |
| E | 1,051 | 1,565 | 0,544 |
| I | 0,980 | 1,020 | 1,873 |
| O | 1,202 | 1,003 | 0,871 |
| U | 9,943 | 2,268 | 0,788 |
| Y | 4,643 | 4,398 | 3,031 |
| Ä | 3,471 | 1,859 | 4,310 |
| Å | 2,210 | 7,432 | 2,852 |
| Ö | 4,010 | 3,212 | 2,233 |
| long A | 1,756 | 5,027 | 1,414 |
| Total time | 30.638 | 28.295 | 19.639 |
| Average time | 3.064 | 2.830 | 1.964 |



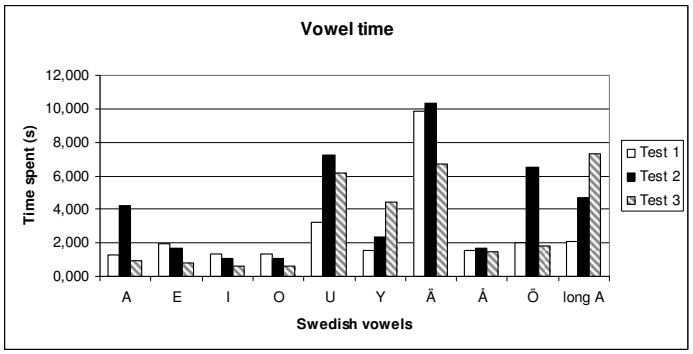

Tester 10. Spanish male.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 0,180 | 3,470 | 0,417 |
| E | 0,931 | 1,090 | 0,667 |
| I | 1,096 | 1,028 | 1,173 |
| O | 9,802 | 7,638 | 3,589 |
| U | 9,557 | 3,641 | 2,534 |
| Y | 1,460 | 2,553 | 1,651 |
| Ä | 3,287 | 1,676 | 2,182 |
| Å | 9,754 | 5,935 | 3,614 |
| Ö | 1,135 | 2,076 | 1,988 |
| long A | 5,834 | 3,073 | 3,166 |
| Total time | 43.036 | 32.180 | 20.981 |
| Average time | 4.304 | 3.218 | 2.098 |



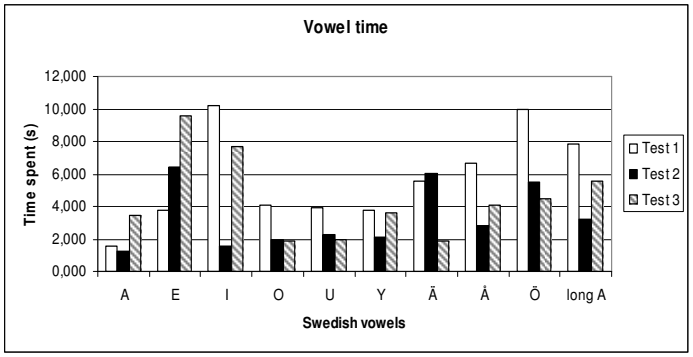

Tester 11. Spanish male.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 1,250 | 4,243 | 0,909 |
| E | 1,972 | 1,708 | 0,832 |
| I | 1,362 | 1,057 | 0,633 |
| O | 1,314 | 1,092 | 0,600 |
| U | 3,242 | 7,244 | 6,178 |
| Y | 1,544 | 2,317 | 4,412 |
| Ä | 9,853 | 10,351 | 6,673 |
| Å | 1,541 | 1,657 | 1,508 |
| Ö | 1,994 | 6,475 | 1,806 |
| long A | 2,068 | 4,695 | 7,288 |
| Total time | 26.140 | 40.839 | 30.839 |
| Average time | 2.614 | 4.084 | 3.084 |



Tester 16. Syrian male.

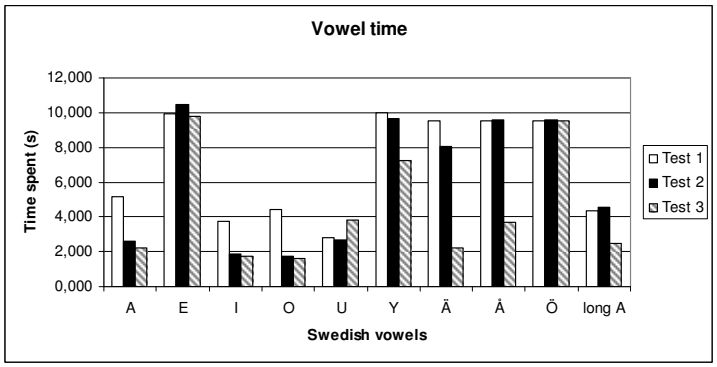
| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 1,584 | 1,278 | 3,478 |
| E | 3,792 | 6,400 | 9,592 |
| I | 10,168 | 1,590 | 7,714 |
| O | 4,089 | 1,936 | 1,852 |
| U | 3,944 | 2,272 | 1,982 |
| Y | 3,754 | 2,136 | 3,617 |
| Ä | 5,581 | 6,018 | 1,855 |
| Å | 6,628 | 2,861 | 4,097 |
| Ö | 9,989 | 5,482 | 4,473 |
| long A | 7,824 | 3,181 | 5,561 |
| | | | |
| Total time | 57.353 | 33.154 | 44.221 |
| Average time | 5.735 | 3.315 | 4.422 |



## 9.4.3. Tests with the special tester

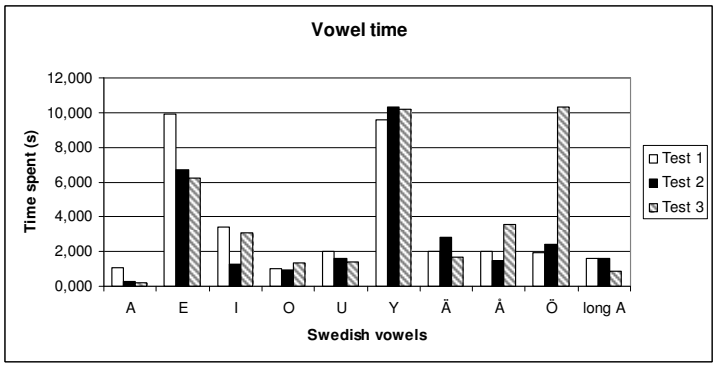
Tester 18. Italian female. Session 1.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 5,135 | 2,621 | 2,241 |
| E | 9,919 | 10,437 | 9,783 |
| I | 3,776 | 1,866 | 1,741 |
| O | 4,441 | 1,753 | 1,612 |
| U | 2,818 | 2,675 | 3,828 |
| Y | 9,970 | 9,666 | 7,237 |
| Ä | 9,512 | 8,047 | 2,199 |
| Å | 9,553 | 9,600 | 3,681 |
| Ö | 9,517 | 9,614 | 9,534 |
| long A | 4,348 | 4,528 | 2,474 |
| Total time | 68.989 | 60.807 | 44.330 |
| Average time | 6.899 | 6.081 | 4.433 |



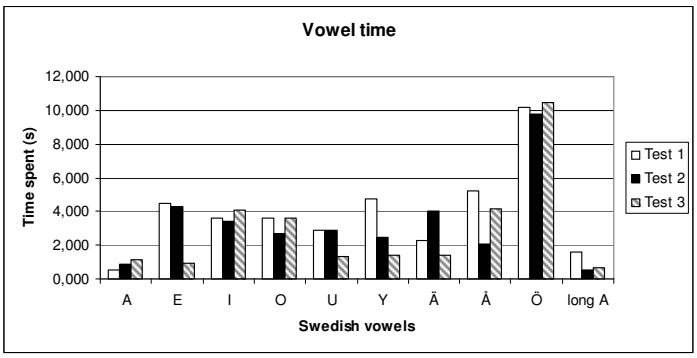

Tester 18. Italian female. Session 2.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 1,075 | 0,269 | 0,215 |
| E | 9,942 | 6,696 | 6,238 |
| I | 3,387 | 1,271 | 3,100 |
| O | 1,020 | 0,907 | 1,342 |
| U | 1,987 | 1,627 | 1,385 |
| Y | 9,608 | 10,292 | 10,190 |
| Ä | 2,026 | 2,783 | 1,646 |
| Å | 1,996 | 1,452 | 3,544 |
| Ö | 1,953 | 2,437 | 10,332 |
| long A | 1,627 | 1,576 | 0,861 |
| Total time | 34.621 | 29.310 | 38.853 |
| Average time | 3.462 | 2.931 | 3.885 |



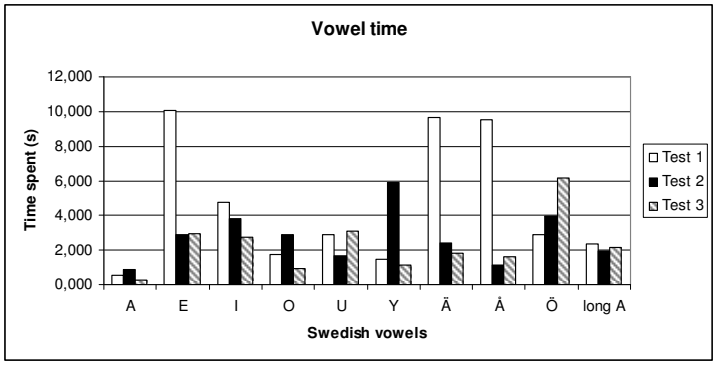
Tester 18. Italian female. Session 3.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 0,503 | 0,900 | 1,128 |
| E | 4,491 | 4,299 | 0,961 |
| I | 3,627 | 3,435 | 4,066 |
| O | 3,625 | 2,674 | 3,629 |
| U | 2,858 | 2,915 | 1,323 |
| Y | 4,743 | 2,451 | 1,407 |
| Ä | 2,264 | 4,010 | 1,428 |
| Å | 5,230 | 2,074 | 4,149 |
| Ö | 10,184 | 9,758 | 10,451 |
| long A | 1,619 | 0,533 | 0,649 |
| | | | |
| Total time | 39.144 | 33.049 | 29.191 |
| Average time | 3.914 | 3.305 | 2.919 |




Tester 18. Italian female. Session 4.

| | Test 1 | Test 2 | Test 3 |
|---|---|---|---|
| A | 0,520 | 0,859 | 0,252 |
| E | 10,039 | 2,882 | 2,927 |
| I | 4,738 | 3,819 | 2,736 |
| O | 1,744 | 2,850 | 0,916 |
| U | 2,911 | 1,700 | 3,113 |
| Y | 1,456 | 5,916 | 1,128 |
| Ä | 9,636 | 2,414 | 1,812 |
| Å | 9,531 | 1,173 | 1,595 |
| Ö | 2,913 | 3,934 | 6,191 |
| long A | 2,376 | 1,917 | 2,114 |
| | | | |
| Total time | 45.864 | 27.464 | 22.784 |
| Average time | 4.586 | 2.746 | 2.278 |



# PRESUPUESTO

1) Ejecución Material

- Compra de ordenador personal (Software incluido) ..... 2.000,00 €
- Material de oficina..... 150,00 €
- Total de ejecución material..... 2.150,00 €

2) Gastos generales

- 16 % sobre Ejecución Material.....352,00 €

3) Beneficio Industrial

- 6 % sobre Ejecución Material.....132,00 €

4) Honorarios Proyecto

- 1000 horas a 15 € / hora..... 15.000,00 €

5) Material fungible

- Gastos de impresión.....60,00 €
- Encuadernación..... 200,00 €

6) Subtotal del presupuesto

- Subtotal Presupuesto..... 17.894,00 €

7) I.V.A. aplicable

- 16% Subtotal Presupuesto ..... 2.863,04 €

8) Total presupuesto

- Total Presupuesto ..... 20.757,04 €

Madrid, Julio de 2008

El Ingeniero Jefe de Proyecto
Fdo.: David Lucas Escribano
Ingeniero Superior de Telecomunicación

# PLIEGO DE CONDICIONES

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, de "*Pronunciation training of Swedish vowels using speech technology, emboided conversational agents and an interactive game*". En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho sistema. Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

Supuesto que la utilización industrial de los métodos recogidos en el presente proyecto ha sido decidida por parte de la empresa cliente o de otras, la obra a realizar se regulará por las siguientes:

## Condiciones generales

1. La modalidad de contratación será el concurso. La adjudicación se hará, por tanto, a la proposición más favorable sin atender exclusivamente al valor económico, dependiendo de las mayores garantías ofrecidas. La empresa que somete el proyecto a concurso se reserva el derecho a declararlo desierto.

2. El montaje y mecanización completa de los equipos que intervengan será realizado totalmente por la empresa licitadora.

3. En la oferta, se hará constar el precio total por el que se compromete a realizar la obra y el tanto por ciento de baja que supone este precio en relación con un importe límite si este se hubiera fijado.

4. La obra se realizará bajo la dirección técnica de un Ingeniero Superior de Telecomunicación, auxiliado por el número de Ingenieros Técnicos y Programadores que se estime preciso para el desarrollo de la misma.

5. Aparte del Ingeniero Director, el contratista tendrá derecho a contratar al resto del personal, pudiendo ceder esta prerrogativa a favor del Ingeniero Director, quien no estará obligado a aceptarla.

6. El contratista tiene derecho a sacar copias a su costa de los planos, pliego de condiciones y presupuestos. El Ingeniero autor del proyecto autorizará con su firma las copias solicitadas por el contratista después de confrontarlas.

7. Se abonará al contratista la obra que realmente ejecute con sujeción al proyecto que sirvió de base para la contratación, a las modificaciones autorizadas por la superioridad o a las órdenes que con arreglo a sus facultades le hayan comunicado por escrito al Ingeniero Director de obras siempre que dicha obra se haya ajustado a los preceptos de los pliegos de condiciones, con arreglo a los cuales, se harán las modificaciones y la valoración de las diversas unidades sin que el importe total pueda exceder de los presupuestos aprobados. Por consiguiente, el número de unidades que se consignan en el proyecto o en el presupuesto, no podrá servirle de fundamento para entablar reclamaciones de ninguna clase, salvo en los casos de rescisión.

8. Tanto en las certificaciones de obras como en la liquidación final, se abonarán los trabajos realizados por el contratista a los precios de ejecución material que figuran en el presupuesto para cada unidad de la obra.

9. Si excepcionalmente se hubiera ejecutado algún trabajo que no se ajustase a las condiciones de la contrata pero que sin embargo es admisible a juicio del Ingeniero Director de obras, se dará conocimiento a la Dirección, proponiendo a la vez la rebaja de precios que el Ingeniero estime justa y si la Dirección resolviera aceptar la obra, quedará el contratista obligado a conformarse con la rebaja acordada.

10. Cuando se juzgue necesario emplear materiales o ejecutar obras que no figuren en el presupuesto de la contrata, se evaluará su importe a los precios asignados a otras obras o materiales análogos si los hubiere y cuando no, se discutirán entre el Ingeniero Director y el contratista, sometiéndolos a la aprobación de la Dirección. Los nuevos precios convenidos por uno u otro procedimiento, se sujetarán siempre al establecido en el punto anterior.

11. Cuando el contratista, con autorización del Ingeniero Director de obras, emplee materiales de calidad más elevada o de mayores dimensiones de lo estipulado en el proyecto, o sustituya una clase de fabricación por otra que tenga asignado mayor precio o ejecute con mayores dimensiones cualquier otra parte de las obras, o en general, introduzca en ellas cualquier modificación que sea beneficiosa a juicio del Ingeniero Director de obras, no tendrá derecho sin embargo, sino a lo que le correspondería si hubiera realizado la obra con estricta sujeción a lo proyectado y contratado.

12. Las cantidades calculadas para obras accesorias, aunque figuren por partida alzada en el presupuesto final (general), no serán abonadas sino a

los precios de la contrata, según las condiciones de la misma y los proyectos particulares que para ellas se formen, o en su defecto, por lo que resulte de su medición final.

13. El contratista queda obligado a abonar al Ingeniero autor del proyecto y director de obras así como a los Ingenieros Técnicos, el importe de sus respectivos honorarios facultativos por formación del proyecto, dirección técnica y administración en su caso, con arreglo a las tarifas y honorarios vigentes.

14. Concluida la ejecución de la obra, será reconocida por el Ingeniero Director que a tal efecto designe la empresa.

15. La garantía definitiva será del 4% del presupuesto y la provisional del 2%.

16. La forma de pago será por certificaciones mensuales de la obra ejecutada, de acuerdo con los precios del presupuesto, deducida la baja si la hubiera.

17. La fecha de comienzo de las obras será a partir de los 15 días naturales del replanteo oficial de las mismas y la definitiva, al año de haber ejecutado la provisional, procediéndose si no existe reclamación alguna, a la reclamación de la fianza.

18. Si el contratista al efectuar el replanteo, observase algún error en el proyecto, deberá comunicarlo en el plazo de quince días al Ingeniero Director de obras, pues transcurrido ese plazo será responsable de la exactitud del proyecto.

19. El contratista está obligado a designar una persona responsable que se entenderá con el Ingeniero Director de obras, o con el delegado que éste designe, para todo relacionado con ella. Al ser el Ingeniero Director de obras el que interpreta el proyecto, el contratista deberá consultarle cualquier duda que surja en su realización.

20. Durante la realización de la obra, se girarán visitas de inspección por personal facultativo de la empresa cliente, para hacer las comprobaciones que se crean oportunas. Es obligación del contratista, la conservación de la obra ya ejecutada hasta la recepción de la misma, por lo que el deterioro parcial o total de ella, aunque sea por agentes atmosféricos u otras causas, deberá ser reparado o reconstruido por su cuenta.

21. El contratista, deberá realizar la obra en el plazo mencionado a partir de la fecha del contrato, incurriendo en multa, por retraso de la ejecución siempre que éste no sea debido a causas de fuerza mayor. A la terminación de la obra, se hará una recepción provisional previo reconocimiento y examen por la dirección técnica, el depositario de efectos, el interventor y el jefe de servicio o un representante, estampando su conformidad el contratista.

22. Hecha la recepción provisional, se certificará al contratista el resto de la obra, reservándose la administración el importe de los gastos de conservación de la misma hasta su recepción definitiva y la fianza durante el tiempo señalado como plazo de garantía. La recepción definitiva se hará en las mismas condiciones que la provisional, extendiéndose el acta correspondiente. El Director Técnico propondrá a la Junta Económica la devolución de la fianza al contratista de acuerdo con las condiciones económicas legales establecidas.

23. Las tarifas para la determinación de honorarios, reguladas por orden de la Presidencia del Gobierno el 19 de Octubre de 1961, se aplicarán sobre el denominado en la actualidad "Presupuesto de Ejecución de Contrata" y anteriormente llamado "Presupuesto de Ejecución Material" que hoy designa otro concepto.

## Condiciones particulares

La empresa consultora, que ha desarrollado el presente proyecto, lo entregará a la empresa cliente bajo las condiciones generales ya formuladas, debiendo añadirse las siguientes condiciones particulares:

1. La propiedad intelectual de los procesos descritos y analizados en el presente trabajo, pertenece por entero a la empresa consultora representada por el Ingeniero Director del Proyecto.

2. La empresa consultora se reserva el derecho a la utilización total o parcial de los resultados de la investigación realizada para desarrollar el siguiente proyecto, bien para su publicación o bien para su uso en trabajos o proyectos posteriores, para la misma empresa cliente o para otra.

3. Cualquier tipo de reproducción aparte de las reseñadas en las condiciones generales, bien sea para uso particular de la empresa cliente, o para cualquier otra aplicación, contará con autorización expresa y por escrito del Ingeniero Director del Proyecto, que actuará en representación de la empresa consultora.

4. En la autorización se ha de hacer constar la aplicación a que se destinan sus reproducciones así como su cantidad.

5. En todas las reproducciones se indicará su procedencia, explicitando el nombre del proyecto, nombre del Ingeniero Director y de la empresa consultora.

6. Si el proyecto pasa la etapa de desarrollo, cualquier modificación que se realice sobre él, deberá ser notificada al Ingeniero Director del Proyecto y a criterio de éste, la empresa consultora decidirá aceptar o no la modificación propuesta.

7. Si la modificación se acepta, la empresa consultora se hará responsable al mismo nivel que el proyecto inicial del que resulta el añadirla.

8. Si la modificación no es aceptada, por el contrario, la empresa consultora declinará toda responsabilidad que se derive de la aplicación o influencia de la misma.

9. Si la empresa cliente decide desarrollar industrialmente uno o varios productos en los que resulte parcial o totalmente aplicable el estudio de este proyecto, deberá comunicarlo a la empresa consultora.

10. La empresa consultora no se responsabiliza de los efectos laterales que se puedan producir en el momento en que se utilice la herramienta objeto del presente proyecto para la realización de otras aplicaciones.

11. La empresa consultora tendrá prioridad respecto a otras en la elaboración de los proyectos auxiliares que fuese necesario desarrollar para dicha aplicación industrial, siempre que no haga explícita renuncia a este hecho. En este caso, deberá autorizar expresamente los proyectos presentados por otros.

12. El Ingeniero Director del presente proyecto, será el responsable de la dirección de la aplicación industrial siempre que la empresa consultora lo estime oportuno. En caso contrario, la persona designada deberá contar con la autorización del mismo, quien delegará en él las responsabilidades que ostente.