Pronunciation Training with Non-Native Automatic Speech Recognition

by

Sofia Lymperopoulou

MSc in Speech and Language Processing

University of Edinburgh

2006

Pronunciation Training with Non-Native Automatic Speech Recognition

by

Sofia Lymperopoulou

Supervisors

Dr. R. Clark

Prof. D. R. Ladd

ABSTRACT

In this thesis I investigate the way that automatic speech recognition can be best implemented so as to be smoothly integrated in the classes of teaching a foreign language. The main goal of this thesis was to implement a speech recogniser who would identify pronunciation mistakes made by Greek native speakers when speaking German. In particular, I am investigating the recognition of duration of vowels. The proposed recogniser shows that there is a difference between conventional recognisers and those who are directed to teaching of pronunciation, given that more principles need to be taken into consideration and more fields need to work together for a satisfactory result.

ACKNOWLEDGEMENTS

Me, Sofia Lymperopoulou, I would like to sincerely thanks all the ones who helped me, in any possible way, throughout the final year of my studies, in order to achieve the completion of this dissertation. Greetings to all my colleagues, to my supervisors Dr. Rob Clark and Prof. Bob Ladd for their constant and very helpful supervision, and to Peter Bell for his invaluable assistance.

Table of Contents

Introduction.....10

1.1 Aims and Objectives.....10

1. 2. Summary.....12

Literature Review.....14

2.1 ASR for CAPT.....15

2.2 State-of-the-art for CAPT systems.....19

2.3 Comparison.....27

2.4 Summary.....32

Corpus design.....33

3.1 Linguistic Background.....33

3.2 The Corpus.....36

3.3 Summary.....37

Data Pre-Processing.....39

4.1 Speech recording.....39

4.2 Data labelling 40

4.3 Data analysis.....43

4.3.1 The importance of the statistical analysis and the assumptions.....43

4.3.2 Analysis of the data containing speech of Greek speakers in Greek.....44

4.3.3 Analysis of the training data.....47

4.3.4 Analysis of the test data 52

4.4 Summary....54

Baseline System....55

5.1 HMM-based recognition....55

5.2 HTK....57

5.3 Components....58

5.4 Training HMMs....59

5.4.1 Initialisation 60

5.4.2 Training....60

....60

5.5 Summary....61

Testing....62

6.1 Testing strategies....63

6.1.1 Testing the recogniser with test utterances containing the junk parts....64

6.1.2 Testing the recogniser with a sequence of phones without junk parts....64

6.2 Results analysis....66

6.2 Summary....67

Improvement 68

7.1 Models with different numbers of states....68

7.2 Embedded training....70

7.3 Results analysis....72

Conclusions.....74

8.1 Future working guidelines.....75

References.....77

Appendices.....80

German Corpus.....80

Greek Corpus.....89

Gaussian distribution of the duration of stressed and unstressed vowels in Greek.....99

Gaussian distribution of the duration of short and long vowels in German.....101

Table of Figures

Figure 1: Feedback provided to the user in form of a waveform and a pitch curve and score assigned to his pronunciation in a scale from 1 to 7. 26

Figure 2: Spectrogram with labels of "junk – < t > – < a_l > – < l > – junk" using the wavesurfer software.....41

Figure 3: Gaussian distributions of the duration of both unstressed and stressed Greek "o" when pronounced by Greek native speakers. 45

Figure 4: Gaussian distributions of the duration of both short and long German "o" when pronounced by Greek native speakers. 48

Figure 5: Short vowels.....51

Figure 6: Long vowels.....51

Figure 7: Three state Hmm model.....56

Figure 8: Training procedure.....59

Figure 9: Initialisation of the models.....60

Figure 10: Re-estimation of the models.....60

Table of Tables

Table 1: Exercise focusing on prosody training (and sentence structure) for the FLUENCY project.....22

Table 2: Comparison of the CAPT systems based on whether they fulfil the pedagogical principles while they train pronunciation.....30

Table 3:Vowels of German.....34

Table 4: Vowels of Greek.....35

Table 5: The label file containing times and labels of the above labelled utterance having the appropriate format that makes it capable of being employed by HTK when training the recogniser. 42

Table 6: Mean and standard deviation of all unstressed vowels in Greek.....45

Table 7: Mean and standard deviation of all stressed vowels in Greek.....45

Table 8: Mean and standard deviation of short vowels in German appearing in stressed syllables. . 47

Table 9: Mean and standard deviation of all long vowels in German appearing in stressed syllables.....47

Table 10: Components of "ReProGreS".....59

Table 11: Part of the grammar used in the initial stage of the testing procedure.....63

Table 12: Accuracy results when recognising with 3 state models and a "loose" grammar.....64

Table 13: The grammar of one utterance of the test data, allowing for all vowels.....65

Table 14: The grammar of one utterance of the test data, allowing either of a long or of a short vowel only.....66

Table 15: Recognition performance with two different restrictive grammars.....66

Table 16: Recognition performance with two different restrictive grammars.....70

Table 17: Recognition performance with two different restrictive grammars, using embedded training.....72

CHAPTER 1

Introduction

With maturing speech technology, the recognition of speech as uttered by non-native speakers of a language is becoming a topic of interest. The study of automatic speech recognition of non-native speech is motivated by several factors. First, any deployed speech recogniser must be able to handle all of the input speech with which it is presented and, in many cases, this includes non-native as well as native speech.

Additionally, speech recognition systems for non-native speech which contain accent specific acoustic and pronunciation models can be used to serve educational purposes. New technological media are introduced in foreign language classes and contribute to better and easier acquisition of the target language. Specifically, speech recognition systems can train students to improve their accent in a foreign language. Effective communication cannot take place without correct pronunciation, since poor phonetics and prosody can distract the listener and impede comprehension of the message [1]. For this reason to obtain good pronunciation in the target language is increasingly focused upon.

1.1 Aims and Objectives

The aims of this thesis are, first, to examine how speech recognition systems can contribute to pronunciation training and be included in foreign language classes. Additionally, through the implementation of building a speech recogniser it will be investigated whether a speech recogniser can capture pronunciation mistakes that concern the duration of vowels. In order to achieve my goals, relevant theory background and recent work findings in the field of

pronunciation training via a speech recogniser will be taken into account. The aims of the project cover the following objectives:

1. Perform a thorough investigation on the state-of-art of automatic speech recognition systems and the way they contribute to pronunciation training.
2. Design a corpus in the target languages containing sounds of interest and make use of it to create training and test data to be employed while building a speech recogniser.
3. Create a system that will recognise errors made by Greek speakers of the German language focusing on the duration of the vowels.
4. Discuss the results and find those parts of the implementation that constitute the most determining factor for building a robust speech recogniser capable of spotting differences in the duration of vowels.

In order to achieve the set of stated requirements, first, software programs employing speech recognition technology have been investigated. Chapter 2 reviews and briefly evaluates commercial systems available for pronunciation training and presents the criteria and the principles which such systems should meet.

In order to fulfil educational purposes and achieve efficient and accurate results it would be beneficial if automatic speech recognition used for pronunciation training focused on specific accents each time and attempted to deal with the pronunciation mistakes expected for each case. This would be of importance, since the phonetic-phonological system of the mother language (L1) plays a significant role when acquiring the phonetic-phonological system of a foreign language (L2). The manner in which the structure of a first language may interfere with learning the sound system of a second language turns out to be complex [2]. This issue is addressed and

discussed in chapter 3.

Furthermore, in this thesis I attempt to build a recogniser which is addressed to Greek speakers of the German language. Based on our knowledge of the phonological system of both languages, we can predict the potential pronunciation mistakes of our speakers. We can somewhat anticipate the interference of L1 in learning and producing L2. Greek speakers of German tend very often to mispronounce German vowels. The question I wish to address is whether the speech recogniser can capture pronunciation mistakes that concern the duration of the vowels pronounced by Greeks speaking in German and, additionally, what are the steps taken in order to fulfil this task. In chapter 3 the vowels of the German and Greek phonetic and phonological systems will be presented and the reason for the expected mistakes when using vowels will be discussed. Additionally, the design of the corpus used for collecting training and test data is described.

The report continues with chapter 4 in which the preprocessing of the data and a statistical analysis of them is discussed and then, in chapter 5 the implementation of the speech recogniser is presented and the baseline system described. Additionally, the initial the results taken after testing the system are displayed. Chapter 6 covers the decisions taken in order to improve the system and the further results are presented. Finally, in chapter 7 the results showing the performance of the recogniser are discussed and the ideas for further expansion and improvement of the delivered system are suggested.

1. 2. Summary

After presenting a brief overview of the structure and the main points of this thesis, we proceed to the following chapter which investigates and evaluates current software systems

employing speech technology. The pedagogical principles and criteria that these systems should fulfil are, first, examined. After a brief description of the implementation of the structure of these software systems, they will be evaluated and compared depending on whether they fulfil pedagogical principles.

CHAPTER 2

Literature Review

As personal computers become more powerful and affordable, they also become more attractive as tools for foreign language teaching and learning. Teachers and learners are increasingly interested in computer assisted language learning (CALL) programs and use them as tools for training pronunciation in L2. By using CALL systems, learners of a foreign language could spend extra learning time and employ extra material to train their pronunciation [3]. With the integration of ASR technology, these systems, henceforth referred to as CAPT (Computer Assisted Pronunciation Training) systems, can be used easily by people without previous usage of similar applications. Therefore, the system processes the students' speech and provides feedback on the quality of the speech, making the learning process more realistic, engaging and educational.

However, CAPT technology still suffers from a number of limitations. This makes a number of researchers and educators sceptical about the usability of ASR technology for pronunciation training in L2, despite the fact that many students generally enjoy learning with speech enabled systems [3]. Two main features of ASR technology are regarded as most important in learning the pronunciation of a second language: the ability to recognise accented or mispronounced speech which means that the recognition performance must be adequate if not efficient, and the ability to provide meaningful evaluation of pronunciation quality which means that the identification of L2 speech errors must resemble to that of native listeners [3]. However, there are cases where these criteria have not been met (or have been met partially) and this has characterised CAPT systems problematic. In the following sections, the problems described above will be considered in order

to establish a set of the criteria for ASR-based CAPT systems through which ASR systems will be efficient enough to handle non-native speech and, at the same time, to meet sound pedagogical criteria. For this reason, several commercial CAPT systems used for pronunciation training in foreign language classes will be examined and briefly evaluated.

2.1 ASR for CAPT

Developing ASR software for CAPT systems is a complex task that relies on a wide range of expertise; it would ideally require software developers, speech technologists and educators to work together. Principles and criteria from all these fields should be taken into account in order for a successful pronunciation L2 acquisition to be guaranteed.

It can be argued that to attain successful pronunciation training when learning a foreign language several principles should be taken into consideration. First, I will describe those principles [1] that are applicable to the automatic language training situation and then I will present the criteria that an ideal ASR-based CAPT system should meet.

1. Learners must produce large quantities of sentences on their own: The learner of a foreign language should be exposed in one-to-one interactive language situations with the teacher of a foreign language in order to be prepared to participate in meaningful conversations later on. This is not always feasible since it is either costly or it is unrealistic to occur in a class consisted from several students.

2. Learners must receive pertinent corrective feedback: Helpful feedback implies that the type of correction offered will give students the tool to deal with other aspects of the same

pronunciation problem, i.e. they will not mispronounce the same phone even when appearing in a different context. Nevertheless, teachers should intervene to correct pronunciation mistakes as soon as it is appropriate for each student's personality. This means that they should not interrupt students' speech that frequently, in order to avoid discouraging them from speaking and, at the same time, they should intervene soon enough to prevent frequent errors from becoming hard-to-break habits.

3. Learners must hear many different native models: Students should be exposed to a wide range of native L2 speech when learning a foreign language, either through different native teachers or through other audio-visual material.

4. Prosody must be emphasized: Prosody is very important since it refers to the intonation, rhythm and lexical stress in speech. These are the so-called supra-segmental features. Correct usage of supra-segmental features has been shown to improve the syntactic and semantic intelligibility of spoken language [4]. Lack of correct prosody can impede the message of the sentence being pronounced.

5. Learners must feel at ease in the language learning situation: It happens very often that speakers of a foreign language loose in self-confidence when producing sounds in L2 in front of their class-mates, etc. In such cases, learners tend to avoid or quit trying to acquire the new sound by using only sounds existing in their mother language. Language instructors should increase students' confidence by correcting only when necessary, encouraging the student when an achievement is made and avoiding negative feedback.

It should be pointed out that the first four principles refer to the external environmental

conditions existing while learning a language, while the fifth principle concerns the learners' behaviour and reaction towards the pronunciation training procedure. In order for an ASR-based CAPT system to be able to achieve successful pronunciation training serving pedagogical purposes it should ideally meet the principles described above [1].

Furthermore, an ideal ASR system is described by [3] as a sequence of five phases. If the criteria contained in these phases are met, then the pronunciation training procedure can successfully produce the expected results. In the following section, I describe the sequence of the five phases, where the first four concern the ASR components, which the user is not aware of when using the system and the fifth has to do with the design and graphical user interface through which the interaction between the user and the system takes place [3].

1. Speech recognition: The incoming speech is recognised by the ASR system. The speech recogniser contains acoustic and language models with which it tries to match the incoming speech signal. Additionally, in this phase the type of the activities is determined, i.e., interactive dialogues with the computer, speech-enabled multiple-choice exercises, etc.

2. Scoring: In this phase the recognition results of the first phase are evaluated. If the student's utterance is close to the native speaker's models used as reference, the system will give a high score otherwise it will give a low score. The comparison between the student's utterance and the native speaker's models takes place on the basis of their temporal (e.g. rate of speech) or/and phonetic properties. Scoring the student's utterance is important for pronunciation training since this can give information to the user about the overall speaking quality and how this can improve over successive attempts.

3. Error detection: The system can locate the errors in an utterance and in this way the learner is made aware of the pronunciation mistakes that he/she makes. Then the learner can focus on the mistakes that he/she made and work on those ones. The errors are detected on the basis of the so-called confidence scoring which shows the certainty of the system that the recognised individually phones matched with the stored native models used as reference.

4. Error diagnosis: The ASR system classifies the types of errors made by the learner and "makes suggestions" on how to avoid them. This is useful for the learner since he is made aware of the nature of his pronunciation problem and then he can follow the suggestions made and improve the speaking skills overall. In order the recogniser to diagnose the type of an error made by the user, it resorts to previously stored models containing errors made by non-native speakers.

5. Feedback presentation: In this phase the results made in the previous phases are presented. This involves design and graphical interface issues. This phase is very important since the results should be given to the learner in a meaningful way in order the feedback to be beneficial to him.

It has been argued by [3] that by fulfilling the above described principles and criteria, effective ASR-based CAPT systems can be developed. However, it is still very difficult -yet challenging- for all these criteria to be met and an ideal system to be developed. It is undeniable that the state-of-art ASR still presents a number of issues which are either due to limitations in ASR technology or to other factors, such as the potential lack of familiarity with ASR technology when teachers and students is concerned. In the following section state-of-the-art ASR technology

is presented and briefly evaluated.

2.2 State-of-the-art for CAPT systems

Business dictation and special needs convenience are fields which have benefited from ASR technology. Therefore, it has been found that there exists a continuous development of language learning systems utilizing ASR technology in recent years. Recently, many companies developing products used to fulfil educational purposes have incorporated speech recognition technology into their products which they claim can facilitate the development of listening and speaking skills. It is of great importance to examine programs featuring speech recognition technologies in order to better comprehend the way this technology can potentially be of importance in the teaching and pronunciation learning procedure. **As far as the technological systems employed for different types of ASR systems are concerned**, early ASR-based software programs adopted template-based recognition systems which perform pattern matching using dynamic programming, where more recent ASR programs have adopted Hidden Markov Models (HMM)-based recognisers. Before introducing CAPT systems using ASR technology, I will first briefly explain the theories by which template-based and HMM-based ASR systems are governed.

In template-based speech recognisers [5, 6] the incoming speech, e.g. the unknown word, is identified by comparing it with a set of reference words known as templates. To achieve this, a copy of the template, of each word to be recognised, is created and then the incoming speech will be compared with each of these templates in turn. When recognising speech, each word is parameterised into frames containing useful information (features) for the recognition task. These are the so-called feature vectors. In order to match the incoming speech with one of the templates, we measure the distance between each frame of the test word (the word which is being

recognised) and each frame of the word in the template. Through appropriate calculations, if the distance is proved to be small or ideally zero, then the frames of the test word and the template will match and accordingly, the word will be recognised. Before measuring the distance between the frames, time alignment is applied. In this way we decide which pairs of frames (one from the test word and one from the template) align with each other. The concept of the time alignment will be extensively analysed in a later chapter.

Template-based recognition has some drawbacks [4, 5]: The templates should be representative of the reference words, otherwise they will not be a good example of the word that should be recognised. In order to deal with this problem many examples of a word can be collected. In this case the recogniser can perform satisfactory for speaker-dependent recognisers. On the other hand speaker independent recognition systems might not give sufficient recognition results since the restrictive number of the templates cannot handle the variability in speech unless multiple templates are used, which makes the computational cost high.

In HMM-based recognition [5, 6, 7] instead of matching a word with a template, the recogniser has models. The recogniser is trained on a sufficient amount of data in order to create models for words that will be recognised in the future. The models consist of states. Each state generates patterns of features (observations) to represent the word in the incoming speech. To generate a sequence of such observations we move from state to state in the model, generating an observation each time we arrive in a state. The best matching word for the recognition task is the one whose model is most likely to produced the observed sequence of feature vectors. A more detailed description of HMM-based recognisers is given in chapter 5.

Exploring CAPT systems that employ different types of speech recognisers is interesting since the advantages and disadvantages of each type of recogniser could be emphasised and additionally the way that these systems fulfil pedagogical principles could be evaluated. In this

thesis I present and compare a small number of commercial CAPT systems. **The decision about which systems would be described in this theses was taken based upon a good and analytical documentation of the systems.** The description of the systems is based mainly on information derived from the documentation of each system.

Some CAPT systems deploying more recent recognition technology, i.e. HMMs, are: Fluency [1, 8], FlueSpeak [9], ISLE [10], PhonePass [11]. A short description of their technical implementation and the way they can be employed by the user is following.

- **Fluency** [1, 8] is a system which aims at training the pronunciation of learners of a foreign language and it also provides correction of both phonetic and prosodic errors. More specifically, it points out errors for which it can provide corrective feedback in order to help the user to work on his pronunciation mistakes and improve it his pronunciation.

  The Fluency system employs the SPHINX II [12] automatic speech recogniser which is a large vocabulary speaker independent, continuous speech recognition system and deploys discrete HMMs. It contains both word dependent phone and triphone models. The recogniser applies forced alignment [6] to match the incoming speech signal with the models that it is trained on.

  It is interesting to mention that the Fluency system measures prosody features such as duration in relative terms. Due to the fact that different speakers have different speech rate the duration of one vowel is compared with the duration of the next one.

  The user does not either repeat sentences that he heard nor choose one of two or three sentences provided to him to read aloud. He has a more active role. This is possible since the sentences that the user will say can be predicted, while giving him the impression that he freely constructs utterances on his own. In order to achieve this, elicitation techniques in carefully

designed exercises are used. Table 1 displays an example of an exercise designed for Fluency which practices the structure of the sentences and prosody. The words in bold show the focus of the exercise:

System: When did you **meet** her? (**yesterday**) – I met her yesterday.
When did you **find it**?
Student: I found it yesterday.
System: **Last Thursday**.
Student: I found it last Thursday.
System: When did **they** find it?
Student: They found it last Thursday.
System: When did they **introduce him**?
Student: They introduced him last Thursday.

Table 1: Exercise focusing on prosody training (and sentence structure) for the FLUENCY project.

● **FlueSpeak** [9] is another software program employing HMMs and it also applies forced alignment. It contains a database of 30,000 phonemic models and 20,000 lexical entries to search for the most probable phoneme to best match an utterance and a scoring system in order to evaluate the user's pronunciation. It gives feedback for both phone and intonation accuracy. A pronunciation accuracy threshold is included under which the pronounced utterance is not recognised and a voice signal is heard "please try again".

The user can listen to sounds or words with an animated video clip showing a native speaker's mouth and tongue movements. Then the user can record his voice repeating the sounds or

words. His pronunciation is then recognised by the program and the pronunciation accuracy is displayed in a spectrogram on the screen. In this way the learner can compare his pronunciation to native speech visually and can try to imitate the animated native speaker's movements of the tongue and lips. Furthermore, by examining the displayed spectrogram the user can see how much he has approximated the native speaker's speech quality. The user obtains feedback through the animated native speaker's face, the spectrogram and furthermore, through other visual aids such as a vocal tract and speech waveform. In order for the user to obtain feedback for his intonation, he listens to the native speaker's pronunciation of a sentence, while seeing the intonation curve shown in yellow on the screen. The user repeats the sentence and he can see the intonation curve of his utterance and then he can compare it with the native speaker's model.

● **ISLE** [10] is a software program one of whose main goals is to provide an appropriate level of specific feedback in order to point out possible ways to improve pronunciation. It attempts to detect not only what error occurred, but also where it occurred and then it provides advice for its correction. ISLE employs an HMM-based recogniser using British English acoustic models. It also contains non native speaker's word and phone models in order to determine which types of errors are reliably detectable. The recognition procedure takes place in two stages: in the first recognition stage the recogniser is tolerant of non-native errors so that it can determine the correctness of the user's response. Then the system will re-recognise the same utterance using forced alignment and less tolerant models in order to determine the quality of the pronunciation. Additionally, a confidence scoring system which determines possible mispronounced words and phones, is included. As mentioned, the system does not merely locate the errors that occurred, but it rather classifies the pronunciation errors and provides the

user with meaningful information to improve his pronunciation. It is important to mention that the diagnosis of the errors is based on the fact that specific types of errors are expected by non-native speakers depending on their mother language. For example German learners make typical "German" mistakes.

- **PhonePass** [11] is a telephone-based test of spoken English that uses automatic speech recognition. It measures students' progress in spoken English. It is a HMM-based speech recogniser that is trained with the speech of native speakers and then adapted for use by non-native speakers. It uses forced alignment to locate the relevant parts of the input speech signal. The recogniser contains a pronunciation dictionary and an expected response network which is constructed from responses collected in over 4000 administrations of the test.

  The user is recorded reading a short text containing all English phonemes. The recogniser can access how the user's pronunciation of each word in the sentence compares with the acceptable pronunciation and it additionally measures the rate at which the user reads. The PhonePass test gives results for some skills including pronunciation correctness and fluency. There are some problems with the test such as the absence of a way to estimate the effects of prosody and the randomness of the presence of particular phonetic pitfalls.

In addition to the systems described, there are also software programs employing template-based recognition systems. These do not provide any feedback on pronunciation accuracy beyond simply indicating which written dialogue choice the user has made, based on the closest pattern match. Learners are not provided with the accuracy of their pronunciation. These programs include ISTRA [13], Tell Me More [14], Talk to Me [15, 16] and in the following section they are briefly described.

● **ISTRA** [13] employs a template-based, speaker-dependent isolated words speech recogniser. This technology is combined with graphic interface containing speech drills and has a game-like format. The graphical interface can depict a BOWLING, a BASEBALL game, etc. In such appealing environments the user is prompt to pronounce a word displayed on the screen. If the interface is modelled, for instance, on a bowling game, the feedback on pronunciation quality is given to the user in the form of number of pins knocked down. Due to the fact that the recogniser is speaker dependent, the templates are made using four tokens of the user's best productions of a target word. The chosen tokens are judged by speech clinicians by using traditional articulation training methods. The recogniser compares the current articulation of a word with the template containing the target word pronounced in an improved manner - since the four best pronounced words produced by the user are employed in the templates - and provides feedback to the user. The speech quality judgements made by human listeners have been shown to be in accordance to the judgement of the recognizer.

● **Tell me More Pro** [14] contains a template-based recogniser and includes among others interactive dialogue-based and pronunciation practice exercises. In the dialogue-based exercises the user first listens to the system and then chooses one of the three acceptable answers provided in the screen to read aloud. The dialogue develops according to the user's responses. If an utterance is not recognised by the system, the user has to repeat it. The recogniser has some helpful settings; the user could adjust the sensitivity of the speech recogniser. If it is adjusted too high, few sentences will be accepted and if it is adjusted too low, most sentences will be accepted. This could lead to distressing the learning experience, given the program does not specify clearly which level is suitable for learners. Additionally, in

the exercises which practise pronunciation the user listens to a word or sentence and then he has to repeat it. The program matches the input word with the template available. Generally, the accuracy of the speech recogniser has been shown to be quite reliable. As feedback the system displays visual voice graphs such as voice waveform. The disadvantage of this feedback is the fact that it is not easily interpreted by the user, so as for him to gather useful information on the types of pronunciation mistakes he made.

- **Talk to Me** [15] deploys a fixed-response, template-based recogniser to provide conversational practice, visual feedback on prosody and scoring pronunciation. Pronunciation is trained at phone, word and sentence level. Algorithms calculate how much the incoming speech deviates from a model and then it assigns a score on pronunciation accuracy. Feedback is provided in the form of waveforms and pitch curves in order for the user to perceive how his prosody deviates from the model. The system also assigns a score for the user's pronunciation in a scale from one to seven. Figure 1 [16] depicts an example of feedback provided to user and the score on the user's pronunciation. In the case that a word is not recognised then the system highlights this word and the user becomes aware of which word he should focus on. Additionally, the user can adjust the level of sensitivity of the recogniser in order the input speech to have a looser or tighter match to the underlying models.

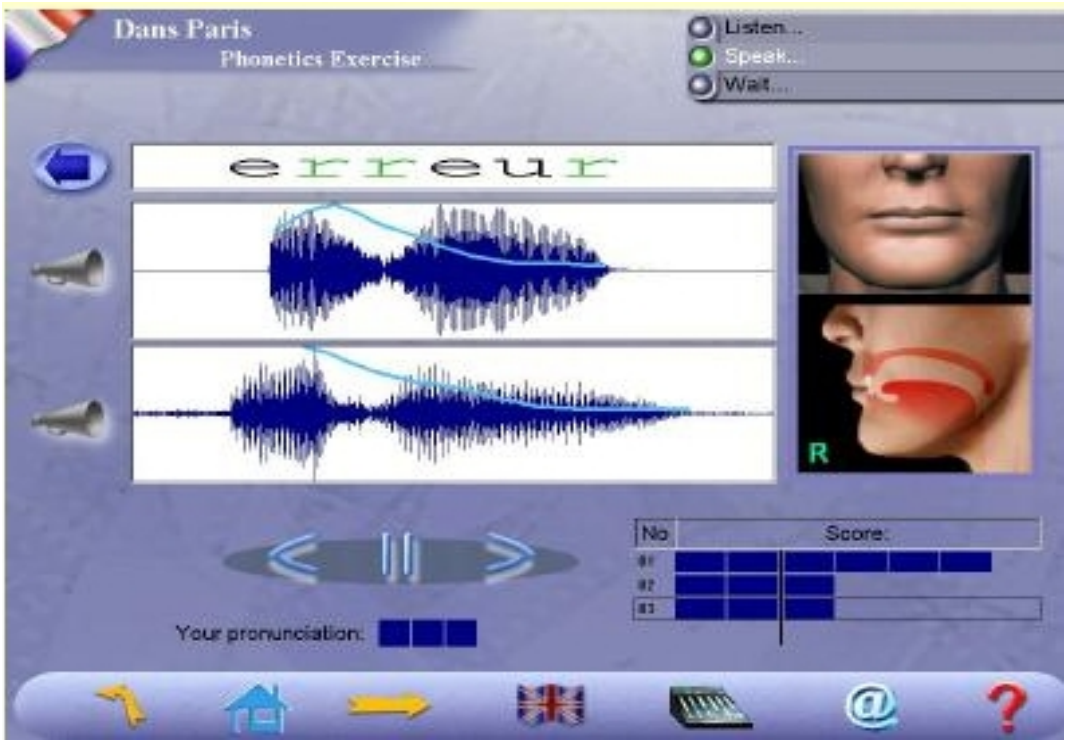


Figure 1: Feedback provided to the user in form of a waveform and a pitch curve and score assigned to his pronunciation in a scale from 1 to 7.

2.3 Comparison

After the investigation of the above CAPT systems it is obvious that several similar elements appear in most of them. Overall, every system employs ASR technology – although different types of ASR technology – which tries to match the incoming speech signal to phonetic models. Moreover, after the speech is recognised, the recognition results are evaluated and the errors in the input utterance are located by scoring the users' pronunciation – usually by setting a threshold under which the input utterance is not recognised. Furthermore, all systems present feedback to the user concerning the results conducted in the scoring phase.

As far as feedback is concerned, it should be noted that not every type of feedback is necessarily beneficial to the user or would help him correct his pronunciation mistakes and improve his overall pronunciation. For instance, in *ISTRA* the feedback is given in the form of number of pens knocked down. *Tell me More* displays a speech waveform of the user's utterance and a speech waveform of the model of reference. However, the displaying of a waveform would only be beneficial for users who know how to read them. But even if this is the case, a waveform could not a;ways be helpful to the user since the waveforms of two words (identical words) might be different, even if they are pronounced by the same user repeatedly. Additionally, the user cannot always interpret the information given through a waveform which means that he cannot use it as corrective feedback to improve his pronunciation. However, all systems either fully or partially provide some feedback concerning the user's pronunciation.

On the other hand, the above mentioned systems could be divided into two fundamentally different design types for speech recognition CAPT systems: these are the *closed response* and *open response* systems [17]. In closed response systems the user must choose one of the utterances that are displayed on the screen. The system anticipates the user's answer. However, in open response systems the user can produce the utterance on their own. The input utterance is hidden and the system should actually predict it. This makes the recognition task more difficult and the accuracy is not always high. Additionally, it is essential to mention that even the different types of recognition technology employed by the systems play a defining role in the recognition performance of the systems. The template-based recognisers contain templates which represent the speech that should be recognised. The number of templates contained in a system is restrictive and cannot give satisfactory results for speaker independent systems due to the big variability in speech produced by several humans, even if such system have efficient performance when they are speaker dependent. However, HMM-based recognition systems contain models that are

trained on several speakers and this makes them more efficient when they recognise the speech of speakers whose speech is not used to train the models.

At this point, having examined the principles judged necessary for a successful ASR system for training pronunciation of L2 speakers, and having presented software packages already available for this reason, I will examine to what extent those CAPT systems fulfil the pedagogical principles [1] necessary. In summary, these principles are:

1. Learners must produce large quantities of sentences on their own.
2. Learners must receive pertinent corrective feedback.
3. Learners must hear many different native models.
4. Prosody must be emphasised.
5. Learners must feel at ease in the language learning situation.

To facilitate better and clearer understanding of the comparison of the systems fulfilling these principles, the results are displayed in the following table.

Systems	*principl 1*	*principle 2*[1]	*principle 3*	*principle 4*	*principle 5*
Fluency	fulfilled	fulfilled	fulfilled	fulfilled	relevant
FlueSpeak	not fulfilled	fulfilled	fulfilled	not fulfilled	relevant
ISLE	not fulfilled	fulfilled	not mentioned	fulfilled	relevant
PhonePass	not fulfilled	not fulfilled	not fulfilled	not fulfilled	relevant
ISTRA	not fulfilled	fulfilled	fulfilled	not fulfilled	relevant
Tell me More pro	not fulfilled	fulfilled	fulfilled	fulfilled	relevant
Talk to Me	not fulfilled	fulfilled	fulfilled	fulfilled	relevant

Table 2: Comparison of the CAPT systems based on whether they fulfil the pedagogical principles while they train pronunciation

Although all programs encourage learners to speak aloud and produce comprehensible utterances, most of them do not satisfy the first principle because learners do not produce their own sentences. Rather, they repeat the utterances suggested by the system. This happens because speech recognisers have a better performance in close response designs. Only Fluency allows the

1 It cannot be argued that the second principle is fully met, it is rather in most cases partially met since despite the fact that most systems provide some type of feedback, it does not mean that this feedback can always be helpful or effective towards the user. However, the table shows that almost all systems fulfil the second principle, given that almost all of them provide some feedback.

user to build his own sentences. This definitely influences the recognition results, but the pronunciation training takes place in accordance with more pedagogical criteria.

The second principle is partially met since all systems give some feedback, but the type of feedback given cannot always be characterised as "pertinent". Template-based recognition systems, such as Tell me More and Talk to Me, display a waveform instead of giving suggestions to the user about how to correct his pronunciation. On the other side, FlueSpeak displays animated video clips showing a native speaker's mouth and tongue movements. Additionally, the pronunciation accuracy is displayed in a spectrogram on the screen. A motivated and experienced user of new technology can take advantage of these means in order to correct potential pronunciation mistakes.

The third principle is generally met by these programs; speakers are exposed to a range of native speakers' models to imitate. Only for ISLE it is not mentioned in the documentation of the system [10] whether the system provides a range of native speakers' models.

The fourth principle is not always met either. Whereas Fluency, FlueSpeak, ISLE and Talk to Me provide correction for prosodic errors by displaying pitch curves aid in order for the user to perceive how his prosody deviates from the model, ISTRA and Tell me More do not fulfil this principle.

Whether the fifth principle is fulfilled by all programs is quite relevant. On the one hand, it is argued [1] that when speaking to a machine and not to a human instructor, the speaker feels at ease. since not speaking and not being "judged" and evaluated by a human instructor reduces the anxieties of the learner in speaking a foreign language. Additionally, using a software program with a pleasant, interesting and user-friendly graphical interface influences the user's mood and enthusiasm positive towards the pronunciation learning experience. On the other hand, it could be argued that there are learners having no great experience using software programs. For this type

of learners, using a CAPT system to improve their pronunciation could be a stressful task since they encounter the use of the computer with hesitation or even fear.

Finally, while the information displayed in table 1 might indicates that the PhonePass system hardly fulfils any of the principles described above, we should bear in mind that PhonePass is designed for testing pronunciation correctness and not for training it. The reason why this paper mentions this system is that it could be used to evaluate the other CAPT systems and to examine whether the user that employed them attained pronunciation improvement.

2.4 Summary

In this chapter I have discussed how CAPT systems employing ASR technology should fulfil some principles and criteria in order to successfully train the user's pronunciation. Capturing pronunciation mistakes is a difficult task for a recogniser to perform. Phonetic differences, such as differences in phonetic duration or even in phone quality, between two languages are sometimes subtle. For this reason the recogniser should be capable of capturing such differences. The efficiency of a recogniser depends on several factors, such as the type of recogniser employed – HMM or template-based –, whether it is speaker dependent or – independent, or whether it is trained on an adequate number of data, etc.

This thesis focuses on the ability of a recogniser to capture such pronunciation errors. After exploring the abilities and performance of pronunciation training systems I will examine how demanding this task is, I attempted to build a speech recogniser and find out which parts of the implementation process would be the trickiest ones. The goal was to make a recogniser which is capable of capturing differences in vowels produced by native speakers of Greek when speaking German as an L2. The following chapters present the procedures undertaken in building such a

system and debate on the success of the recognisers' performance.

CHAPTER 3

Corpus design

In order for a recogniser to be built, a corpus is needed for the training and test data. In this chapter I describe the design of this corpus. Since my goal was to examine whether a speech recogniser can detect pronunciation mistakes concerning the duration of vowels in German and Greek, I designed a corpus containing sentences that focus on vowels pronounced in specific contexts in the two languages. First, it is essential to make a linguistic comparison of Greek and German and describe the phonetic system of the two languages with respect to vowels, in order to clarify the idea on which the design of the corpus was based. Furthermore, the procedure of the design of the corpus is outlined.

3.1 Linguistic Background

The phonetic-phonological system of the mother language (L1) plays an important role in the acquisition of the phonetic-phonological system of a foreign language (L2). The manner in which the structure of a first language may interfere with learning the sound system of a second language turns out to be complex. There are two sources of interference: the phonological and the

phonetic [13]. We can predict certain types of L1 interference in L2 phonological learning using knowledge of the phonological structures of the two languages. Phonological rules existing in L1 can be wrongly "borrowed" when using L2, leading to pronunciation mistakes. On the other hand, it is less obvious that phonetic similarity of L1 and L2 can contribute to the difficulty of L2 learning. According to [2], it is more difficult to produce a phone in L2 that is closer in formant space to a phone in L1 due to "equivalence classification". This means that in the case that a phone in L2 is perceived as being sufficiently close enough to a phone in L1, the learner will use the phone that he already knows rather than learn a new sound. This is precisely the case I will be investigating in this thesis.

Based on the above mentioned, I assume that the Greek speakers of German will mispronounce the German vowels since, in each case, they will use the vowel that they already know rather than learn the German vowels. This could happen due to the fact that the German vowels are sufficiently close enough in formant space to the Greek vowels despite the fact that they actually differ in quality and duration. What follows is a phonetic comparison of the vowels of the two systems in order to point out the degree that the phonetic systems of the two languages deviate from each other. Tables 3 and 4 show the vowels of the two systems and makes the difference between the vowels distinct.

	front	central	back
high	[i:] [y:] [I] [Y]		[u:] [U]
mid	[e:] [ø:] [ɛ] [œ] [ɛ:]	[ə]	[o:] [ɔ]
low		[ɑ:] [a]	

Table 3:Vowels of German

	front	central	back
high	[I]		[U]
mid	[ɛ]		[ɔ]
low		[a]	

Table 4: Vowels of Greek

German has seventeen vowels. The vowels are either both long and tense, meaning that they are articulated with the muscles of the lips being tensed, or short and lax. The tense and lax vowels differ not only in duration but also in quality [18]. Only the vowel [ɛ:] is an exception to this rule since it has a long duration but it is pronounced with lax lips [19].

On the other hand, Greek has only five vowels. They are all short and lax. It has been established, though, that short vowels in Greek are pronounced with a longer duration only when they appear in stressed syllables [20, 21]. Finally, [o] and [u] are the only rounded vowels of Greek.

From the above description it is shown that German contains vowels which do not exist in Greek. These are all the long vowels, the front-rounded vowels [y:],[Y], [ø:], [œ] and the schwa [ə]. For this reason, it is expected that a Greek speaker of German will pronounce the vowels with a short duration and not tensed – when needed –as there are no long and tense vowels in Greek [19].

3.2 The Corpus

Building a recogniser and then testing its recognition performance in order to establish whether it can predict the duration of the vowels, requires a set of training and a set of test data. A well designed corpus should be used to create the training and test data. This section describes the design of the corpus employed to build the recogniser. The corpus consists of sentences in Greek and in German. First, I will describe the set with the Greek sentences. The five vowels of the Greek system are as mentioned above: [a], [ɛ], [I], [ɔ], [U]. First, I found words containing each vowel in the same context -the same phone at the right and left side of the vowel of interest. The vowels are all, according to the phonetic rules described above, short and lax. As already mentioned above, when a vowel appears in a stressed syllable, it has a longer duration. Because of this parameter I used the vowels in both stressed and unstressed syllables, keeping the context the same. In the case that the vowels appear in an unstressed syllable, the phone at the left is the [t] and the phone at the right is the [l]. When the vowels appear in stressed syllables the context is [m] and [n] respectively. For example, the unstressed vowels appear in the form t < a > l, t < ɛ> l,

t < I > l, t < ɔ > l, t < U > l, and the stressed vowels in m<a>n, m < ɛ > n, m < I > n, m < ɔ > n, m < U > n.

The words constructed were then embedded in sentences, which would sound natural and could belong in the every-day speech of a Greek native speaker. The word containing each vowel in stressed and unstressed syllable, keeping the context the same, was repeated 5 times in different sentences occupying a different position in each sentence. This was due to the fact that differences in position can cause differences in the stress or the intonation of the word. In addition, the vowels both in stressed and unstressed syllables are used in five different contexts since the effects of co-articulation are such that the acoustic realisation of any one phoneme can vary greatly depending on the acoustic context [5]. In total, I created 100 sentences since words with each vowel in a stressed syllable with the same context are repeated five times and five more sentences containing words with the same vowel in different contexts. The same is the case where the vowel is contained in unstressed syllables.

For the design of the set of the German sentences I focus, first, on vowels being long – and in some cases also tense – which appear in stressed syllables and have the same context. The long vowels have the form: t < a: > l, t < e: > l, t < i: > l, t < o: > l, t < u: > l. The word containing the vowel of interest with the same context is repeated five times in five different sentences each. Then, the long vowels are used appearing in five different contexts, for the reason mentioned above. In the same way, I concentrate on vowels being short and lax in stressed sentences, as well. The reason for this is the following: it is expected that the Greek speaker of the German language will probably pronounce the vowels in the stressed syllables with a longer duration due to the parameter existing in the Greek phonetic system. The form of the short and lax vowels in the same context and in stressed syllables is: m < a > n, m < ɛ > n, m < I > n, m < ɔ > n, m < U > n. Then the vowels of interest are used in five different contexts in five different sentences. In this

case I created 100 sentences, as well. Both the Greek and German corpora can be found in the appendices.

The assumptions made while designing the corpus are tested in chapter 4 where the data are analysed in order to find out how much the duration of the German vowels differs to the duration of the Greek data.

3.3 Summary

Taking into account linguistic phenomena existing in the target languages and then making a comparison of the phonetic systems of the languages is fundamental, when someone wishes to examine whether a machine can be so well trained and designed in order to capture their differences and to achieve similar results like those achieved by human judgements. It should be always borne in mind, though, that only carefully designed data to train and to test the machine i.e. the speech recogniser, can contribute to the construction of a capable for this task recogniser and can bring the desired results. The following chapter describes how these data were collected and processed. Finally the data are statistically analysed in order to reassure whether they fulfil the linguistic theory concerning the vowels of Greek and German, described in this chapter.

CHAPTER 4

Data Pre-Processing

Building a speech recogniser from scratch involves a number of tasks that concern a preliminary data processing. The preparation of both training and test data is essential. This chapter contains details concerning the collection of the data. Furthermore, the appropriate pre-processing of the data is described, since they were assigned with the appropriate format that enabled them to be used by the speech recogniser in a future stage. Finally, the data were analysed in order to support the assumptions concerning the things that the recogniser was expected to do.

4.1 Speech recording

The first main issue encountered in this project was collecting data. In order to achieve this the corpus described in chapter 3 was designed. The corpus consisted of two set of sentences: the German and the Greek. Each set contained 100 utterances. In order to create training and test data, the utterances were recorded by 15 different speakers. For collecting the training data 6 German native speakers, 3 males and 3 females adults, were recorded reading the German corpus, resulting in a set of 600 utterances. The recording took place in the sound-proof booth of the recording studio of the Linguistics Department of the University of Edinburgh, using AKG CK98 Hypercardoid microphone and SONAR software. For collecting the test data 3 Greek speakers (all females) of German were recorded reading the German script in the same recording studio as mentioned above and using the same microphone and the SONAR software. The utterances used for test data were 300.

Additionally, 6 female Greek native speakers were recorded reading the Greek corpus and

resulting in a set of 600 utterances, as well. The recordings of the Greek speakers did not take place in a recording studio but in the living room of a house trying to keep the environmental conditions as quiet as possible. This means that it had been taken into account to keep the background noise levels of the recorded data low; in order not to have an negative impact on the following phases of development of our recogniser. The recording took place using a DAT recorder and the microphone was a headset-type.

Whereas all utterances spoken by each speaker and used for training data were stored in one wav file (6 files for the recordings made by Germans and 6 files for the recordings made by the Greek speakers), in the recordings of the test data each utterance was stored in a different wav files (300 wav files).

4.2 Data labelling

The recorded speech should be pre-processed for the training of the HMM models and furthermore, in order to create test data having the appropriate form for testing the performance of the recogniser later on. The waveform files were manually labelled and the speech was segmented into the phones of interest.

Speech sounds can differ in pitch, loudness and quality. When labelling the training data, I should distinguish sounds based on their quality. Sounds depicted on a spectrogram have some specific characteristic which distinguish the one sound from the other. For instance, all voiced sounds are distinguishable from one another by their formant frequencies, voiceless sounds are evident either by the high frequencies occurring when they are spoken or by the formant transitions occurring in the previous or in the following voiced sound [18].

The goal when labelling was to find the phones which would be used for training HMM

models and to determine where the boundaries between one phone from the other lie. Particularly, in each utterance the vowel of interest was, first, found and labelled and then the phones comprising the context of the vowel were labelled. Moreover, the remaining left and right part of the sentence was labelled as "junk". (see for example Figure 2). However, determining the boundaries of phones was not always an easy task. Whereas it was easy to determine the boundaries between nasal sounds, fricatives or voiceless stops, and vowels, it was not always very clear where to set the boundaries between a vowel and a lateral or approximative consonant due to the fact that these consonants have characteristics, i.e. formants, not unlike those of vowels. In cases like that, the formant transitions play a significant role since they show the transitions from one sound to the other without being clear where the one sound stops and the other begins. Thus, in all those cases I tried to be as consistent as possible in my segmentations.

Therefore, this way of labelling i.e. “junk – consonant – vowel of interest – consonant – junk”, was not enough for the test data, since different strategies of testing the recogniser were applied (the strategies used for testing the recogniser are described in detail in chapter 5). For this reason, I created three different formats of label files for the test data. In the first format of the label files, the data were segmented into the phones of interest .i.e. a label for the vowel of interest and two labels for each surrounding phoneme. This set of labels was used for analysis of the test data, where the duration of each vowel was examined in order to come up with some statistical results. This is described in detail in the following section. In the second set, the data were segmented into labels containing a sequence of phones, i.e. the vowel of interest and the surrounded context and of labels of junk. In the third set, the labels containing a sequence of phones were split. In this way, the junk parts of each utterance of the test data were removed. The two last label sets were used for testing the performance of the recogniser.

The segmentation took place by using the *wavesurfer* software program. When in each wav

file the phones of interest – including the labels of the junks – were labelled, *wavesurfer* created automatically a new file (*label* file) containing the time space of each label in each utterance. Figure 2 displays an utterance labelled by *wavesurfer*.

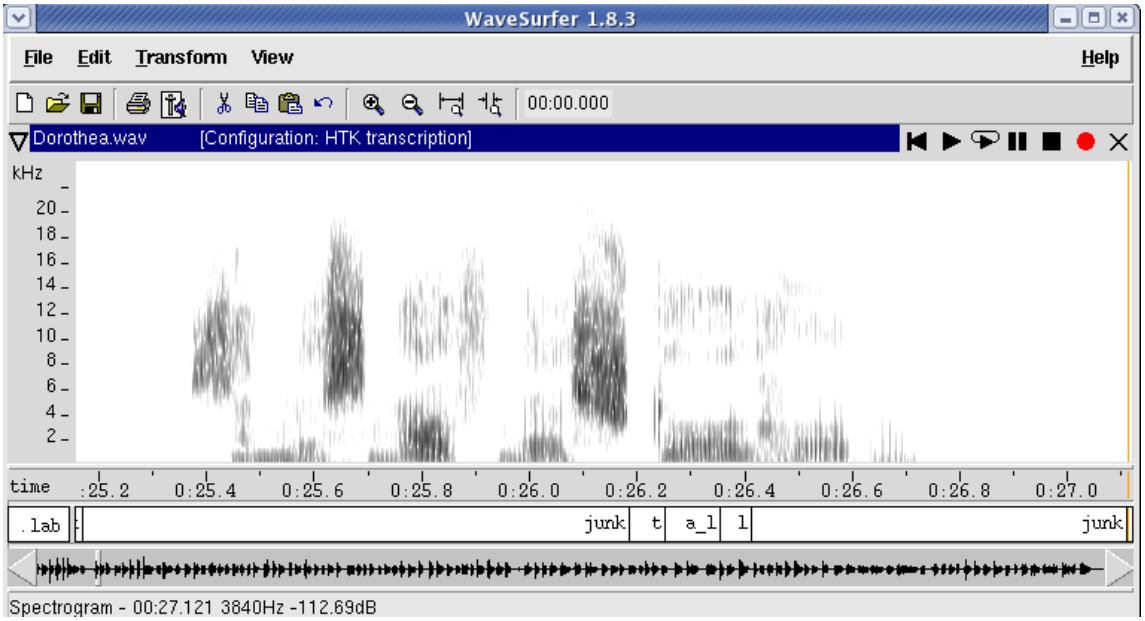


Figure 2: Spectrogram with labels of "junk – < t > – < a_l > – < l > – junk" using the wavesurfer software.

Wavesurfer was used due to the fact that HTK – HTK is mentioned in chapter 5 – requires a particular format of label files for labelled data. When using Wavesurfer for labelling, it automatically creates label files which could directly be deployed by HTK. Table 5 shows an example of a label file.

```
0 262010000 junk
262010000 262510000 t
262510000 263570000 a_l
263570000 264080000 l
264080000 287160000 junk
```

Table 5: The label file containing times and labels of the above labelled utterance having the appropriate format that makes it capable of being employed by HTK when training the recogniser.

4.3 Data analysis

After collecting and labelling all the data, it was necessary to analyse them phonetically since the results of a phonetic analysis would be useful during the implementation of the speech recognition system which will be described in chapter 5. In this section, firstly, we discuss the necessity of analysing the speech data and present the initial assumptions. Furthermore, I will describe the statistical analysis of the data. Finally, the results I came up with through the analysis are presented. The analysis concerns the duration of all vowels involved in the data I have collected. Additionally, the data was statistically analysed using the SPSS software [22].

4.3.1 The importance of the statistical analysis and the assumptions

There are several reasons that made a statistical analysis of my phonetic data necessary:

- It will be examined whether the assumption I made regarding the differences of the linguistic characteristics of the phonetic systems of German and Greek will be duplicated in my data. At this point, I am only referring to measurement of duration.
- Comparing the vowels of the training data i.e. the vowels produced by German native speakers, with the vowels contained in the test data i.e. the vowels produced by non-native speakers of German, would contribute to giving a definition of what is a "pronunciation mistake", i.e. which duration length defines a long or short vowel in German, and then according to this, the utterances of the test data containing a pronunciation mistake (in reference to the duration) will be revealed.
- A statistical analysis of the vowels in the training data will indicate the outliers contained in them. Examining the outliers is useful since it establishes the reliability and homogeneity of the training data. Additionally, if it is proved that the training data contain many outliers,

then the utterances containing the outliers could be removed when the data will be used to train the recogniser. In this way it is expected that the recogniser will contain better and stricter acoustic models. Having strict models is important when the recogniser comes up with the task of predicting duration of vowels since duration is a very refined detail contained in the acoustic models. Only strict models could "capture" differences in duration of a vowel and when recognising speech they could discern which vowel is short or long according to the data they are trained on.

Based on the linguistic attributes presented in chapter 3 which concerns the phonetic systems of German and Greek, I assumed that vowels produced by Greek speakers of German would differ in duration to vowels produced by native speakers of German. More precisely, it was expected that the analysis would prove that the Greek speakers would pronounce long German vowels with a shorter duration. Moreover, the short but stressed German vowels were expected to be pronounced with a longer duration by the Greek speakers due to the parameter existing in the Greek phonetic system stating that despite the fact that the Greek language has only short vowels, when a vowel appears in a stressed syllable, then it is pronounced with a long duration.

4.3.2 Analysis of the data containing speech of Greek speakers in Greek

First, I analysed the data containing speech in Greek pronounced by native Greek speakers. In this way, I wanted to find out whether there is indeed a difference in the duration of Greek vowels of my data depending on whether they appear in stressed or unstressed syllables. I calculated mean and standard deviation of each vowel in the Greek data. The vowels which were examined were the stressed /a, e, i, o, u/ and the unstressed /a, e, i, o, u/. The following tables depict the values for mean and standard deviation of first unstressed and then stressed vowels.

unstressed vowels	*mean*	*standard deviation*
a	61.82	13.9
e	40	11.29
i	36.98	13.24
o	61.14	15.92
u	39.96	8.87

Table 6: Mean and standard deviation of all unstressed vowels in Greek

stressed vowels	*mean*	*standard deviation*
a	112.28	30.45
e	79.88	28.19
i	88.88	24.93
o	105.22	35.67
u	90.56	24.76

Table 7: Mean and standard deviation of all stressed vowels in Greek

The values of the means presented in the tables show the average duration of each vowel. From the mean values it is established that stressed and unstressed vowels have a considerable difference in duration. Moreover, the values of the standard deviation show the dispersion the data from the mean value. The more spread apart the data is, the higher the deviation. In Figure 3, the Gaussian distributions of the unstressed and stressed Greek /o/ are illustrated and the way the data are distributed around the mean of each Gaussian is depicted. The time in the horizontal axis is measures in milliseconds (ms).

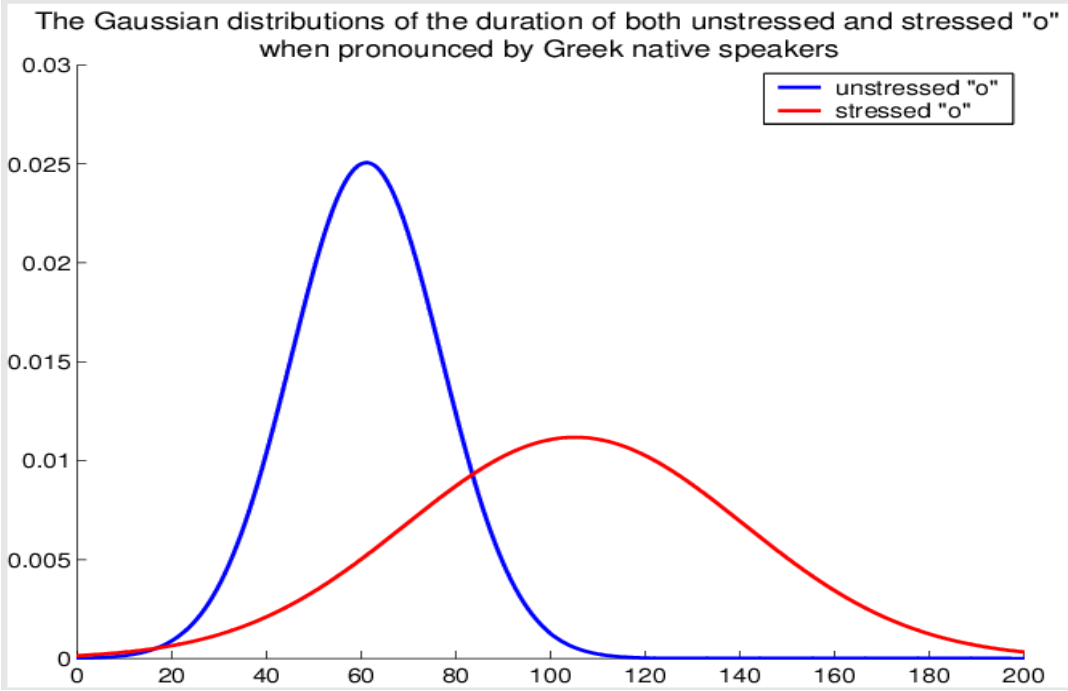


Figure 3: Gaussian distributions of the duration of both unstressed and stressed Greek "o" when pronounced by Greek native speakers.

It is clear that the data of the stressed vowels are more spread apart from the mean value than the data of the unstressed vowels. This means that there are no big variations of the duration values of the unstressed vowels, whereas the duration of the stressed vowels seems to be more unstable since it takes values from a greater range.

In order to make sure that I indeed have the differences presented in the above mentioned distributions, I ran statistical tests, investigating the hypothesis that the stressed and unstressed vowels are significantly different.

For that purpose I ran One-Way ANOVAs for each of the Greek vowels using the duration of the vowel as dependent variable, and the stressed against unstressed distinction as independent variable. All vowels were found to be significantly shorter in their unstressed renditions, than in the stressed ones. For example, the vowel o was found significantly longer when stressed with

p<.001 (F= 63.678, df= 98). The same significant difference was also found for all other vowels.

4.3.3 Analysis of the training data

In the second part of my analysis I focused on the training data containing speech of the native speakers of German. I calculated the mean and standard deviation of each vowel – contained in my data – in the German language. The vowels which were examined were the short /a, e, i, o, u/ and the long /a, e, i, o, u/. The values for mean and standard deviation of, first, short, and then, long vowels are presented in the following tables:

short vowels	***mean***	***standard deviation***
a	69.87	12.6
e	63.03	14.71
i	48.28	19.05
o	57.55	15.71
u	58.15	15.05

Table 8: Mean and standard deviation of short vowels in German appearing in stressed syllables.

long vowels	***mean***	***standard deviation***
a	134.57	31.33
e	103.12	27.82
i	81.57	21.43
o	111.73	24.82
u	81.42	28.36

Table 9: Mean and standard deviation of all long vowels in German appearing in stressed syllables.

Analysing the values of the means presented in the tables it is established that long and short German vowels do have a remarkable difference in duration. Moreover, Figure 4 depicts the

Gaussian distributions of the short and long German /o/ in order to make clear how much the data are spread from the mean of each distribution and to see which values of the duration represent both the short and long /o/. The Gaussian distributions comparing the rest vowels are included in the part with the appendices.

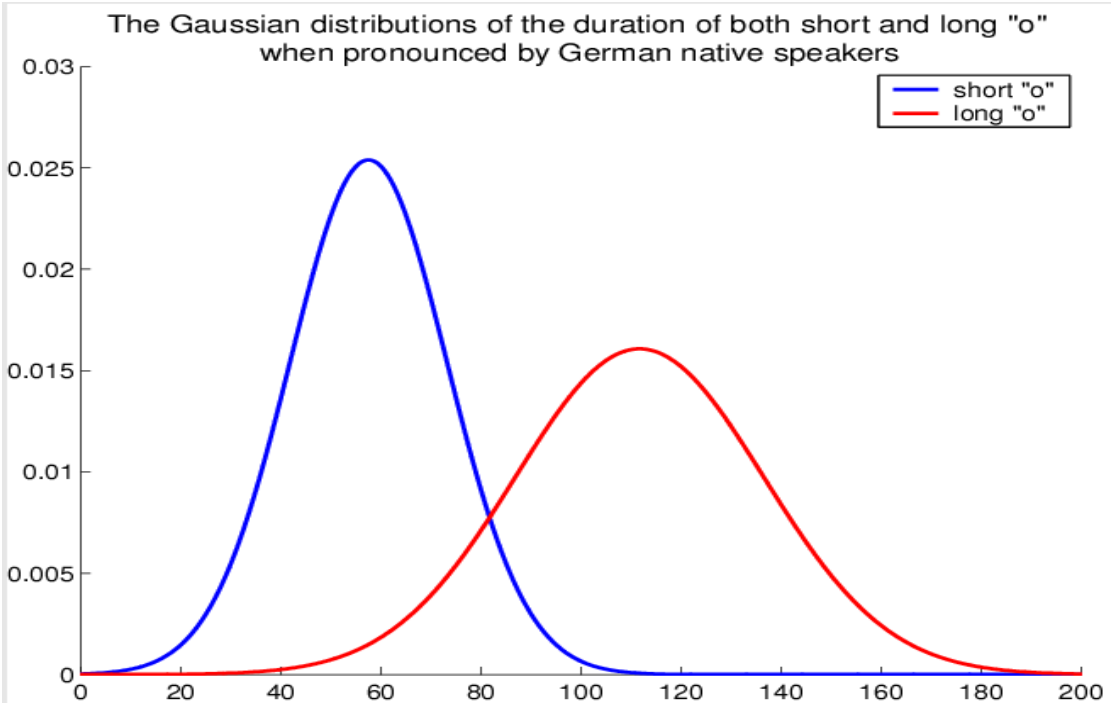


Figure 4: Gaussian distributions of the duration of both short and long German "o" when pronounced by Greek native speakers.

Analysing the distribution, it can be deduced that the data containing the long vowels are more spread apart from the mean than the data containing the short vowels. This means that long vowels tend to take a greater range of duration values. On the other hand, there are no big variations in the duration values of the short vowels. Additionally, it is of great significance to mention that the duration of the short vowels overlap with the duration of the long vowels for some values. This can create an indecisiveness for the recogniser when it should be decided which

duration values are representative for short and which are representative for long German vowels in order to give a definition of a short and a long German vowel respectively. This issue will be discussed and clarified later on. Given that I wanted to make sure that there are indeed significant differences between the short and the long German vowels, once more One-Way Anovas were performed for each vowel comparing the duration of the long to that of the short vowels. All vowels were found to be significantly longer in their long renditions than in their short ones (for example for the vowel o: $p<.001$, $F=204.451$, $df=(118)$ – with the same significance differences for all vowels).

Although a direct comparison of the Gaussian distributions of the Greek and German vowels cannot be made due to the fact that the vowels in the Greek data are either stressed or unstressed, whereas the vowels of the German data are either long or short appearing only in stressed syllables, we can observe that there are some interesting similarities and differences between the Greek pairs stressed-unstressed and the German pairs short-long. Comparing the unstressed /o/ in Greek with short /o/ in German, and the stressed /o/ in Greek with the long /o/ in German, it is observed that the values of their means are very close to each other. Additionally, comparing the standard deviation of the unstressed /o/ in Greek with the short /o/ in German, it is observed that the data in both cases deviate in the same way from their mean value.

On the other hand, comparing the standard deviation of the stressed /o/ in Greek with the long /o/ in German, it is observed that they differ considerably. This means that the values which could take a long /o/ in German are more limited than the values which could take a stressed /o/ in Greek. This could make us think that the duration that defines a long vowel in German is more distinct than the duration that could define a stressed vowel in Greek. Another indication for this conclusion could be the fact that, after comparing the stressed-unstressed pair with the short-long pair, it is observed that the means of the stressed-unstressed pair, see Figure 3, are closer to each

other than the means of the short-long pair, Figure 4. The closer the means of the stressed and unstressed vowels, the bigger the overlapping duration values of the stressed and unstressed vowels that are represented by the Gaussian distribution. On the other hand, the distance between the means of the short-long vowels (Figure 4) is greater than the distance between the means of the stressed-unstressed vowels (Figure 3); for this reason, as it can been seen from Figure 4, the overlapping area of the Gaussian distributions of the short-long pair in German is smaller.

Furthermore, based on the statistical analysis of the training data I attempted to detect eventual outliers i.e. data points being further away from their expected values. Detecting the outliers in the training data was important since outliers can be an indication of faulty data. Figures 5 and 6 depict outliers in the set of short and long vowels respectively. As it is clear in Figure 5, the set of data containing short vowels has only two outliers. The first outlier is in the set of data containing the short /a/s which has a much shorter duration that the other short /a/s. The second outlier comes from the data-set containing the short /u/s, which has a much longer duration that expected. On the other hand, the set of training data containing the long vowels has more outliers. The data-set of the long /e, i, u, o/ have either two or three outliers having duration values bigger that expected and only the data-set containing the long /o/ has an outlier which has a much shorter duration than the other long /o/ vowels in the data. In overall, the number of the outliers existing in the training data is small and, however, a small number of outliers is usually expected in normal distributions [23].

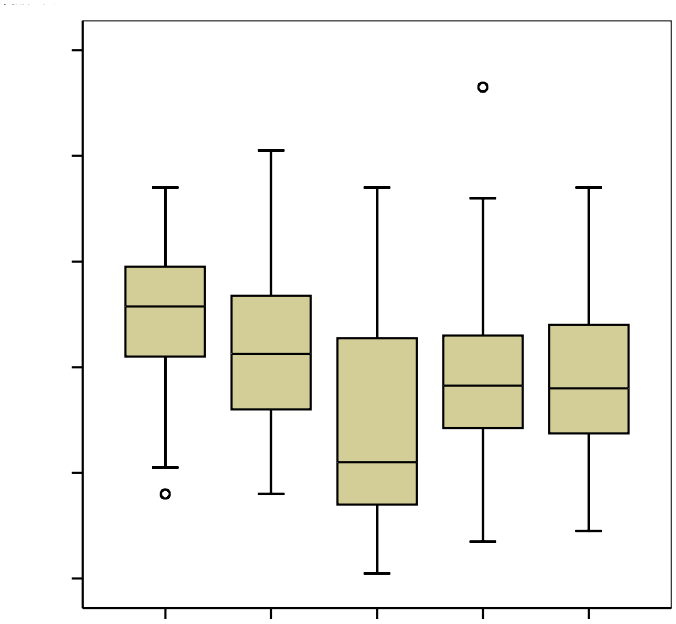


Figure 5: Short vowels

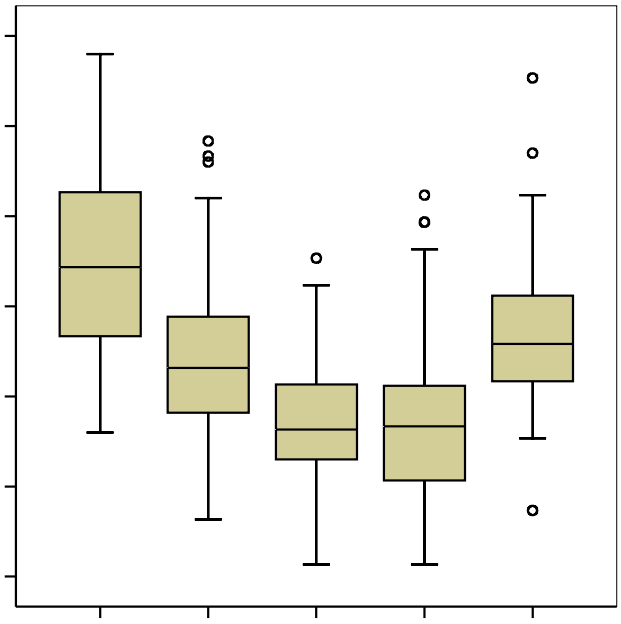


Figure 6: Long vowels

Since the training data were used for training a recogniser capable of predicting differences in the duration of vowels, it is very important that reliable data were employed. In the analysis of the short and long German vowels above it is mentioned that the problematic area of the data was the part where the duration of the short vowels was overlapping with the duration of the long ones. In order to decrease the range of the overlapping values of the duration of the short and long vowels, I omitted from the training data only those outliers which occurred in the overlapping space. This was an attempt to train the recogniser with more reliable data and, furthermore, to create a more robust system since the less the outliers in the data-set, the better the robustness of the speech recogniser [23].

Finally, it is important to point out than none of the values of the duration of the vowels analysed so far were normalised. This could be a sensible attempt since individual features of the speakers (e.g. speaking rate) would be eliminated from the data. The data that I analysed would become more representative for describing the duration of the vowels either in Greek or in German. Definitely, this attempt would be of a great phonetic interest and would come up with more and more detailed phonetic results. However, I preferred not to normalise the values of the data, since this data having these values were used to train the recogniser. Since the analysis of the data occurs in order to come up with results concerning the construction of the recogniser, the data was analysed without normalising their values.

4.3.4 Analysis of the test data

The third part of the analysis concerns the set of the test data. Conventional recognition systems employ a large set of data, one part of which is used as training data, whereas the other part is used for testing the performance of the recogniser. This is not the case of the data employed for the building of this recogniser. As already described, the training data employed is

speech of native speakers containing vowels, the duration of which is considered to be representative for defining the duration of a short and the duration of a long vowel. On the other hand, the set of the test data contain the same German utterances pronounced by Greek speakers of German. This means that the utterances might contain pronunciation mistakes concerning the duration of the vowels or they might be correct. In order to know whether the recogniser predicts the duration of the vowels we must know first which utterances in the test data contain pronunciation mistakes. To achieve this, I should decide to which Gaussian distribution of the short and long German vowels correspond each utterance of the test data. The overlapping duration values of the short and long German vowels made this task more complicated, since it was not clear to which distribution corresponded the vowels of the test data with a duration value belonging to the overlapping area. To overcome this problem a probabilistic method was employed. The equation $z=x-\mu\ /(\sigma-1)$ is applied, where x is the duration of each separate vowel in the test data, that we are trying to find out the Gaussian where is came from, μ is the mean of each Gaussian and σ the standard deviation. First, the distance from the duration value of each vowel of the test data, belonging to the overlapping area, to the mean value is calculated. Then, it is concluded that the vowel belongs to that distribution which is closer to its mean.

It was proved that the short but stressed German vowels were pronounced with a longer duration by the Greek speakers. Moreover, Greek speakers pronounced long German vowels with a shorter duration. However, this mistake was not as frequent as it was expected to be due to the parameter existing in the Greek phonetic system stating that despite the fact that the Greek language has only short vowels, when a vowel exists in a stressed syllable, then it is pronounced with a longer duration.

After it was defined which utterances of the test data contained a pronunciation mistake, the test data were split into two sets: the set containing correctly pronounced vowels and a set

containing vowels pronounced with a wrong duration.

4.4 Summary

This chapter presented the pre-processing of the data used to construct a speech recogniser. Good recording conditions guarantee a high quality of the speech data available for building a speech recogniser. Moreover, labelling the data is proved to be a tricky task, since not only correctness but also consistency should be taken into account while labelling in order to avoid, on the one hand, creating insufficient models when the data are used to train the recogniser and in order to test the performance of the recogniser with accurate test data. Finally, information derived from statistical analysis of the data used to train and test the performance of the recogniser, is necessary in order to obtain deep prior knowledge of the data fed to the recogniser, involving features that the recogniser should be capable and robust enough to predict. The following chapter describes the steps taken to build a speech recogniser employing the already pre-processed available data.

CHAPTER 5

Baseline System

This chapter focuses on the decisions taken during the implementation phase of a recognition system, capable of spotting duration mistakes in vowels made by Greek speakers of German. The system I built is called "ReProGreS" which is the acronym for "Recognising Pronunciation of Greek Speakers" and it is a non-native automatic speech recogniser which takes a speech waveform as its input, and extracts from it feature vectors or observations which represent the information required to perform recognition. The recognition stage is performed using Hidden Markov Models (HMMs). Phone-level acoustic models are formed and they are then combined with a language model. The language model constrains the recogniser to recognise only valid phone sequences. The outcome of the recogniser was stored in a Master Label File (mlf)[2]. In this chapter I will present those processes, as well as the theory on which the recogniser is based and the implementation tools used. Finally, the methodology followed when training the recogniser and the recognition results will be discussed.

5.1 HMM-based recognition

HMM-based automatic speech recognition [5, 6, 7] is a probabilistic method used for speech recognition. In order to take into account speech variability, a stochastic model of a token – a model governed by a set of probabilities – is used. The stochastic models are Hidden Markov Models (HMM) and they represent each token as a sequence of states. Each state emits a set of

2 The mlf files contain the time needed for the recognition of the input speech, the phone which was recognised and the log probability associated to it.

feature vectors which contain useful information to represent the token that is to be recognised. The model either moves from state to state at regular intervals of time, or it is allowed to stay at the same state for successive frames of time. Figure 7 shows an example of an HMM.

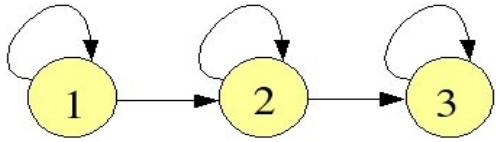


Figure 7: Three state Hmm model

In order for an HMM to represent a token, some probabilities are taken into account: these are, first, the transition probabilities between each state in the model and its permitted successors, and, second, the probability distribution defining the expected observed features for each state.

When a model is activated, a set of feature vectors is emitted in the same way as might be observed when a word is spoken. Observing the feature vectors does not actually determine the sequence of states activated that emitted the feature vectors. The sequence of states is determined by the transition probabilities and it is not influenced by the state where the model is at a particular frame of time. This means that the states themselves are hidden from the observer and this is the reason that the model is characterised “Hidden Markov Model”. However, it is possible to calculate the probability associated with the most probable sequence of states using the Viterbi algorithm (reference) which is a dynamic programming algorithm applied to probabilities.

As mentioned above, when a model of a particular token is activated, it emits a set of feature vectors which contain observations to represent the token that is to be recognised. The best matching token in speech recognition is the one whose model is most likely to produce the observed feature vectors. In order to achieve the best matching between a model and a token some probabilities must be calculated. Particularly, the probability that has to be calculated is the a posteriori probability P (t | Y) of the token t having been uttered during the emission of a set of

feature vectors Y. To calculate this probability the Bayes' rule is applied which breaks this probability down into a likelihood and an a priori probability as follows:

$$P(t \mid Y) = P(Y \mid t)\, P(t) / P(Y)$$

This equation states that in order to calculate the probability of the token given the observations we can use the model for the token t to calculate the likelihood P (Y | t) which is the probability of the observations P given the token t and then multiply it by the a priori probability of the token P(t). Then, this is divided by the probability of the observations P(Y). Actually, the observation likelihood P (Y | t) is the acoustic model whereas the a prior probability P (t) is the language model. The significance of the acoustic and language models is addressed in the Chapter 7 (*Improvements*).

The recogniser I built employs HMM models and is based on the theory described in this section. In the following sections I present the implementation of building the recogniser and its components are briefly described. Additionally, the tools employed for the construction of the recogniser are addressed.

5.2 HTK

ReProGreS is an HMM-based speech recogniser. In order to build the recogniser I used the third version of the Hidden Markov Model Toolkit (HTK) constructed in Cambridge University Engineering Department, which is used for creating and operating Hidden Markov Models. HTK is mainly used for building recognition systems and contains a great number of tools which provide speech analysis, HMMs training, testing and results analysis. The only reference to the HTK documentation is the toolkit itself and the HTK Book [24].

5.3 Components

"ReProGreS" consists of some directories that were used either while training the recogniser or during the recognition task. A list of the various components of the system with a short description of their functionally, can be found in the following table (Table 10).

Components of "ReProGreS"	*Description of the components*
Recordings	This directory contains all the recordings for both training and test data in wav format files.
Label files	In this directory the label files of the recorded speech are stored containing each phoneme with its start and end time.
Resources	Under this directory, the dictionary which holds a list of the required phonemes including "junk" and their pronunciation is defined – covering both training and test data – and, additionally, a list containing the required phonemes is stored. Furthermore, the configuration file, which contains information concerning parameters and their values for customising the HTK working environment, is included.
Models	This directory contains the HMM models, i.e. hmm0 which were created during the initialisation of the models and the hmm1 which were created while training the models. Additionally, a directory with prototypes of HMMs states (until 9 states) is included.
Grammar	This directory contains a grammar of the expected speech input in the recognition stage. HTK provides a grammar definition language, which fulfils this task. For ReProGreS, several grammars were created in order to find out which one would guarantee better recognition results. Moreover, using the HTK tool HParse, the grammar was converted into a word network containing phones and transitions according to the grammar.
Recognition output	Under this directory recognised tokens are stored.

Table 10: Components of "ReProGreS"

After presenting the basic components of the recogniser, the training procedure of the HMMs follows.

5.4 Training HMMs

HMM training was accomplished through the standard Baum-Welch training procedure [5, 7]. For the baseline system, HMM training was carried out exploiting the data of the German speakers mentioned in section 4. The training procedure is described on Figure 8:

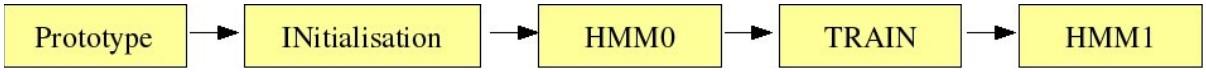


Figure 8: Training procedure

5.4.1 Initialisation

In order to store the speech files in the parameterised Mel Frequency Cepstral Coefficient (MFCC) form, the models were initialised by using the HTK tool HInit [24]. During the initialisation, Figure 9, the initial parameters for the models were calculated, such as the MFCCs features and the HMM states. In this stage, HMMs with three active states were used.

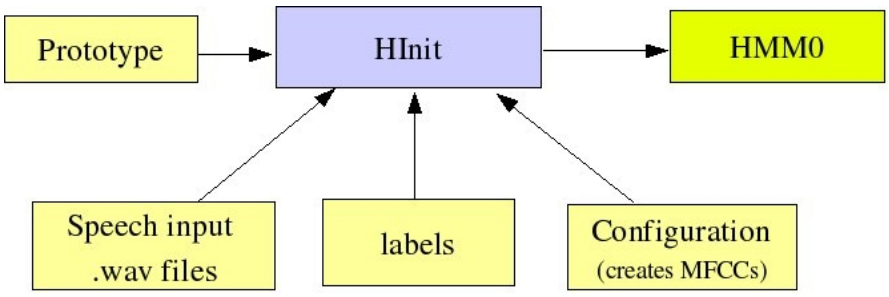


Figure 9: Initialisation of the models

5.4.2 Training

Furthermore, the models were trained using the HRest tool [24], which uses the Baum-Welch algorithm for the optimisation of the parameters for the HMM models with respect to the training data. This means that the weight of the probabilities for each state was calculated. In Figure 10 a re-estimation iteration is depicted.

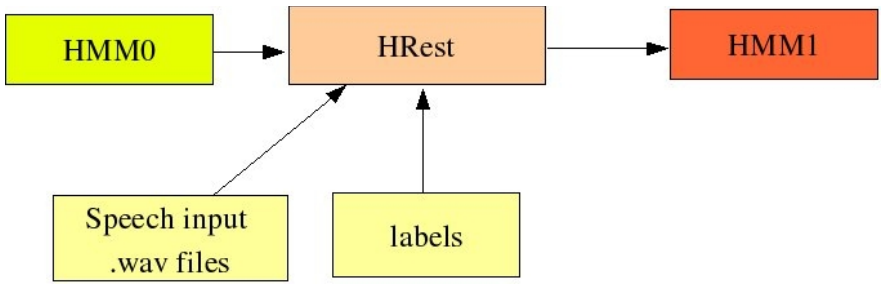


Figure 10: Re-estimation of the models

5.5 Summary

In this chapter the implementation of building a speech recogniser employing HMMs is described and the components of the recogniser were presented. It was also focused on the training procedure by explaining briefly the algorithms involved. After "ReProGreS" was trained, it was tested by following different testing strategies. The testing procedure and the results of the recognition performance are presented in the following chapter.

CHAPTER 6

Testing

The testing phase of the speech recogniser which was developed is crucial, since it provides the points in which the system results in erroneous behaviour. In this section, the testing approaches used in order to identify any operation flows in the implemented system, will be discussed. Moreover, the results occurring after each testing approach will be presented.

After the models were trained, they were tested using the HVite tool of the HTK, Figure 9, which employs the Viterbi algorithm. Viterbi algorithm computes the most likely sequence of hidden states in each HMM given a sequence of observed events [5, 7].

The system was, first, tested under different conditions. I decided to apply a number of different testing approaches in order to identify which are the best conditions that would contribute to a satisfactory recognition outcome

Before using the trained HMMs to perform recognition, the basic architecture of the recogniser was created. For this reason, the grammar that the recogniser would employ to perform recognition was defined. The grammar built at this initial stage of the testing procedure had the form displayed in Table 11.

```
   $vowel = a_l | e_l | i_l |o_l | u_l |
a_s | e_s | i_s | o_s | u_s;
   (junk (t $vowel l) junk) |
   (junk (z $vowel g) junk)|
   (junk (k $vowel m) junk)|
   (junk (t $vowel g) junk)|
   (junk (r $vowel t) junk)|
   (junk (v $vowel l) junk)|
   (junk (v $vowel g) junk)|
   ...
```

Table 11: Part of the grammar used in the initial stage of the testing procedure

According to this grammar, the input utterance is expected to start and end with a "junk". Between the junks all contexts and all vowels are possible to occur. The recogniser has to decide which vowel and which context corresponds better to the input specification

6.1 Testing strategies

The testing phase took place by using the set of test data as a whole, including utterances containing a pronunciation mistake and utterances that contained no pronunciation mistake, since my first concern was to find out whether the recogniser was capable enough to predict differences in the duration of vowels in overall.

6.1.1 Testing the recogniser with test utterances containing the junk parts

Before testing the recogniser with utterances containing the phones of interest, including junks, I was expecting that the recogniser would be capable of identifying the sequence of phones defined by the grammar.

After testing each utterance, the output mlf file was compared with the label file of the test utterance. During the analysis, I realised that the recogniser could not make the correct time alignment. For this reason the outcome of the recogniser cannot be considered accurate since the recogniser was looking at the wrong place of the test utterance when it was trying to recognise the vowel.

This drawback of the recogniser is the reason why I do not display any recognition results of this testing attempt.

To overcome this problem and be able to focus on my experimental question, whether a recogniser can capture pronunciation mistakes concerning the duration of vowels, the recogniser was fed only with the phonemes that it should recognise omitting the junk parts. I attempted this because of the assumption that the bad quality of the junk models might influence the ability of the recogniser to do the correct time alignment.

6.1.2 Testing the recogniser with a sequence of phones without junk parts

The recogniser was fed with utterances containing only the sequence of phones of interest, keeping the grammar the same. The results of the recognition performance were:

	Accuracy
Set of test data using a "loose" grammar	20.55%

Table 12: Accuracy results when recognising with 3 state models and a "loose" grammar.

The results were low and this could be due to the way that the grammar was implemented. Different structures of the grammar signify a different anticipation of the input expected by the recogniser. In order to test this hypothesis, the recogniser was tested employing more restrictive structures of grammar in order to perform better recognition of the input speech.

In order to make the grammar more restrictive towards the expected input speech, several grammars were created, one for each utterance of the test data. Table 13 depicts an example of the structure of the grammar of one utterance in the test data.

```
    $vowel =  a_l | a_s | e_l | e_s | i_l | i_s | o_l | o_s | u_l |
u_s;
    (m $vowel n)
```

Table 13: The grammar of one utterance of the test data, allowing for all vowels.

This grammar states that the input utterance is expected to start with an /m/, then a vowel follows, and then the consonant /n/. The vowel is defined as a variable and can have several possible values. The values of the vowel can be all short and long vowels defined above.

Additionally, an even more restrictive structure of grammar (Table 14) was used to test the performance of the recogniser. Instead of expecting all possible vowels, the recogniser was expecting one specific vowel being either long or short. The probability of finding the correct vowel was expected to increased. An example of this grammar is depicted in the following table:

```
$vowel =  a_l | a_s;
(m $vowel n)
```

Table 14: The grammar of one utterance of the test data, allowing either of a long or of a short vowel only.

As shown in table 8, the recognition results were improved considerably. When using the grammar that allows for all possible vowels the recognition performance increased. Moreover, when using the grammar that allows for one specific vowel and the recogniser had to predict only whether it was short or long, a significant increase of the recognition performance occurred.

Grammar	***Accuracy***
allowing for all vowels	71.33%
allowing for a short or long specific vowel	74,67%

Table 15: Recognition performance with two different restrictive grammars

6.2 Results analysis

It is shown that when the input speech contained parts of junk, the speech recogniser could not do the correct time alignment and it failed to recognise the vowel of interest. Time alignment is proved to be a very hard but significant task for the recogniser. In this case, bad alignment occurred due to the fact that the junk models were of a very bad quality. Junk models were trained on any kind of speech appearing on the left and right hand side of the sequence of phones of interest. To overcome this problem, it would be recommended that instead of labelling the speech around the phones of interest as junk and create one model, to segment that speech into words or ideally into phones and, then, train on more but better models. Furthermore, the language model represented by the grammar has a significant influence on the recognition accuracy. The more

restrictive the language model, the better the recognition results. However, the overall performance of the recogniser was satisfactory, indicating that the recogniser was sensitive enough to capture differences in the duration of vowels when recognising.

6.2 Summary

In this chapter I focused on the testing phase of the recogniser, first, by explaining briefly the algorithms involved and, furthermore, by presenting the results derived after applying each testing strategy. The strategies I used focused either on the form of the input utterances fed to the recogniser or on the language model employed while testing. It was proved that the language model itself would not guarantee adequate recognition results, when pronunciation mistakes concerning duration of vowels are to be detected and recognised. However, the quality of the acoustic models are of great significance since they not only guarantee capability of the recogniser to predict correctly the duration of a vowel but also to do correct time alignment and identify the vowel of interest in a whole utterance. The following chapter focuses on finding other techniques to improve the performance of the recogniser. Instead of focusing exclusively on the language model, the improvement of the acoustic models will be my first concern.

CHAPTER 7

Improvement

"ReProGreS" was tested employing different language models but its performance was problematic since it could not make an efficient time alignment of the models to the testing utterances. This procedure is shown in the previous chapter. It was assumed that in order to succeed a better time alignment and, in effect, better recognition performance, the acoustic models should be improved. This was made by changing some conditions when training. This chapter describes the implementation of improving the recogniser and discusses the decisions made for every improvement step. Finally the results are presented.

7.1 Models with different numbers of states

Until now I have trained 3 state HMMs. Since the recogniser should detect differences in the duration of the vowels I assumed that the number of states could contribute to a better recognition performance. It is already mentioned in chapter 5 that each state of the models emits a set of feature vectors. Additionally, each state has transition probabilities which are the probabilities of moving from this state to a new state. Observing the feature vectors does not actually determine the sequence of states activated that emitted the feature vectors. The transition probabilities describe the linear order in which we expect the states to occur [5, 6, 7]. The model thus generates two strings of information. One is the underlying state path (the labels), as we transition from state to state. The other is the observed sequences (the waveform), each feature being emitted from one state in the state path. The more the states of the model of the long vowels the more the information that the model can generate in order to find the best match with the incoming speech (testing utterance). Additionally, it is expected that a long vowel in the incoming speech will carry

more information that a short vowel. When the HMMs can generate this information then it is more likely to match the incoming utterance with the models correctly. The same is the case with the junk models. In my data the junk models contain a great amount of information which could represent any sequence of phonemes or words. Since the junk model does not contain specific features it is considered to be better if it has more states which makes it capable of generating more feature vectors when trying to match the incoming speech.

7.1.1 Training and testing

I initialised and trained 9 state junk models since this model is considered to carry too much information, 7 state models for long vowels and 5 state models for all the rest phonemes. After testing with the first set of utterances containing junk parts, the time alignment was still a problem. Due to this, I do not display any recognition results since they are considered to be inaccurate. It is assumed that in order to overcome the time alignment problem a stronger training algorithm should be employed. The following section describes the further improvement done to the recognition system. However, when testing with the test utterances which did not contain junk parts the results were much better. The test data set was fed to the recogniser split in two sets of utterances, one set of utterances containing only correctly pronounced vowels by Greek speakers and the other set of utterances containing only mispronounced vowels. Then, the performance of the recogniser was tested on both sets together. It was expected that the recogniser would perform better when recognising utterances including only correctly pronounced vowels since the recogniser contains models trained on those vowels. On the other hand, it was assumed that its performance would be lower when it should recognise utterances containing a different in duration vowel than the one it was expecting. Table presents the results after testing, using different types of grammars.

Set of test data	*grammar allowing all vowels*	*grammar allowing only for a short or long specific vowel*
set of test data containing utterances without pronunciation mistakes	76.86%	84.57%
set of test data containing utterances with mispronounced vowels	71.11%	82.78%
both data set together	71.67%	85.00%

Table 16: Recognition performance with two different restrictive grammars

7.2 Embedded training

After training HMMs with a different number of states, I attempted to apply a different training algorithm, in order to find out whether that would contribute to a further improvement of the recogniser. Embedded training of the models was expected to bring even better recognition results.

In the training procedure I applied so far, a training set of speech signals of speech which were already manually labelled with correct phonetic labels, were used to set the parameters of the recogniser. When performing recognition, the labels were assigned to the new input signal. It is, though, possible that the boundaries of the labels that I set manually are not the same with the boundaries the recogniser will set. This could have only negative effects on the recognition procedure. In order to avoid this mismatch and, in effect, the negative consequences when recognising speech, embedded training [24] is applied. In embedded training the recogniser re-labels the training data and the training takes place. Then the new models were re-trained starting

from the relabelled training data, and this carries on for a few iterations. Embedded training was combined with adding Gaussian Mixture components [5, 25]. I assumed that the more Gaussian components the better the recognition accuracy, bearing always though in mind not to add too many components and overfit the models the training data. After a few iterations of embedded training, two Gaussian components were added while training. I stopped training after 7 iterations of embedded training and after having trained with with eight Gaussians. The recogniser was tested using both types of grammars; the grammar that allows for all vowels and the grammar that allows for either a short or long specific vowel. The results (Table 17) were interesting.

The test data set was fed to the recogniser split in two sets of utterances, one set containing only correctly pronounced vowels by the Greek speakers and the other set containing only mispronounced vowels. Then, the performance of the recogniser was tested on both sets together. It was expected that the recogniser would perform better when recognising utterances including only correctly pronounced vowels since the recogniser contains models trained on those vowels. On the other hand, its performance was lower when it should recognise utterances containing a different in duration vowel than the one it was expecting.

Set of test data	*grammar allowing all vowels*	*grammar allowing only for a short or long specific vowel*
set of test data containing utterances without pronunciation mistakes	73.78%	80.58%
set of test data containing utterances with mispronounced vowels	70.94%	82.77%
both data set together	69.56%	78.33%

Table 17: Recognition performance with two different restrictive grammars, using embedded training

7.3 Results analysis

It can be observed that the recognition performance was slightly increased when training models with different states. The assumption was verified that more states would improve the recognition performance. However, the recognition performance dropped slightly when applying embedded training and adding the Gaussian components. My assumption that this training algorithm would improve the recognition performance was not proven. My speculation is that the worse performance of the recogniser when applying embedded training is due to the limited number of training data I had available. For such techniques a greater number of training data is required in order to achieve a better clustering of the data, where each cluster has adequate parameters to use for recognition. It has become evident throughout this dissertation that each system needs to find its balance between having too few and too many data. Although it is desirable to build a recogniser with as few training data as possible, it is still important that we have enough data for a good accuracy result. Furthermore, different techniques might require a

different amount of data. These are all questions that a researcher in the field is called to answer.

Conclusions

One of the major goals of this dissertation was to develop a speech recogniser capable of predicting pronunciation mistakes of Greek speakers of German concerning the duration of vowels. Such a recogniser is intended to serve educational purposes by being integrated in CAPT systems. The recogniser was maintained with respect to its sensitivity against predicting mistakes concerning the duration of vowels. The development of both the recogniser and its research report, consisting the delivered dissertation, was indeed a rewarding experience. Through the research done in order to complete this dissertation, it was proved that building a recogniser is not an easy task. Constructing a recogniser for being integrated in CAPT systems and serve educational purposes consists of various tasks that require the co-operation of linguists, speech technologists and teachers.

During the implementation of the system it was found out that there are some tricky parts that should be taken into account in order for the recogniser to fulfill the expected requirements. Conventional recognisers focus mainly on the language model in order to achieve successful results. It was, though, proved in this thesis, that a recogniser whose further goal is the pronunciation training, should contain sufficient acoustic models in order to be capable of predicting the potential pronunciation mistakes with high confidence. A good acoustic model could be the result of the use of a sufficient training algorithm, good quality of training data combined with a sufficient number of training data. A balance of all these should be achieved, in order for the recogniser to fulfil expected goals and requirements that guarantee high recognition results, and at the same time all these to be attained with a low computational cost.

In conclusion, it is essential to mention that whether a recogniser constructed for accomplishing educational purposes could be characterised as ideal and capable of fulfilling pedagogical requirements and have successful results, depends not only on the recogniser as such

but also on other factors, such as the goals set when teaching and learning pronunciation of L2 by making use of CAPT systems, the way the recogniser is introduced in class, e.g. as a tool or as an instructor, or even the familiarity of the learners with employing technological media for learning and practising pronunciation.

8.1 Future working guidelines

Concerning the recogniser implemented in this thesis, it should be mentioned that several points might be subject of future work. One of the extensions concerning the training procedure could be the creation of more models describing the speech around the phones that are expected to be recognised. This would eventually lead to better time alignment. Additionally, a greater number of data than those available for training ReProGreS would contribute into creating more sufficient models when training. It would also be very interesting to train the recogniser with Greek data containing speech of Greek native speakers. The recogniser could contain models not only of the long and short German /e/ but also of the stressed and unstressed Greek one. In this way it is assumed that under the appropriate conditions i.e. adequate training, good acoustic models, etc., the recogniser would not only be capable of spotting duration differences of vowels but additionally, it could predict differences in the quality of vowels being very close to formant space in two languages. Another suggestion for further improvement of the recogniser could refer not only to its capability of predicting successfully utterances containing pronunciation mistakes, but also after recognising the mispronounced vowel of the utterance, to be capable of returning the correct vowel to the user.

As future work guidelines, we should not focus only on the improvement of the implementation procedure of ASR technology and CAPT systems alone. It is of vital importance that teachers and second language learners are taught how to use such systems at school, so that

such systems are actually used as tools for obtaining a good pronunciation in a foreign language and , in effect, attaining a fluent communication.

References

[1] Eskenazi, M and Hansma, S. (1998) The Fluency Pronunciation Trainer. *Proc. Speech Technology in Language Learning 1998*, Marholmen, Sweden, May 1998.

[2] Flege, J.E. (1987). The Production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics*, 15, 47-65.

[3] Neri, A; Cucchiarini, C; Strik, W: A Automatic Speech Recognition for second language learning: How and why actually works. Proceedings from the 15th ICPhS, Barcelona, Spain, 1157-1160.

[4] Crystal, D. (1981). *Clinical linguistics*. New York: Harper Press.

[5] Holmes, J. & W. (2001). *Speech Synthesis and Recognition.* Second Edition. London and New York: Taylor and Francis.

[6] King, S. (2005). *Speech Processing*. University of Edinburgh.

[7] Jurafsky, D. and Martin, J.(2000). *Speech and Language Processing,* Upper Saddle River, NJ, USA: Prentice-Hall.

[8] Eskenazi, M. (1999) Using Automatic Speech Processing for Foreign Language Pronunciation Tutoring: Some Issues and a Prototype. *Language Learning anad Technology.* 2 (2), 62-72.

[9] Kim, I,-S. (2006) Automatic Speech Recognition: Reliability and Pedagogical Implication for Teaching Pronunciation. *Education Technology & Society*, 9 (1), 322-334.

[10] Herron, D; Menzel, W; Atwell, E; Bisiani, R; Daneluzzi, F; Morton, R; Schmidt, J. A. (1999). Automatic localization and diagnosis of pronunciation errors for second-language learners

of English. *Paper presented at the 6th European Conference on Speech Communication and Technology*, September 5-9, 1999, Budapest, Hungary.

[11] Hinks, R. (2001) Using speech recognition to evaluate skills in spoken English. *Working Papers*, 49. Lund University Department of Linguistics, 58-61.

[12] Lee, K; Hon, H; Reddy, R. (1990). AN Overview of the SPHINX Speech Recognition System. *IEEE Transaction of Acoustics Speech and Signal Processing*. 38 (1)

[13] Dalby, J and Kewley-Port, D. (1999). Explicit pronunciation training using automatic speech recognition. CALICO, 16 (3), 425-445.

[14] Chen, H. (2001) Evaluating five speech recognition programs for ESL learners. Paper presented at ITMELT'2001, Hong-Kong, 9 Nov 1999.

[15] Hinks, R. (2002): Speech Recognition for Language Teaching and Evaluating: a Study of Existing Commercial Products. Proceedings of ICSLP, 733-736.

[16] Talk to Me (Auralog), http://www.camsoftpartners.co.uk/ttm.htm, URL , last accessed: 01/09/2006

[17] Ehsani, F and Knodt, E. (1998) Speech Technology in Computer-aided Language Learning: Strengths and Limitations of a new call paradigm. *Language Learning and Technology.*2(1), 45-60.

[18] Ladefoged, P. (2006). *A Course in Phonetics*. Fifth Edition. Boston: Thomson Wadsworth.

[19] Balassi, E. (2002). *Phonetik/ Phonologie und Ausspracheschulung.* Band B. ΕΑΠ ΠΑΤΡΑ.

[20] Petrounias, E. Β (2002). Νεοελληνικη Γραμματικη Συγκριτικη αντιπαραθετικη αναλυση. Thessaloniki:Ziti.

[21] Botinis, A.; Bannert, R.; Fourakis, M.; Pagoni-Tetlow, S. (2002). *Crosslinguistic Segmental Durations and Prosodic Typology.* International Conference of Speech Prosody 2002 183-186 Aix-en-Provence, France.

[22] Field, A. (2000). *Discovering Statistics Using SPSS*. Second Editon. London: Sage Publications.

[23] Outliers, http://en.wikipedia.org/wiki/Outliers, URL, last accessed: 01/09/2006

[24] HTK Speech Recognition Tool, http://htk.eng.cam.ac.uk, URL, last accessed: 01/09/2006

[25] L. Rabiner, B. H. Juang (1993). Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs, NJ

Appendices

German Corpus

Short [a] :

Keeping the context the same: /m a n/

Vitamin D-M**a**ngel führt zur "Ausdünnung" der Knochenmasse.

Der M**a**ngel an Lebensmittel war sehr groß.

Nicht der M**a**ngel an Liebe, sondern der Mangel an Freundschaft macht die unglücklichsten Ehen.

Der Ausgeklagte wurde aus M**a**ngel an Beweisen freigesprochen.

Es m**a**ngelt mir an nichts.

In different context:

S**a**lz und Brot macht Wangen rot

Er fuhr so l**a**ngsam,dass er nicht rechtzeitig ankam.

Es ist k**a**lt und ganz Deutschland friert, nachdem die eisigen Temperaturen aus Russland zu uns herüber gezogen sind. In diesem Dorf gab es kein trinkbares Wasser.

In diesem Dorf gab es kein trinkbares W**a**sser.

Man sollte Probleme zwischen G**a**stgeber und Besuchern vermeiden.

Short [e]:

Keeping the context the same: /m e n/

Wenn es um Ihre Vorsorge geht, müssen Sie eine ganze M**e**nge rechnen.

Ich muss noch eine M**e**nge lernen!

Er hat noch eine M**e**nge Arbeit übrig.

Wir haben nur noch eine begrenzte M**e**nge dieser Ware vorhanden.

In den Ferien wird er jede M**e**nge Zeit haben.

In different context:

Im Mittelalter war die N**e**lke ein Symbol für die Gottesmutter Maria.

Ein einfaches M**e**sser besteht aus einer Klinge, einem Griff und dem dazwischen liegenden Handschutz.

Nicht in jeder Verwendung sind die V**e**rben „haben“, „werden“ oder „sein“ Hilfsverben.

Die W**e**llen brechen sich am Felsen.

Solche Diäten sind s**e**lten langfristig erfolgreich.

Short [u]:

Keeping the context the same: /m u n/

Atmen durch die Nase ist vor allem deshalb so gesund, weil man dabei den M**u**nd hält.

Kannst du endlich deinen M**u**nd halten?

Das Kind stand mit dem Finger im M**u**nd da.

Es gibt viele Wörterbücher für diese M**u**ndart.

Warum antwortest du nicht? Hast du etwa deinen M**u**nd verloren?

In different context:

Oft gilt Ged**u**ld als eine Tugend.

Das Geheimnis der K**u**nst liegt darin, dass man nicht sucht, sondern findet.

Sie hatte keine L**u**st auszugehen.

Die L**u**ftverschmutzung ist ein wichtiges Umweltproblem.

Ich habe auf dem B**u**s fast über eine Stunde gewartet.

Short [i]:

Keeping the context the same:

Häufig werden M**i**nderheiten auf Grund von Vorurteilen ausgegrenzt.

Ein Faktor für die Entstehung von M**i**nderheiten ist die Besiedlung eines Landes.

Es gibt unzählige Versuche, M**i**nderheiten zu kategorisieren und typologisieren.

Russlanddeutschen sind die größte Minderheit in Deutschland.

Eine M**i**nderheit sind Menschen, die aufgrund ihrer ethnischen, sozialen oder religiösen Zugehörigkeit oder wegen ihrer sexuellen Orientierung, Diskriminierung erfahren.

In different context:

Meine Schwester hat drei K**i**nder.

S**i**ngen liegt nordwestlich vom Bodensee.

Diese Hose sitzt dir gar n**i**cht.

Gegen Mittag saß die ganze Familie am T**i**sch.

Die Nacht war ganz st**ill**.

Short [o]:

Keeping the context the same:

„Die M**o**nster AG“ ist ein bekannter Zeichentrick-Film.

Im engeren Sinn bezeichnet M**o**nster eine phantastische Kreatur.

Das M**o**nster als fiktive Kreatur hat seinen Ursprung meist in der Phantasie der Menschen, in Alpträumen oder Mythen.

Die Heavy Metal-Band „Lordi“, die stets in M**o**nstermasken und -kostümen auftritt, nahm den ersten Platz.

Seit 40 Jahren jagt Robert Braun das geheimnisvolle M**o**nster der Tiefsee.

In different context:

Es gab fast keine W**o**lken in den blauen Himmel.

Sie war ein freundliches und hübsches Mädchen, tr**o**tzdem liebte er sie nicht.

Der Jäger f**o**lgt dem Wildschwein.

Wieviel k**o**stet dieses Buch?

P**o**tsdam liegt eine halbe Stunde weit von Berlin.

Long [a]:

Keeping the context the same: /t a l/

Sie muss Nerven aus St**a**hl haben!

Wir brauchen noch fünf Meter St**a**hldraht

Die erste bekannte Herstellung von St**a**hl wurde in Europa von Benjamin Huntsmann durchgeführt.

Dieses Messer ist aus St**a**hl.

Heute werden ungefähr 2500 verschiedene St**a**hlsorten hergestellt.

In different context:

So was würde ich dir nie s**a**gen.

Die letzten Gäste k**a**men zu spät.

Sie geht ins Fitness-Studio fast jeden T**a**g.

Vielleicht kannst du mir einen guten R**a**t geben

Ich würde gern mal einen W**a**l sehen.

Long [e]:

Keeping the context the same: /m e n/

Wir st**e**hlen unsere Beute.

Wir müssen das Mädchen st**e**hlen.

Ich st**e**hle ein Stück Kuchen.

Geld st**e**hlen ist gar nicht gut.

Ich konnte nicht das Geld st**e**hlen.

In different context:

W**e**gen eines Herzinfarktes durfte er nicht Tennis spielen.

Ich habe eine Freundin zum T**ee** eingeladen.

Für dieses Rezept werden wir 1kg M**e**hl brauchen.

Viele Leute haben dieses Buch noch nicht gel**e**sen.

Kann ich bitte ein Stück H**e**fekuchen haben?

Long [i]:

Keeping the context the same: /t i l/

Dieser Schriftsteller hat einen eigene St**il**.

Dieser Mensch hat überhaupt kein St**i**lgefühl.

ieser Barockschrank gehört zu den eleganten St**i**lmöbel.

Man muss St**i**l haben, um ihn sich kaufen zu können.

Nur der maßvolle St**i**l ist der klassische.

In different context:

Darf ich Ihnen etwas anb**ie**ten?

Kannst du mir ein Beisp**ie**l geben?

Sie beschäftigt sich jetzt ausschl**ie**ßlich mit ihrer Familie.

Unsere Mannschaft wurde gestern bes**ie**gt.

D**ies**e Tasche gehört ihm.

Long [u]:

Keeping the context the same: /t u l/

Ich möchte so einen St**uh**l kaufen. Er ist sehr bequem.

Das St**uh**lbein ist gebrochen.

Wenn du müde bist, setz dich auf diesen St**uh**l.

Der Mörder wurde durch den elektrischen St**uh**l hingerichtet.

Ich glaube wir werden noch einen St**uh**l brauchen.

In different context:

Er hat einen guten R**u**f als Zahnarzt.

Mein Br**u**der kommt morgen zum Besuch.

Dieser H**u**t steht dir sehr gut.

Ich bin auf der S**u**che nach einer neuen Wohnung.

Beeile dich, sonst wirst du deinen Fl**u**g verpassen.

Long [o]:

Keeping the context the same: /t o l/

Ich habe nie in meinem Leben etwas gest**oh**len.

Bist du sicher, dass dein Koffer gest**oh**len worden ist?

Im Falle eines Diebstahls sollte das Handy als gest**oh**len gemeldet werden

Ein sehr berühmtes Kunstwerk wurde gest**oh**len.

Ein sehr berühmtes Kunstwerk wurde gest**oh**len.

In different context:

Das Auto braucht einen neuen M**o**tor.

Sie hat von ihrer Tante einen V**o**gel als Geschenk bekommen.

Diese Frisur ist nicht mehr in M**o**de.

Das Blech ist im **O**fen.

Ich habe für mein Auto ein sehr gutes Angeb**o**t bekommen.

Greek Corpus

[a] in stressed syllable:

Keeping the context the same: /m a n/

1. Η **μάνα** άναψε τη φωτιά και μαγείρεψε.
2. Όταν έχασε τη **μάνα** του βυθίστηκε στη μελαγχολία.
3. Το απόγευμα η **μάνα** πήγε να τον επισκεφτεί.
4. Η '**μάνα**' της Περλ Μπάκ είναι το αγαπημένο μου βιβλίο.
5. Η **μάνα** είναι αυτή που πονάει περισσότερο τα παιδιά της.

In different context:

26. Για πρωινό πίνω ένα ποτήρι **γάλα**.
27. Το κρεβάτι μου το στρώνω **κάθε** πρωί πριν φύγω για το σχολείο.
28. Μου άρεσαν ανέκαθεν οι διακοπές δίπλα στη **θάλασσα** αν και με τους γονείς μου πηγαίναμε συχνά διακοπές στο βουνό.
29. Όταν πηγαίνουμε διακοπές παίρνω **πάντα** το καλάμι μου μαζί και πάω για ψάρεμα.
30. Κάθε φορά που βρέχει μου αρέσει να κοιτάω έξω από το **παράθυρο** και να χαζεύω τη βροχή που πέφτει.

[o] in stressed syllable.

Keeping the context the same: /m o n/

6. Ήταν **μόνη** στο σπίτι εκείνη τη μέρα.

7. Λυπάται τους ανθρώπους που είναι **μόνοι**. Ολομόναχοι.

8. Η γειτόνισσα μένει **μόνη** εδώ και πολλά χρόνια.

9. Το μάθημα το πέρασαν **μόνο** πέντε μαθητές.

10. Ήταν ο **μόνος** που με βοήθησε στο πρόβλημά μου.

In different context:

31. Η παραλία είχε πολύ **κόσμο**, γι' αυτό προτιμήσαμε να πάμε σε μια πιο ήσυχη.

32. Πιστεύω πως δεν είχε **λόγο** να μου φερθεί τόσο άσχημα.

33. Το σκιουράκι ανέβηκε τον **κορμό** του δέντρου και έφτασε στο πιο ψηλό κλαδί.

34. Το να διατηρήσουμε τις παραλίες καθαρές είναι υποχρέωση **όλων** μας.

35. Τράκαρε με το μηχανάκι και έσπασε το **πόδι** του.

[ε] in stressed syllable:

Keeping the context the same: /m ε n/

11. Στον τέταρτο όροφο στην πολυκατοικία μου **μένει** μια πενταμελής οικογένεια.

12. Στον δεύτερο όροφο **μένει** μια γυναίκα μόνη της και έχασε τον άντρα της πρόσφατα.

13. Στο διπλανό διαμέρισμα δε **μένει** κανείς

14. Μια γειτόνισσα απέναντι **μένει** μόνη.

15. Μερικά βράδια **μένει** μαζί της.

In different context:

36. Το εστιατόριο ήτανε γεμάτο κόσμο και δεν βρήκαμε ούτε **καρέκλα** να κάτσουμε.

37. Έκανε πολλή **ζέστη** τα μεσημέρια και γι' αυτό προτιμούσαμε να μένουμε σπίτι εκείνες τις ώρες.

38. Πήρε από το **χέρι** τη γιαγιά και την πέρασε απέναντι.

39. Κάθε καλοκαίρι παίρνω άδεια για ένα μήνα και πηγαίνω **διακοπές**.

40. Κάθε απόγευμα **παίζω** μπάλα με τους συμμαθητές μου στη γειτονιά.

[I] in stressed syllable.

Keeping the same the context:/ m I n/

16. Το **αλουμίνιο** είναι υλικό που ανακυκλώνεται.

17. Τα κουτάκια από **αλουμίνιο** τα πετάμε σε ειδικούς κάδους.

18. Η δουλειά του είναι να φτιάχνει έργα τέχνης από **αλουμίνιο**.

19. Με πολλά κουτάκια από **αλουμίνιο** κατάφερε να φτιάξει ποδήλατο.

20. Εκτός από το **αλουμίνιο** υπάρχουνε και άλλα ανακυκλώσιμα υλικά.

In different context:

41. Έβαψα τους τοίχους του δωματίου μου **κίτρινους** για να ταιριάζουν με τις πράσινες κουρτίνες.

42. Τα καλοκαίρια ο **ανεμιστήρας** λειτουργεί όλη μέρα αλλά σκοπεύω να πάρω σύντομα κλιματιστικό.

43. Τον ελεύθερό μου χρόνο μου αρέσει να **ζωγραφίζω**.

44. Το **μίσος** του για τους οπαδούς του Παναθηναικού το θεωρώ ανεξήγητο.

45. Έχει βαθιά **πίστη** μιας και κάθε Κυριακή πηγαίνει στην εκκλησία.

[U] in stressed syllables:

21. Το σπίτι είναι γεμάτο με μικρά **μαμούνια**.

22. Κάθε φορά που πηγαίνουμε στο χωριό μας αρέσει να βρίσκουμε **μαμούνια**.

23. Το Λα- **μαμούνια** είναι μεγάλο κλαμπ στην Αθήνα.

24. Η μητέρα μου όποτε δει **μαμούνι** το σκοτώνει.

25. Εγώ, όμως, βρίσκω τα **μαμούνια** πολύ χαριτωμένα και για αυτό τα φυλάω σε γυάλες.

In different context:

46. Είναι τόσο ιδιότροπος άνθρωπος που δεν είναι περίεργο το γεγονός ότι έχει μείνει **μπακούρι**.

47. Στον κήπο του έχει κάθε λογής λαχανικά όπως ντομάτες, **αγγούρια**, πιπεριές.

48. Όποτε κάνει πολύ κρύο τρώει μια ζεστή **σούπα** για να νιώσει καλύτερα.

49. Εξαιτίας της μετακόμισης το σπίτι ήτανε γεμάτο **κούπες.**

50. Δυστυχώς χάλασε η ηλεκτρική **σκούπα** και πρέπει να πάρουμε καινούρια.

[U] in unstressed syllables:

Keeping the context the same:

71. Υπολόγιζε πως θα ξόδευε **τουλάχιστον** 300 Ευρώ εκείνη την εβδομάδα.

72. Θα μπορούσε **τουλάχιστον** να μου τηλεφωνήσει.

73.Πρέπει να γράψω **τουλάχιστον** 1 0 για να περάσω το μάθημα.

74. **Τουλάχιστον** πάρε αυτή τη ζακετούλα μαζί σου! Κάνει κρύο!

75.Πρέπει να ήπιαμε **τουλάχιστον** 30 μπύρες εκείνο το βράδυ.

In different context:

96. Καθημερινά **δουλεύω** μόνο 4 ώρες.

97. Έφερε ένα **κουτί** γεμάτο γλυκά.

98. Η γιαγιά μας φτιάχνει πάντα **λουκουμάδες** όποτε την επισκεφτόμαστε.

99. Οι **κουραμπιέδες** είναι το αγαπημένο μου χριστουγεννιάτικο γλυκό.

100. Τα **σουτζουκάκια** της μητέρας μου δεν συγκρίνονται με τίποτα άλλο.

[ε] in unstressed syllable:

61. Η δίαιτα που μου έδωσε ο διαιτολόγος ήταν **αποτελεσματική**.

62. Το απορρυπαντικό είχε **αποτελεσματική** δράση κατά των λεκέδων.

63. Τα φάρμακα που μου έδωσε ο γιατρός ήταν **αποτελεσματικά**.

64. Δεν ήταν **αποτελεσματικός** ο τρόπος που αντιμετώπισε το πρόβλημά του.

65. **Αποτελεσματική** θεραπεία θα έχεις μόνο αν ακολουθήσεις τις συμβουλές μου.

In different context:

81. Θέλω να πάω στην **έκθεση** βιβλίου το απόγευμα.

82. Πάρε **μερικά** ζεστά ρούχα μαζί σου σε περίπτωση που θα χιονίσει.

87. Πρέπει να πάρετε ένα **τετράδιο** τριών θεμάτων.

89. Πρέπει να πάω στην **τράπεζα** να βγάλω χρήματα.

90. Πιες ένα **ζεστό** τσάι και θα νιώσεις αμέσως καλύτερα.

[I] in unstressed syllable:

Keeping the context the same: /t I l/

56. Πάρε με **τηλέφωνο** το απόγευμα.

57. Το **τηλέφωνο** δε σταμάτησε να χτυπά όλο το πρωί.

58.Είναι όλη μέρα με το **τηλέφωνο** στο χέρι και μιλάει.

59. Ξέχασα να πληρώσω το λογαριασμό και σήμερα μου έκοψαν το **τηλέφωνο**.

60.Δε με έχει πάρει **τηλέφωνο** από τον Ιανουάριο.

In different context:

82. Πάντα **συμβουλεύομαι** το λεξικό όταν γράφω στα Γερμανικά.

83. Ο **εκτυπωτής** μου δεν έχει μελάνι.

84. Ξάπλωσε στην **πολυθρόνα** και παρακολούθησε τον αγώνα.

85. Μα γιατί δεν φέρνεις ποτέ **στυλό** στο μάθημα;

88. Το κασετόφωνο χάλασε και **μάσησε** την κασέτα.

[a] in unstressed syllables:

Keeping the context the same:/ t a l/

51. Όλοι τον αποκάλεσαν **ταλέντο** όταν είδαν τις επιδόσεις του.

52.Είχε **ταλέντο** στη ζωγραφική και γι' αυτό γράφτηκε στη σχολή καλών τεχνών.

53. Με **ταλέντο** στη ζωγραφική γεννιέσαι, δεν γίνεσαι.

54.Ο διαγωνισμός αφορούσε **ταλέντα** στο τραγούδι.

55. Η αγάπη και το **ταλέντο** του για τη μουσική έβαλε το σχολείο του στη δεύτερη θέση των προτεραιοτήτων.

In different context:

76. **Μακάρι** να έχει καλό καιρό αύριο.

77. Πάρε μια **καρέκλα** και κάτσε μαζί μας.

78.**Αγόρασε** καινούργια παπούτσια και τελικά το μετάνιωσε.

79. Να νοικιάσουμε μια **βιντεοκασέτα** να δούμε το βράδυ.

80. Το **καλοκαίρι** είναι η αγαπημένη μου εποχή.

[o] in unstressed syllables:

Keeping the context the same:

66.Η αποκριάτικη **στολή** του κέρδισε τις εντυπώσεις στο πάρτυ.

67.Άφησε τη **στολή** στο καθαριστήριο ελπίζοντας να την πάρει ως την Πέμπτη.

68.Η **στολή** που φορούν στο πεζικό έχει διαφορετικό χρώμα από αυτήν της αεροπορίας.

69. **Στολίστηκε** και βγήκε με τις φίλες της.

70. Τα Χριστούγεννα **στολίζουμε** όλοι μαζί το δέντρο.

In different context:

91. Τα σώματα του καλοριφέρ είναι καινούργια, τα αλλάξαμε το **φθινόπωρο**.

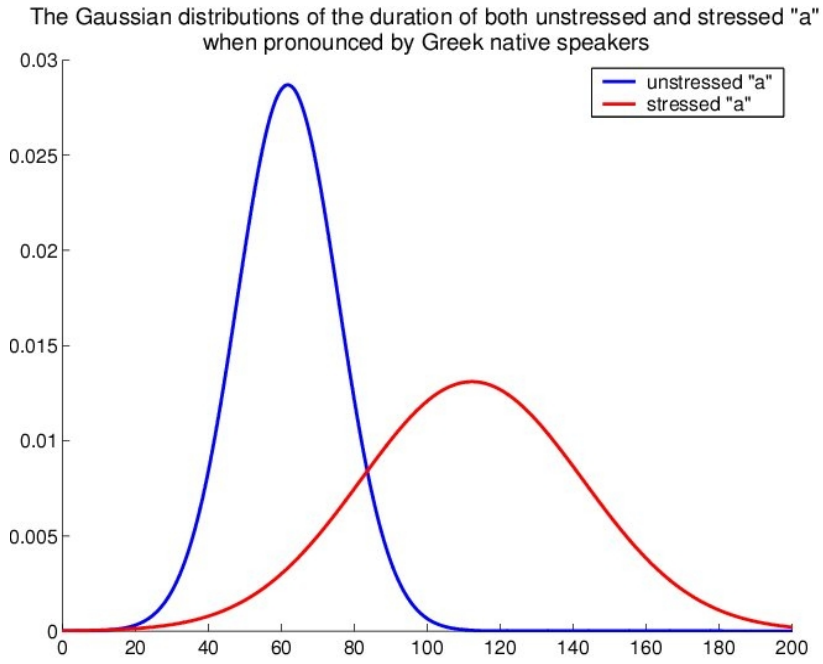
92. Τα παιδιά του είναι γεμάτα **ζωντάνια** και ενέργεια.

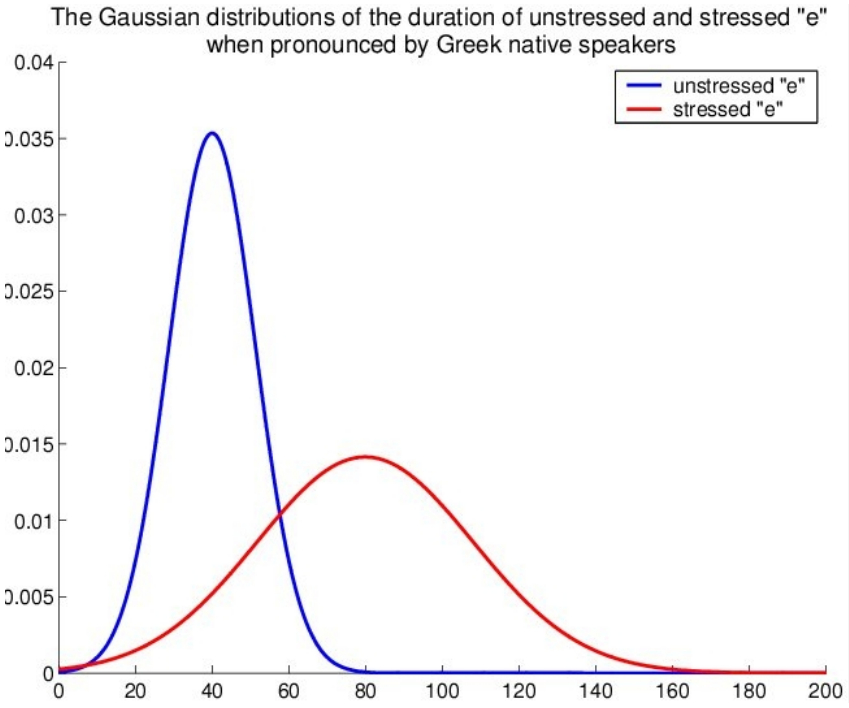
93.Το καλοκαίρι πηγαίνουμε στο **εξοχικό** μας.

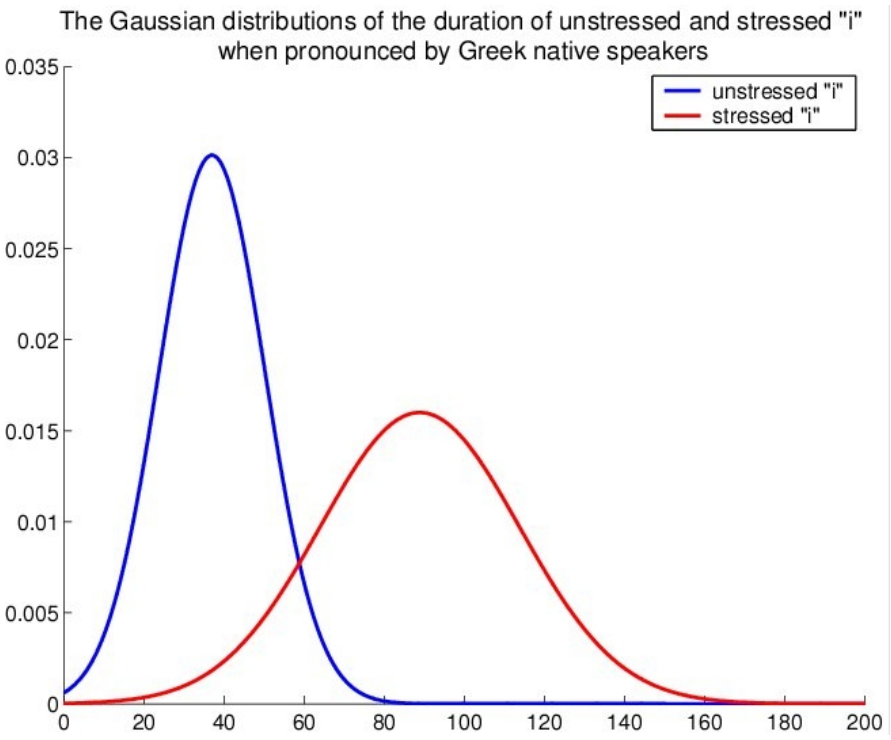
94. Ξέχασα το φορτιστή σπίτι και ήμουν όλη τη **βδομάδα** χωρίς κινητό.

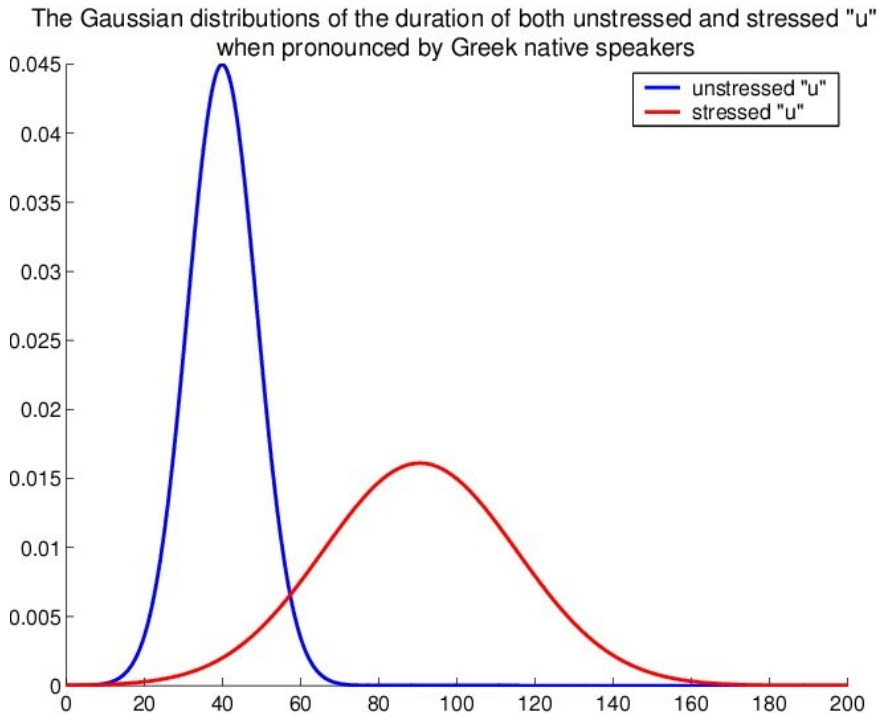
95. Το **μωρό** κλαίει κάθε βράδυ και δεν μπορούμε να κοιμηθούμε.

Gaussian distribution of the duration of stressed and unstressed vowels in Greek.

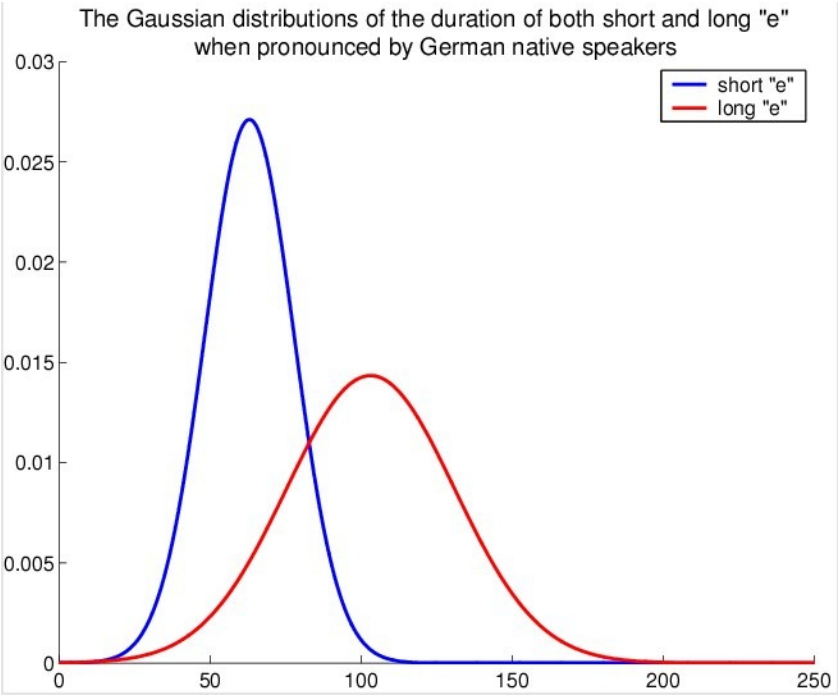


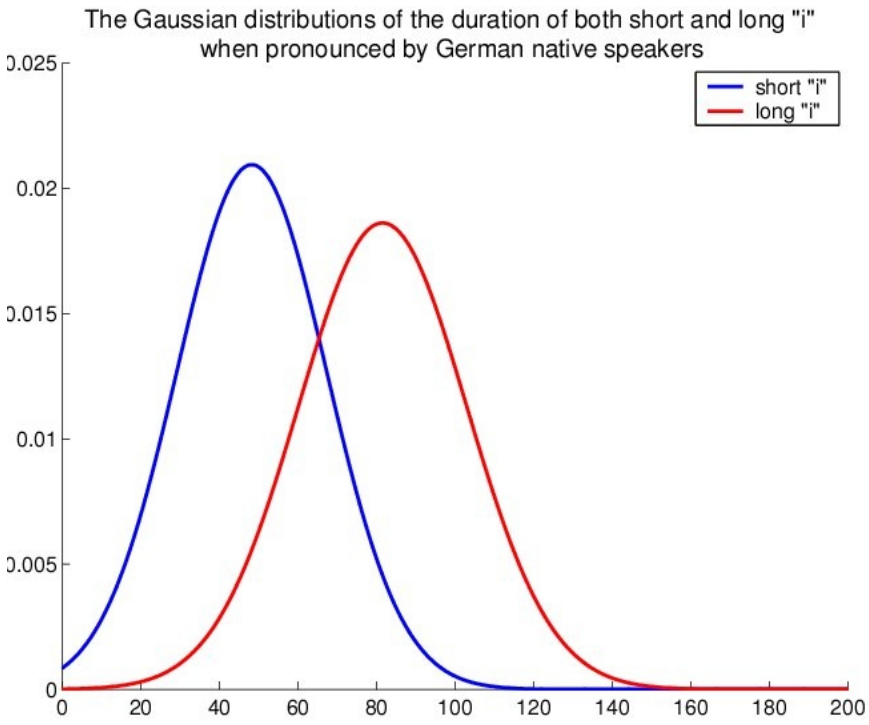


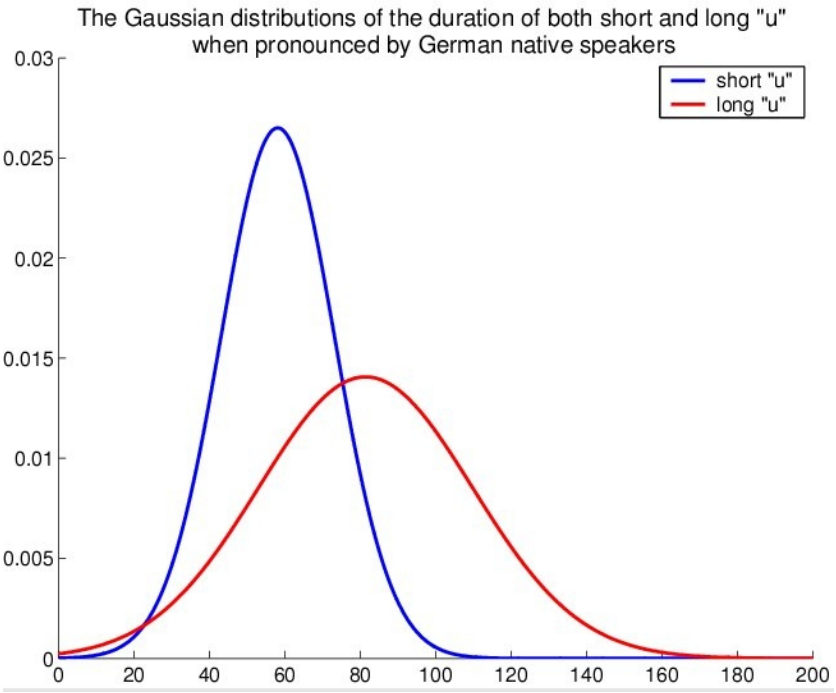

The Gaussian distributions of the duration of unstressed and stressed "i"
when pronounced by Greek native speakers
unstressed "i"
stressed "i"
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
20
40
60
80
100
120
140
160
180
200


The Gaussian distributions of the duration of both unstressed and stressed "u"
when pronounced by Greek native speakers
unstressed "u"
stressed "u"
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
20
40
60
80
100
120
140
160
180
200

Gaussian distribution of the duration of short and long vowels in German.






The Gaussian distributions of the duration of both short and long "u"
when pronounced by German native speakers
short "u"
long "u"
0.03
0.025
0.02
0.015
0.01
0.005
0
0
20
40
60
80
100
120
140
160
180
200