



UPPSALA  
UNIVERSITET

Department of Linguistics and Philology  
Språkteknologiprogrammet  
(Language Technology Programme)  
Master's thesis in Computational Linguistics

# A Prototype of an Arabic Diphone Speech Synthesizer in Festival

Maria Moutran Assaf

Supervisor:  
Harald Berthelsen, STTS  
Beata Megyesi, Uppsala University

## **Abstract**

This master's thesis describes the construction of an Arabic voice using the diphone concatenation method within the Festival Speech Synthesis framework. We show that recognizable automatic Arabic speech can be constructed by using 200 read sentences as basis to the speech synthesis. In addition, we define a phone set for Arabic to be compatible with Festival's framework. Tools for quick diphone collection are also developed and the problems with the Arabic language when building a diphone database are illustrated. The test results indicate that the majority of the words and sentences are recognizable. In fact, 80 to 85 % of the words and 75 % of the sentences were correctly and completely recognized by the listeners in the test. When it comes to the recognition rate of the sentences that were written down by the listeners, 70 to 95 % were recognized. The sentence completely recognition rate increased from 30% after the first time of listening to 45%, partially recognition rate increased from 40% to 50% after the second time of listening.

# Contents

<b>Abstract</b>	<b>ii</b>
<b>Contents</b>	<b>iii</b>
<b>List of Figures</b>	<b>vi</b>
<b>List of Tables</b>	<b>vii</b>
<b>Acknowledgment</b>	<b>viii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Purpose . . . . .	1
1.2 Outline . . . . .	2
<b>2 Text-to-Speech Systems</b>	<b>3</b>
2.1 The structure of Text-to-Speech systems . . . . .	3
2.1.1 The Natural Language Processing module . . . . .	3
2.1.2 The Digital Signal Processing module . . . . .	4
2.2 TTS applications . . . . .	5
2.3 Festival . . . . .	6
2.3.1 The Festival Speech Synthesis System . . . . .	6
2.3.2 Festvox . . . . .	8
<b>3 Testing and evaluating TTS-systems</b>	<b>9</b>
3.1 Different evaluation methods of TTS-systems . . . . .	10
3.1.1 Testing intelligibility . . . . .	10
3.1.2 Testing naturalness . . . . .	12
3.1.3 Testing Front-end processing . . . . .	12
3.1.4 Overall quality evaluation . . . . .	12
<b>4 The Arabic Language</b>	<b>14</b>
4.1 Introduction to the Arabic Language . . . . .	14
4.2 The Alphabet . . . . .	14
4.3 Arabic morphology . . . . .	16
4.4 Arabic prosody . . . . .	16
<b>5 Arabic Speech Synthesis</b>	<b>18</b>
5.1 Existing Arabic voices . . . . .	18
5.1.1 MBROLA-project . . . . .	18
5.1.2 Acapela-group . . . . .	18

5.2	Special challenges of the Arabic language for TTS systems . . . . .	19
5.2.1	The diacritization problem . . . . .	19
5.2.2	Dialects . . . . .	19
5.2.3	Differences in gender . . . . .	20
5.3	Transliteration vs transcription . . . . .	20
<b>6</b>	<b>Diphone databases in Festival</b>	<b>22</b>
6.1	Specifying a phone set in Festival . . . . .	22
6.1.1	The production of vowels . . . . .	23
6.1.2	The production of semi-vowels . . . . .	24
6.1.3	The production of Arabic consonants . . . . .	24
6.1.4	The production of additional consonants . . . . .	24
6.2	Letter-to-Sound rules . . . . .	24
6.3	Diphone Database Construction . . . . .	26
6.3.1	The script for extracting all possible diphones . . . . .	26
6.3.2	Recording the diphones . . . . .	27
6.3.3	Segmentation and labelling of the diphones . . . . .	27
6.3.4	The database construction . . . . .	27
<b>7</b>	<b>Testing the Arabic voice</b>	<b>29</b>
7.1	Test group . . . . .	29
7.2	Method . . . . .	29
7.3	Test and evaluation results . . . . .	30
7.3.1	Perception of the words . . . . .	30
7.3.2	Perception of the sentences - first part . . . . .	31
7.3.3	Perception of the sentences - second part . . . . .	32
7.3.4	Naturalness . . . . .	33
7.3.5	Speed . . . . .	34
7.3.6	Sound quality . . . . .	34
7.3.7	Pronunciation . . . . .	35
7.3.8	Intelligibility . . . . .	38
7.3.9	Stress/Intonation . . . . .	39
<b>8</b>	<b>Discussion</b>	<b>42</b>
<b>9</b>	<b>Summary and future development</b>	<b>44</b>
9.1	Future work . . . . .	45
9.2	Conclusion . . . . .	45
	<b>Bibliography</b>	<b>46</b>
<b>A</b>	<b>Phone set</b>	<b>48</b>
<b>B</b>	<b>Letter to Sound rules</b>	<b>50</b>
<b>C</b>	<b>Questionnaire</b>	<b>52</b>
C.1	Background information (Bakgrundsfrågor) . . . . .	52
C.2	Testing the Arabic voice (Testa den arabiska rösten) . . . . .	53
C.2.1	Words (Ord) . . . . .	53
C.2.2	Sentences (Meningar) . . . . .	56

C.2.3 Sentences 2 (Meningar 2) . . . . .	57
C.3 Evaluating the Arabic voice (Evaluering av den arabiska rösten) . .	59

# List of Figures

2.1	The structure of a TTS system . . . . .	3
7.1	Perception of the words . . . . .	31
7.2	Perception of the first part of the sentences . . . . .	31
7.3	Perception of the second part of the sentences - the first listening . . . . .	32
7.4	Perception of the second part of the sentences - the second listening . . . . .	33
7.5	Naturalness of the voice . . . . .	33
7.6	The speed of the speech . . . . .	34
7.7	The sound quality of the voice . . . . .	35
7.8	Pronunciation mistakes . . . . .	36
7.9	The pronunciation's effect on understanding . . . . .	36
7.10	The concentration needed to hear the pronunciation . . . . .	37
7.11	The annoying level of the pronunciation . . . . .	37
7.12	Understanding the voice . . . . .	38
7.13	The level of difficulty in understanding the voice . . . . .	39
7.14	The intonation of the system . . . . .	40
7.15	The stress of the system . . . . .	40
7.16	The intonation difference of the system . . . . .	41

# List of Tables

3.1	Possible evaluating attributes . . . . .	13
4.1	The Arabic alphabet . . . . .	15
6.1	Phoneme mappings from SAMPA . . . . .	23
6.2	Vowels in Arabic . . . . .	23
6.3	Semi-vowels in Arabic . . . . .	24
6.4	Consonants in Arabic . . . . .	25
6.5	Additional consonants in Arabic, in loan words . . . . .	25
6.6	Some examples of Letter-to-Sound rules . . . . .	25
6.7	Example of the diphone index . . . . .	28
7.1	Age distribution of the listeners . . . . .	29
7.2	The most problematic sounds to understand . . . . .	38

# Acknowledgment

I would like to express my sincerest gratitude towards a few persons that have made this project possible to achieve. First I would like to thank my supervisor, Harald Berthelsen, at STTS in Stockholm, for his support and for his patience when guiding me through the Festival Speech Synthesizer system. I would like to thank Alan W. Black for his quick replies when trivial questions were asked about the Festival system.

I would also like to thank professor Bo Isaksson and Ablahad Lahdo for their help with the Arabic language and especially Ablahad for his help with gathering the student group that have participated in testing and evaluating the system. I would like to seize the opportunity to thank the students at the Orientalistic programme for participating in the questionnaire. I would also like to thank Per Starbäck with his help with Latex guidance and with writing Arabic in Latex.

I would like to thank, Sarah Roxström, director for foreign languages at "Sveriges radio", for giving me the rights to use their news texts on their website. A big thanks to the editorial and production manager, Alison Green, at Lund Humphries publishing house in London, for giving me the rights to use the book "A New Arabic Grammar of the Written Language".

I am very grateful to my supervisor at the department of Linguistics and philology, Beáta Megyesi, for her help, advices and encouragement. Without this encouragement I would still be working on this thesis.

Finally I would like to thank my husband and my son for standing out with me during the period the project was done.

“Tell me and I’ll forget. Show me and I’ll remember. But involve me and I’ll understand.”

A Chinese proverb

# 1 Introduction

One of the major areas in Language Technology is Speech Technology; this area contains techniques for speech synthesis, speech recognition and dialog systems. Speech recognition technologies allow computers to interpret human speech for example to interact with a computer. Text-to-Speech synthesis is the process of converting a written text into speech. The major purposes of speech synthesis techniques are to convert a chain of phonetic symbols into artificial speech, to transform a given linguistic representation and to generate speech automatically with information about intonation and stress i.e. prosody. Speech synthesis, i.e. artificial language, is not a new technology. The first attempts to build mechanical speech synthesizer started already in 1700. The first real speech synthesizer, called VODER<sup>1</sup>, was presented at the World's Fair in New York in 1939 (Black and Lenzo, 2003b). This synthesis could produce connected speech and whole sentences.

According to Black and Lenzo (2003b), research in speech synthesis before 1980, was limited to the large laboratories that could afford investment in both time and money for hardware. Today, with faster machines and larger disk space, people have began to improve synthesis by using larger and more varied techniques for developing synthesized voices. These improvements do not only consider speech synthesis but also consider developments in speech technology, such as speech recognition.

The interest in building speech synthesis and in building voices is increasing, therefore the demand on the technology to deliver good and acceptable quality of speech applications. This leads to the availability of free and semi-free synthesis systems, such as the Festival Speech Synthesis System<sup>2</sup>, that has been developed at the CSTR (Center for Speech Technology Research) in Edinburgh, Scotland and the MBROLA-project<sup>3</sup> that has been started by TCTL Lab, Faculté Polytechnique, Mons, Belgium. These two systems reduce the cost of entering the field of speech synthesis. However, although the quality of the Text-to-Speech synthesizers has significantly improved with result of highly intelligible speech, the synthesized speech does not sound as natural as human voice does, and the construction of synthetic voices is still difficult.

## 1.1 Purpose

The area of Arabic Text-to-Speech system is still in its early development stage. The aim of this thesis is to create an Arabic speech synthesis based on a diphone database. Diphone synthesis is a kind of concatenative speech synthesis, which uses

---

<sup>1</sup><http://www.haskins.yale.edu/haskins/HEADS/SIMULACRA/voder.html>

<sup>2</sup><http://www.cstr.ed.ac.uk/>

<sup>3</sup> <http://tcts.fpms.ac.be/synthesis/mbrola.html>

diphones as basic speech units to concatenate. The focus will be put in the development of guidelines for Arabic speech synthesis, that make future construction of Arabic voices easier. These tools consider tools for examining the Arabic language and will also consider a development of a diphone collection of the language. This project has been carried out at STTS in Stockholm which specialises in research and development in speech and language technology. The Festival speech synthesis toolkit has been used for speech generation. This thesis will contain attempts to create techniques for easier processing of the Arabic language.

## 1.2 Outline

Chapter two, Text-to-Speech systems, gives a short introduction to speech synthesis and its techniques. Two components of Text-to-Speech synthesis, the Natural Language Processing (NLP) and the Digital Signal Processing, are discussed.

Chapter three introduces some of the most common used test and evaluation methods used to evaluate text-to-speech systems.

For a better understanding of the challenges faced when building the Arabic voice in Festival, chapter four introduces the Arabic language, its alphabet, structure and prosody.

Chapter five introduces two existing speech synthesizer systems that have an Arabic voice. This chapter addresses also the challenges with the Arabic language when building a synthetic voice.

In chapter six the focus is put on the work that was done during this project, i.e. building an Arabic voice in Festival. This chapter introduces the specified phone-set in Festival for Arabic, as well as letter-to-sound rules, the construction of the diphone database and the recordings of the database.

In chapter seven, a test of the diphone synthesis is performed, where informants with knowledge of the Arabic language test the system, in order to investigate the intelligibility, correctness of the synthesis and its naturalness. The results are summarized and a discussion is raised.

Chapter eight brings up a general discussion on TTS-systems and a summary of the work done in this project. Finally in this chapter future research is discussed.

## 2 Text-to-Speech Systems

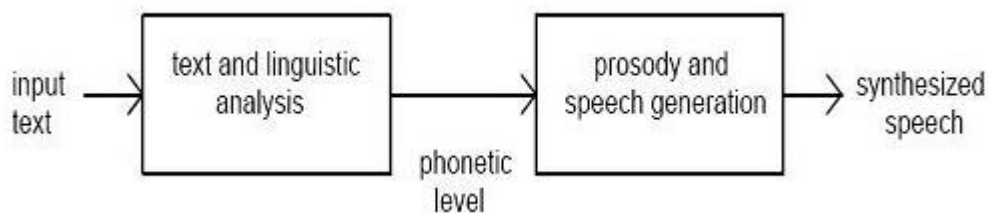
Text-to-Speech synthesis is the process of converting a written text into artificial speech. It is a computer-based program in which the system processes the text and repeats it aloud. The system contains two modules, the Natural Language Processing Module and the Digital Signal Processing Module. The following section aims to give a brief introduction and overview of the structure of Text-to-Speech synthesizers in general.

### 2.1 The structure of Text-to-Speech systems

In TTS systems, the process of converting written text into speech contains a number of steps. In general cases, a TTS system contains the two modules:

- The Natural Language Processing Module (NLP) is able to produce files with a phonetic transcription of the text, together with the desired intonation and rhythm.
- The Digital Signal Processing Module (DSP) transforms the symbolic information it receives from the NLP module into speech.

These modules are discussed in the following section, where an overview of each model is given. The general structure of a Text-to-Speech system is shown in Figure 2.1.



**Figure 2.1:** The structure of a TTS system

#### 2.1.1 The Natural Language Processing module

The NLP module consists of three processing stages: *Text Analysis*, *Automatic Phonetization* and *Prosody Generation*. The first stage, *the text analysis*, consists of four categories:

- A pre-processing module, where the input sentences are organized into lists of words. The system must first identify these words or tokens in order to find their pronunciations. In other words, the first step in *the text analysis* is to make chunks out of the input text - *tokenizing* it. There are many tokens in a text, that appear in a way where their pronunciation has no obvious relationship with their appearance. Such as abbreviations, acronyms and numbers. Apart from tokenization, *Normalization* is needed where a transformation of these tokens into full text is done.
- A morphological analysis module, where all possible part-of-speech categories for each word are proposed on the basis of their spelling. For example inflected, derived and compound words are decomposed into their morphs by simple grammar rules.
- A contextual analysis module that considers word in their contexts. This is important to be able to reduce possible part-of-speech categories of the word by simple regular grammars by using lexicons of stems and affixes.
- A syntactic-prosodic parser, where the remaining search space is examined and the text structure is found. The parser organizes the text into clause and phrase-like constituents. After that the parser tries to relate these into their expected prosodic realization (Dutoit, 1996).

Step two and step three can be done in one step, i.e. the Part-of-Speech tagging. According to Dutoit (1996) performing this in two steps will give a more reduced list of the possible part-of-speech categories for a word.

The second module is *the Letter-to-Sound module (LTS)*, where the words are phonetically transcribed. In this stage, the module also maps sequences of grapheme into sequences of phonemes with possible diacritic information, such as stress and other prosodic features, that are important to fluency in speech and natural sounding speech.

The last module, *the Prosody Generator*, is where certain properties of the speech signal such as pitch, loudness and syllable length are processed. Prosodic features create segmentation of the speech chain into groups of syllables. This gives rise to the grouping of syllables and words in larger chunks (Dutoit, 1996).

## 2.1.2 The Digital Signal Processing module

In this module a transformation of the received symbolic information, from the NLP module, into speech is done. There are three different categories of waveform generation or so called types of synthesis, *Articulatory Synthesis*, *Formant Synthesis* and *Concatenative Synthesis*. These approaches are used to process the textual input into a sound output.

*Articulatory Synthesis* is a method of synthesizing speech by controlling the speech articulators. This method determines the characteristics of the vocal tract filter by means of a description of the vocal tract geometry and places the potential sound sources within this geometry (Fant, 1960).

*Formant Synthesis* specifies directly the formant frequencies and bandwidths as well as the source parameters (Fant, 1960). Formant Synthesizers are also referred to as Rule-Based Synthesizers, where generalized rules are extracted from the filtered

information. The input is then tested on the rules. The vocal tract transfer function can be modeled by simulating formant frequencies and formant amplitudes. The model transfer function for the vocal tract makes the formants much more evident (Johnson). The speech output is determined by phonetic rules to determine the parameters that are necessary to synthesize a desired utterance using a formant synthesizer. Examples of the input parameters are "Voicing fundamental frequency" (F0) and "Voiced excitation open quotient" (OQ).

*Concatenative Synthesis* is the most used technique of today, where segments of speech are tied together to form a complete speech chain. The speech output is produced by coupling segments from the database to create the sequence of segments. This technique requires a bit of manual preparation of the speech segments. There are two categories within this method, diphone and unit-selection synthesis:

- A diphone synthesis uses a minimal speech database containing all the diphones occurring in a given language. Diphones are speech units that begin in the middle of the stable state of a phone and end in the middle of the following one. The number of diphones depends on the possible combinations of phonemes in a language. In diphone synthesis, only one example of each diphone is contained in the speech database. The quality of the resulting speech is generally not as good as that from unit selection but more natural-sounding than the output of formant synthesizers. Diphone synthesis suffers from the robotic-sounding quality of formant synthesis (Black and Lenzo, 2003b). In order to build a diphone database, the following questions have to be answered and determined: What diphone pairs exist in a language and what carrier words should be used? The answer for these questions are very language independent.
- The unit-selection synthesis uses large speech databases. More than one hour of recorded speech is usually used. During database creation, each recorded utterance is segmented into some or all of the following: individual phones, syllables, morphemes, words, phrases, and sentences. The division into segments can be performed using a number of techniques, for instance clustering, using a specially modified speech recognizer, it can also be done by hand, using visual representations such as the waveform and spectrogram. The unit-selection technique gives naturalness due to the fact that it does not apply digital signal processing techniques to the recorded speech, which often make recorded speech sound less natural. However, maximum naturalness often requires unit selection speech databases to be very large (Black and Lenzo, 2003b).

The main difference between the two types of *Concatenative Synthesis* lies in the size of the units being concatenated. Both methods store the pre-recorded speech units in a database, from which the concatenation originates. Those parts of utterances that have not been found, processed and stored in the database are built up from smaller units (Black et al., 2002).

## 2.2 TTS applications

As it has been described in this thesis and other articles, a text-to-speech synthesizer is a computer based system that can convert text into speech. Over the last few

decades, extensive work has been done on text-to-speech synthesis for the English language. Other languages such as Arabic have had limited testing mentioned in section 5.1. Concatenative speech synthesis can be achieved in two different ways, either by concatenating a fixed number and size of units, such as phones or diphones, or by concatenating a variable size units, which is called unit selection. Comparing to unit selection, diphone synthesis are more challenging in terms of signal processing, since only one example of each unit exists in the database. On the other hand diphone synthesis are to prefer when building applications like mobile devices, where the memory size is the main concern.

Today we have Text-to-Speech systems with a very high intelligible level and an adequate level for numerous applications. These high quality TTS systems have numerous applications like the examples below:

- Aid to handicapped people

By the help of an especially designed keyboard and a fast sentence assembling program, synthetic speech can be produced in a few seconds to remedy the voice handicaps. Also blind people can benefit from TTS systems which gave them access to written information.

- Language Education

They provide a helpful tool to learn a new language known as computer aided learning system.

- Talking books and toys

A language learning development tool, i.e a tool for children to learn a new language or their own language.

- Telecommunication services

In these systems textual information can be accessed over the telephone. Mostly they are used when the requirement of interactivity is little and texts range from simple messages. Queries can be given through the user's voice (needs speech recognition) or through the telephone keyboard.

- Multimedia

Man-machine communication that can help people with their work and other things, for example voice activation systems in the car.

## 2.3 Festival

### 2.3.1 The Festival Speech Synthesis System

The Festival Speech Synthesis System was developed at the center for Speech Technology Research at the University of Edinburgh in the late 1990s. The system has been developed by Alan Black, Paul Taylor and Richard Caley. The Festival Speech Synthesis System is designed for three particular users of speech synthesis: speech synthesis researchers, speech application developers, and the end user.

Festival is designed to allow the addition of new modules in an easy and efficient way. Another aspect which makes the festival system more compatible, is that Festival is not simply used for researching into new synthesis techniques, but also as a platform for developing and testing Text-to-Speech systems as well as a fully usable Text-to-Speech system. This is good for embedding into other projects that require speech output (Black et al., 2002).

Not only is Festival considered as a multilingual system suitable for research, development and general use, Festival is also used to transcribe unrestricted text to speech. The system is freely available for research and educational use and its range of languages has been recently expanded by the Center for Speech Technology Research: there are now available TTS systems for English, Spanish and Welsh.

The Festival System is implemented in c++, but in order to give parameters and specify flow of control, Festival also offers a scripting language based on the Scheme programming language. Scheme has a very simple syntax but is at the same time powerful for specifying parameters and simple functions. Scheme is chosen because it is restricted and is considered as a small language and would not increase the size of the Festival system (Black et al., 2002).

Festival is a general text-to-speech system that uses the residual-LPC (Linear Predictive Coding) synthesis technique, and is able to transcribe unrestricted text-to-speech. With LPC method, the residuals and LPC coefficients are used as control parameters. LPC methods are the most widely used in speech coding, speech synthesis, speech recognition, speaker recognition and verification and for speech storage. LPC methods provide extremely accurate estimates of speech parameters, and does it extremely efficiently.

Black and Lenzo (2003b) offer a list with basic processes to be able to build a new voice in Festival. These basic processes can be done technically. The most significant processes are:

- To define a phoneme set for a language that is well defined and that contains a detailed description of their articulatory features.
- To create a lexicon and/or Letter-to-Sound rules which is not as simple as it sounds. Lexicon construction can be considered to be hard and it is important to be consistent with the entries. Therefore, Black and Lenzo (2000) have developed techniques that aid in the construction of new lexicons. For those languages where the relationship between orthography and phonetics is close, such as the Arabic language, it is possible with a hand written set of Letter-to-Sound rules. For other languages, where pronunciation is different from the orthography, automatic learning techniques exist where Letter-to-Sound rules are built from existing words with existing pronunciations.
- To provide text analysis
- To build prosodic models and to build a waveform synthesizer

Though we have today very good TTS-synthesizers the quality of the text-to-speech synthesizers is still in need of improvements. The problem area in the text analysis and prosodic models steps in speech synthesis are very wide. There are several problems in text pre-processing, such as numerals, abbreviations, and acronyms. Another major problem today is to have a correct prosody and pronunciation analysis from written texts. Written text contains no explicit emotions. Pronunciation of

proper and foreign names are also a major problem of today. At the low-level synthesis, the discontinuities and contextual effects in wave concatenation methods are the most problematic. Speech synthesis has difficulties with female and child voices. Female voices (200 Hz) have a pitch almost twice as high as male voices (100 Hz) and with children it may be even four times as high (400 Hz). The higher fundamental frequency makes it more difficult to estimate the formant frequency locations (Klatt, 1987).

### 2.3.2 Festvox

The Festvox package is developed in 2003 by Alan Black and Kevin Lenzo at the Language Technologies Institute at Carnegie Mellon University, Pittsburgh (Black and Lenzo, 2003a). This package is intended to be used together with the Festival Speech Synthesis System. The purpose with this project was to build synthesized voices in a more systematic way and to make the documentation process better. The goal with this package is to make it easier for anyone with little knowledge of the subject to be able to build an own synthetic voice. Basic skeleton files are included in the Festvox distribution, that will facilitate the building of a new voice. The package can be downloaded and used freely.

### 3 Testing and evaluating TTS-systems

How to carry out an effective evaluation of a TTS-system is not always a simple task. Some of the main contributing factors that Klatt (1987) believes affect the overall quality of a TTS system are:

- *Intelligibility*, that is how much of the spoken output the user understands, as well as how quickly a listener gets fatigue by only listening?
- *Naturalness*, that is how much like real speech does the output of the TTS system sound?
- *Suitability for used application*, different applications have differing needs for a TTS system. For example a system for the blind requires higher rates considering intelligibility, more than the naturalness.

Depending on what kind of information is needed, the evaluation can be made on several levels; phoneme, word, or sentence level. There are many tests that help to address these three issues and others. Several individual test methods for synthetic speech have been developed during the last decades. There are even some researchers that complain that there are too many existing methods which make the comparisons and standardization procedures difficult. On the other hand, it is clear that there is still not a single test method that gives a final correct result. In this chapter I will give a short introduction of the most commonly used methods. This will be the foundation of the test and the evaluation questionnaire introduced in chapter 7.

There are many different methods to evaluate TTS systems. Not all methods are presented and discussed in this chapter only the main tests. There are for example tests used to test intelligibility of proper names, tests for prosody evaluation, comprehension tests and much more. Those tests that are found of importance to the questionnaire done for the evaluation of the Arabic diphone voice are chosen. One can perhaps say that the most valuable test does not exist yet. Therefore, the most suitable way to test a speech synthesizer is to mix several test and evaluating methods (SpeechWorks International). For example using tests for phoneme, sentence level, prosody tests, front-end tests and overall tests would provide many useful and important information. On the other hand, it is important that the tests do not give the same information, due to the fact testing is extremely time consuming. A factor that can effect the results is that the results may increase significantly when the listeners get familiar with the synthetic voice and when repeating the test procedure to the same listening group. The listeners hear and understand the voice better after every listening. One problem with this is that the listeners may lose some of their concen-

tration. Therefore, the decision of using naive or pro listeners in the listening tests was important.

One wish there would be a simple method to test and evaluate TTS-systems but the evaluation and assessment of synthesized speech is not a simple task. As mentioned before synthesized speech can be evaluated by many methods and at several levels. Each method will provide some kind of information. To have a good test it is required to select several methods to assess each feature separately. The evaluation method must be chosen carefully to achieve desired results. There is no sense in having the same results from two tests. It is important to get feedback from users to improve and optimise the system.

I would also like to mention that there are some computer softwares that have been developed for making the test procedure easier to perform. I will not discuss this in my thesis. For further information I refer you to read Howard-Jones and Partnership (1991).

## 3.1 Different evaluation methods of TTS-systems

### 3.1.1 Testing intelligibility

Segmental methods are often used to test a single segment or phoneme intelligibility. The most common method that is used for testing the intelligibility of synthetic speech is the rhyme tests. One of the most advantages with rhyme tests is that the test is easy and the test procedure does not take too much time. The most famous segmental tests are the Diagnostic and Modified Rhyme Tests described below (The DISC Best Practice Guide).

#### On word level

- *Diagnostic Rhyme Test (DRT)*

DRT is a test of the intelligibility of word-initial consonants. Subjects are played pairs of words with a different first consonant, and asked to identify which word they heard (e.g., thing vs. sing). Six contrasts are represented, namely: voicing, nasality, sustention, sibilation, graveness, and compactness. The system that performs best is the one that produces the lowest error rate. Usually, only total error rate percentage is given, but also single consonants and how they are confused with each other can be investigated. This is a very effective method and can be used since it is an easy and reliable method. However, this method does not test any vowels or prosodic features (SpeechWorks International).

- *Modified Rhyme Test (MRT)*

MRT is like DRT, but includes tests both for word-initial and word-final intelligibility (e.g., bath vs. bass). The test consists of 50 sets of 6 one-syllable words which makes a total set of 300 words. The 6 words set is played one at the time and the listener marks which word he thinks he hears. The first half of the words are used for the evaluation of the initial consonants and the second one for the final ones. Results are summarized as in DRT, but both final and initial error rates are given individually (SpeechWorks International).

- *SAM Standard Segmental Test*  
The SAM Standard Segmental Test uses open vocabulary, mostly meaningless but also (by chance) meaningful, of the structure CV, VC, and VCV. All phonotactically possible combinations of initial, medial, and final consonants and the three vowels, /i/, /u/, and /a/ are the basic items. Examples: pa, ap, apa. The vowels are not tested at all since only the missing consonant must be filled to the response sheet (The DISC Best Practice Guide).
- *Diagnostic Medial Consonant Test (DMCT)*  
Diagnostic Medial Consonant Test is of the same kind of test like rhyme tests described before. Word pairs like "stopper - stocker" are examined. These words were selected to differ only with their intervocalic consonant. As in DRT, the listeners task is to choose the correct word from two possible alternatives in the answer sheet (Pisoni and Hunnicutt, 1980).
- *Cluster CLuster IDentification Test (CLID)*  
Consonant clusters, sequences of one or more consonants and vowel clusters, sequences of one or more vowel, are the basic items to test the intelligibility of a system with this test. The test is based on statistical approach. The test vocabulary is not predefined and it is generated for each test sequence separately. The test consists of three main phases. The first phase is the word generator, that generates the test material in phonetic representation. The second phase is the phoneme-to-grapheme converter, that convert phoneme strings to graphemic representations. The reason for this is that most of the synthesizers do not accept phoneme strings. The last phase the automatic scoring module does fetch the error rates automatically. Initial, medial, and final clusters are scored individually (The DISC Best Practice Guide).

### **On sentence level**

- **Harvard Psychoacoustic Sentences**  
The Harvard Psychoacoustic Sentences test contains a fixed set of 100 meaningful sentences. The test was developed to test the word intelligibility in sentence context. The test is easy to perform, no training of the subjects is needed and the scoring is simple (The DISC Best Practice Guide).
- **Haskins Syntactic Sentences**  
As in Harvard sentences, a fixed set of 100 semantically unpredictable sentences is used. Unlike in Harvard sentences, the disadvantage with the Haskins Syntactic Sentences is that the missed items cannot be concluded from context as easily as with use of meaningful sentences (Pisoni and Hunnicutt, 1980).
- **SAM Semantically Unpredictable Sentences (SUS)**  
This test is also an intelligibility test on sentence level. The test contains five grammatical structures listed below (The DISC Best Practice Guide).
  - Subject - Verb - Adverbial, e.g., The table walked through the blue truth
  - Subject - Verb - Direct object, e.g., The strong way drank the day
  - Adverbial - Transitive verb - Direct object (imperative), e.g., Never draw the house and the fact

- Q-word - Transitive verb - Subject - Direct object, e.g., How does the day love the bright word?
- Subject - Verb - Complex direct object, e.g., The place closed the fish that lived.

Fifty sentences, with ten of the above mentioned structure, are generated and played in random order to test subjects. Subjects are asked to transcribe sentences that have no context, and therefore it is not possible to derive phonetic information from any source but the signal.

### 3.1.2 Testing naturalness

- *Mean Opinion Score (MOS)*

Mean Opinion Score method is the simplest method to evaluate speech quality in general. It is also suitable for overall evaluation of synthetic speech. MOS is a five level scale from bad (1) to excellent (5) and it is also known as Absolute Category Rating (ACR). The test procedure is that listeners are asked to rate the speech quality of different systems, usually synthesizing the same set of sentences. It is important, when using MOS, to use a large set of test persons, at least 10 persons and to use at least 50 sentences per TTS system (SpeechWorks International).

- *Forced-Choice Ranking*

In this kind of tests, persons are asked to play the system with other TTS systems and to compare it to another systems. The same set of sentences is used and the persons are asked to rank the renditions of each sentence.

### 3.1.3 Testing Front-end processing

- *Functionality Test*

Sentences that contain multiple examples should be gathered from multiple contexts that the system is supposed to encounter. The sentences should be synthesized with the TTS system in consideration and simply mark how correct or incorrect the output is (SpeechWorks International).

### 3.1.4 Overall quality evaluation

This subsection will present methods for an overall quality evaluation of a TTS system.

- *Mean Opinion Score (MOS)*

This method is presented in subsection 3.1.2.

- *Categorical Estimation (CE)*

In this method the speech is evaluated by several aspects independently (Kraft and Portele, 1995). The possible attributes that can be used for evaluating can be seen in Table 3.1.

- *Pair comparison (PC)*

These methods are to test system's overall acceptance. The idea with this test

**Table 3.1:** Possible evaluating attributes

Attributes	Ratings levels + ... -
Naturalness	very natural ... very unnatural
Speed	much too fast ... much too slow
Sound quality	very good ... very bad
Pronunciation	not annoying ... very annoying
Intelligibility	very easy ... very hard
Stress/Intonation	not annoying ... very annoying

is that the listener will listen to artificial speech for perhaps hours per day so the small and negligible errors may become very annoying because of their frequent occurrences. Some of this effect may be apparent if few sentences are frequently repeated in the test procedure (Kraft and Portele, 1995).

## 4 The Arabic Language

In the previous sections, a brief introduction to the speech synthesis mechanism is presented. There was also a short introduction of the toolkit which was used to build an Arabic voice in Festival. Several test and evaluation methods were also introduced in the previous section. In this section, the Arabic language is introduced, where the alphabet is represented as well as the Arabic morphology and prosody, that is relevant for this thesis.

### 4.1 Introduction to the Arabic Language

The Arabic language belongs to the Semitic group of languages which also includes Hebrew and Amharic, which is the main language of Ethiopia. The Arabic alphabet can be traced back to the alphabet used to write the Nabataean dialect of Aramaic. The Aramaic language is descended from Phoenician (Lebanon today) which gave rise to the Greek alphabet, which is the alphabet used by ancient Greeks. Arabic is ranked as number four among the world's major languages by First Language, with 225 million native speakers of all dialects (Comrie, 1987).

The Arabic language is spoken throughout the Arab world and is the liturgical language of Islam. This means that Arabic is known widely by all Muslims in the world. Arabic either refers to Standard Arabic or to the many dialectal variations of Arabic. Standard Arabic is the language used by media and the language of Qur'an. Modern Standard Arabic is generally adopted as the common medium of communication through the Arab world today. Dialectal Arabic refers to the dialects derived from Classical Arabic. These dialects differ sometimes which means that it is hard and a challenge for a Lebanese to understand an Algerian and it is worth mentioning there is even a difference within the same country.

As mentioned before there are many varieties of the Arabic language, many dialects that reflect social diversity of its speakers. Arabic can be sub-classified as Classical Arabic, Eastern Arabic, Western Arabic and Maltese. Western Arabic encompasses the Arabic spoken colloquially in the region of northern Africa, often referred to as the Maghreb. Eastern Arabic includes the Arabic dialects spoken in North Africa (Egypt, Sudan), the Middle East (Lebanon, Syria). Arabic speakers use Modern Standard Arabic (MSA) to communicate across dialect groups. It is used in a situation where the native dialect will not be understood.

### 4.2 The Alphabet

The Arabic language contains 29 letters. There are 6 vowels in Arabic, 3 short and 3 long and there are 2 semi-vowels, which are diphthongs. Arabic short vowels are

written with diacritics placed above or below the consonant that precedes them. Table 4.1 shows the alphabet. The Arabic alphabet is written from right to left and there is no difference between upper and lower case. Most of the letters are attached to one another and they vary in writing whether they connect to preceding or following letters.

**Table 4.1:** The Arabic alphabet

Arabic		Transcription	
أ	أَلِف	'a	'alif
ب	بَاء	b	ba:'
ت	تَاء	t	ta:'
ث	ثَاء	T	Ta:'
ج	جِيم	Z	Zi:m
ح	حَاء	X	Xa:'
خ	خَاء	x	xa:'
د	دَال	d	da:l
ذ	ذَال	D	Da:l
ر	رَاء	r	ra:'
ز	زَاي	z	za:y
س	سِين	s	si:n
ش	شِين	S	Si:n
ص	صَاد	s.	s.a:d
ض	ضَاد	d.	d.a:d
ط	طَاء	t.	t.a:'
ظ	ظَاد	z.	z.a:d
ع	عَيْن	H	Hain
غ	غَيْن	G	G
ف	فَاء	f	fa:'
ق	قَاف	q	qa:f
ك	كَاف	k	ka:f
ل	لَام	l	la:m
ن	نُون	n	nu:n
ه	هَاء	h	ha:'
و	وَاو	w	wa:w
ي	يَاء	y	ya:'

Arabic short vowels are not written in Arabic. Therefore the reader must have some knowledge of the language. In sacred texts, vocalized texts and children's book the short vowels are generally written. Otherwise, short vowels are marked where

ambiguity appears and cannot be resolved simply from the context, the two examples below show a vocalized sentence and a non vocalized sentence.

- (1) أَنَا جَمِيلَةٌ جَدًّا  
 "I am very beautiful"
- (2) أَنَا جميلة جدًا  
 "I am very beautiful"

### 4.3 Arabic morphology

Most Arabic words can be reduced to a root, which as a rule, consists of three consonants. They are called Radicals. The majority of these Arabic roots are trilaterals. Trilaterals consist of three radical letters or consonants. These trilateral roots express a certain conceptual content. For example, the Form VI تَفَاعَلَ "tafa3ala" often have the meaning of a reciprocal action, such as تَكَاتَبَ "takAtaba" (to write to each other). Similarly, verbs of Form II are often causative or intensive in meaning.

A large number of word patterns can be formed from each root by modifying the root, by adding prefixes and/or suffixes, and by changing the vowels (short or long vowels). To simplify this, one can say that the root consonants fulfill a semantic function and the vowels fulfill a grammatical function in the Arabic word. Consider the root سَلِمَ "salima" (to be safe) we can derive سَلَّمَ "sallama" (to deliver or to say hi); اسْتَلَمَ "istalama" (to receive); اسْتَسَلَّمَ "istaslama" (to surrender), سَلَامٌ "salAmun" (peace). To be able to obtain a good command of the language one must have knowledge of the patterns that occur frequently in Arabic. These can be estimated to be ten roots out of fifteen roots (Haywood and Nahmad, 2003).

Standard Arabic and the modern dialects use different strategies to form the passive of a verb. The passive form only differs formally from the active voice in sequence of vowels. Standard Arabic verbs form their passives by changing the vowel pattern inside the verb stem, as in دَفِنَ "dafana" (he buried) > دُفِنَ "dufina" (he was buried).

For more information about the morphology of the Arabic language, I advice to read (Haywood and Nahmad, 2003).

### 4.4 Arabic prosody

Haywood and Nahmad (2003) describe in their book that the written Arabic is a language of syllable length, rather than accent or stress furthermore, all syllables should be given their fully length without slurring any letter. The authors mean that one should not emphasis any syllable at the expense of another. In the Arabic language there are two kinds of syllable, short and long ones. All syllables have a single onset C followed by a long or a short vowel. Short vowels are denoted by "V" and long vowels are denoted by "V :". The short syllable, CV, consists of a consonant with a short vowel. For example the word كَتَبَ "ka-ta-ba" (he wrote) contains three syllables. These syllables should be pronounced in an even and equal way. The long syllables contain a vowelled consonant followed by an unvowelled letter. This can

either be a consonant with vowel followed by a long vowel CVV : as in the word كَاتَبَ “kA-ta-ba” (he corresponded with) or it can be followed by a vowelled consonant followed by a consonant CVC for example the word كَلْبُهُ “kal-bu-hu” (his dog). Super heavy syllables such as CVV:C or CVCC are not possible in Arabic (Haywood and Nahmad, 2003).

Loudness in pronunciation of single words in Arabic differs in levels. Youssef and Emam (2004) mention three levels of loudness that can be found in the same word. Stress is the only cause of these differences. According to the authors, stress is the intensity or energy of the time domain signal of the syllable. The three levels are: primary stress, secondary stress and tertiary stress. The types of the syllables, their distributions and their numbers are factors that effects the stress in a word. These rules are well known by all Arabic speakers and are as easy as they sound (Youssef and Emam, 2004). These three kinds of syllables can be divided as follows:

- For instance, the first syllable gets the main stress and the rest of the syllables get the weak stress when a word consists of sequence of short CV syllables.
- However if a word contains one long syllable this syllable will than take the main stress. The remaining will get the weak stress.
- In those cases where a word contains two or more long syllables, than the nearest long syllable to the end will get the main stress, the one in the middle will take the secondary stress and last the first long syllable will be the weakest.

## 5 Arabic Speech Synthesis

Some of the World languages such as English, Spanish, French, German have been extensively studied and processed, unlike Arabic. In this chapter two existing speech synthesizer systems that have an Arabic voice are discussed as well as the fundamentals of building an Arabic voice from scratch. To be able to map these parts, the challenges with the Arabic language for building a synthetic voice must be addressed. So from now on, all focus will be on the different steps in building the Arabic voice that were undertaken during this project.

### 5.1 Existing Arabic voices

#### 5.1.1 MBROLA-project

MBROLA-project<sup>1</sup> is one of the two main systems that has an Arabic voice. The MBROLA project was initiated by the TCTS Laboratory in the Faculté Polytechnique de Mons, Belgium. The main goal of the project is to have a speech synthesis for as many languages as possible. MBROLA is used for non-commercial purposes. Another purpose with it is to increase the academic research, especially in prosody generation.

The MBROLA speech synthesizer is based on diphone concatenation. MBROLA produces speech samples on 16 bits (linear) if it is provided with a list of phonemes as input together with prosodic information.

MBROLA uses the PSOLA (Pitch Synchronous Overlap Add) method that was originally developed at France Telecom (CNET). It is actually not a synthesis method itself but allows prerecorded speech samples to be concatenated and provides good controlling for pitch and duration.

The diphone databases are currently available for US English/UK English/Breton English, Brazilian Portuguese, Dutch, French, German, Romanian, Spanish, Greek, Welsh, Indian languages, Venezuelan Spanish, Hungarian, Turkish and Arabic. Some of these languages exists with male and female voice (MBROLA).

#### 5.1.2 Acapela-group

Acapela group<sup>2</sup> constitutes all speech technologies that have been developed over the last 20 years. Speech synthesis and speech recognition have been created and improved by Acapela. Acapela Group evolves from the strategic combination of three major European companies in vocal technologies: "Babel Technologies" created in

---

<sup>1</sup><http://tcts.fpms.ac.be/synthesis/mbrola.html>

<sup>2</sup><http://www.acapela-group.com/>

Mons, Belgium, "Infovox" created in Stockholm, Sweden and "Elan Speech" created in Toulouse, France. Acapela owns currently three technologies, TTS by di-  
phone, TTS by Unit Selection and Automatic Speech Recognition. Acapela is currently available for US English, UK English, Arabic, Belgian Dutch, Dutch, French, German, Italian, Polish, Spanish and Swedish.

## 5.2 Special challenges of the Arabic language for TTS systems

### 5.2.1 The diacritization problem

As has been mentioned in section 3.2. The Arabic written text does not contain vowels and other markings that make the orthography easy to understand. In their article, Mayfield Tomokiyo et al. (2003) compare the Arabic language with English. They mean that the correct pronunciation of an English word is not often obvious from its spelling and that there are many words with multiple pronunciations. This problem can be solved by relying on electronic lexicons that provide the correct pronunciation for an orthographic string. Mayfield Tomokiyo et al. mean that Arabic has no such electronic solution.

To be able to use the language, we must know what the correct vowel is. The authors mention two approaches to solve the vowelizing problem for spoken language; either inferring the vowels or enumerating the lexicon. Other authors that have approached this specific problem are Al-Muhtaseb et al. (2003). They describe in their article an Arabic Text-to-Speech System (ATTS) for classical Arabic where they solved the problem with vowelization by implementing a processor for automatic vowelization of the text before applying it to speech rules. To be able to generate vowels automatically, the processor requires integration of morphological, syntactical, and semantic information (Al-Muhtaseb et al., 2003).

### 5.2.2 Dialects

Arabic is spoken in more than 15 countries and as mentioned before by 225 million people. There are many varieties of the Arabic language, many dialects that reflect social diversity of its speakers. Mayfield Tomokiyo et al. (2003) concern these varieties in dialects as a problem for speech synthesis for many reasons. First what dialect is to be generated? The Modern Standard Arabic (MSA) or one of the dialects should be generated? The second problem would be that a limitation of listeners will rise because MSA is understood only by people with high education and/or with a good level in reading and writing. The third problem considers the transcription of spoken Arabic. Spoken Arabic has very little occasions to be written down. News and Newspapers are delivered in Modern Standard Arabic to some extension. For example nunations are not fully said in news and this can be considered as being influenced by the dialect one speaks. Mayfield Tomokiyo et al. mean that speakers of the same dialect can differ significantly in their choice of which vowel is being used in the spoken language. The reason for this is that the vowelizing of the Modern standard Arabic is only learned in school.

### 5.2.3 Differences in gender

Mayfield Tomokiyo et al. (2003) bring up the issue of gender differences in speech. There are in Arabic inflectional components that reflect the gender of the speaker and of the listener. For example consider the word تَكَلَّمَ “takallama” which means “he spoke” or “he has spoken”. If the listener is male then the imperative form تَكَلِّمْ “takallam” is said. In cases where the listener is female we say تَكَلِّمِي “ta-kal-lamī” is said where the long vowel “ī” indicates the female gender. Therefore when speech is the final product in systems such as a translation system or a synthesizer, an appropriate gender marking becomes more obvious and should be done correctly.

## 5.3 Transliteration vs transcription

To avoid the problem with vowelling and to find the vowels it was decided to use a romanized version of Arabic text, though realistically one would want to do automatic vocalization. The automatic vocalization requires integration of morphological, syntactical and semantic information. Since Arabic is written in a very different alphabet than English or any other language written with Latin alphabet, it is difficult for people with no knowledge of the Arabic alphabet to understand Arabic texts. It is than helpful to transliterate this alphabet into Latin alphabet. Transliterating is a representation of an alphabet with letters from a different alphabet. The translation is done character by character, syllable by syllable. With other words, transliteration is used to reproduce the Arabic writing system into Latin alphabets. When trying to find a standard transliterator for Arabic, it was found that the writing is similiar to its pronunciation. Therefore, it was decided to find a system for phonetic transcription that both reproduce speech and writing.

In conjunction with Professor, Bo Isaksson, at the Department of Linguistics and Philology in Uppsala a new mapping table of transcription was created. There are a few different fonts on the market to serve the purpose of transliteration. It was important to discuss these possible fonts with a competent person in the Arabic language. One of the font packages used by the Department is *Semitic Transliterator*<sup>3</sup>. These fonts follow the keyboard layout of US Windows fonts for standard English punctuation marks. The characters in the font cover the following transliteration methods for: Hebrew, Arabic, Aramaic<sup>4</sup>, Ugaritic<sup>5</sup>, Greek and the Turkish Language. Along with the fonts, it follows a Keyboard Switcher (keyboard driver) which allows access to four characters per key (instead of the normal two) and alternate keyboard layouts. This makes the transliterating or transcription work easy to type and less ambiguous than other fonts. But since Festival framework does not accept the notation for some phonemes used by the font package *Semitic Transliterator*, a transformation had to be performed. Earlier work on Arabic voice done for the MBROLA-project did use the Speech Assessment Methods Phonetic Alphabet (SAMPA), which will also be used in this project. This transformation will be discussed in section 6.1.

<sup>3</sup><http://www.linguistsoftware.com/st.htm>

<sup>4</sup>Aramaic is a Semitic language. It was probably the language of Jesus, and it is still spoken today as a first language by numerous small communities.(Comrie, 1987)

<sup>5</sup>Ugaritic is a Semitic Language, used from around 1300 BC for the Ugaritic language, an extinct Canaanite language discovered in Ugarit (Comrie, 1987)

Once the system for phonetic transcription was decided, the next step was to make a list of the Arabic vowels and consonants. This list contains four concepts for vowels. For consonants the place and manner of articulation as well as the voicing concept are included in the list. These concepts will be described in the next chapter.

## 6 Diphone databases in Festival

In Festival, the natural language modules require two types of analysis. The first type is the language specific analysis such as phones, lexicon, tokenization and others. The second type is the speaker specific analysis, where prosodic analysis such as duration and intonation are the main issues.

The phone-set definition is the first text analysis module in which every phoneme of the alphabet is classified according to phone features like consonant voicing and vowel height. How the phone-set was defined for Arabic is discussed in section 6.1.

The second text analysis module is the lexicon module. This module covers methods for finding the pronunciation of a word. This is done either by a lexicon, i.e. a large list of words and their pronunciations or by some letter to sound rules from grapheme to symbols. For some languages the syllabic structure is very simple and well defined and can be unambiguously derived from a phone string, i.e. Arabic. However, in English or Swedish this is not always the case. With other words these letter to sound rules (LTS-rules) are able to cover almost all Arabic words, except maybe for loan words. An example of loan words is the name فُولْفُو “volvo” This is discussed further in section 6.2.

The third module is the tokenization module, where numeral expressions such as date, numbers, clock, abbreviations are converted into word sequences. This module has not been processed in this thesis.

Prosodic analysis is the second module of analysis that are required by Festival. The major components of the prosodic analysis are pitch, amplitude and the duration of the speech segments. These components were not processed. Therefore the durations of the speech segments were decided roughly, only by listening to the different speech segments.

### 6.1 Specifying a phone set in Festival

As mentioned in section 2.3.1, when building a voice, a phoneme set of a language has to be well defined. Many languages have fixed phonological studies and a well defined phoneme set. Even in Arabic, where allophones of a phoneme are rare, if not to say do not exist, there are choices. For example should phonemes that cover loan words in Arabic be added to the phoneme list? Considering the phoneme /v/, which is not one of the Arabic phonemes, this phoneme does often appear in words such as names and places. Black and Lenzo (2000) mention that it is always hard and deceptive to decide the best phoneme set, which will cover the language.

The speech synthesizer should include all phones in a language. As was mentioned before in section 5.3, the Festival framework does not accept the notation for some phonemes used by the font package *Semitic Transliterator*. Therefore, these

notations have to be transformed. Earlier work with an Arabic voice done for the MBROLA-project did use the Speech Assessment Methods Phonetic Alphabet SAMPA, which is a computer-readable phonetic script using 7-bit printable ASCII characters, based on the International Phonetic Alphabet (IPA) (IPA). In the MBROLA-project, they used a modified version of SAMPA, where not all phonemes in SAMPA were used. Diphthongs were not considered since the sound is made by the combination of the vowels in question. The transformed phonemes are listed below in Table 6.1.

**Table 6.1:** Phoneme mappings from SAMPA

SAMPA symbol	New symbol
aa	A
ii	I
uu	U
s'	s.
d'	d.
t'	t.
D'	z.
ç	H
j	y

Since the speech synthesis system, Festival, requires a specification of articulatory features to be included within the phone set, these were settled according to the doctoral dissertation of Lahdo (2003). The complete phone set is included in Appendix A.

### 6.1.1 The production of vowels

The vowel system of Standard Arabic makes use of the vowels listed in Table 6.2 with information about whether it is a long or a short vowel. Information about the tongue's height, that can be either "high", "mid" or "low" and information about the location of the highest point of the tongue, what vowels can be regarded as "front", "central" or "back" are also indicated in the table. Finally, there is also the quality of a vowel which can be affected by the shape of the lips. They can be either "rounded" (+) or "unrounded" (-) (Katamba, 1989). The vowel /aa/ is an allophone of the long vowel /a/ and therefore it was added to the vowels.

**Table 6.2:** Vowels in Arabic

vowel	length	height	front	round
aa	long	low	front	-
A	long	low	back	-
I	long	high	front	-
U	long	high	back	+
a	short	low	back	-
i	short	high	front	-
u	short	high	back	+

### 6.1.2 The production of semi-vowels

Diphthongs can be seen as a complex speech sound or glide that begins with one vowel and gradually changes to another vowel, or semi-vowel within the same syllable, as /ay/ in بَيْت “bayt” (house) or as in يَوْم “yawm” meaning (day). Table 6.3 shows the diphthongs of the Arabic language with their production information. Diphthongs will not stand as phonemes since their sound can be collected by concatenating the two vowels in question.

**Table 6.3:** Semi-vowels in Arabic

semi-vowel	length	height	front	round
ay	diphthong	low	front	-
aw	diphthong	low	back	+

### 6.1.3 The production of Arabic consonants

Modern Standard Arabic uses the consonants listed in Table 6.4. Information about the consonants manner of articulation was decided and about the consonants manner of articulation that shows the ways in which articulation can be accomplished, for example the articulators may close off the oral tract for an instant or a relatively long period. The table also contains information about the place of articulation where the obstruction takes place, and the organs involved when producing a consonant. It contains information about their voicing concept, where consonants may be either voiced or voiceless (Katamba, 1989).

### 6.1.4 The production of additional consonants

Additional characters were created that are used in various Arabic-speaking countries to represent sounds not found in standard Arabic. An example of such additional consonants are the phoneme /v/ that is used for names such as “Vivianne”, “Volvo”. The consonants, that were chosen, are listed in Table 6.5, which are /p/, /g/ and /v/, and are exclusively attested in loan-words. These three additional consonants, complete the western consonant characters.

## 6.2 Letter-to-Sound rules

Letter-to-Sound rules were created where phones map to themselves. The reason for this is that the standard Arabic language has a very close relationship between orthography and phonetics, and there exists one-to-one relationship between the written form and the pronunciations. However, confusions may appear between the written Arabic and the many varied dialects. In this case the use of the standard Arabic helped to overcome these problems and confusions. Some examples of letter-to-sound rules for Arabic are shown in Table 6.6. The rest of the Letter-to-Sound rules (LTS) are listed in Appendix B.

**Table 6.4:** Consonants in Arabic

consonant	manner	place	voicing
ʔ	stop	glottal	-
b	stop	bilabial	+
t	stop	dental	-
T	fricative	dental	-
Z	affricate	palato-alveolar	+
X	fricative	laryngeal	-
x	fricative	velar	-
d	stop	dental	+
D	fricative	dental	+
r	trill	dental	-
z	fricative	dental	+
s	fricative	dental	-
S	fricative	palatal	-
s.	fricative	dental	+
d.	stop	dental	+
t.	stop	dental	-
z.	fricative	dental	+
H	fricative	laryngeal	+
G	fricative	velar	+
f	fricative	labio-dental	-
q	stop	uvular	-
k	stop	velar	-
l	lateral	dental	+
m	nasal	labial	+
n	nasal	dental	+
h	fricative	glottal	-
w	vocalic	labial	-
y	vocalic	palatal	-

**Table 6.5:** Additional consonants in Arabic, in loan words

consonant	manner	place	voicing
p	stop	bilabial	-
g	stop	velar	+
v	fricative	labio-dental	+

**Table 6.6:** Some examples of Letter-to-Sound rules

( [b] = b )	all b.s are b
( [a] = a )	all a's are a
( [d] = d )	all d's are d
( [f] = f )	all f's are f
( d. [a] = a.)	a after d. is a.
( s. [i] = i.)	i after s. is i.

## 6.3 Diphone Database Construction

The first step in constructing a diphone database for Arabic is to determine all possible diphone pairs of Arabic. In general, the typical diphone size is the square of the phone number for any language (Black and Lenzo, 2003b). As mentioned in section 3.2, Arabic has 28 consonant phonemes, four of these consonants are the emphatic ones. Two semi-vowels and six vowels and three additional consonants. This results in 39 possible phonemes. Since we are interested in possible phoneme combinations, i.e. diphones, we get  $39 \times 39 = 1521$  diphone pairs. However not all phone-phone pairs occur physically in a language. The Arabic language has few allophonic variations. These variations consider only short and long vowels when they are preceded by an emphatic consonant<sup>1</sup>. This is discussed more in detail below.

The diphone list will be categorized in different categories:

1. Vowels - vowels
2. Vowels - consonants - vowels
3. Vowels - emphatic consonants - emphatic vowels
4. Consonants - consonants
5. Emphatic consonants - consonants
6. Consonants - emphatic consonants
7. Silence - vowels - carrier word - vowels - silence
8. Silence - consonants - carrier word - consonants - silence
9. Silence - emphatic consonants - carrier word - emphatic consonants - silence
10. Double emphatic consonants

The reason why the emphatic consonants are placed in their own category is that both long and short vowels appear to take on something of the /o/ sound when they appear after an emphatic consonant. The idea for this categorization is to be able to manage the 1600 diphone pairs. This categorization is even proposed by (Black and Lenzo, 2003b).

Fabricated words were chosen to be recorded where only one occurrence of each diphone is recorded. For best result, the words should be pronounced with little prosodic variation, as monotone as possible. The advantage with recording nonsense words according to Black and Lenzo, is that one does not need to search for natural examples that have the desired diphone. The diphone list can then be easily checked and the presentation is less prone to pronunciation errors than if real words were presented (Black and Lenzo, 2003b).

### 6.3.1 The script for extracting all possible diphones

A perl script, to extract all possible diphone in Arabic, and to construct the carrier word have been developed. As mentioned before these carrier words are to help the reader to keep constant prosody during recording by letting the stress fall on these carrier words on not on the diphone pairs to be recorded.

---

<sup>1</sup>The Arabic emphatic consonants are: "s." (ص), "d." (ض), "t." (ط) and "z." (ظ)

### 6.3.2 Recording the diphones

The recordings were read by a fluent Arabic speaker. The recordings were done in a professional recording studio at Stockholm. The diphone database was completed in two hours. Two hundred sentences of different length were recorded. The reason for recording the sentences was to start building a Unit selection database to be able to re-synthesize the sentences with the standard method for resynthesis in the system, LPC (Linear Predictive Coding) in Festival. Re-synthesizing is used in this project as part of the test and evaluation methods when evaluating the Arabic voice. The sentences were taken from two sources. Firstly, news in Arabic from Sveriges radio (Sweden Radio) website. This source was chosen due to its use of modern formal language content (SR). Approximately ninety sentences were chosen from this source. These sentences have an average word length of 20 words. The recordings were tough to achieve because of the sentences length. The second source for the remainder of the sentences is from the book “A new Arabic grammar - of the written language” (Haywood and Nahmad, 2003). These sentences are short and easy to use since the vowelling is already done. The language and grammar within the book is modern, therefore a good starting point for testing the system. The reason for choosing the news in Arabic from Sveriges radio is so modern Arabic is used. The sentences were originally not vocalized, therefor vocalizing them in order to transliterate them correctly was undertaken.

### 6.3.3 Segmentation and labelling of the diphones

Once the blocks of speech were recorded, the words that contain diphones were cut out and stored in separate files, and the diphone boundaries are marked. There are numerous ways to do this. One of the ways is to Label the database by hand and to compute the diphone boundaries automatically. This method has been used in this project.

The segmentation tool of Festival was used to segment automatically the diphone list. By convention, the prompts are saved in “wav” files and their labels in “lab” files. After listening to the Arabic voice, numerous mistakes were found, a manual checking and correcting the labels was required. When a problem occurred it was traced back and checked the entry in the diphone index, then the label for the fabricated word. Whether the label matches the actual waveform file itself was checked. This was done by checking the waveform with the label file with spectrogram to see if the label was correct. A high percentage of hand-correction was made, then the basic database was built. The program Wavesurfer (KTH)<sup>2</sup> was used to check the segmentations and to correct the mistakes.

### 6.3.4 The database construction

Once the nonsense words have been labeled, building a diphone database was needed. A diphone database consists of a dictionary file, a set of waveform files and a set of pitch mark files. The dictionary file, also called the diphone index, identifies which diphone comes with which files, and from where. The index consists of a simple header, followed by a single line for each diphone: the diphone name, the file name

---

<sup>2</sup><http://www.speech.kth.se/wavesurfer/>

without any extension, a point start position in seconds, a mid position and an end position also in seconds (Black and Lenzo, 2003b). Example of the list is listed in Table 6.7.

**Table 6.7:** Example of the diphone index

The diphone name	the file name	the start position	the mid position	the end position
z.-a.	diph_1202	2.0525	2.08	2.11
t.-U.	diph_1201	0.815464	0.869588	0.980542
y-m	diph_0879	1.16109	1.22197	1.26363
h-v	diph_0832	0.804181	0.855	0.904169
n-q	diph_0795	0.931408	0.996838	1.05553
n-f	diph_0794	0.899539	0.964943	1.02371

Waveform files may be in any form as long as every file is the same type. The format must be supported by the speech tools wave reading functions. In this project no processing of this was needed, everything was done automatically by the speech synthesis system.

The pitch mark files consist a simple list of positions in seconds in order, one per line of each pitch mark in the file. This was also done automatically.

## 7 Testing the Arabic voice

To test the intelligibility, naturalness and overall quality of the Arabic Text-to-Speech system developed in this project, a test for the Arabic voice was designed. In this chapter, test parameters plus design of the test and results are discussed.

### 7.1 Test group

The only concern when choosing the test group is that they should have a good command of the Arabic language. Since it is difficult to decide what a good command is, it was decided that the participators should have the Arabic language as their mother tongue. The group consists of 9 people. The age distribution is showed in Table 7.1.

**Table 7.1:** Age distribution of the listeners

	15 - 25	26 - 35	36 - 45	46 +	Total
Female	2	3	1	0	6
Male	3	0	0	0	3
Total	5	3	1	0	9

In background questioning, a question about the listeners' occupation was asked. This was to obtain information about their education level that may affect their use of computers and their patience and tolerance. The majority of the participators are students at the Orientalistic programme at the Department of Linguistics and Philology. One of the participants is an Arabic teacher at the same department.

### 7.2 Method

The main goal of this evaluation test, which was designed according to existing methods for evaluations of the TTS-systems is to determine how much of the spoken output one can understand is.

The essential methods used to test the intelligibility are the Rhyme Tests, "the Dignostic Rhyme test" (DRT) and "the Modified Rhyme test" (MRT). The test can be divided into four parts. The first part contains twenty pairs of words with different levels of confusability. The participants heard one word at a time and marked on the answering sheet which one of the two words they think is correct.

The second part contains four pairs of sentences, also with different levels of confusability. The listeners repeated the procedure and exactly the same task is given to them. That is to mark which of the two sentences they think is uttered by the voice.

The results of both the first and the second section of the questionnaire are summarized by calculating the average error rate of the answer sheets. Usually, only total error rate percentage is given, but also single consonants and how they are confused with each other can be investigated.

The third part of the questionnaire is to listen to twenty different sentences and to write them down on a sheet. It was believed that the test will give information of the intelligibility of the system. The sentences played in this part are resynthesized from real human speech (see section 6.3.2).

The fourth part is to evaluate the system with respect to naturalness, speed, sound quality, pronunciation, intelligibility and stress/intonation. The participant is asked a few questions about these aspects and is asked to mark how well the voice performs. The method used to test the overall quality of the system is "Categorical Estimation" (CE). This method will well indicate individual strong and weak points in the system. This simple exercise will asses the overall assessment of this synthetic speech. I have though changed some scales and I have considered other aspects I would like to know about the system developed in this thesis.

The participators listened to audio files on a created website. A naive listening test was undertaken with the participators. This involved participators listening to the files then collecting their initial answers. This exercise was then repeated a second time. This was undertake to show if participants become familiar with the synthesized voice and therefore a different result is achieved.

## 7.3 Test and evaluation results

Now that the test is done according to the method mentioned in section 7.2, a summary of the results is presented in this section. The results are presented in diagrams with percentage values. I will also compare the results from the first listening with the results from the second listening. As already mentioned the test and the evaluation was done twice by each listener.

### 7.3.1 Perception of the words

This part contains 20 pairs of words with different levels of confusability. Examples of these are words that may differ in the first consonant or middle consonant or the last one. The listener hears one word at a time and marks on the answering sheet which one of the two words he/she thinks is correct.

Figure 7.1 shows the word recognition rate after both listenings. That is the results of how the listeners caught or heard each word after each listening.

The diagram shows how many words were understood by the listeners after each hearing. One can see that almost all words were recognized from the first hearing, except three words. Those words have the letters "ص" "s.", "ش" "S" and "و" "w". That means that 85 % of these words were correctly understood. 15% were correctly recognized with 89%.

Significante look at the results from the second listening one can see that there is no huge difference. The problematic words or sounds are the same after the second listening. 80 % of the words were correctly perceived. The final 20% were correctly recognized with 89%. One person did not recognize the same word as he did recognize in the first listening.

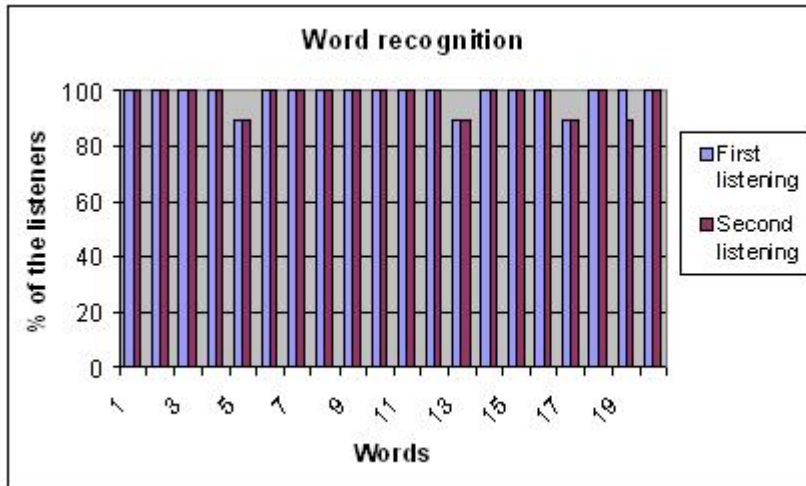


Figure 7.1: Perception of the words

### 7.3.2 Perception of the sentences - first part

This part contains four pairs of sentences with different levels of confusability, both declarative sentences and questions. Again the participator hears one sentence at a time and marks on the answering sheet which of the two sentences he/she thinks is the one uttered by the voice.

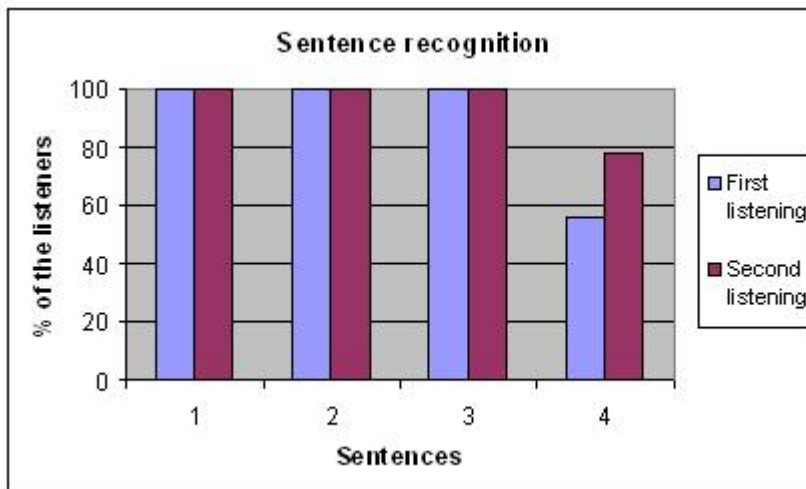


Figure 7.2: Perception of the first part of the sentences

We can see in the Figure 7.2 how the sentences were understood by the participants after the first and the second time of listening. All listeners perceived the first three sentences in the questionnaire. The last sentence had a high level of confusability because one of the sentences in this sentence pair contains the definite article "ال" "al" which is not pronounced in some cases. One of these cases is one of the so called "sun letters" where the "l" in the article changes to the initial letters in question. This sentence was correctly recognized by only 56% of the listeners.

After the second time of listening, the sentences were perceived by all subjects,

the same as in the first listening. The only difference is that 78% recognized the last sentence as opposed to 56% after the first time of listening.

### 7.3.3 Perception of the sentences - second part

In this part of the questionnaire, the participants were supposed to write down on the answer sheet what they think they hear. One sentence is played at a time and they were only allowed to listen to the audio file one time. The sentences are resynthesized by Festival. Short sentences were chosen in the hope that a person's memory would not affect the end result.

The answers are divided into three categories: correct answers, partially correct answers and incorrect answers. The correct-answers category consists only of complete and correctly recognized sentences.

The partially-correct-answers category consists of the answers where only one word was missrecognized or not recognized at all and in those cases where I suspect there have been misspellings of the words in question.

The third part, the incorrect-answers part, consists of those sentences where the participants could not recognize the meaning of the sentence or of those sentences where the participants recognize two or more words.

The Figure 7.3 lists the results of the first time of listening. 30% of these sentences were recognized correctly by all listeners. 40% were partially correctly recognized and the rest, 30%, were not correctly recognized by the listeners.

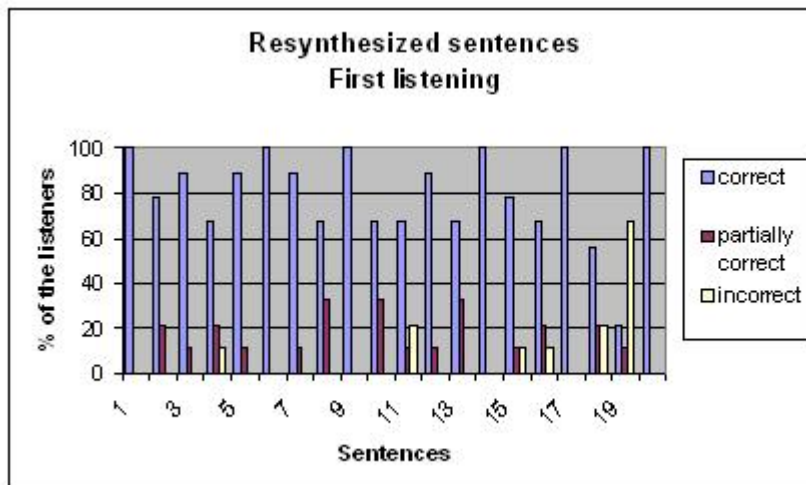


Figure 7.3: Perception of the second part of the sentences - the first listening

After the second time of listening 45% of these sentences were recognized correctly by all participants which is shown in Figure 7.4, (compared to 30% after the first listening). 50% were partially correctly recognized, 5% were not recognized correctly. This indicates that if a listener listens to the voice and gets more familiar with it the level of understanding will increase.

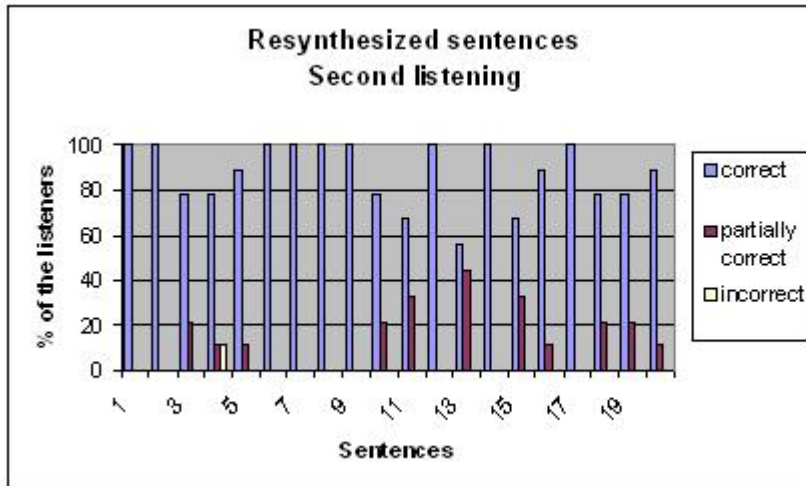


Figure 7.4: Perception of the second part of the sentences - the second listening

### 7.3.4 Naturalness

Regarding the question whether the voice is nice to listen to or not, 33% considered the voice natural, 45% thought that the naturalness of the voice was acceptable and 22 % considered the voice unnatural.

The results changed slightly after the second time of listening. Compared to the first results of 33%, 45% believed the voice sounded natural after the second time of listening. 33% thought it was acceptable or understandable and 22% considered the voice unnatural.

The results of the first time of listening compared to the second time of listening are shown in Figure 7.5 below.

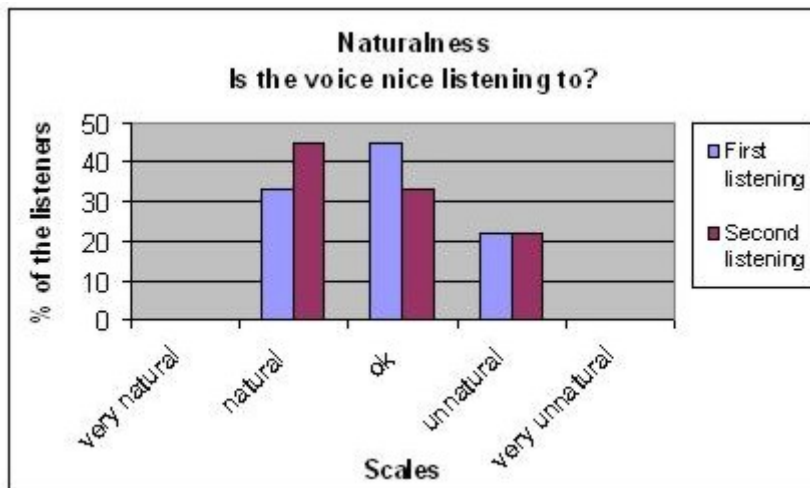


Figure 7.5: Naturalness of the voice

Another task that was given to the participants was to give a grade from 1 to 10 for the system's naturalness. Of 90 possible points, the naturalness gained 47 points after the first time of listening, that is 52%. The grade after the second listening increased.

The naturalness of the voice got 53 out of 90, that is 59%.

### 7.3.5 Speed

The speed of a system is a major concern, if the system speaks too fast or too slow this may have a negative effect on the concentration of the subjects. They might give up listening if it is too fast and the speech would not sound as natural as possible if the speech is too slow.

78% of the listeners thought that the system speaks adequately fast after the first time of listening. In other words, they considered the voice to have normal speech speed. 11% thought it was much too fast and another 11% thought it was too slow.

There is very little difference between the two listenings. Still 78% of the participants considered the system speaks adequately fast or normal after the second listening. While 11% thought the speed of the speech was too slow and another 11% thought it was much too slow. Figure 7.6 below shows the results for the first and the second listening.

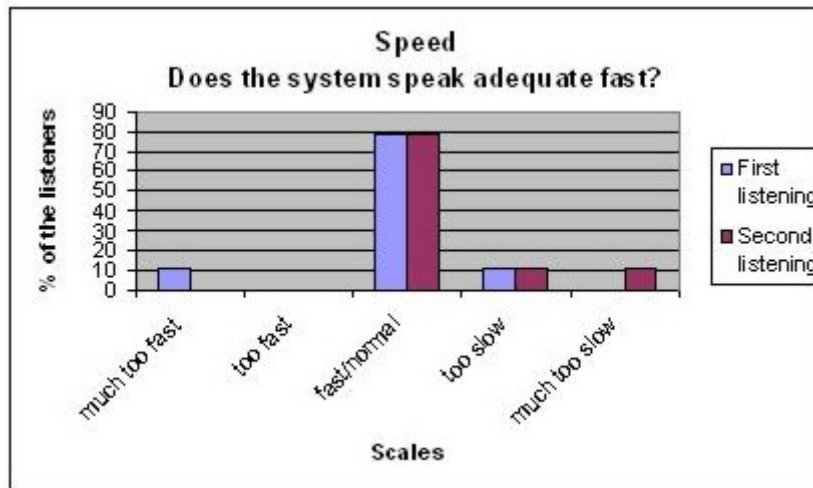


Figure 7.6: The speed of the speech

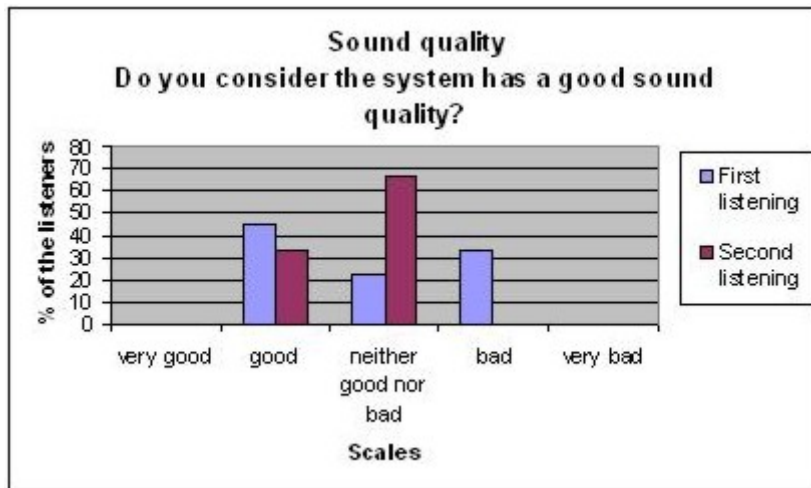
The grade for the first time of listening is 58 points of possible 90, that is 64%. After the second time of listening the grade did increase with 1 point, 59 out of 90 possible points, that is 66%.

### 7.3.6 Sound quality

The question for this part is "Do you consider the system to be of good sound quality?". After the first time of listening 45% considered the voice has good quality; 22% thought the sound quality of the voice was neither bad nor good and the remaining 33% considered that the sound quality of the system bad.

After the second listening 67% considered the sound quality of the system as neither good nor bad compared to the results of the first listening which was 22%. 33% considered the sound quality good. The results are shown in Figure 7.7 below.

When grading the sound quality from scale 1 to 10 after the first time of listening, the subjects gave the sound quality 42 out of 90, that is 47%. The grading did not



**Figure 7.7:** The sound quality of the voice

change drastically after the second listening; the listeners gave the sound quality 49 of 90, that is 54%.

### 7.3.7 Pronunciation

The pronunciation part consists of six questions addressed to the participants to be able to get an idea of how difficult the speech uttered by the system is to understand and to be able to decide what sounds are the most difficult ones to catch and gradually process these sounds in some way and improve them.

On the question whether the participants think that the voice makes many pronunciation mistakes, 33% considered that the system makes neither many nor few mistakes. 22% of the listeners believed that the mistakes were many and another group of the listeners, 22% believed that the mistakes were few. 11% considered the mistakes to be too many and another 11% thought that the mistakes were too few. This means that the subjects are able to understand what is being uttered by the system without any difficulty. Figure 7.8 shows the results.

The results of the same question after the second time of listening did not differ much from the first time of listening. 33% considered that the system makes neither many nor few mistakes. Also 33% of the listeners thought that the mistakes were few and for each of the remaining scales too many, many and too few 11% have marked them.

The second question in this category is if the listeners found it was very hard to understand some of the words. I hoped to get some information about what words were considered hard to understand and what sounds these words contained.

56% of the listeners thought it was easy to understand, while 33% thought it was neither hard nor easy, and 11% thought it was hard to understand some of the words.

The results, after the second time of listening, do not differ from the first listening. 56% of the listeners in this test thought it was easy to understand some of the words, while 33% believed it was neither hard nor easy. 11% believed it was hard to understand some of the words and they also referred to which words they found

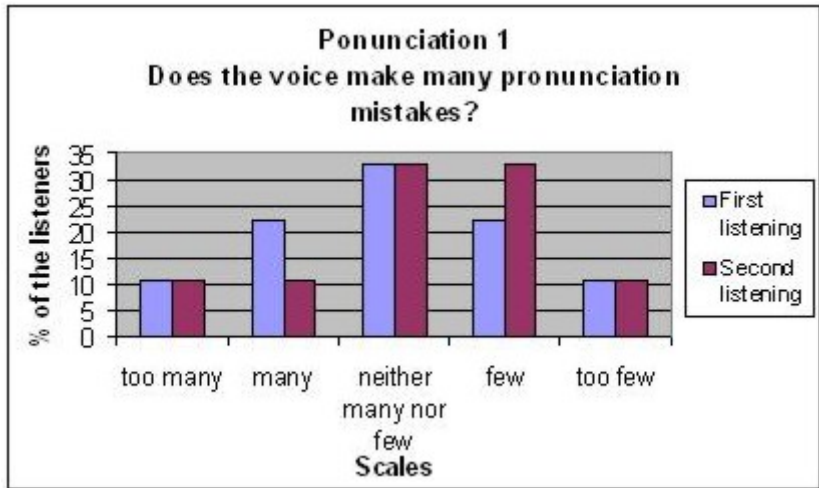


Figure 7.8: Pronunciation mistakes

hard to understand. These words contain the problematic sounds listed in Table 7.2. Figure 7.9 shows the results of the first and the second time of listening.

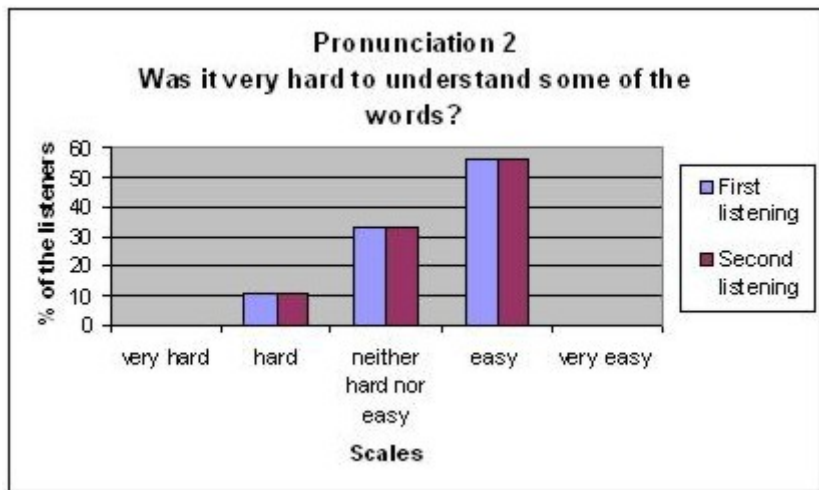


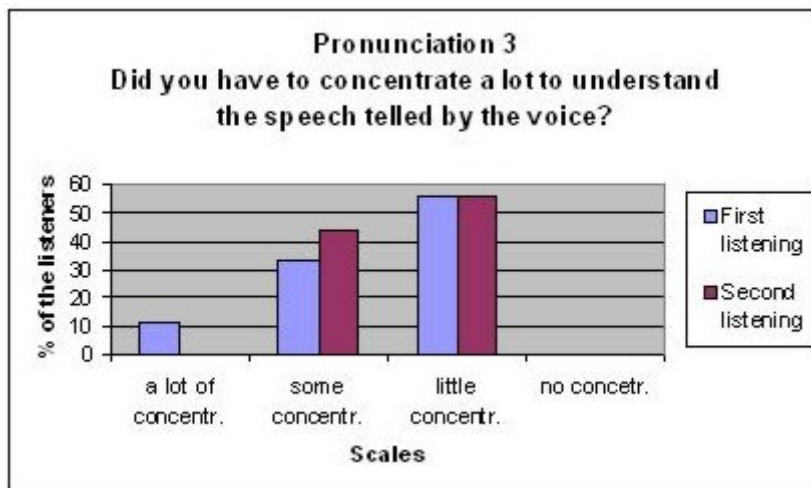
Figure 7.9: The pronunciation's effect on understanding

The third question, in the pronunciation part, is intended to investigate if the participators had to concentrate hard to be able to understand the speech uttered by the system. This question can give information about how difficult the voice is to understand and how much the participators had to concentrate to understand the voice. The results are summarized according to the subjects' own estimations.

The results from the first listening shows that 56% of the participants did concentrate a little. For 33% some concentration was needed with some specific sounds, which are listed later in this chapter. The remaining 11% had to concentrate a lot.

After the second time of listening 56% of the participants had to concentrate a little, while 44% needed some concentration with some words. The results are shown in Figure 7.10.

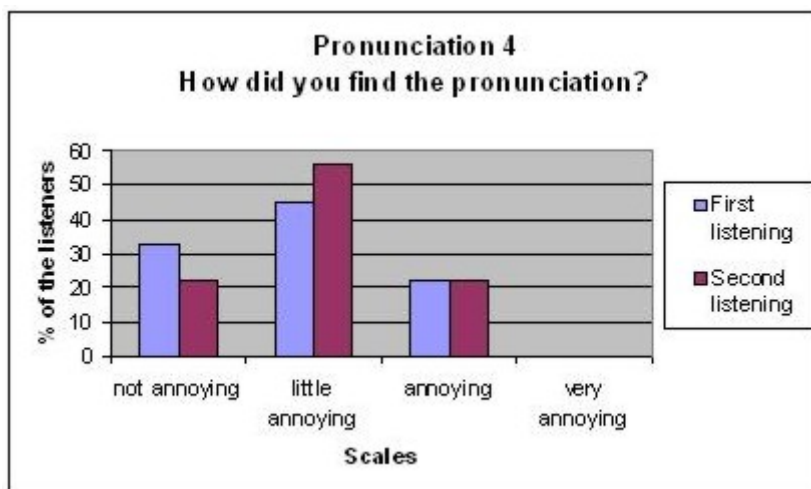
The fourth question, considering pronunciation, is how annoying the participants



**Figure 7.10:** The concentration needed to hear the pronunciation

found the voice. 45% of the participants found the voice slightly annoying, 33% of the participants thought the voice was not annoying and 22% found it annoying.

After the second time of listening 56% of the participants found the voice slightly annoying, 22% found it not annoying and 22% found the voice annoying. The results of the first and the second time of listening are shown in Figure 7.11.



**Figure 7.11:** The annoying level of the pronunciation

The fifth question in the pronunciation part was designed to extract information and to recognize which of the letters in the alphabet are problematic for this specific system. The question where the participants’ task was to write down which sounds were most difficult to understand gave the following results.

The results after the first and the second time of listening are listed in Table 7.2. The most problematic letters, after the first listening are “ش” “S”, “ج” “Z”. It is obvious that the letter “ش” “S” seems to be the most problematic letter found after the first listening.

The same problematic letters and sounds are found after the second time of listening. The problematic sounds increased from 33% to 56% after the second time of listening for the letter “ش” “S”. For the letters “ج” “Z”, all emphatic letters such as “ص” “s.”, “ض” “d.”, “ط” “t.” and “ظ” “z.” and the letter “ق” “q” the results increased from 22% to 33% after the second time of listening.

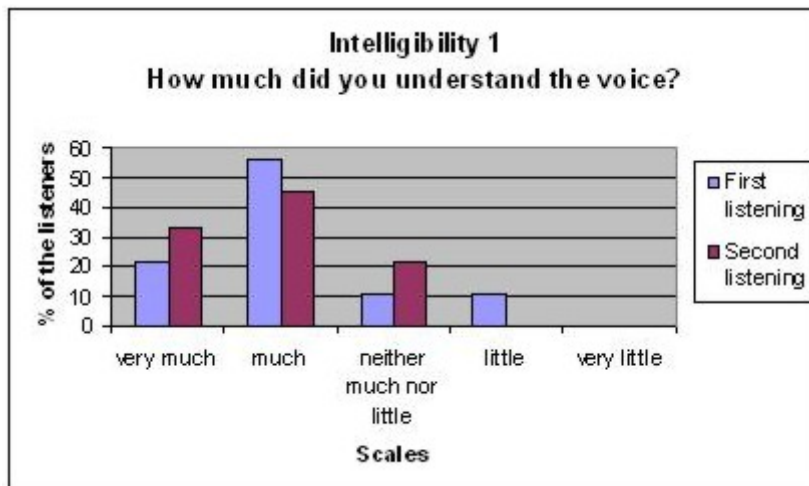
**Table 7.2:** The most problematic sounds to understand

First listening			Second listening		
Letter	Ortography	Percentage	Letter	Ortography	Percentage
S	ش	33%	S	ش	56%
Z	ج	22%	Z	ج	33%
s., d., t. and z.	ظ, ط, ض, ص	22%	s., d., t. and z.	ظ, ط, ض, ص	33%
q	ق	22%	q	ق	22%
w	و	11%			

When grading this part of the questionnaire, the pronunciation part, the listeners gave 57 out of 90 as a grade. That is 63% and I find this very satisfying at this stage. Grading after the second listening gave the same result as after the first listening, 56 of 90, that is 62%.

### 7.3.8 Intelligibility

Two questions were asked concerning the intelligibility of the system. The question how much the participants understood the voice or how much of what the voice said the participants understood, 56% of the participants understood much (well). 22% did understand the voice very much (very well), 11% neither much nor little and another 11% understood a little, i.e. not very well. As mentioned before these are the subjects' own estimations.



**Figure 7.12:** Understanding the voice

The results of the second time of listening are as follows; 45% of the participants understood the voice much (well). 33% understood very much (very well) and 22% understood a little, i.e. not very well. The results are shown in Figure 7.12.

The second question of this part is "Was the voice easy to understand?". The reason for this question was to establish if the difficulty in understanding is in the voice or in the listeners' lack of knowledge and lexicology.

After the first time of listening 67% of the listeners found the voice easy to understand, while 22% found it neither hard nor easy. 11%, which is the same percentage that understood little, considered it hard to understand the speech.

The results after the second time of listening differ slightly. 33% of the participants found the voice easy to understand and another 33% thought the voice was very easy to understand, while 23% found it neither hard nor easy. 11% considered that it was hard to understand the speech. Figure 7.13 shows the results of the first and the second time of listening.

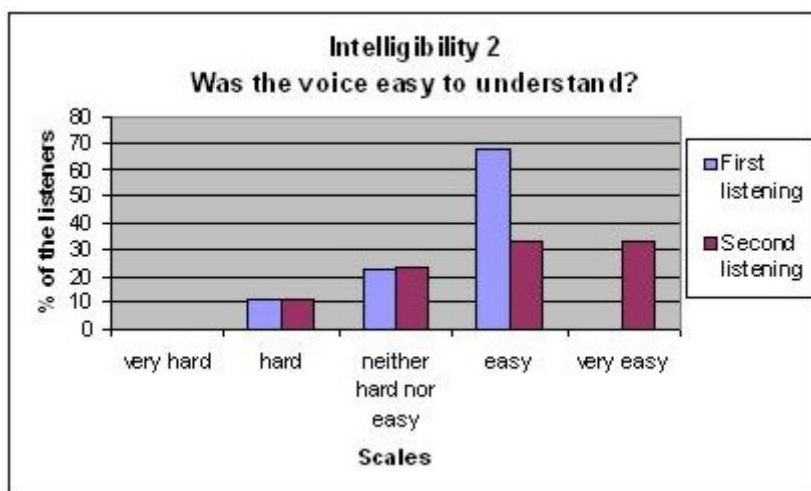


Figure 7.13: The level of difficulty in understanding the voice

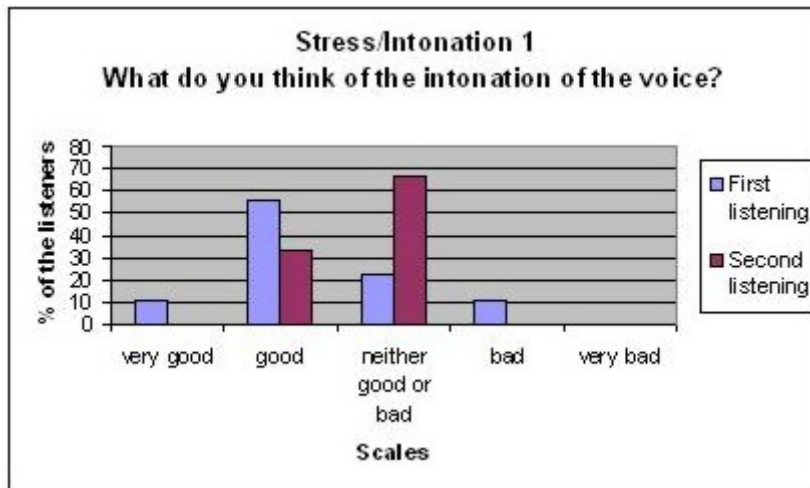
When the listeners were asked to grade the intelligibility of the system after the first listening, they gave it the grade 58 of 90. This is 64%. When the listeners were asked to grade the intelligibility of the system after the second time of listening, they gave it the grade 65 of 90. This makes 72%.

### 7.3.9 Stress/Intonation

Though no process concerning the stress and the intonation has been undertaken on the system. It was decided to survey the participants concerning the voice aspects.

The first question in the stress and intonation part is what the participants think of the intonation of the voice. The results from the first time of listening are as follows: 56% considered the intonation good. 22% thought it was neither good nor bad. 11% thought it was very good and another 11% thought it was bad.

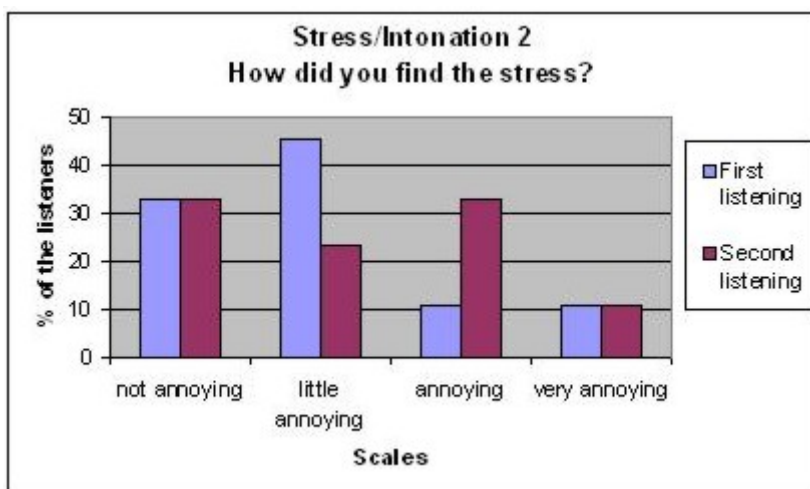
The results of the same question after the second time of listening are as follows: 67% considered the intonation neither good nor bad. The remaining 33% thought the intonation was good. The results of both first and second time of listening are shown in percentages in Figure 7.14.



**Figure 7.14:** The intonation of the system

”How do you find the stress?” is the second question in this part of the evaluating questionnaire. 45% of the subjects found the voice little annoying, 33% found it not annoying at all. 11% considered it annoying and another 11% very annoying.

33% of the listeners did not find the stress annoying after the second listening. The same percentage of the listeners, i.e. 33%, found it annoying; 23% considered it a little annoying and the remaining 11% considered the stress very annoying. The results are listed in Figure 7.15 in percentage of the number of subjects.



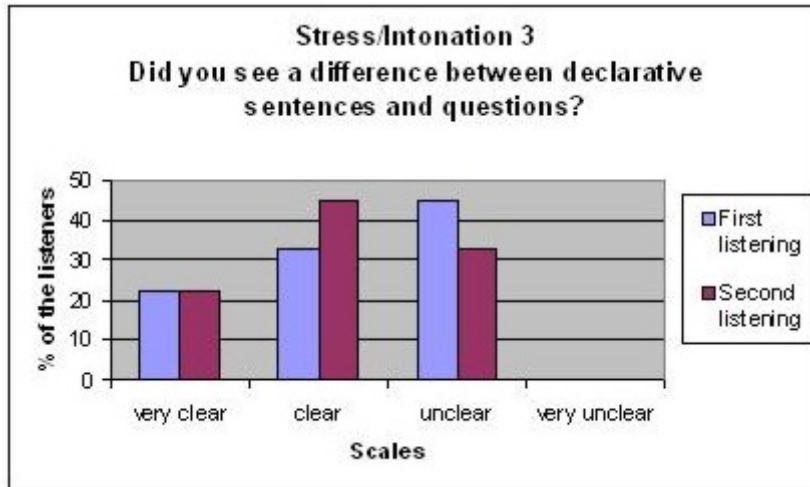
**Figure 7.15:** The stress of the system

The last question in this part is whether the listeners can see a difference between declarative sentences and questions.

After the first time of listening, 45% thought that the difference between these two types of sentences is unclear. 33% found the difference clear, and the remaining listeners, i.e. 22%, thought that the difference is very clear.

On the question whether the listeners can see a difference between declarative sentences and questions after the second listening, 54% found the difference clear,

while 33% found it unclear. The remaining 22% found the difference very clear. There is a little difference from the first listening. Figure 7.16 shows the results of both listenings.



**Figure 7.16:** The intonation difference of the system

When grading the sound quality with a grade from 1 to 10, the listeners gave the sound quality 62 out of 90, that is, 69%. This is good considering that the stress and the intonation part of the system were not processed at all. One can maybe say that the voice has really good stress and intonation. The grading did differ in somewhat after the second listening. It decreased by 7%.

## 8 Discussion

This thesis shows that the creation of synthetic speech covers a whole range of processes and that extensive work has to be done in order to build a voice in Festival. The availability of free and semi-free synthesis systems, such as the Festival Speech Synthesis System, makes the building of a speech synthesis easier and the costs lower.

This pilot project has fulfilled its purpose, through creating a fully working Arabic voice. The results of this voice are very promising, with high level of intelligibility. Although the questionnaire with the test and the evaluation of the system is quite simple and on a small scale, guidelines and information of the intelligibility, naturalness, speed and overall quality of the system can be identified. In hindsight a larger and diverse group of participants would have enabled a better evaluation of the system. The small number of participants have affected the test results and the evaluation of this project negatively. The small number of subjects do not allow to compare the subjects and the results taking age and gender into consideration, which would have been interesting to see if it would be any differences. Therefore a better division of the group and a larger amount of people is recommended.

The persons listened to 20 pairs of words, 4 pairs of sentences and 20 sentences that had to be written down. Simple and specific feedback, such as scales and rankings, was provided so processing the results was not difficult and it was easy to derive statistics from the answers. Two answering sheets were collected for each participants, one after the first time of listening and the second was collected when the participants turned it in by themselves. This means that at the second time of listening, the participants were allowed to listen as many times as they wanted. The reason for this is to see if the results would be effected by how many times the listeners are exposed to the voice. This can either have positive or negative effects on the results.

As mentioned before the test results are promising. After the first time of listening, 85% of the words were completely recognized. 15% were correctly recognized with 89%. These words contain the problematic sounds such as “س” “S”, “ج” “Z”, “ص” “s.”, “ض” “d.”, “ط” “t.”, “ظ” “z.” and the letter “ق” “q”. 75% of the sentences were successfully recognized. The final 25% had a 56% successful recognition rate. Among the written sentences, 30% were completely recognized, 40% were partially recognized and the final 30% were not successfully recognized.

It seems worthwhile to check the speech or diphone segmentation of the problematic sounds since they are considered hard to understand. A manual checking and correcting the labels is required. One has to trace back and check the entry in the diphone index and compare it to the label for the fabricated word. Data gathered shows possible areas where upgrading and improvement of the system can be carried out.

After the second time of listening the words recognition decreased to 80%. 20% were correctly recognized with 89%. 75% of the sentences were recognized and the final 25% increased to 78% successful recognition rate. Written sentence recognition

rate increased from 30% after the first time of listening to 45%, partially recognition rate increased from 40% to 50% after the second time of listening. A fact that may explain the reason why the written sentences are more difficult to recognize and to perceive is the sentences length. The longer the sentence is the more difficult it is to perceive all words.

This implies that, in general, when it comes to the intelligibility of the system, the Arabic TTS system is successful. The participants can hear what is being said and recognize changes with the synthesized speech. However, there is a plan for improvements which will be described next. The majority of both words and sentences were correctly recognized and perceived by the majority of the listeners and the evaluation of the overall quality of the system is satisfying at this stage. The words and sentences could not be predicted by the listeners. The fact that the sentences' results increased while the words results decreased, after the second time of listening, can be explained by the fact that in general words without their contexts are more difficult to recognize. The context is a main factor that effects the recognition rate.

It should be mentioned that a difference between the original recorded sentences and the synthesized sentences was noticed but apparently this difference did not affect the results to a great extent. The synthesized speech was successfully evaluated.

A remark done by the listeners is that the artificial speech was annoying and monoton. As mentioned, the project did not involve prosodic and post-lexical processing. It is then clear that the absence of these two processes effected the quality of the system in a negative way.

This thesis is considered to give guidelines and offer help when building new voices in Festival. Creating an Arabic voice is after all done though it seemed impossible at the start phase.

## 9 Summary and future development

The purpose of this thesis is to build a Festival based Text-to-Speech system for Arabic. The Arabic speech synthesizer is built with the diphone concatenation method. The challenges with the Arabic language when building TTS systems are addressed. Examples of these problems and challenges are the diacritization problem, the existing of many dialects in the Arabic language, the differences in gender and the transliteration vs transcription problem. This mapping of the problems would be helpful for others who wish to build a TTS-synthesizer in Arabic and other languages who have not been extensively studied and processed.

The implementation part considers the construction of the Arabic diphone database. First a phone set, that suits the Festival framework, was settled for the Arabic language that uses 7-bit printable ASCII characters, based on the International Phonetic Alphabet (IPA). Once a phone set was settled, all diphone pairs in the language were extracted, to be able to build the diphone database. The different steps when building the diphone database have been explained, such as the recording process, the segmentation of the diphones and the database construction itself. As a result, speech from about 200 sentences are possible to be automatically generated with the Festival system toolkit. The resulting system is evaluated using Diagnostic Rhyme test (DRT) and the Modified Rhyme test (MRT). Testing and evaluating, where Arabic speakers listened to the synthesized voice and decided whether the Arabic synthesized voice produced good, intelligible and understandable Arabic.

This pilot project has fulfilled its purpose, through creating a fully working Arabic voice. The results of this voice are very promising, with high level of intelligibility. Although the questionnaire with the test and the evaluation of the system is quite simple, guidelines and information of the intelligibility, naturalness, speed and overall quality of the system can be identified.

The test results are very promising. After the first time of listening, 85% of the words and 75% of the sentences were completely recognized. Among the written sentences, 70% were recognized. Compared to the second time of listening the words recognition decreased to 80%. The same percentage, 75%, of the sentences were recognized. Written sentence recognition rate increased from 70% after the first time of listening to 95%.

As a summary it can be said that the system provides satisfactory results after this initial testing but extensive and continued work is required to develop the system further and to get a high level TTS-system.

## 9.1 Future work

Future considerations to improve the quality of the system should be addressed. An initial task is to check the speech or diphone segmentation of the problematic sounds since they are considered hard to understand. A manual checking and correcting the labels is required. One has to trace back and check the entry in the diphone index and compare it to the label for the fabricated word.

Another important issue is signal processing to obtain the required prosody. Speaker specific intonation and speaker specific duration have to be considered when building a new voice. The major components of the prosody that can be recognized are pitch, amplitude and the duration of the concatenated speech. These components have to be processed more in order to get better speech, for example, to extract pitchmarks from an EGG signal instead of extracting them from the raw waveform would give better results.

Another issue that will make the system more widely used is to have data that contain expressions such as numbers, dates, times etc. and to ensure that Festival is able to process and generate those. Having a limited domain is different from having a standard text where such abbreviations and numbers occur.

As mentioned before, stress and intonation were not processed in this project. Since Arabic relies more on stress rather than intonation, writing syllable rules in Festival is needed to get the correct stress of syllables in words and sentences.

Other future developments, besides the above mentioned issues, is to build new voices in Arabic or any other non-European language. One can use different types of waveform generation such as Unit Selection concatenative synthesis, Articulatory synthesis or Formant synthesis.

Future development of the system, for instance, is to send the diphone recordings to the MBROLA-group and let them build the diphone database, since MBROLA uses the signal processing method Pitch Synchronous OverLap and Add (PSOLA) which often gives a better quality of the synthesized voice than LPC (Linear Predictive Coding) which is used by Festvox.

## 9.2 Conclusion

We have presented in this project that building an Arabic voice is possible. This voice was built in an easy and efficient way. We showed that the system, when provided accurate phone sets and good recording quality, can achieve speech with high intelligibility. A future use of the work done in this project is to help other persons who wish to build their own voices in Arabic or any other non-European language.

## Bibliography

- Husni Al-Muhtaseb, Moustafa Elshafei, and Mansour Al-Gamdi. Techniques for high quality arabic speech synthesis. College of Computer Science and Engineering, King Fahd University of Petroleum and Minerals, 2003.
- Alan Black, Paul Taylor, and Richard Caley. *The Festival Speech Synthesis System, Edition 1.4, for Festival version 1.4.3*. University of Edinburgh, Scotland, UK, 2002.
- Alan W Black and Kevin A Lenzo. Multilingual text-to-speech synthesis. Language Technologies Institute, Carnegie Mellon University and Cepstral, LLC, 2000.
- Alan W. Black and Kevin A. Lenzo. Language Technologies Institute, Carnegie Mellon University and Cepstral, LLC, 2003a.
- Alan W Black and Kevin A Lenzo. *Building Synthetic Voices, For FestVox 2.0 Edition*. Language Technologies Institute, Carnegie Mellon University and Cepstral, LLC, 2003b.
- Bernard Comrie. *The World's Major Languages*. Oxford University Press, 1987.
- Thierry Dutoit. *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, Dordrecht, 1996.
- Gunnar Fant. *Acoustic theory of speech production*. Mouton, the Hague, 1960.
- J A Haywood and H M Nahmad. *A new Arabic grammar*. Lund Humphries, London, 2003.
- P. Howard-Jones and SAM Partnership. 'SOAP' - a speech output assessment package for controlled multilingual evaluation of synthetic speech. Proceedings of Eurospeech 91 : 281-283., 1991.
- IPA. The international phonetic association. <http://www.arts.gla.ac.uk/IPA/ipa.html>.
- Don Johnson. Connexionx - sharing knowledge and building communities. <http://cnx.rice.edu/content/m0088/latest/>.
- Francis Katamba. *An Introduction to Phonology*. Longman Group UK Limited, 1989.
- D. Klatt. Review of text-to-speech conversion for english. Journal of the Acoustical Society of America, JASA vol. 82, pp.737-793., 1987.
- V. Kraft and T. Portele. Quality Evaluation of Five German Speech Synthesis Systems. Acta Acustica 3 (351-365), 1995.

- Ablahad Lahdo. *The Arabic Dialect of Tillo in the Region of Siirt (South-eastern Turkey)*. PhD thesis, Uppsala University, 2003.
- Laura Mayfield Tomokiyo, Alan W Black, and Kevin A Lenzo. Arabic in my hand: Small-footprint synthesis of egyptian arabic. Cepstral LLC, Pittsburgh, USA, 2003.
- MBROLA. The MBROLA project towards a freely available multilingual speech synthesizer. <http://tcts.fpms.ac.be/synthesis/mbrola.html>.
- D. Pisoni and S. Hunnicutt. Perceptual evaluation of mitalk: The mit unrestricted text-to-speech system. Proceedings of ICASSP 80 : 572-575., 1980.
- SAMPA. computer readable phonetic alphabet. <http://www.phon.ucl.ac.uk/home/sampa/home.htm>.
- INC SpeechWorks International. <http://www.tmaa.com/tts/>.
- SR. Sweden radio. [www.sr.se](http://www.sr.se).
- The DISC Best Practice Guide. A survey of existing methods and tools for developing and evaluation of speech synthesis and of commercial speech synthesis systems. <http://www.disc2.dk/tools/SGsurvey.html>.
- Amr Youssef and Ossama Emam. An arabic tts system based on the ibm trainable speech synthesizer. Le traitement automatique de l'arabe, JEP-TALN, 2004.

## A Phone set

Phoneme	SAMPA	Keyword	English gloss	Orthography
aa	-	?AlAm	pain	آلام
a	a	Xallun	solution	حل
i	i	Xilmun	dream	حلم
u	u	kurhun	detestation	كره
I	i:	fi:lun	elephant	فيل
A	a:	ma:lun	money	مال
U	u:	fu:lun	beans	فول
?	?	?aklun	food	أكل
b	b	baladun	country	بلد
t	t	tura:bun	soil	تربة
T	T	Tala:Tun	three	ثلاث
Z	Z	Zabalun	mountain	جبل
X	X	Xali:bun	milk	حليب
x	x	xita:bun	letter	خطاب
d	d	darsun	lesson	درس
D	D	Dakarun	male	ذكر

Phoneme	SAMPA	Keyword	English gloss	Orthography
r	r	rabiHun	spring	رَبِيعٌ
z	z	za:ʔirun	guest	زَائِرٌ
s	s	sala:mun	peace/hello	سَلَامٌ
S	S	Samsun	sun	شَمْسٌ
s.	s'	s.aGi:run	small	صَغِيرٌ
d.	d'	d.ayyiqun	narrow	ضَيِّقٌ
t.	t'	t.awi:lun	tall	طَوِيلٌ
z.	D'	z.ala:mun	darkness	ظَلَامٌ
H	ħ	Ha:lamun	world	عَالَمٌ
G	G	Gari:bun	weird	غَرِيبٌ
f	f	faraXun	joy	فَرَحٌ
q	q	qalbun	heart	قَلْبٌ
k	k	kalbun	dog	كَلْبٌ
l	l	lubna:na	Lebanon	لُبْنَانٌ
m	m	marad.un	sickness	مَرَضٌ
n	n	namlun	aunts	نَمْلٌ
h	h	hirrun	cat	هَرَّةٌ
w	w	wa:din	valley	وَادٍ
y	j	yaumun	day	يَوْمٌ

## B Letter to Sound rules

```
(lts.ruleset
stts_ar
( (EmphCons s t d z) )
(
;; LTS rules
;; Emphatic vowels
( EmphCons % [ aa ] = aa. )
( EmphCons % [ a ] = a. )
( EmphCons % [ i ] = i. )
( EmphCons % [ u ] = u. )
( EmphCons % [ A ] = A. )
( EmphCons % [ I ] = I. )
( EmphCons % [ U ] = U. )

;; Vowels
( [ a ] = aa )
( [ a ] = a )
( [ i ] = i )
( [ u ] = u )
( [ A ] = A )
( [ I ] = I )
( [ U ] = U )

;; Emphatic Consonants
( [ s % ] = s. )
( [ t % ] = t. )
( [ d % ] = d. )
( [ z % ] = z. )

;;Consonants
( [ ? ] = ? )
( [ b ] = b )
( [ t ] = t )
( [ T ] = T )
( [ Z ] = Z )
( [ X ] = X )
( [ x ] = x )
( [ d ] = d )
( [ D ] = D )
```

( [ r ] = r )  
( [ z ] = z )  
( [ s ] = s )  
( [ S ] = S )  
( [ H ] = H )  
( [ G ] = G )  
( [ f ] = f )  
( [ q ] = q )  
( [ k ] = k )  
( [ l ] = l )  
( [ m ] = m )  
( [ n ] = n )  
( [ h ] = h )  
( [ w ] = w )  
( [ y ] = y )  
( [ g ] = g )  
( [ p ] = p )  
( [ v ] = v )  
)

## C Questionnaire

The aim of this questionnaire is to help me test and evaluate the Arabic voice build in this project. I would appreciate if you answer freely and as honest as possible. I would like to thank you for your help and participation.

Maria Moutran Assaf  
maro@stp.ling.uu.se

### C.1 Background information (Bakgrundsfrågor)

- Gender (kön)
  - Woman (kvinna)
  - Man (man)
- Age (ålder)
  - 15 - 25
  - 26 - 35
  - 36 - 45
  - 45 +
- Occupation (sysselsättning)  
\_\_\_\_\_ (specification)

## C.2 Testing the Arabic voice (Testa den arabiska rösten)

Listen to the following audio files and tick off what you consider is the right answer  
Lyssna på den arabiska rösten och bocka för vad du anser är det rätta svaret/vad som sägs

### C.2.1 Words (Ord)

#### 1. Audio 1

قَلْبٌ (heart)

كَلْبٌ (dog)

#### 2. Audio 2

فَيْلٌ (champion)

وَحْلٌ (clay)

#### 3. Audio 3

بَعْلٌ (husband)

بَعْلٌ (mule)

#### 4. Audio 4

رَعْدٌ (thunder)

وَعْدٌ (promise/vow)

#### 5. Audio 5

بَصَلٌ (onion)

بَطْلٌ (champion)

#### 6. Audio 6

لَحْمٌ (meat)

فَحْمٌ (coal)

#### 7. Audio 7

جَمَلٌ (camel)

جَبَلٌ (mountain)

#### 8. Audio 8

شَمْعَةٌ (candle)

جَمْعَةٌ (gathering)

9. Audio 9

قَلَمٌ (pen)

كَلَامٌ (speech)

10. Audio 10

تَفَاحٌ (apples)

سُفَّاحٌ (murderers)

11. Audio 11

مَلْعَبٌ (play field)

تَعَلَّبٌ (fox)

12. Audio 12

سَيَّارَةٌ (car)

طَيَّارَةٌ (aircraft)

13. Audio 13

شَعْرٌ (hair)

شِعْرٌ (poetry)

14. Audio 14

تَارِيخٌ (history)

مَرِّيخٌ (space)

15. Audio 15

أَثَاثٌ (furniture)

أَسَاسٌ (foundation)

16. Audio 16

أَلَمَ (to pain)

أَلَمَ (he pained someone)

17. Audio 17

بَسْطَاءٌ (simple)

وَسْطَاءٌ (mediator)

18. Audio 18

بُعِدَ (to be far)

بَعْدَ (after)

19. Audio 19

بَاَعَ (to tell)

بَاعَ (to sell)

20. Audio 20

مَتَاجِفٌ (museums)

مَلَاَحِفٌ (sheets)

## C.2.2 Sentences (Meningar)

Listen to the following audio files and tick off what you consider is the right answer

Lyssna på den arabiska rösten och bocka för vad du anser är det rätta svaret/vad som sägs

### 1. Sentence 1

- كَيْفَ أُمِضِي نَهَارِي فِي الْبَيْتِ؟ (how will I spend my day at home?)
- سَوْفَ أُمِضِي نَهَارِي فِي الْبَيْتِ (I will spend my day at home)

### 2. Sentence 2

- يَوْمَ أَمْسٍ ذَهَبْتُ إِلَى الْمَدْرَسَةِ (yesterday I went to school)
- الْيَوْمَ أَذْهَبُ إِلَى الْمَدْرَسَةِ (today I will go to school)

### 3. Sentence 3

- لِمَاذَا طَلَبُونَا؟ (why did they send for us?)
- لَمْ تَعْرِفْ لِمَاذَا صَرَبُونَا؟ (you didn't know who hit us?)

### 4. Sentence 4

- أَيْنَ الذَّهَبُ؟ (where is the gold?)
- أَيْنَ ذَهَبُوا؟ (where did they went?)

### C.2.3 Sentences 2 (Meningar 2)

Listen to the following audio files and write down what you think you hear.  
(Lyssna på ljudfilerna nedan och skriv ner vad du tror rösten säger).

1. Sent. 1

---

---

2. Sent. 2

---

---

3. Sent. 3

---

---

4. Sent. 4

---

---

5. Sent. 5

---

---

6. Sent. 6

---

---

7. Sent. 7

---

---

8. Sent. 8

---

---

9. Sent. 9

---

---

10. Sent. 10

---

---

11. Sent. 11

---

---

12. Sent. 12

---

---

13. Sent. 13

---

---

14. Sent. 14

---

---

15. Sent. 15

---

---

16. Sent. 16

---

---

17. Sent. 17

---

---

18. Sent. 18

---

---

19. Sent. 19

---

---

20. Sent. 20

---

---

### C.3 Evaluating the Arabic voice (Evaluering av den arabiska rösten)

- Naturalness (naturlighet)
  - Is the voice nice listening to? (är rösten trevlig att lyssna på?)
    - very natural
    - natural
    - ok
    - unnatural
    - very unnatural
  - Please give a grade from a scale 1 to 10 for the systems naturalness (ge ett betyg från 1 till 10 när det gäller röstens naturlighet)  
\_\_\_\_\_
- Speed (hastighet)
  - Does the system speak adequate fast? (pratar systemet lagom fort?)
    - much too fast
    - too fast
    - fast/normal
    - too slow
    - much too slow
  - Please give a grade from a scale 1 to 10 for the systems speed (ge ett betyg från 1 till 10 när det gäller talhastigheten)  
\_\_\_\_\_
- Sound quality (röstkvaliteten)
  - Do you consider the system has a good sound quality? (Anser du att systemet har en bra ljudkvalitet?)
    - very good
    - good
    - neither good nor bad
    - bad
    - very bad
  - Please give a grade from a scale 1 to 10 for the sound quality (ge ett betyg från 1 till 10 när det gäller röstkvaliteten)  
\_\_\_\_\_
- Pronunciation (uttal)
  - Does the voice make many pronunciation mistakes? (Gör systemet uttalsfel?)
    - too many
    - many

- neither many nor few
- few
- too few
- Was it very hard to understand some of the words? (Var det svårt att förstå några ord?)
  - very hard
  - hard
  - neither hard nor easy
  - easy
  - very easy
- Did you have to concentrate a lot to understand the speech told by the voice? (Krävdes det stor koncentration för att förstå?)
  - a lot of concentration
  - some concentration at some words
  - little concentration
  - no concentration was needed
- How did you find the pronunciation? (Vad tyckte du om uttalet?)
  - not annoying
  - little annoying
  - annoying
  - very annoying
- What sounds were most difficult to understand? (Vilka ljud var svårast att förstå?)
 

\_\_\_\_\_
- Please give a grade from a scale 1 to 10 for the pronunciation (ge ett betyg från 1 till 10 när det gäller uttalet)
 

\_\_\_\_\_

- Intelligibility (begriplighet)

- How much did you understand the voice? (Hur mycket gick det att förstå rösten?)
  - very much
  - much
  - neither much nor little
  - little
  - very little
- Was the voice easy to understand? (Var rösten lätt att förstå?)
  - very hard
  - hard
  - neither hard nor easy
  - easy
  - very easy

- Please give a grade from a scale 1 to 10 for the intelligibility  
(ge ett betyg från 1 till 10 när det gäller begripligheten)
- 

- Stress/intonation (betoning/intonation)

- What do you think of the intonation of the voice? (Vad tycker du om röstens intonation?)
    - very good
    - good
    - neither good nor bad
    - bad
    - very bad
  - How do you find the stress? (Vad tycker du om betoningen?)
    - not annoying
    - little annoying
    - annoying
    - very annoying
  - Did you see a difference between declarative sentences and questions?  
(Tycker du att systemet gör skillnad på deklarativa meningar och frågor?)
    - very clear
    - clear
    - unclear
    - very unclear
  - Please give a grade from a scale 1 to 10 for the stress/intonation  
(ge ett betyg från 1 till 10 när det gäller betoning/intonation)
-