Segmental and prosodic improvements to speech generation

E.A.M. Klabbers

The work described in this thesis has been carried out at
IPO, Center for User-System Interaction,
Eindhoven, the Netherlands.

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN

Klabbers, Esther A.M.

Segmental and prosodic improvements to speech generation /
by Esther Anna Maria Klabbers. -
Eindhoven: Technische Universiteit Eindhoven, 2000. -
ISBN 90-386-0563-3
NUGI 832
Keywords: Speech synthesis

Segmental and prosodic improvements to speech generation

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
Rector Magnificus, prof.dr. M. Rem, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op woensdag 7 juni 2000 om 16.00 uur

door

Esther Anna Maria Klabbers

geboren te Nijmegen

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr. R.P.G. Collier
en
prof.dr. L.W.J. Boves

Copromotor:

dr.ir. R.N.J. Veldhuis

ACKNOWLEDGMENTS

Many people have helped and supported me while I was working on this thesis. I would like to take this opportunity to thank a few of them:

René Collier, my thesis supervisor, for giving me all the support I needed. Our weekly meetings were truly inspirational. Raymond Veldhuis, my co-promotor, for the cooperation on audible discontinuities and for being so patient with me in explaining the technical details of signal processing. Jan van Santen, for making my time at Bell Labs so worthwhile and pleasant. Jan-Roelof de Pijper and Leo Vogten, for supervising me especially at the outset of this project. Alice Dijkstra, Jan Landsbergen and Jacques Terken for coordinating the NWO-TST Programme. Mariët, my roommate at IPO, for putting up with me day after day. Loe Boves and Louis Pols as members of my thesis committee for providing me with useful comments and suggestions.

Everybody of the Spoken Language Interface group, the IPO choir, and everyone else at IPO for making it such a nice place to work. Kim Koppen, Ineke Burema and all the participants in my experiments for contributing to this thesis.

My parents Martien and Truce Klabbers and my brothers Ruben and Lucas, for supporting and loving me in everything that I do.

Contents

Chapter 1

General introduction

1.1 Introduction

The research presented in this thesis was carried out in the framework of the 'Language and Speech Technology' Priority Programme of NWO (the Netherlands Organisation for Scientific Research). In the context of this programme, a spoken dialogue system was developed, called OVIS, a Dutch acronym for *Openbaar Vervoer Informatie Systeem* (Public Transportation Information System), which provides information on train timetables in the Netherlands. The scientific goal of the programme was to improve each of the components involved in the development of such a system for Dutch, and to integrate them seamlessly. The research described in this thesis is only concerned with the speech generation component of this system. In Section 1.2, all the components of the OVIS spoken dialogue system will be briefly explained. Section 1.3 explains the goal of this thesis and gives a brief outline of the body of this thesis.

1.2 The OVIS Spoken Dialogue System

Figure 1.1 shows the components that make up the OVIS spoken dialogue system. The process starts at the top with the user issuing a request. This speech signal is input to the *automatic speech recognition* module (Strik et al., 1996), based on Hidden Markov Modelling (HMM). It outputs a word graph containing a network of word hypotheses, as a result of uncertain-

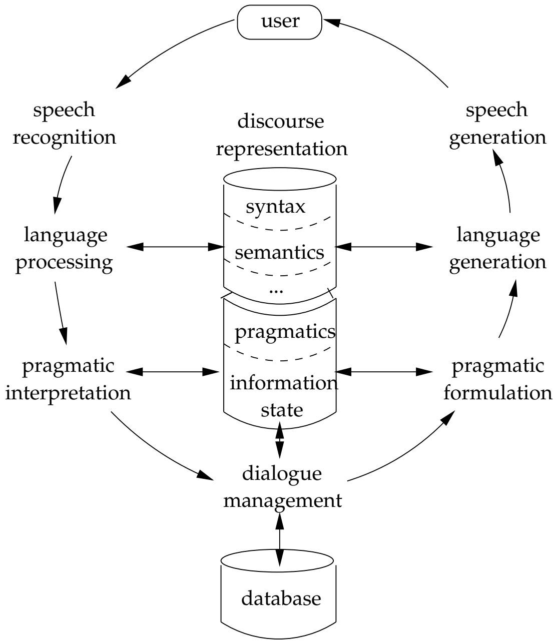


Figure 1.1: *Global architecture of the OVIS spoken dialogue system.*

ties arising from acoustic-phonetic analysis errors and linguistic ambiguities. The *natural language processing* module determines which path in this word graph most likely represents the user's utterance. It can do so on the basis of syntactic, semantic and pragmatic information. Syntactic information can be used to check whether a hypothesised word is likely to occur in a particular position in the sentence, which is usually checked in relation to the preceding and/or the following word. Semantic information can determine if the sentence is actually meaningful and pragmatic information enables the system to determine if meaningful sentences are plausible or appropriate in the context of the ongoing dialogue.

There are two alternative approaches to natural language processing in OVIS, one is knowledge-based (Nederhof et al., 1997), the other is corpus-based (Bonnema et al., 1997). In the knowledge-based approach, a grammar is constructed that covers the entire domain, whereas in the corpus-based approach a database of syntactic subtrees is created from a large training corpus consisting of real information dialogues. In either case, the natural language processing module outputs an abstract representation of the sentence to the dialogue management module, which interprets this representation in the context of the dialogue carried out so far.

The *dialogue management* module (Veldhuijzen van Zanten, 1998) determines what an appropriate system reaction could be on the basis of changes in the information state. The dialogue management module sends a command to the natural language generation module, signalling that it needs additional information, that it needs to verify recently obtained information, or that it is ready to output the requested information found in the database. An important aspect of dialogue management is the amount of control that resides with the system. The control can lie with the system in a menu-based application, where the structure of the dialogue is fixed and the system guides the user through the questions the answers of which provide the relevant information to the system. This approach constrains the user to a large extent and, if not properly implemented, can be perceived as very unpleasant. Alternatively, the control can lie with the user. This can be undesirable when the user doesn't have a clear picture of the system's capabilities. The key is to design a mixed-initiative system that guides the user when necessary, but gives the experienced user the opportunity to take control whenever possible.

The *natural language generation* module (Theune, 2000) converts a message (dialogue act) from the dialogue management module into an enriched text representation, i.e., with accent and phrase boundary markers added, which it passes onto the *speech generation* module. The natural language generation module uses syntactic templates, consisting of fixed phrases with slots in which other templates can be inserted. With information about syntax, semantics and pragmatics being readily available, accent and phrase boundary locations can be computed with much greater accuracy than when a text-to-speech synthesis system computes them from unknown text.

The *speech generation* module converts the enriched text representation into speech. Since the speech output is the part of the system that is most prominently perceived by the user, good quality is imperative. Users should at least be able to easily understand the system's utterances and preferably they should find it pleasant to listen to.

1.3 Goal of this thesis

The goal of this thesis is to improve the quality of IPO's speech generation algorithms. Speech generation at the level closest to the signal is achieved in three steps: duration computation, intonation (F_0) computation and sound generation. In the restricted domain of concatenative synthesis, sound generation involves concatenating pre-recorded units and imposing computed duration and F_0 values onto the speech signal. The quality of the output speech can be decomposed into two components: the segmental quality and the prosodic quality. These crucially depend on the size of the concatenative units. This has led to our first research question.

1. What is the optimal size of the concatenation unit?

In concatenative speech generation, there is a trade-off between output quality and flexibility, which is directly related to the size of the concatenative units. We did not explore the full range of possibilities, but rather restricted ourselves to two techniques that represent the top and bottom end of the spectrum, phrase concatenation and diphone synthesis, respectively. Diphones (Dixon and Maxey, 1968) are recorded speech segments that represent transitions between any two sounds in a language. These segments are stored in a coded form so that their timing and pitch can be modified after concatenation. The advantage of diphone synthesis is that it is extremely flexible in that only a limited number of diphones are required (approximately 2000 for Dutch) to synthesise unlimited text. Diphone synthesis can be adopted for a wide range of applications without much adaptation. The drawback is, however, that the output quality often leaves a great deal to be desired. A number of factors influence the segmental quality. First, the context in which the diphones have been recorded is important. It makes a difference whether the diphones are excised from nonsense words or from real sentences and whether they occur in accented words versus unaccented words or in stressed syllables versus unstressed syllables. Second, in order for the duration and intonation values to be imposed on the signal, it has to be coded. Dependent on the speaker's voice characteristics, modifications of the signal can deteriorate the segmental quality considerably. Ultimately, the prosodic quality depends on the adequacy of the rules determining the duration and intonation values, assuming that the positions of accents and phrase boundaries are accurately determined by the NLG module.

Phrase concatenation involves the concatenation of pre-recorded words and phrases. It is often used in commercial applications for a limited domain, e.g., weather reports. The segmental quality depends for a large part on the context in which the words and phrases have been recorded. If they have been recorded in isolation, there is a risk of a mismatch in pitch, loudness and tempo that will make the speech sound disfluent. Often no modification of the speech signal takes place, so in that respect the segmental quality is very high. The prosodic quality depends on the ability to deal with the prosodic variation introduced by the NLG module. This is achieved by recording units in the proper context, providing multiple instances of otherwise identical words and phrases, which differ only in their duration and intonation.

The implementation of these two techniques is discussed in Chapter 2, *Generating high-quality speech*. It will be shown that for applications like OVIS, in which the domain is restricted and has a fairly stable vocabulary, speech can be generated using pre-recorded words and phrases with a quality close to that of natural speech, provided that a couple of conditions are met. For instance, it is essential that the prosodic realisations can comply with the instructions provided by the natural language generation module. This requirement is by no means trivial to implement.

However, whenever unrestricted text-to-speech is required, i.e., in applications such as e-mail reading, phrase concatenation is no longer feasible. Therefore, the remaining chapters are devoted to answering two additional questions:

2. How can we improve the segmental quality of diphone synthesis?
3. How can we improve the prosodic quality of diphone synthesis?

In Chapter 3, *Reducing audible spectral discontinuities*, a common problem in diphone synthesis is discussed, viz., the occurrence of audible discontinuities at diphone boundaries. Informal observations show that spectral mismatch is most likely the cause of this phenomenon. The problem is most obvious in vowels and semi-vowels. Chapter 3 first sets out to find an objective spectral measure for discontinuity. To this end, several spectral distance measures are compared to the results of a listening experiment. Then, a solution is proposed to reduce the occurrence of audible discontinuities by extending the diphone database with context-sensitive diphones. The number of additional diphones is limited by clustering consonant contexts that have a similar effect on the surrounding vowels on the basis of the best performing distance measure.

As mentioned earlier, the prosodic quality is determined by the duration and intonation characteristics of the speech. We assume here that the assignment of accent and phrase boundary positions is accurately performed by the NLG module, so that only the realisation of duration and pitch affect the prosodic quality. IPO has a long history of intonation research, which resulted in the Grammar of Dutch Intonation (GDI, Collier and 't Hart (1981);'t Hart, Collier, and Cohen (1990)). We therefore concentrate on another aspect of prosody, namely duration.

Chapter 4, *Modelling segmental duration*, discusses the development of a new duration model using the sums-of-products approach of Van Santen (1992a), because the rule-based model in Calipso is not deemed satisfactory. The main drawback of the old model is that some important higher-level factors, such as position of the syllable in the word and of the word in the phrase, are not taken into account and that interactions among factors are not sufficiently modelled. Moreover, this system is not specifically modelled after the speaker of the diphones.

The new duration module is specifically designed to approximate the temporal behaviour of one female speaker, the same as was used for the diphone recordings. Phonemes that are affected similarly by the various factors are grouped into subclasses. The decisions concerning this grouping are based on exploratory data analysis and phonetic/phonological literature. For each subclass of phonemes, a separate sums-of-products model is trained. Exploratory data analysis is also required to appropriately decide which factors are important and how many levels on a factor should be distinguished. For each subclass, this can lead to different results. Thus, a great deal of phonetic and phonological knowledge can be incorporated in the model. This approach has already resulted in well performing duration modules for a number of languages, including American English, French and German.

This study ends with Chapter 5, *Summary and conclusion*. Some of the examples that are presented in this thesis are made available as sound files on the web site listed below. The content of these sound files is described in Appendix C.

http://www.ipo.tue.nl/homepages/eklabber/audio.html

Chapter 2

Generating high-quality speech

2.1 Introduction

This chapter describes research work done on the speech generation module in OVIS. Because the speech output is the part of the spoken dialogue system that is most prominently perceived by the user, good quality is imperative. Users should be able to easily understand the system's utterances (prompts) and preferably they should find the voice of the system pleasant to listen to. This means that both the *intelligibility* and *naturalness* must be high.

In concatenative speech generation, there is a trade-off between output quality and flexibility, which is directly related to the size of the concatenative units. The maximum amount of flexibility is obtained by using diphones (Dixon and Maxey, 1968) as the basic concatenative unit. Diphones are recorded speech segments that represent transitions between any two sounds in a language. These segments are stored in a coded form so that their timing and pitch can be modified after concatenation. With a limited number of diphones, unlimited text can be synthesised. Diphone synthesis can be adopted for a wide range of applications without much adaptation. Unfortunately, there is a price to be paid: the quality of the speech generated by such systems still leaves a great deal to be desired. Current speech synthesis systems are capable of generating speech which has a high degree of intelligibility, but, in general, the speech still sounds quite unnatural. The diphone synthesis system developed at IPO will be discussed in more detail in Section 2.2.

At the other end of the spectrum, the best quality can be achieved by playing back digitally stored natural speech. The quality of the speech output is then limited only by the medium, e.g., a standard telephone channel, through which it is transmitted. However, this approach is only practical in the simplest of applications. The key is to find a balance in the tradeoff between naturalness and flexibility. In that respect, concatenating prerecorded units like words and phrases appears to be a good alternative. With this approach, a large number of utterances can be produced on the basis of a limited set of pre-recorded phrases, saving memory space and increasing flexibility. This technique is practical only if the application domain is limited and remains rather stable, as is the case with train timetables.

The use of concatenated words and phrases in limited-domain applications such as OVIS is quite common. It is used in many commercial applications such as the talking clock, telephone banking systems, market research teleservices and travel information services. But often the method is so straightforward that it is not even mentioned. In the German train timetable system (Aust, Oerder, Seide, and Steinbiss, 1995) and in the first version of OVIS which is based on this German system, speech output was obtained by simply recording the necessary words and phrases and playing back the concatenated sentences when required.

This approach (which will be referred to as conventional phrase concatenation) has two major problems:

- First, recordings are often not carefully controlled. The concatenative units are usually recorded in isolation, which causes mismatch in loudness, tempo and pitch between concatenated units, leading to disfluent speech. Phrases seem to overlap in time and create the impression that several speakers are talking at the same time, from different locations in the room. In order to disguise these prosodic imperfections, pauses are often inserted, which are very conspicuous and make the speech sound even less fluent. This causes the *segmental quality* of the speech output to be suboptimal.

- Second, in natural speech, the prosody of words in an utterance varies depending on several factors such as their position in the utterance, the syntactic structure of the utterance, and the discourse context. For instance, words expressing information that has been previously mentioned tend to be deaccented. By recording all words and phrases in one prosodically neutral version, such contextual variation is not taken into account. This causes the *prosodic quality* of the speech output to be suboptimal.

One simple application that does take some prosodic properties into account is the telephone number announcement system described in Waterworth (1983). In order to increase the naturalness of the telephone number strings that are output by the system, digits are recorded in three versions with different intonation contours. There is a *neutral form*, a *continuant*, with a generally rising pitch, and a *terminator*, with a falling pitch contour. Most digits in a telephone number, e.g., *010 - 583 15 67*, are pronounced using the neutral form. However, the numbers occurring before a space, viz. *0, 3* and *5*, are pronounced using the continuant form to signal a boundary and to indicate that the utterance has not yet finished. The final *7* is pronounced with a terminator to signal the end of the string. Experiments showed that people preferred this method over the simple concatenation method.

Another application, a computer-assisted language learning program called Appeal, uses a more sophisticated form of word concatenation to deal with prosodic variations (De Pijper, 1996). When making the recordings, the words were embedded in carrier sentences to do justice to the fact that words are shorter and often more reduced when spoken in context. Only one version of each word was recorded, but when during generation the words are concatenated to form a text, the duration and pitch of the words are adapted to the context using the Pitch-Synchronous Overlap-and-Add technique (PSOLA, Charpentier and Moulines (1989)). This ensures a natural prosody, but the coding scheme may deteriorate the quality of the output speech to some extent.

Our approach to phrase concatenation can be seen as an extension to the simple concatenation approach. It is discussed in more detail in Section 2.3. It is different from conventional concatenation in that a) all concatenative units have been recorded embedded in carrier sentences, and b) like Waterworth, it takes prosodic variation into account by recording different prosodic versions for otherwise identical units. No manipulation of the speech signal is required, thus retaining a natural speech quality. The resulting output quality has been evaluated in a listening experiment, reported on in Section 2.4, where it is compared to natural speech, conventional phrase concatenation and diphone synthesis. This chapter ends with a discussion in Section 2.5.

2.2 Diphone synthesis

2.2.1 Introduction

Using full-fledged speech synthesis provides maximum flexibility. In spoken dialogue systems like OVIS the linguistic analysis part can be skipped, as the NLG module can determine the prosodic annotations for accents and phrase boundaries with much more accuracy using syntactic, semantic and pragmatic information (Theune, Klabbers, Odijk, and De Pijper, 1997). The phonetic transcription can be obtained via a lookup table.

The following section will discuss the synthesis routines in IPO's diphone synthesis system Calipso (previously known as SPENGI (Collier and Houtsma, 1995; Terken, 1996)). For excellent introductions on concatenative speech synthesis in general, the reader is referred to Dutoit (1997) and Sproat (1998). In the course of the NWO-TST Programme, interest has increased in new variants of concatenative synthesis using units of variable length (Campbell and Black, 1997; Black and Taylor, 1997; Balestri et al., 1999). This novel approach will be briefly dealt with in the discussion in Chapter 5.

2.2.2 Synthesis routines

The synthesis routines can be split up into three steps: a) Duration assignment, i.e., computing temporal structure, b) Intonation assignment, i.e., computing melodic structure, and c) Sound generation, i.e., concatenating the diphones and overlaying the durational and intonational values.

Duration assignment: Duration prediction in Calipso is done via a rule-based duration model. The rules in the model have either been taken from the literature or are derived from small-scale experiments. First, the phonemes receive a base duration value which is then altered by applying a series of rules. For rules relating to vowels, the dissertation of Nooteboom (1972) served as a good source of information. Some additional rules were borrowed from the dissertation of Van Coile (1989). In general, lengthening or shortening occurs as an effect of the immediately preceding or following segment or as an effect of stress and accentuation. Hardly any higher-level effects are taken into account, and interactions between different factors cannot be modelled easily in this approach. For several years, it has been experienced that this rule set does not yet produce satisfactory results. Therefore, a new duration module was developed using the sums-of-products approach of Van Santen (1992a). It is described in Chapter 4.

Intonation assignment: The intonation module in Calipso is an implementation of the Grammar of Dutch Intonation as developed over the years at IPO. It is described extensively in Collier and 't Hart (1981) and 't Hart, Collier, and Cohen (1990). The GDI is characterised by a distinction between involuntary and voluntary pitch variations. The latter carry linguistic information (e.g., about accented words, phrase boundaries, speaker attitude). Only the voluntary actions of the speaker are considered to be relevant for perception. This makes it possible to stylise a pitch contour, i.e., to reduce it to its perceptual essentials. A pitch contour is assumed to be composed of a series of pitch movements alternating between two reference declination lines, a bottom line and a top line. A perceptually relevant pitch movement is characterised by its *direction* (rise vs. fall), its *rate of change* (slow vs. fast), its *size* (full vs. half) and its *timing* (early vs. late vs. very late in the syllable). These pitch movements are grouped into pitch configurations which are then combined into pitch contours at the clause level. The permissible sequences of the pitch movements are specified in a language-specific grammar. In Calipso, the intonation module reads its input from two files, one specifying the acoustic definition for the declination line and each of the pitch movements, the other specifying the allowed sequences of pitch movements on all possible combinations of accents and phrase boundaries.

Sound generation: The diphone database contains 1955 units. They have been excised from 1355 nonsense words. Table 2.1 lists the diphone types the database contains. Consonant-Vowel (CV) and Vowel-Consonant (VC) diphones come from the same symmetrical nonsense word *C@CVC@*. For instance, /ta/ and /at/ come from the Dutch nonsense word *tetaate* (/t@tat@/). See Table A.1 for a listing of all phonemes in the database. Other types of diphones are VV, CC, silence-V, silence-C, V-silence and C-silence. All diphones are recorded in a position where they have word stress. Thus, there are no reduced diphones in the database, which may cause the speech to sound over-articulated at times.

The actual synthesis takes place on the basis of TD-PSOLA (Moulines and Charpentier, 1990), which allows pitch and duration modification by direct manipulation of the waveform. In the analysis stage, pitch-markers are set at a pitch-synchronous rate on the voiced portions of the signal and at a constant rate on the unvoiced portions. Analysis windows, called *bells*,

Diphone type	Nonsense words	Example	Number in database
CV	C@CVC@	tetaate	595
VC	C@CVC@	tetaate	595
VV	tVVt@	teate	287
CC	C@CCe	tetree	324
silence-V	.Vt@	.aate	23
silence-C	.C9t@	putte	21
V-silence	p@pV.	pepee	23
C-silence	C@CaC.	pepaap	14
Total:			1955

Table 2.1: *Description of all diphone types in the database; V = vowel, C = consonant, @ = schwa, . = silence.*

are centred around these pitch markers. The bells span approximately two periods, thus overlapping each other in time. During synthesis, the pitch is changed by altering the time distance between successive bells and then adding them. The duration is changed by inserting or deleting bells, i.e., the bell is the smallest unit of durational modification.

Although TD-PSOLA works very well for manipulating natural speech, some problems occur when it is used for modifying diphones. Repeating identical bells to lengthen a sound can lead to unnatural sounding speech. Phase mismatches can occur when the windows are not placed at the same position within the period. Pitch mismatches can occur if the speaker was unable to keep the pitch reasonably constant during recording. Spectral envelope mismatches can also occur. This is one of the major drawbacks of concatenative synthesis. It is mostly due to a failure to account for the predictable coarticulation effects in speech production. These mismatches can result in audible discontinuities at diphone boundaries. This problem is described in Chapter 3.

2.2.3 Speech output quality

Ultimately, the aim of speech synthesis systems is to generate speech of high quality from unrestricted text. This is not impossible, but many problems are encountered in text preprocessing, for instance at the level of grapheme-to-phoneme conversion, phrasing and stress assignment, and in synthesis, for instance the spectral continuity and duration/intonation assignment mentioned earlier. All these errors occur at different levels of analysis or synthesis and lead to a degradation of the speech output quality. Assessing the performance of a speech synthesis system is not as simple a task as assessing the performance of an automatic speech recognition system, where the performance of the system is captured in one measure: the Word Error Rate (WER).

In speech synthesis, it is possible to quantify the errors made at some levels, e.g., segmental intelligibility, using a whole range of diagnostic tests (Steeneken, 1992; Pols, 1994). Some other levels, such as the prosody computation level, cannot be evaluated so easily. Moreover, there is little standardisation in terms of which tests to use, and few proper benchmarks are being performed (except for a recent initiative by Van Santen, Pols, Abe, Kahn, Keller, and Vonwiller (1998)). The overall quality of the system can be assessed by a mean opinion score (MOS) test, which is usually carried out in the context of a specific application. These scores give an indication of the overall quality of a system, but since MOS tests are relative, two or more systems must be evaluated in the same test with the same set of users in order to give comparable results. Gibbon, Moore, and Winski (1997) provide an overview of tests on all levels and give some guidelines for their employment.

In a recent evaluation by Rietveld et al. (1997), three Dutch synthesis systems (one of which was Calipso) were compared. Forty-four participants took part in an intelligibility test and another group of forty-four participants took part in a subjective evaluation. In both groups, twenty-two participants listened to standard telephone (PSTN) coded materials, and twenty-two others to GSM-coded materials. In the intelligibility test, the participants were asked to transliterate a semantically unpredictable text of 147 words, excluding articles. The percentage of correct transliterations was lower in the GSM condition (between 52 and 59%) than in the PSTN-condition (between 66 and 72%). In the subjective evaluation, participants had to rate their preferences for pairs of systems on a scale ranging from -3 to +3. They also had to rate 16 questions, such as overall quality, listening effort, and voice pleasantness, on a five-point scale. The texts represented three text types: email messages, stock exchange information and public transport information. The overall quality of the systems ranged between 2.34 and 2.73 on a 5-point scale. This score was the same in the GSM and the PSTN condition. In general, speech is considered acceptable when the MOS-score is at least 3 on a 5-point scale. Therefore, we can conclude that the quality of the speech synthesis systems under investigation is still insufficient for use in commercial applications.

2.3 IPO's phrase concatenation

Many of these applications do not require unrestricted text-to-speech synthesis. In cases where the domain is limited and remains rather stable, pre-recorded words and phrases can be used. We will now describe the phrase concatenation approach that was developed in the context of the NWO-TST Programme.

2.3.1 Introduction

This section describes the development of an advanced approach to phrase concatenation which results in high quality speech. The concatenation of words and phrases requires no manipulation or coding of the recordings. This phrase concatenation approach fits well with the NLG module, because the syntactic templates that form the core of this module indicate which parts are carrier sentences and which are slots. All concatenative units are recorded embedded in carrier sentences. This reduces the chance of mismatch in loudness, tempo and pitch occurring after concatenation, resulting in more fluent speech. A natural intonation is achieved by using several prosodic variants of otherwise identical words and phrases that serve as slot fillers. In order to determine which units have to be recorded and how many different prosodic realisations are required, a thorough analysis of the material to be generated is a necessary phase in the development of a phrase database.

This technique was first used in an application called GoalGetter (Klabbers, Odijk, De Pijper, and Theune, 1996), which is a data-to-speech system generating spoken soccer reports on the basis of tabular information available via Teletext[1]. (See Appendix C for an example.)

2.3.2 Database definition

Once the content and the prosodic properties of the messages to be generated are known, a phrase database can be developed. For the slot fillers, e.g., station names and time expressions, it was decided to use six different prosodic realisations, which are depicted in Figure 2.1. They are styli-

[1] Teletext is a system with which textual information is broadcast along with the television signal and decoded in the receiver. It is also available via the internet.

sations of pitch contours on monosyllabic words (according to the IPO Grammar of Dutch Intonation ('t Hart, Collier, and Cohen, 1990)) of the most common realisations occurring in each of the prosodic contexts indicated.

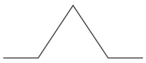
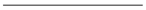
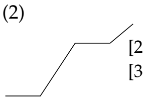
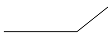
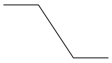
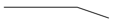
Accent Boundary	Yes	No
None	(1) 	(4) 
Minor / Major continuation	(2)  [200 ms]/ [300 ms]	(5)  [200 ms]/ [300 ms]
Finality	(3)  [500 ms]	(6)  [500 ms]

Figure 2.1: *Stylised examples of the different prosodic versions that are needed. Two factors determine their pitch and pausing: the accentuation and the position relative to a minor/major/final phrase boundary. The pauses are indicated between brackets.*

1. An accented slot filler which does not occur before a phrase boundary is produced with the most frequently used pitch configuration, the so-called (pointed) *hat pattern*, which consists of a rise and fall on the same syllable. This contour corresponds to the prosodically neutral version often used in many other phrase concatenation techniques.

2. An accented slot filler which occurs before a minor or a major phrase boundary is most often produced with a rise to mark the accent and an additional continuation rise to signal that there is a non-final boundary. A short pause follows the constituent, which is 200 ms in length in case of a minor boundary (/) and 300 ms in case of a major boundary (//).

3. An accented slot filler which occurs in final position receives a final fall. It is followed by a longer pause of 500 ms.

4. Unaccented slot fillers are pronounced on the declination line without any pitch movement associated with them.

5. Unaccented slot fillers occurring before a minor or a major phrase boundary only receive a small continuation rise. This prosodic situation does not occur very often. The NLG module usually puts a minor or major phrase boundary immediately after an accented word. Again, a 200-ms or 300-ms pause is inserted.

6. Unaccented slot fillers in a final position are produced with final lowering, i.e., a declination slope that is steeper than in other parts of the utterance. It is followed by a 500-ms pause.

2.3.3 Recording

When recording the material for the phrase database, the slots in the carrier sentences were filled with dummy words so that the fixed phrases to be stored in the database could be excised easily. In this way, the effect of coarticulation at the word boundaries was minimised. Fade-in and fade-out was applied to all material in the phrase database to avoid clicks in concatenation. The slot fillers, such as station names and time and date expressions, were embedded in dummy sentences that provide the right prosodic context. The sentences were constructed in such a way as to make the speaker produce the right prosodic realisation naturally. She received no specific instructions about how to produce the sentences. The intonation in the carrier phrases is not so critical, so the speaker could use her own intuitions about how to pronounce them. The recordings were made in a sound-treated room using two high-quality microphones which were positioned on either side of the speaker, a fixed distance away from the mouth. The speech was recorded on a DAT-tape using a 48 kHz sampling frequency. The speech signal was stored on an SGI workstation in mono with sampling frequency of 16 kHz. The concatenative units were excised manually and sentences were generated to check for large differences in loudness to be corrected.

2.3.4 Generating speech

To concatenate the proper words and phrases, an algorithm has been designed that performs a mapping between the enriched text, i.e., text with accentuation and phrasing markers, as provided by the NLG module, and the pre-recorded phrases that have to be selected. The different prosodic variants are chosen on the basis of the prosodic markers. The algorithm recursively looks for the largest phrases to concatenate into sentences. It works from left to right. First, it tries to find the string of N words that contains the entire sentence. If it is present, it is retrieved and can be played. If not, the string comprising the first $N - 1$ words is looked up. This process continues until a matching phrase is found. Then the remaining part of the sentence undergoes the same procedure, until the entire sentence can be played. It is perhaps not the most optimal search algorithm, but it provides the correct phrases to be concatenated.

As an example, consider the sentence in Figure 2.2 (English: 'On which day would you like to travel from Groningen to Paris?'). The sentence consists of 5 pieces of the carrier phrase *op welke dag*, *wilt u*, *van*, *naar* and *reizen?*. The two slot-filling station names *Groningen* and *Parijs* are both accented but *Groningen* is realised with a continuation rise because of the minor phrase boundary following it.

The OVIS speech database takes up approximately 93 MB. There are 2987 units in the database that can be divided into different categories as listed in Table 2.2. As can be seen, the station names (382 different names that cover all Dutch train stations and some foreign stations) form the bulk of the data, especially since they have been recorded in six prosodic versions. The days of the week have been recorded in only two prosodic versions. This is because in OVIS the NLG module always generates them followed by a number and a month, e.g., 'Maandag 24 januari' (Monday January 24th), and no phrase boundary can be inserted immediately after them. Thus, the two versions only distinguish accented from unaccented days of the week. The years are only recorded in accented position. There are three versions corresponding to the position relative to a phrase boundary, i.e., before no boundary, before a minor/major boundary or at the end of the sentence. In conventional phrase concatenation, the size of the database would be 667 units, the total number of types as indicated in Table 2.2. Recording additional prosodic variants increases the size of the database with a factor 4.5 to 2987, the total number of tokens in the table. Figure 2.3 shows the unequal distribution of the unit length in the database. The vast

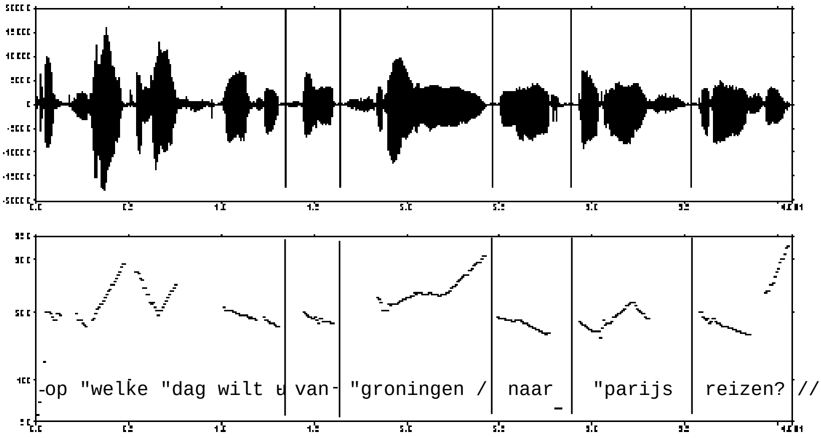


Figure 2.2: *Example sentence "Op welke dag wilt u van Groningen naar Parijs reizen?", with two different prosodic realisations of a station name* Groningen *(+accent, minor boundary)* , Parijs *(+accent, no boundary).*

Concepts	Number of types	Number of prosodic variants	Number of tokens
Carrier sentences	195	1	195
Station names	382	6	2292
Numbers	60	6	360
Months	12	6	72
Implicit dates	7	6	42
Years	4	3	12
Days	7	2	14
Total:	667		2987

Table 2.2: *Composition of the OVIS phrase database. For each concept it is indicated how many types there are in the database, how many prosodic variants are recorded of each concept, and the resulting number of tokens.*

majority of units is only one word in length (62.8%), whereas 89.4% of the units consists of maximally two words. Since the number of concatenative boundaries in a sentence of N words can, in the worst case, be $N - 1$ (if all units contain just one word), minimisation of mismatch in loudness, pitch, tempo is important. This once more indicates that careful recording is required. The OVIS speech output quality is evaluated in a listening experiment which will be discussed in the next section.

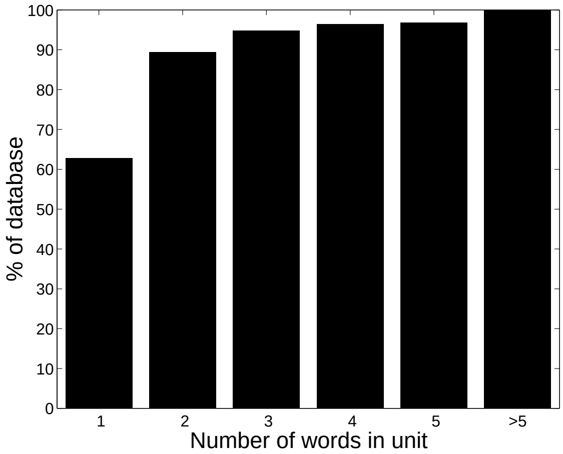


Figure 2.3: *Cumulative distribution of the size of the concatenation units in the OVIS phrase database.*

2.4 Evaluation

The IPO phrase concatenation method has been evaluated in a formal listening experiment, in which it was compared to 1) natural speech output, 2) a conventional concatenation approach and 3) IPO's diphone synthesis system Calipso. The sentences generated with the conventional concatenation approach have been provided by a telephone company, which uses it in a commercial train timetable application. The concatenative units have been recorded in isolation and all words and phrases have been recorded in one neutral version only.

The listeners' task was to evaluate only the overall output quality, so accentuation and phrasing were kept constant over all conditions. The experiment consisted of three parts: an intelligibility test, a fluency test and a final test about the overall quality and suitability to the application. In the final part, the natural speech condition was excluded, since it is not a realistic option for actual use in applications like OVIS.

2.4.1 Procedure

All stimuli, twenty-three in total, consisted of information about train connections that can occur in OVIS. First, the participants were presented with three example fragments (see the example in (1)), to get a general indication of the different speech output conditions. The natural speech output condition was not included in these examples.

(1) ik heb de volgende verbinding gevonden / / /
met de sneltrein / vertrek vanuit enkhuizen / om elf uur eenenvijftig / aankomst in oosterbeek / om drieentwintig uur twee / / /
daar verder met de stoptrein / vertrek om dertien uur zeventien / aankomst in stavoren / om drie uur achtentwintig / / /
wilt u nog een andere verbinding weten ? / / /

Translation:
I found the following connection ///
with the express train / departing from enkhuizen / at 11:51 / arrival in oosterbeek / at 23:02 ///
there continue with the local train / departing at 13:17 / arrival in stavoren / at 03:28 ///
would you like to have a different connection ? ///

The experiment started with the intelligibility test. All participants listened to twenty fragments, five different fragments for each speech condition. The task of the listeners was to recognise and write down the two station names that occurred in each fragment. The fragments were shorter than the example fragments (see the example in (2)), containing only two station names, and no stop-over, so as to make sure that it was an intelligibility test and not a memory test.

(2) met de sneltrein / vertrek vanuit vierlingsbeek / om negen uur zevenenvijftig / aankomst in koudum molkwerum / om drieentwintig uur dertig / / /

with the express train / departing from vierlingsbeek / at 09:57 / arriving in koudum molkwerum / at 23:30 ///

The station names were balanced with respect to familiarity (correlated with the number of inhabitants of the town or city) and number of syllables. Some station names were easily confused (Heiloo vs. Heino and Oss vs. Olst). The fragments of conventional phrase concatenation were obtained from a commercial train timetable information system. Since these fragments were only available in telephone bandwidth, the fragments representing the other speech output conditions were filtered to a bandwidth of 300 Hz to 3400 Hz. The IPO phrase concatenation, diphone synthesis and natural speech fragments all originated from the same semi-professional female speaker, whereas the conventional phrase concatenation made use of a different female speaker.

The order of the speech output conditions was balanced over all participants, so that one speech condition occurred five times in the first block, five times in the second block, etc. This was done to compensate for order effects. Each station name occurred equally frequently in each speech condition. Each subject listened to 40 different station names. So for instance, the station name 'Heiloo' was presented in the DS condition for one subject and in the IC condition for another subject. Due to the elaborate format in which the train connections were presented to the participants, it was impossible to ask them to also write down the time information, let alone the entire sentence. After listening to the five fragments that made up one speech output condition, participants had to rate on a 7-point scale how they judged the overall speech intelligibility. Listeners were instructed to take into account whether they had problems recognising individual words, or whether it took a lot of effort to understand the message.

In the next part of the test, listeners had to rate the fluency of each fragment on a 7-point scale. The fragments were the same 20 fragments that were used in the intelligibility test, but the speech output conditions were presented in a different order, and the fragments were again randomly divided over the conditions. We did not ask to evaluate "naturalness", as this attribute is a composite measure for many aspects of the speech signal, including segmental quality and prosody, each in turn having multiple dimensions (such as speech melody, accentuation, phrasing). Instead, we asked listeners to evaluate the fluency of the fragments, which we consider a better criterion with fewer dimensions, that relates directly to an inherent flaw of concatenative synthesis. The participants were told they could take into account whether the speech had a faltering speaking style, or contained audible jumps in pitch.

In the final part of the experiment, the participants were again presented with the longer example messages that were introduced at the start of the experiment. Per message they had to rate two questions on a 7-point scale: one concerning the overall quality of the speech and the other concerning its suitability for the given application. Here, participants had to imagine actually calling such an information service. How would they evaluate the speech output quality of the system in that case?

2.4.2 Subjects

Twenty participants took part in the experiment. They were students from the Eindhoven University of Technology, with no prior knowledge of speech technology. They were not paid for their participation.

2.4.3 Results

In the discussion of the results, the speech output techniques will be referred to by codes. N stands for the natural speech condition, IC stands for IPO's phrase concatenation, CC for conventional phrase concatenation and DS for diphone synthesis. In the first part of the intelligibility test, where the participants had to transcribe 40 station names, recognition (i.e., transcription) errors are the dependent variable. Table 2.3 lists the word error rate, i.e., the percentage of items that were transcribed incorrectly. A station name was considered to be transcribed incorrectly when at least one letter was deleted, inserted or changed, such that the name sounded differently. The table shows that the intelligibility of the station names is identical for the N and IC conditions. CC is slightly less intelligible, but since the station names have been produced as whole words in both the IC and CC condition, this tells us more about the articulatory properties of the different speakers than about the method. Diphone synthesis scores worse than any other speech output technique.

Table 2.4 displays the average values for each of the four subjective quality ratings and each of the four speech output conditions. In Figure 2.4, these are displayed graphically together with the standard deviations to make the differences better quantifiable. The ratings for each speech condition for intelligibility, overall quality and suitability for the application have 20 data points, one for each subject. For fluency, the number of data points is 100, as each subject gives a value for 5 fragments in each speech condition.

Speech output techniques	Word error rate (%)
N	3.5
IC	3.5
CC	6.0
DS	14.5

Table 2.3: *Word error rate (percentages of items transcribed incorrectly) in the transcription of station names. The total number of station names transcribed per speech output condition is 200 (5 fragments x 2 stations x 20 participants); N = natural speech, IC = IPO's phrase concatenation, CC = conventional phrase concatenation, DS = diphone synthesis.*

	Number of observations	N	IC	CC	DS
Intelligibility	20	5.95	5.85	4.80	2.75
Fluency	100	6.16	5.41	3.19	2.72
Overall quality	20	n/a	6.05	3.90	2.25
Suitability for application	20	n/a	6.60	4.10	2.45

Table 2.4: *Average quality ranks on a 7-point scale for all speech output conditions; 1 = low, 7 = high.*

Overall quality and suitability for application were not tested for natural speech, but we may safely assume that this form of speech output would receive near maximal scores on these dimensions. IPO's phrase concatenation clearly is the best alternative. Diphone synthesis performs worst on all dimensions.

T-tests for independent samples were conducted for pairs of speech conditions to test whether the difference in means was significant. This was done for all four questions. The motivation for choosing this type of test lies in the fact that the fragments in each speech condition are not identical. All but two contrasts were significantly different. The difference between N and IC in terms of intelligibility proved to be insignificant (t(38) = 0.303, $p > 0.05$). Additionally, the difference between CC and DS in terms of fluency proved to be insignificant (t(38) = 1.158, $p > 0.05$).

2.4.4 Conclusion

This experiment has given quantitative evidence that the quality of IPO's phrase concatenation approximates that of natural speech. It clearly pro-

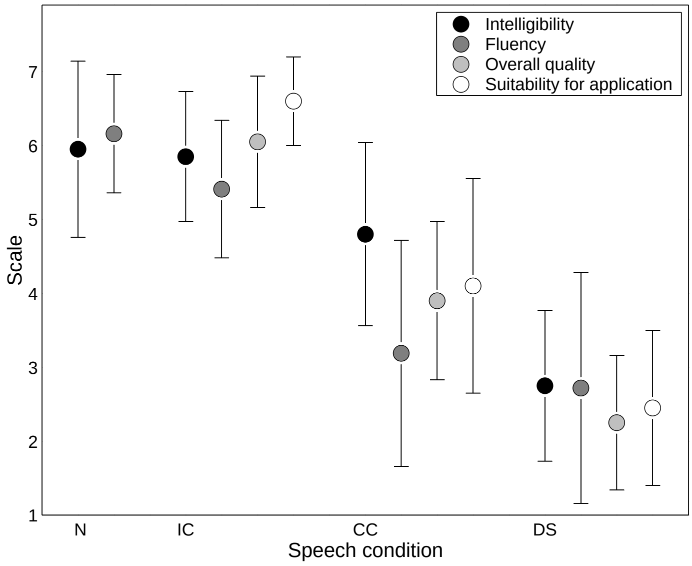


Figure 2.4: *Graphical display of average quality ranks from Table 2.4; N = natural speech, IC = IPO's phrase concatenation, CC = conventional phrase concatenation, DS = diphone synthesis.*

vides a significantly better output quality than the conventional concatenation technique. The extra ”manual” effort needed to construct a prosodically sophisticated phrase database, is definitely worthwhile. But admittedly, if the procedure cannot be substantially automated, it may not be a viable alternative.

Although the current example of conventional phrase concatenation sounded quite intelligible, it scored considerably less on fluency and overall quality. This is mainly due to the fact that all words and phrases were recorded in isolation. The mismatch in loudness between consecutive units was most remarkable. This can be solved by putting more effort into careful recording, and recording in context.

On all dimensions, diphone synthesis still has an inferior quality compared to the other techniques. However, if the intended application requires more variability than can be handled with phrase concatenation, full-fledged speech synthesis is the only technique available.

2.5 Discussion

In this chapter, it was shown that IPO's advanced technique for concatenating words and phrases while taking the prosodic context into account results in a quality of speech output that is hardly inferior to natural speech. Indeed, it appeared significantly superior to a simple concatenation scheme as used in many commercial information services. For the speech output in the German train timetable system (Aust, Oerder, Seide, and Steinbiss, 1995) an alternative version was implemented using our phrase concatenation approach. Although no formal evaluation took place, the improvement was clear (see Appendix C for examples).

This advanced approach to phrase concatenation that relies on a large number of prosodic variants is only feasible for applications in which the vocabulary is medium-sized and relatively stable, so that recordings only have to be made once. The risk with making additional recordings is that recording conditions may change, in terms of the voice characteristics of the speaker, the recording level, etc. At present, the largest drawback of our approach is that constructing the recording script and excising all necessary words and phrases manually is very time-consuming. In order to automate the design of the script for recording the necessary units in the appropriate context, an expert system is needed. This system has to make sure that coarticulation at unit boundaries is minimised such that units can be easily excised. But more importantly, it should know about rule-based variation in intonation that determines what prosodic variants have to be recorded and what the best carrier sentences are. Automation of the segmentation task is possible, but is often inaccurate in its predictions, so that manual correction has to be performed afterwards.

In the present version of OVIS, phrase concatenation is still a realistic option. But if the domain were extended to include, for instance, all bus stations in the Netherlands, this technique would become impracticable. The main bottleneck lies in the amount of speech a speaker can utter in a constant manner. Some domains, such as cinema or theatre information, have a vocabulary that is constantly changing. In these cases, speech synthesis is the only available alternative. Diphone synthesis, however, is still far removed from the quality expected by customers of an information system. Therefore, we concentrate our efforts on improving this technique.

Chapter 3

Reducing audible spectral discontinuities[1]

3.1 Introduction

One well-known problem with concatenative synthesis is the occurrence of audible discontinuities at concatenation points, which are most prominent in vowels and semi-vowels. It is due to variability in the pronunciation of these sounds which is caused by the phonetic/prosodic context.

Discontinuities are caused by mismatches in F_0, phase or spectral envelopes across concatenation points (Dutoit, 1997). In Calipso, F_0 mismatches are avoided by monotonising the diphones before storing them in the database. Phase mismatches are avoided by using a method called *phase synthesis* for re-synthesis of the nonsense words (Gigi and Vogten, 1997). Phase synthesis is based on accurate measurements of the mixture of periodic and noise information in speech. The input speech is analysed pitch-synchronously like in TD-PSOLA, but the pitch periods are estimated more precisely by means of 'first-harmonic filtering'. This forms the basis of a Discrete Fourier Transform (DFT), providing exact amplitude and phases for all harmonics. It uses overlap-and-add over two pitch periods. The signal is reconstructed by means of an amplitude and a phase value for each harmonic. Because the harmonics are added with coherent phases, phase mismatches are avoided.

[1] Part of this chapter is presented in Klabbers and Veldhuis (1998)

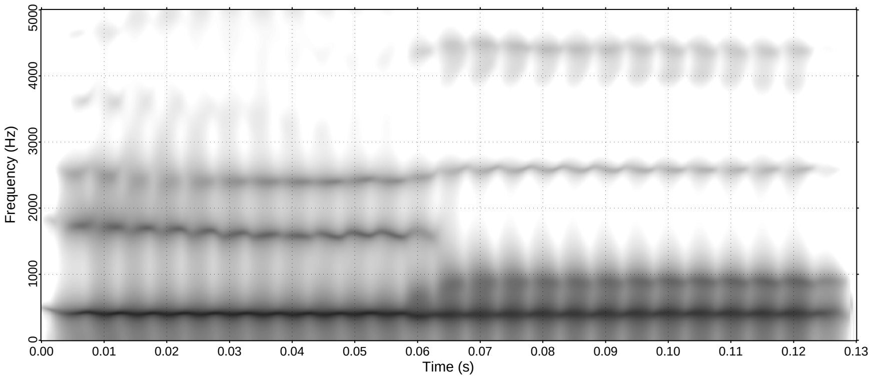


Figure 3.1: *Spectrogram for the vowel /u/ in the synthesised Dutch word* doek. *A considerable mismatch in F_2 between the left and right half of the phoneme is visible. The sudden transition at the concatenation point causes an audible discontinuity.*

Spectral mismatch is still a major problem, though. As an example, consider Figure 3.1, which shows the spectrogram for the vowel /u/ in the synthesised Dutch word *doek* (consisting of the diphones /du/ + /uk/). It reveals a considerable mismatch in F_2 between the first and second half of the phoneme. An audible discontinuity was clearly present (see Appendix C). This, along with other informal observations, suggests that spectral mismatch is the main cause for the occurrence of audible discontinuities.

In order to solve the problem of spectral mismatch, several solutions have been proposed. One approach is to use larger units such as demi-syllables or triphones. However, this does not solve the problem, as discontinuities continue to occur albeit less frequently. Moreover, the inventory size increases drastically. As Olive et al. (1998) point out, in American English assuming a 43-phone alphabet, at least 70,000 of the theoretical maximum of 79,507 triphones actually occur in the language. Even when incorporating all these units, smooth joins are not guaranteed as all triphones can occur in different contexts with strong coarticulatory effects that can even span word boundaries.

Another approach is to vary the location of the cutting point in the nonsense words dependent on the context (Conkie and Isard, 1997). This calls for a spectral distance measure that correctly represents the amount of spectral mismatch. Moreover, the short-term spectral envelopes are not constant over time, resulting for instance in non-flat formant trajectories. Figure 3.1, along with many other observations in our database, shows that formant trajectories can be fairly flat throughout a vowel when they are embedded in symmetrical nonsense words.

A third approach is to perform smoothing by means of waveform interpolation, spectral-envelope interpolation or formant trajectory smoothing. It requires specific signal representations that allow these types of operations. The disadvantage of formants as a representation is that they are very difficult to estimate reliably. Waveform and spectral envelope interpolation have the disadvantage that smooth transitions are often achieved at the expense of naturalness (Dutoit, 1997). Examples of signal representations that allow waveform interpolation are Multi-Band Resynthesis Overlap-and-Add (MBROLA, Dutoit (1997)) and Harmonic plus Noise Modeling (HNM, Stylianou, Dutoit, and Schroeter (1997)). Chappell and Hansen (1998) present several different techniques for spectral smoothing, none of which they found really satisfactory.

A fourth approach is to include context-sensitive or specialised units in the database (Olive et al., 1998). This implies that one knows which contexts can be clustered so as to keep the inventory size within bounds. Our investigation is aimed at gaining insight in this approach. In this chapter, we first present a detailed analysis of the occurrence of audible discontinuities in our diphone database (Section 3.2). The aim of the study was to find an objective spectral distance measure that best predicts when discontinuities are audible. Therefore, we correlated the results of a perceptual experiment with several distance measures. In Section 3.3 we propose a solution to reduce the occurrence of audible discontinuities by extending the diphone database with context-sensitive diphones. The number of additional diphones is limited by clustering similar contexts on the basis of the best performing distance measure.

3.2 Analysis of the problem

3.2.1 Perceptual experiment

The first step in our analysis was to find out to what extent audible discontinuities occur in our diphone database. This was established via a perceptual experiment. IPO's speech-synthesis system Calipso currently uses diphones as concatenative units from a professional female speaker. They have been excised from nonsense words. For instance, consonant-vowel (CV) and vowel-consonant (VC) diphones are excised from symmetrical nonsense words of the form C@CVC@. In order to reduce the data set to manageable proportions, this study was restricted to five Dutch vowels in this database, i.e., the vowels /a/, /i/, /A/, /I/, /u/. The vowels /a/, /i/, and /u/ were chosen because they cover the extremes in the vowel space. The vowels /A/ and /I/ are chosen because they are the short counterparts for /a/ and /i/. A study by Van den Heuvel, Cranen, and Rietveld (1996) has shown that coarticulation is speaker-specific. Therefore, it should be noted here that the results presented in this chapter only reflect the coarticulatory behaviour of the speaker of our diphone database.

Material

The stimuli consisted of concatenated left $\mathrm{C_lV}$ and right $\mathrm{VC_r}$ diphones, which were excised from the symmetrical nonsense words $\mathrm{C_l@C_lVC_l@}$ and $\mathrm{C_r@C_rVC_r@}$. The stimuli consisted of five vowel conditions in the context of all consonant pairs that can occur in $\mathrm{C_l}$ and $\mathrm{C_r}$ position (See Table 3.1). The total number of stimuli is 23 × 5 × 21 = 2415. So for instance, the diphones /du/ and /uk/ that form the stimulus /duk/ were extracted from the nonsense words *d@dud@* and *k@kuk@*. The diphones were created using the phase synthesis technique mentioned in Section 3.1. No spectral smoothing was applied at the boundary.

$\mathrm{C_l}$:	p	t	k	b	d	g	f	s	x	S	v	z	G	Z	m	n	l	L	r	j	w
	c	h																			
V:	a	A	i	I	u																
$\mathrm{C_r}$:	p	t	k	b	d	g	f	s	x	S	v	z	G	Z	m	n	l	L	r	j	w

Table 3.1: *Composition of material for the perceptual experiment; The total number of $\mathrm{C_lVC_r}$ stimuli is 2415 (23 $\mathrm{C_l}$ × 5 V × 21 $\mathrm{C_r}$).*

In the stimuli, the consonant portions were cut off to prevent them from influencing the perception of the diphone transition in the middle of the vowel. Fading was used to smooth the transition from silence to vowel and vice versa. Because all stimuli were presented in isolation, the stimulus duration had to be long enough to be able to perceive the transition at the diphone boundary. The duration of the vowels was fixed to 130 ms with the diphone boundary located exactly in the middle of the vowel. The signal power of the second diphone was scaled to match that of the first diphone. In some cases the influence of the first consonant persisted in a large part of the vowel. It proved impossible to remove it without losing too much of the vowel. These cases were discarded, leaving 2284 stimuli to be judged in the perceptual experiment.

Procedure

Five participants with a background in psycho-acoustics or phonetics participated in the perceptual experiment. It was a within-subjects design meaning that each subject received all stimuli in random order. For each stimulus, the participants had to judge the transition at the diphone boundary as either smooth (0) or discontinuous (1). The experiment was divided into three hourly sessions which were held on different days, with a short break halfway through each session. The session order was different for all participants. The experiment started with a training phase in which two stimuli were presented for each vowel, one being clearly smooth and the other being clearly discontinuous (see Appendix C).

Results

The participants found the task difficult, but felt they had been able to make consistent decisions after the training phase. As a consistency check, we presented two stimuli, one clearly smooth, the other clearly discontinuous, ten times at random positions in the total stimulus list. All participants were 100% consistent in their scoring of these two stimuli. Between participants there was more variability, as some participants applied a stricter threshold than others. In order to reduce the variability between participants, a majority score was calculated, i.e., a stimulus was marked as discontinuous when four out of five listeners perceived it as such. Summing the majority scores obtained in the experiment for each of the vowels, we get the percentage of perceived discontinuities as presented in Table 3.2.

The results show that the number of audible discontinuities is particularly high for /u/ and comparatively low for /a/. Our results additionally reveal a slightly better score for the long vowels /a/ and /i/ than for the short vowels /A/, /I/ and /u/. This is partly in line with findings by

Vowel	Percentage of perceived discontinuities	Number of observations
/a/	17.1%	474
/i/	43.1%	445
/A/	52.1%	468
/I/	55.5%	449
/u/	73.9%	448

Table 3.2: *Percentage of perceived discontinuities for each vowel. The percentages are computed from the sum of the majority scores.*

Van den Heuvel, Cranen, and Rietveld (1996). They investigated speaker variability in the coarticulation of /a/, /i/ and /u/. Their results show that the /u/ has the greatest amount of coarticulation and the smallest amount of coarticulation was found for /i/, closely followed by /a/.

3.2.2 Spectral distance measures

The second step in our investigation was to correlate the results from the perceptual experiment with several spectral distance measures in order to obtain an objective measure for predicting audible discontinuity. In speech recognition and speech coding, spectral distance measures have been widely used. In automatic speech recognition, one of the earliest studies comparing several distance measures was conducted by Gray and Markel (1976). They investigated measures based on spectral and cepstral coefficients, log area ratios and the Itakura-Saito distance. They obtained the best performance with the Root-Mean-Squared (RMS) Log Spectral Distance. Hermansky and Junqua (1988) and Krishnan and Rao (1996) showed that using warped frequency scales (such as Mel-scale or Bark-scale) improved the performance of speech recognisers even further. The most commonly used distance measure in automatic speech recognition is the Euclidean distance between Mel-Frequency Cepstral Coefficients (MFCC).

In speech synthesis, this Euclidean MFCC distance has also been adopted in order to select optimal units or segment diphones at the optimal cutting point (among others by Conkie and Isard (1997) and Carvalho et al. (1998)). The question is whether a measure used in speech recognition is equally suitable for use in speech synthesis, as it serves a different purpose. In speech recognition, the task is to classify different instances of one and the same phoneme as belonging to the same target phoneme, whereas in speech synthesis the task is to distinguish these instances when their spectra are perceptually different. Therefore, it should be investigated whether some distance measures can be found that correspond to human perception in that they are able to distinguish perceptually relevant differences in spectra (Van Santen, 1997).

An investigation that ran parallel to ours (Wouters and Macon, 1998; Macon, Cronk, and Wouters, 1998), also aimed at performing a perceptual evaluation of distance measures in the context of speech synthesis. In their study, listeners had to judge the difference between a pair of stimuli on a scale from zero to five. One stimulus was the reference stimulus produced by a diphone synthesis system, the other stimulus was altered in that the first (c.q. second) half of the vowel phoneme was replaced by a different instance of the vowel preceded (c.q. followed) by a consonant from the same class as the original. Five feature representations were studied: FFT-based cepstra, LPC-based cepstra, Line Spectral Frequencies (LSF), Log Area Ratios (LAR) and a symmetrised Itakura distance. All but the FFT-based cepstra were computed from LPC coefficients. The feature representations were computed in three ways: 1) using the FFT amplitude spectrum, 2) using a perceptual spectrum (PLP, Hermansky (1990)), 3) using a Mel-warped spectrum. Correlations between the average of the listeners' responses and the distance measures were computed and then combined into a population correlation using Fisher's z-transform. Correlations were not particularly high, reaching from 0.28 for the linear log area ratio to 0.50 for the linear Itakura distance. PLP and Mel-scale improved the correlations, but the improvement from PLP to Mel was not significant. Using a weighted Euclidean distance improved the linear measures, but only slightly for PLP and Mel. Delta features gave only a small increase in correlation (0.02). The best correlation was obtained for Mel cepstra with delta features (0.66), where it did not make a difference whether these were computed from FFT or LPC coefficients.

The measures used in this thesis are taken from various fields of research. They were used to determine distances between spectral envelopes across diphone boundaries. The following spectral distance measures were used. They will be explained in more detail below.

1. The Euclidean distance between (F_1, F_2) pairs, or the Euclidean Formant Distance (D_{EFD}), which is often used in phonetics.

2. The Kullback-Leibler distance (D_{KL}), which originates from the field

of statistics.

3. The Partial Loudness D_{PL}, which comes from the area of sound perception.

4. The Euclidean distance between Mel-Frequency Cepstral Coefficients (D_{MFCC}), which comes from automatic speech recognition.

5. The Likelihood Ratio (D_{LR}), which is used in speech coding and automatic speech recognition.

6. The Mean-Squared Log-Spectral Distance (D_{MSLSD}), which also comes from automatic speech recognition.

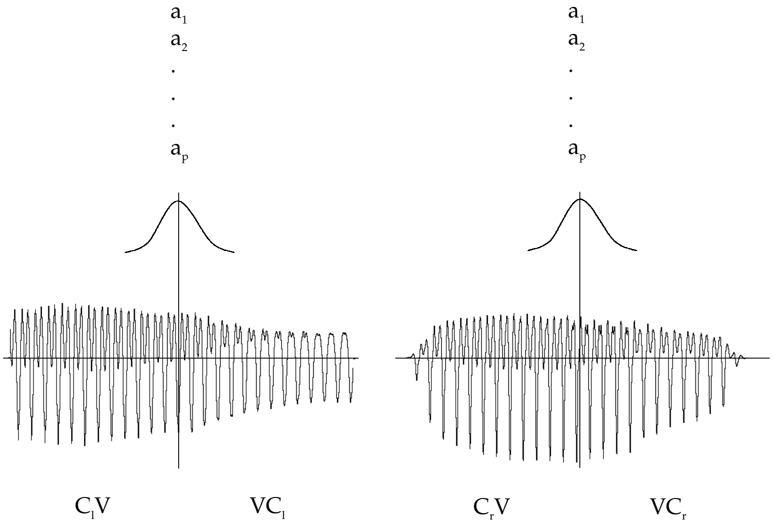


Figure 3.2: *Computation of LPC coefficients at the diphone boundary in the CVC-part of the nonsense words, using a 40 ms Hanning window.*

All spectral distances excepting the Euclidean Formant Distance, were calculated from LPC-spectral envelopes. First, two sets of LPC coefficients $a_1, \ldots, a_{\mathrm{p}}$ (p = 14) were computed (see Figure 3.2). One set is measured at the right diphone boundary of the $\mathrm{C_lV}$ diphone in the nonsense word $\mathrm{C@C_lVC_l@}$, which also produces the diphone $\mathrm{VC_l}$. The other set of LPC coefficients is measured at the left diphone boundary of the VC$_r$ diphone in the nonsense word C@C$_r$VC$_r$@, which also produces the diphone C$_r$V. From those LPC coefficients, two power spectra were computed at sufficiently many equidistant points (in our case 512), which were arranged in a vector. These two vectors were normalised by dividing them by the sum of the elements. This gives two power-normalised spectral envelopes P(ω) of C$_l$V and Q(ω) of VC$_r$. Equation 3.1 displays the computation of the power-normalised spectral envelope P(ω), where K = constant, such that $\int P(\omega)\mathrm{d}\omega = 1$. The distance measures will be discussed in more detail in the following paragraphs.

$$\mathrm{P}(\omega) = \frac{K}{|1 + a_1 e^{-j\omega} + a_2 e^{-2j\omega} + \cdots + a_p e^{-pj\omega}|^2}. \tag{3.1}$$

Euclidean Formant Distance

In phonetics, it is quite common to describe coarticulation in terms of changes in the formants F$_1$ and F$_2$ (Van den Heuvel et al., 1996; Olive et al., 1998). In this investigation, the formants were measured automatically at the diphone boundary in the vowel using Praat (Boersma and Weenink, 1996). Close inspection of the stimuli reveals that most formant trajectories are fairly stationary throughout the vowel, except when the surrounding consonants are the alveolars /j/, /J/, /c/, /S/ and /Z/.

Figure 3.3 displays the F$_1$ and F$_2$ values for the five vowels measured in our diphone database at the indicated locations. It shows that the /a/, /i/ and /I/ have small variations, whereas the /A/ and /u/ seem to be affected to a greater extent by their surrounding consonants. For the /u/, differences in F$_2$ are considerable. They can be as large as 700 Hz. For the /A/, differences in F$_1$ are also considerable. The /i/ and /I/ are very close to each other in the vowel space. All formant frequencies were transformed to a Mel-scale which is more in line with the hearing process than a linear frequency scale. For the Mel transformation we used Equation 3.2 (Makhoul and Cosell, 1976).

$$m = 2595 log(1 + \frac{f}{700}) \tag{3.2}$$

The Euclidean Formant Distance is calculated by Equation 3.3.

$$D_{\mathrm{EFD}}(l, r) = \sqrt{(F_{1,l} - F_{1,r})^2 + (F_{2,l} - F_{2,r})^2}. \tag{3.3}$$

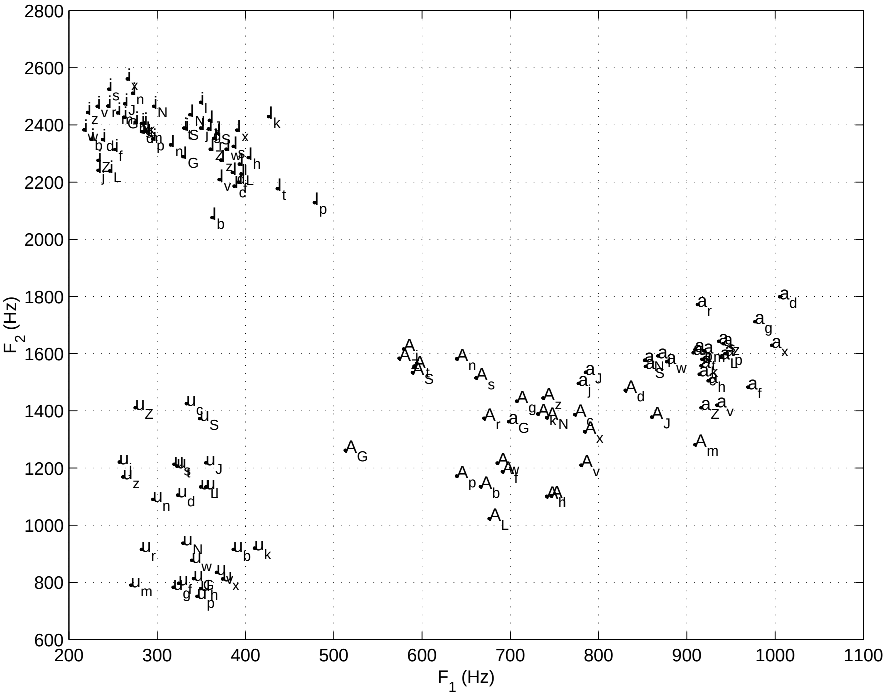


Figure 3.3: *Vowel space in terms of F_1 and F_2 for the five vowels /a/, /A/, /i/, /I/, and /u/ in different (symmetric) consonantal contexts as indicated by the subscripts.*

Kullback-Leibler distance

The Kullback-Leibler (KL) distance or *relative entropy* is a measure taken from statistics (Kullback and Leibler, 1951), where it is used to compute the distance between two probability distributions. Here, it is calculated from the two power-normalised spectral envelopes $\mathrm{P}(\omega)$ and $\mathrm{Q}(\omega)$ that were explained earlier. The original asymmetrical definition of the KL distance is changed into a symmetrical version. The main reason is that $D_{\mathrm{KL}}(P(\omega), Q(\omega))$ then equals $D_{\mathrm{KL}}(Q(\omega), P(\omega))$, which is convenient because all vowels are recorded in nonsense words with the same left and right consonant context. This means that the diphones $\mathrm{C_lV}$ and $\mathrm{VC_r}$, where $l = r$, come from the same nonsense word, so concatenating them should result in a zero spectral distance.

The KL distance has the important property that it emphasises differences in spectral regions with high energy more than differences in spectral regions with low energy. Thus, spectral peaks are emphasised more than valleys between the peaks and low frequencies are emphasised more than high frequencies, due to the 6 dB/octave declination in spectral energy that results from the combination of the damping of the high frequency components in the signal (-12 dB/octave) and the radiation at the mouth (+6 dB/octave). The definition for $D_{\mathrm{KL}}(P, Q)$ is given in Equation 3.4.

$$D_{\mathrm{KL}}(P, Q) = \int (P(\omega) - Q(\omega)) \log \left(\frac{P(\omega)}{Q(\omega)} \right) \mathrm{d}\omega. \tag{3.4}$$

Partial loudness

The partial loudness comes from the area of sound perception. In a study by Van Dinther, Rao, Veldhuis, and Kohlrausch (1999), partial loudness was shown to be a reasonably good predictor for audibility discrimination thresholds. Therefore, it was decided to include this measure to see how well it would predict audible discontinuity. The partial loudness of a signal is the loudness of the signal when presented in a background sound (Moore, Glasberg, and Bear, 1997). The background sound generally reduces the loudness of the signal. This effect is called partial masking. The loudness of a signal in a background sound is therefore called partial loudness. In this study, the excitation patterns E_{l} and E_{r} at the $\mathrm{C}_{\mathrm{l}}\mathrm{V}$ and VC_{r} diphone boundaries were computed. These excitation patterns are decomposed into excitation patterns representing the background sound ($min(E_{\mathrm{l}}, E_{\mathrm{r}})$), the total sound ($max(E_{\mathrm{l}}, E_{\mathrm{r}})$) and the absolute difference ($|E_{\mathrm{l}} - E_{\mathrm{r}}|$). The resulting excitation patterns are then fed into Moore's partial loudness model.

Mel-Frequency Cepstral Coefficients

In the field of speech recognition, Mel-Frequency Cepstral Coefficients (MFCC) are currently the most commonly used type of signal representation. They provide a successful and efficient way to represent the signal for the purpose of recognition. The MFCC coefficients were computed as described in Rabiner and Juang (1993), Chapter 4, except that samples of the LPC power spectrum (see definition for P(ω) above) were used instead of the squared magnitudes of the DFT spectrum. Equation 3.5 computes the Euclidean distance between cepstral coefficients.

The cepstral coefficients can be interpreted in the following way. The c_0 coefficient represents the average energy in the speech frame. It is not included in the distance measure. c_1 reflects the energy balance between low and high frequencies (or *spectral tilt*), higher values indicating sonorants and low values suggesting frication. Higher order cepstral coefficients reflect increasing spectral detail, but no simple relationship exists between these parameters and formants (Hunt, 1995). In this study, the order p is set to 22.

$$D_{\mathrm{MFCC}} = \sum_{k=1}^{p} (c_{\mathrm{k,l}} - c_{\mathrm{k,r}})^2. \tag{3.5}$$

Delta features are estimations of the time derivatives of the static features, thus capturing more of the speech dynamics. Because our stimuli consist of two parts that are both fairly stationary, the addition of delta features will not make much difference to the result. In the study presented in Wouters and Macon (1998) and Macon, Cronk, and Wouters (1998), the added effect of delta features was also negligible. The correlation between the distance measure and the perceptual judgements increased with just 0.02. Therefore, it was decided not to include delta features in our investigation.

Likelihood ratio

The likelihood ratio or *Itakura distance*, is a measure of spectral similarity between two LPC vectors a_L and a_R, which represent the left and right diphones (Itakura, 1975). It indicates how well the analysis filter of the left diphone matches that of the right diphone. It is defined in terms of an autocorrelation function. V_R represents the signal autocorrelation matrix that gave rise to a_R. The likelihood ratio is computed with Equation 3.6 taken from Rabiner and Juang (1993), Chapter 4.

$$D_{\mathrm{LR}}(a_L, a_R) = \frac{a'_L V_R a_L}{a'_R V_R a_R} - 1. \tag{3.6}$$

Mean-Squared Log Spectral Distance

The Mean-Squared Log Spectral Distance (MS LSD) is derived from the Log Spectral Distance presented in Rabiner and Juang (1993), Chapter 4 and is computed by Equation 3.7. It is similar to the Root-Mean-Squared

(RMS) Log Spectral Distance that performed best in the automatic speech recognition experiments by Markel and Gray (1976). By taking the logarithmic differences between P(ω) and Q(ω), it is expected to better reflect the hearing process.

$$D_{\text{MSLSD}} = \int (logP(\omega) - logQ(\omega))^2 \mathrm{d}\omega. \tag{3.7}$$

3.2.3 Correlating the results

In order to find out how well the different spectral distance measures could predict audible discontinuities, it was decided to use Receiver Operator Characteristic curves (Luce and Krumhansl, 1988), coming from signal detection theory. The procedure works as follows. In order to correlate the measures with the scores of the listeners, two probability density functions $p(D|0)$ and $p(D|1)$ are estimated from the data, representing the probability of a spectral distance (D) given that the transition was marked by the listeners as continuous (0) or discontinuous (1), respectively. For a certain threshold β, the probability of a *false alarm*, the case that a transition is wrongly classified as discontinuous, is $\mathrm{P_F}(\beta)$ (see Equation 3.8) and the probability of a *hit*, the correct detection of a discontinuity, is $\mathrm{P_D}(\beta)$ (see Equation 3.9).

$$P_{\mathrm{F}}(\beta) = \int_{\beta}^{\infty} p(D|0)\mathrm{d}D. \tag{3.8}$$

$$P_{\mathrm{D}}(\beta) = \int_{\beta}^{\infty} p(D|1)\mathrm{d}D. \tag{3.9}$$

The probability of a *miss*, the case that a discontinuity goes undetected, is 1 - $\mathrm{P_D}$ and the probability of a *correct rejection*, the case that a transition is rightly classified as being smooth, is 1 - $\mathrm{P_F}$. Since these are directly derivable from the hit and false alarm probabilities, they are not relevant here.

A plot of pairs ($P_{\mathrm{D}}(\beta)$, $P_{\mathrm{F}}(\beta)$) for all values of β constitutes a Receiver Operating Characteristic (ROC) curve. See Figure 3.4 for a schematic representation of ROC curves. In this experiment, we assumed that the participants were correct in their judgements. The question is then how well a spectral distance measure can predict the participants' judgements. ROC curves are always upward concave. The straight line represents the chance level meaning that a measure gives no information. The further the curve extends to the upper left corner, the better the measure serves as a predictor. This indicates that the two probability density functions $p(D|0)$ and $p(D|1)$ are moving away from each other, thus increasing the hit rate and

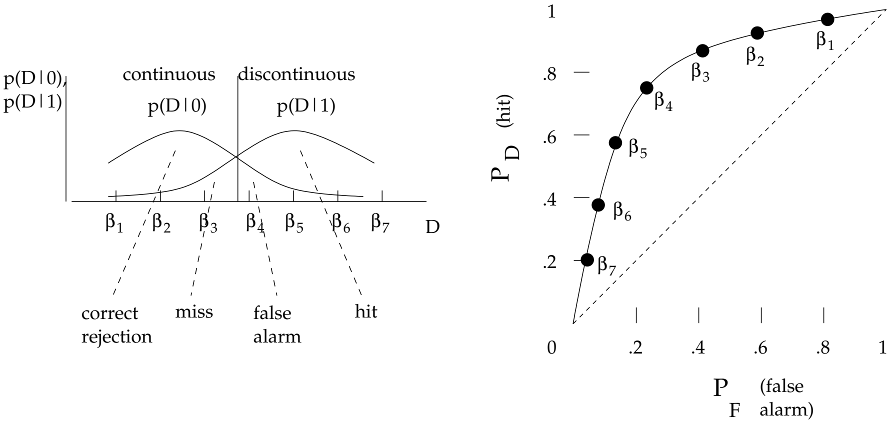


Figure 3.4: *Principle of Receiver Operating Characteristic curves. The left panel displays two probability density functions $p(D|0)$ and $p(D|1)$, for the distribution of distances given that the participants have judged a stimulus as smooth and discontinuous, respectively. The right panel displays a Receiver Operator Characteristic curve, formed by plotting the false alarm probabilities against the hit probabilities for different threshold values β.*

decreasing the false alarm rate. We do not have to decide on an appropriate threshold in this study, since the purpose of the analysis is solely to determine the best performing distance measure relative to the other measures.

3.2.4 Results

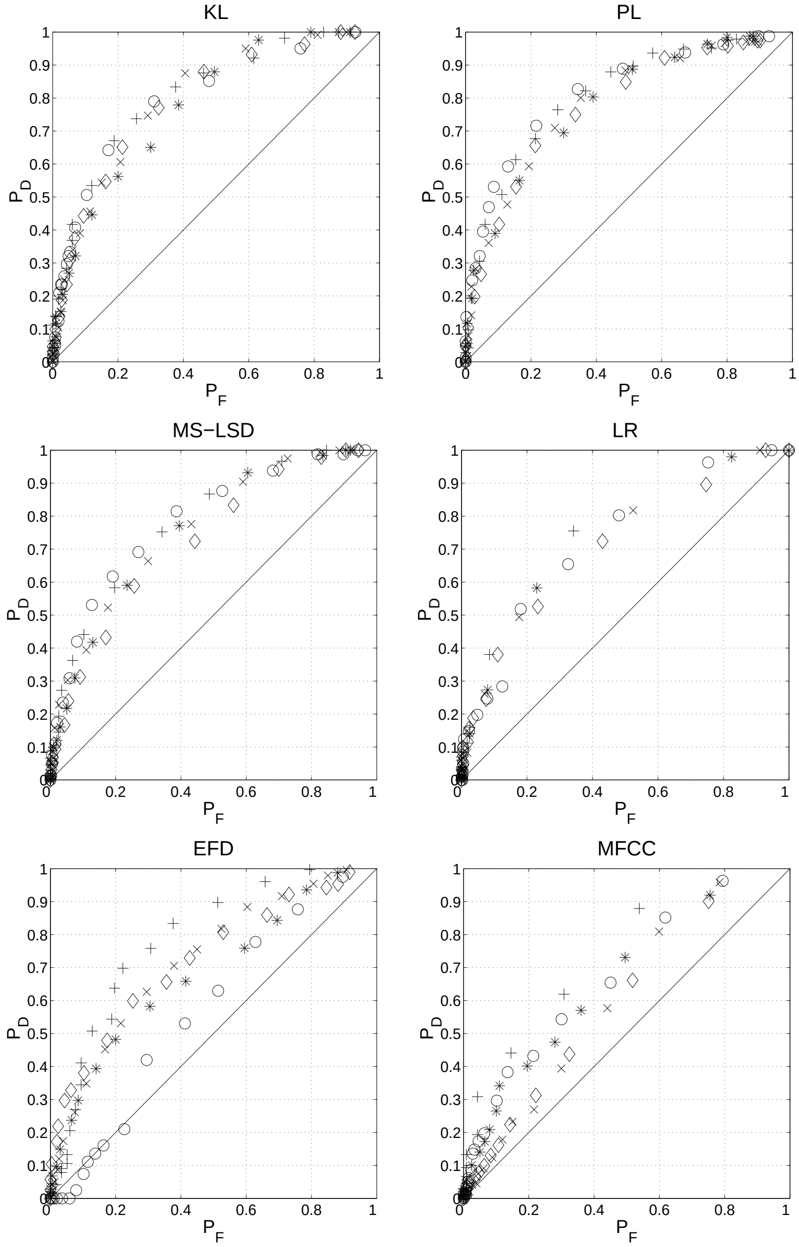
Figure 3.5 displays five ROC curves per distance measure, one for each vowel. Their inspection leads to a number of interesting observations. First, it can be observed that the KL and PL distances perform equally well for all vowels, whereas the divergence between vowels is greater for the other measures.

Second, it can be seen that the Euclidean Formant Distance performs well for /u/, bad for /a/ and moderately well for the other vowels. This is understandable as Figure 3.3 showed that /u/ had the largest degree of variation in the second formant. Extending the distance with F_3 and F_4 did not enhance the measure in any way. It was decided not to include formant bandwidths in the distance measure. As listed in Rabiner and Juang (1993) the just-noticeable-difference (jnd) for formant bandwidths is much larger (20-40%) than for formant frequencies (3-5%). It was therefore not expected to add much to the performance of the EFD.

Figure 3.6 displays six ROC curves per vowel, one for each distance measure. Here, it can clearly be seen that all spectral distance measures perform almost equally well for the /u/, whereas the divergence between them is much larger for the other vowels.

The Euclidean distance between Mel-Frequency Cepstral Coefficients is consistently among the worst predictors of audible discontinuity. In some cases, it is barely above chance level. This is a surprising result, because until now it was a commonly used distance measure for this task. However, its bad performance is understandable when we consider that this measure is almost standard as a distance measure in automatic speech recognition, where its task is to group together different allophones of a phoneme instead of distinguishing them when their spectral characteristics lead to perceptual differences.

Two measures are always positioned in the most upper left corner, meaning that they always performed best, the Partial Loudness and the Kullback-Leibler distance. Apparently, they correlate well with human perception. The fact that the relatively simple Euclidean Formant Distance only performs well for /u/ but not as well for other vowels, indicates that coarticulation affects the vowels differently. It could be that coarticulation has a larger effect on the formants for /u/, whereas in the other vowels it manifests itself in other ways, e.g., via changes in spectral tilt. Additionally, the fact that the KL and PL distances *are* good predictors of audible discontinuity for all vowels, shows that there is systematic variation in the signal due to coarticulation that these measures capture. Because the KL distance is much easier and faster to compute than the PL distance, it will be used in the remainder of this study.



○ = /a/, × = /A/, ◇ = /i/, * = /I/, + = /u/.

Figure 3.5: *ROC curves grouped per vowel for the measures KL, PL, MSLSD, LR, EFD, MFCC.*

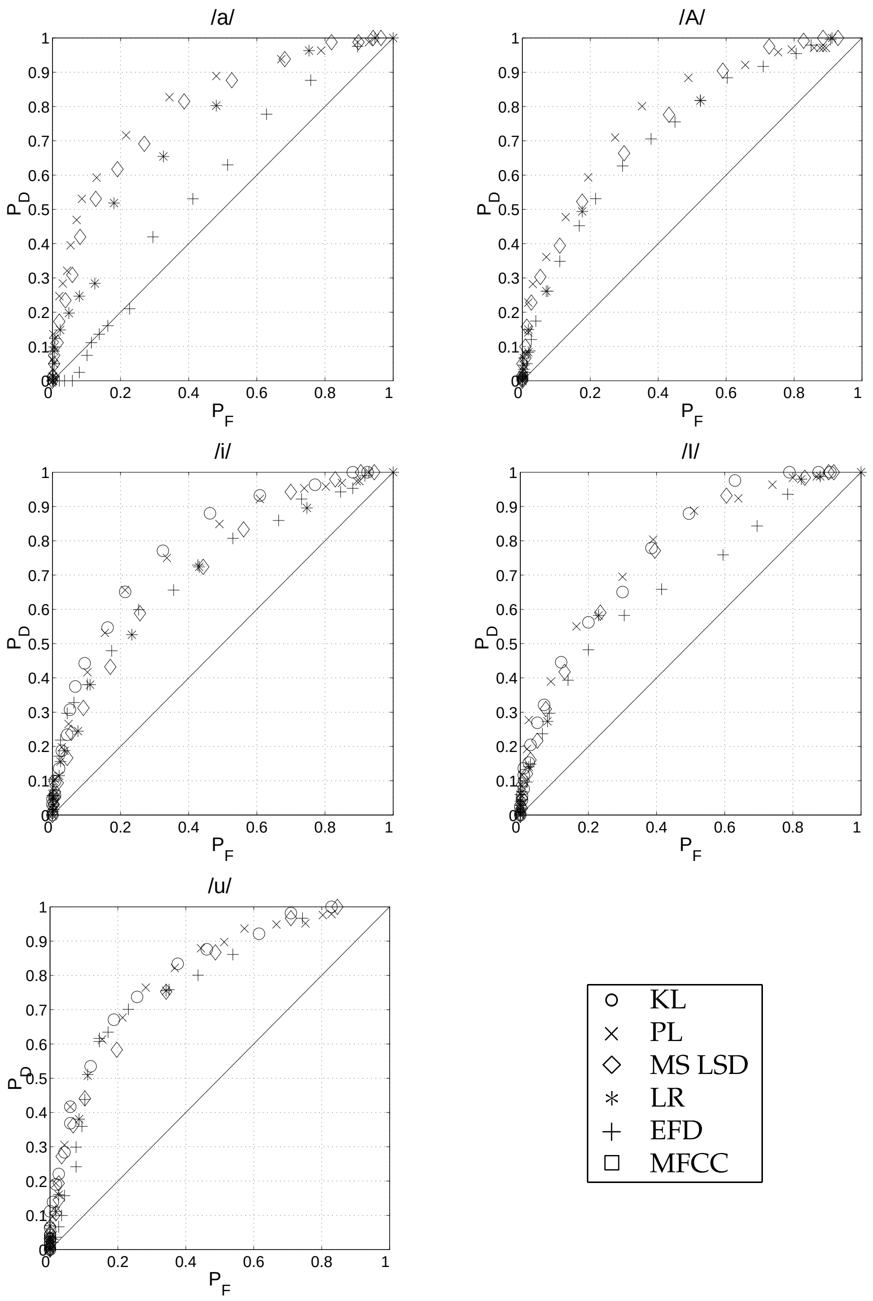


Figure 3.6: *ROC curves grouped per measure for the vowels /a/, /A/, /i/, /I/, and /u/.*

3.3 A solution to the problem

3.3.1 Clustering procedure

In order to reduce the number of audible discontinuities, we propose to extend the diphone database with context-sensitive diphones. One way of keeping the inventory size within bounds is to cluster contexts that are spectrally alike according to a distance measure. In this part of the study, the KL distance is used for this purpose. Suppose we divide the diphone sets $\mathrm{C_lV}$ ($l = 1, \ldots, M$) and $\mathrm{VC_r}$ ($r = 1, \ldots, M$), for a particular vowel V into two sets of N clusters $\{L(V)_1, \ldots, L(V)_\mathrm{N}\}$ and $\{R(V)_1, \ldots, R(V)_\mathrm{N}\}$, such that the maximum KL distance across diphone boundaries in corresponding clusters $L(V)_\mathrm{k}$ and $R(V)_\mathrm{k}$ ($k = 1, \ldots, N$) is below a threshold β. The maximum distance between non-corresponding clusters $L(V)_\mathrm{k}$ and $R(V)_\mathrm{l}$ ($k \neq l$) will then not be limited to β. We now construct additional clusters $R(V)_\mathrm{l,k}$ ($k \neq l$), which contain the diphones of $R(V)_\mathrm{l}$, but which are recorded with a left-side context consisting of a representative diphone in $L(V)_\mathrm{k}$, e.g., the diphone closest to the centroid of $L(V)_\mathrm{k}$. Instead of concatenating a diphone from $L(V)_\mathrm{k}$ with one from $R(V)_\mathrm{l}$, a diphone from $R(V)_\mathrm{l,k}$ will be used, which will reduce the maximum KL distance across diphone boundaries, which is hopefully lower than β, although a guarantee cannot be given in advance. This procedure will increase the number of VC diphones for a particular vowel by a factor $\mathrm{N}(\beta)$, which is equal to the number of clusters.

The number of clusters $\mathrm{N}(\beta)$ can under certain assumptions statistically be related to the quality improvement. If the maximum KL distance between corresponding clusters is β, then the total number of transitions between all corresponding clusters is less or equal than the total number of diphone combinations with maximum KL distance β. This is expressed in the right-hand inequality of Equation 3.10

$$\frac{M^2}{N(\beta)} \leq \sum_{i=1}^{N(\beta)} m_i^2 \leq \int_0^\beta p(D)\mathrm{d}DM^2, \tag{3.10}$$

with $\int_0^\beta p(D)\mathrm{d}D$ being the probability of a distance smaller than β and M^2 being the total number of transitions. It can furthermore be shown that the total number of transitions between corresponding clusters is always larger than $M^2/N(\beta)$. This is expressed in the left-hand inequality of Equation 3.10. In fact, the minimum is attained when all clusters have equal sizes.

This leads to the following inequality for $N(\beta)$:

$$N(\beta) \geq \frac{1}{\int_0^\beta p(D)\mathrm{d}D\mathrm{d}D} \tag{3.11}$$

$$\geq \frac{1}{\int_0^\beta (p(D|0)P\{0\} + p(D|1)P\{1\})\mathrm{d}D}. \tag{3.12}$$

This factor can be used as a measure of cost of improvement. After the extension of the diphone inventory, the probability of an audible discontinuity occurring, $\mathrm{P}_{\mathrm{click}}$, can be computed. It constitutes two independent probabilities, one representing the probability of a discontinuity occurring in a cluster that is not detected (given by $P\{1\}[1 - P_{\mathrm{D}}(\beta)]$), and the other representing the probability of a detected discontinuity occurring between the original left diphone and the newly added right diphone, which is estimated to be maximally $P_{\mathrm{D}}(\beta)(1 - P_{\mathrm{D}}(\beta))$, assuming that the distance between the new cluster and the original one is less than β. Before actual extension of the database, we can compute $\mathrm{P}_{\mathrm{click}}$ by Equation 3.14.

$$P_{\mathrm{click}}(\beta) = P\{1\}[1 - P_{\mathrm{D}}(\beta) + P_{\mathrm{D}}(\beta)(1 - P_{\mathrm{D}}(\beta))] \tag{3.13}$$

$$= P\{1\}[1 - P_{\mathrm{D}}^2(\beta)]. \tag{3.14}$$

Figure 3.7 plots the number of clusters $N(\beta)$ against the probability of a click $\mathrm{P}_{\mathrm{click}}$. This represents an estimate of the maximum improvement that can be obtained by adding a certain number of clusters. The threshold β can now be chosen according to cost or performance constraints. Figure 3.7 shows that when using three clusters, the probability of an audible discontinuity occurring is predicted to maximally decrease from 0.17 to 0.04 for /a/, from 0.43 to 0.12 for /i/, from 0.52 to 0.13 for /A/, from 0.55 to 0.17 for /I/ and from 0.74 to 0.29 for /u/.

Figure 3.8 illustrates the clustering procedure for the vowel /u/. In our investigation, the maximum number of clusters is restricted to three, which contain the same consonantal contexts for left and right diphones. Adding more than three clusters is not likely to improve synthesis much more. Concatenating /bu/ and /uk/ to make *boek* is expected to be unproblematic, as both consonants come from the same cluster with a small KL distance. However, for the words *doek* and *hoek* an audible discontinuity is likely to occur as they come from non-corresponding clusters. In order to remedy this, additional right diphones are recorded with a representative from a non-corresponding left cluster. In recording this means that for the /uk/ diphone which was originally recorded in the symmetrical nonsense

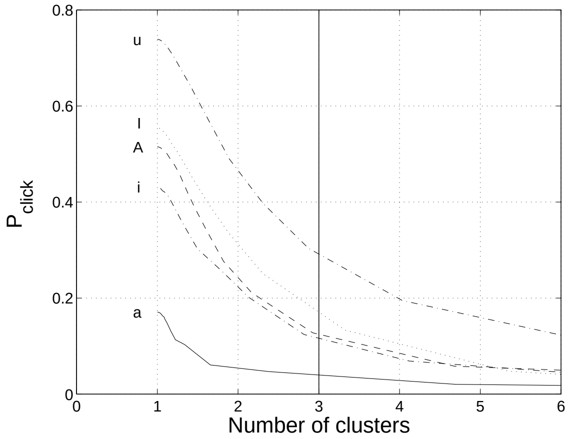


Figure 3.7: *Number of clusters versus P_{click}. The lines represent the lower bound, i.e., the maximum improvement that might be obtained by using additional diphones.*

word *k@kuk@*, two additional diphones are recorded in the asymmetrical nonsense words *l@luk@* and *f@fuk@*. Then, the word *doek* can be created by concatenating the original left diphone /du/ with the new diphone /uk/ taken from *l@luk@* and the word *hoek* is created by concatenating the same /du/ diphone with the new right diphone /uk/ coming from *f@fuk@*.

The clusters are constructed according to a classification algorithm, derived from the LBG algorithm (Veldhuis and Breeuwer, 1993), which is commonly used for codebook generation for the purpose of vector quantisation. The KL distance is used as a criterion for the division. A Distance Matrix (DM) is constructed with $\mathrm{C_lV}$ diphones in the rows and $\mathrm{VC_r}$ diphones in the columns. Since the KL distance is symmetrical, the diagonal, where $\mathrm{C_l}$ equals $\mathrm{C_r}$, contains zeros. The clustering procedure works as follows:

1. Three $\mathrm{C}_l\mathrm{V}$ diphones are chosen as the initial representatives of the clusters.

2. The distance matrix is reduced to a cluster matrix with KL distances between the three representatives and the $\mathrm{VC_r}$-diphones. Each $\mathrm{VC_r}$-diphone is added to the cluster to which representative it has the lowest KL distance.

3. The initial representative does not necessarily have the lowest av-

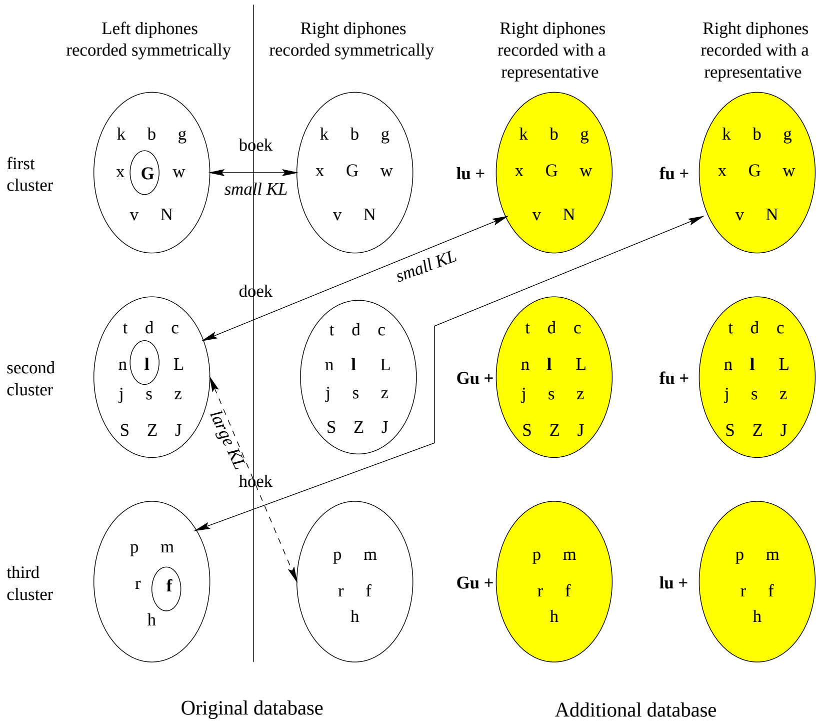


Figure 3.8: *The principle of the construction of additional diphone clusters. The context-sensitive diphone-clusters are indicated in grey. They consist of VC*$_\mathrm{r}$ *diphones that were recorded with a representative from a non-corresponding left cluster. The representative of each cluster is circled.*

erage KL distance to all other diphones in the cluster, so for each cluster a new representative is chosen that does adhere to this criterion. Then steps 2 and 3 are repeated until the cluster configuration converges.

All possible combinations of initial representatives were tried. The best ones, i.e., the ones that lead to the lowest maximal distance in a cluster, are displayed in Table 3.3. Weighting the average KL distance with the frequency of occurrence of each diphone as measured in a large text corpus (27,000 sentences) did not have any effect on the configuration. It was decided to allow the occurrence of just one diphone in a cluster, to be able to separate an outlier that has a great spectral distance to all other diphones.

For the /a/, the /S/ ends up in a cluster on its own. There is no clear pattern related to manner or place of articulation of the consonants, except for the /u/ where all alveolars end up in the same cluster. We will come back to this issue in the discussion. The cluster configuration for /u/ was already visualised in Figure 3.8

Vowel	Consonants in cluster	Maximum KL	Average KL
/a/	1: v bfwz	1.08	0.45
	2: S	0.00	0.00
	3: x GJLNcdghjklmnprstz	1.31	0.47
/i/	1: k GNfgnpstx	2.40	0.90
	2: b JLZdjmrvwz	2.79	0.86
	3: S chl	1.85	0.89
/u/	1: G Nbgkvwx	2.36	1.08
	2: l JLSZcdjnstz	2.02	0.80
	3: f hmpr	2.39	0.90

Table 3.3: *Cluster configurations for /a/, /i/, and /u/. The representatives in each cluster are the first consonants in each row.*

3.3.2 Second perceptual experiment

In order to measure the improvement that results from the addition of context-sensitive diphones new recordings were made with which a new perceptual experiment was performed. In order to make comparison possible, the new recordings were contained both the old symmetrical and the new asymmetrical nonsense words.

Material

Again, stimuli were created consisting of concatenated C_lV and VC_r-diphones, except that now for each C_lV and VC_r combination there were two versions, one with a right diphone from the symmetrical nonsense word C_r@C_rVC_r@ (database without clustering) and one with a right diphone from the asymmetrical nonsense word C_{rep}@$C_{rep}VC_r$@ (database with clustering). In order to reduce the total number of stimuli, it was decided to focus on just three vowels /a/, /i/, and /u/. The total number of stimuli used in the experiment is 2254, of which 1449 were constructed according to the original concatenation method (23 C_l × 3 V × 21 C_r) and 805 diphone combinations were obtained using diphones from the context-sensitive database (202 for /a/, 295 for /i/ and 308 for /u/, based on three clusters).

Procedure

The perceptual experiment was repeated, this time using six participants with a background in psycho-acoustics or phonetics. They had not taken part in the previous experiment. The participants again had to judge whether the diphone boundary in the middle of the vowel was either smooth or discontinuous. The stimuli were presented in three hourly sessions which were held on three different days. Each session was split into two 30-minute blocks by a 15-minute break. The session order was different for all participants.

Results

Table 3.4 lists the percentage of perceived discontinuities for the new database with and without clustering. Again, these are based on the majority scores (4 out of 5 listeners agree). Since we had six participants in this experiment, one randomly chosen subject was left out to keep the results comparable to the old situation. The results for the new database without clustering is better than for the original database. This can be a result of more careful pronunciation on the part of the speaker. Nevertheless, it can be seen that clustering does reduce the number of audible discontinuities.

Vowel	New database without clustering	New database with clustering
/a/	7.7%	6.0%
/i/	19.3%	17.0%
/u/	53.4%	26.3 %

Table 3.4: *Percentage of perceived discontinuities for each vowel. The percentages are computed from the sum of the majority scores.*

Its significance is demonstrated by a paired-samples t-test which was performed on the KL distance and on the summed participants' scores. The results are presented in Table 3.5. When looking at the results for the KL distance one can observe that the distance has significantly decreased for context-sensitive diphones for both /i/ and /u/, but is not significant for /a/. However, in the judgement of the participants, the number of detected discontinuities has significantly decreased for all three vowels.

Vowel	KL distance	Sum of participants' scores
/a/	t(482) = 0.524, p = 0.601	t(482) = 2.772, $p < 0.05$
/i/	t(482) = 6.326, $p < 0.05$	t(482) = 2.618, $p < 0.05$
/u/	t(482) = 8.121, $p < 0.05$	t(482) = 12.47, $p < 0.05$

Table 3.5: *Paired samples t-test on new database with and without clustering for KL distance and summed participants' scores.*

The effects found for the KL distance are nicely illustrated in Figure 3.9, which displays the distribution of KL distances in rank order. For /a/ indeed it can be seen that clustering does not have any effect on the distribution of the KL distances, probably because the average distance was low to begin with. For instance, when taking a threshold β of 1.5, 400 out of 500 stimuli lie below this threshold. For /i/ and /u/, the improvement is clear. Taking the same β it can be observed that the number of stimuli below threshold increases from approximately 250 to 320 for /i/ and from 280 to 350 for /u/.

When comparing the improvement prediction for the new database without clustering with the actual improvement obtained by clustering (Figure 3.10), one can see that clear improvements are obtained although they are not as good as the maximally predicted improvement. The deviation from the optimal line is 0.07 for /u/ and /a/ and 0.15 for /i/. One reason for this is that the maximum improvement is estimated assuming that each cluster contains an equal amount of contexts, which is not the case here.

When again considering the specific example of /duk/, it can now be seen that adding an additional /uk/ diphone that has been recorded in the appropriate context makes a well visible difference (see Figure 3.11). Instead of an abrupt and large jump in the F_2 as observed in Figure 3.1, the F_2 descends more gradually. These examples are described in Appendix C.

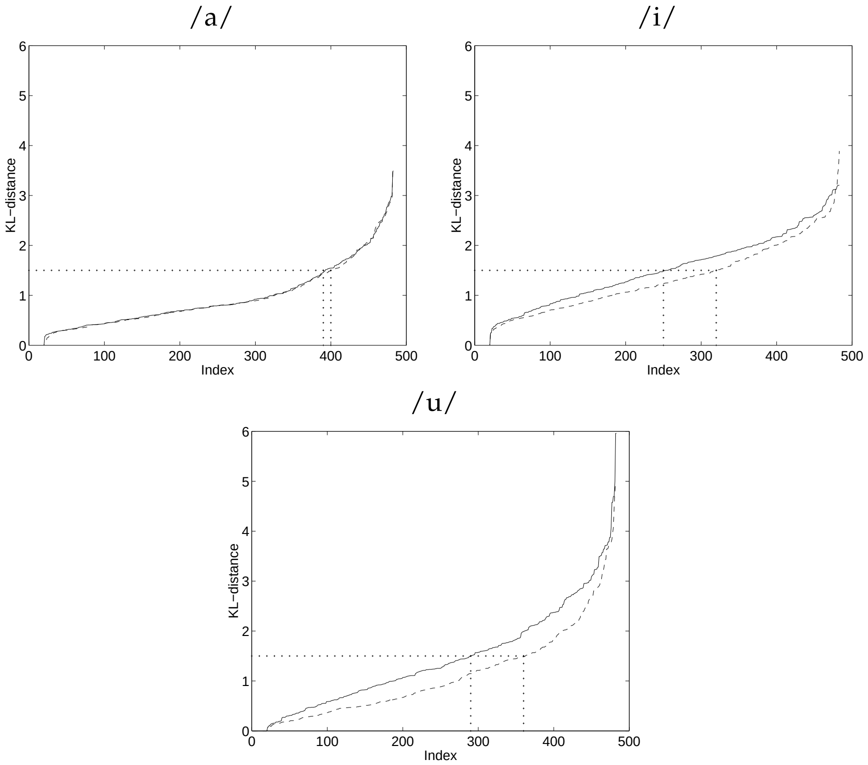


Figure 3.9: *Distributions of KL distances for /a/, /i/, and /u/. The solid line is for the database without clustering, the dashed line is for the database with clustering.*

3.4 Discussion

The findings of this investigation lead to a number of interesting observations. First, the differences in results between the three vowels /a/, /i/ and /u/ shows that /a/ is least affected by coarticulation, whereas /i/ is more and /u/ is most affected by it. The alveolars account for most of the coarticulation in /u/. There, the slow movement of the tip of the tongue causes the frontal mouth cavity to be small throughout the pronunciation of the vowel, which leads to a relatively high F_2 value. For /a/ and /i/, this does not make much difference since the locus of alveolars is much nearer to their customary F_2 value than for /u/.

Second, the finding that audible discontinuities still occur for the /a/ and

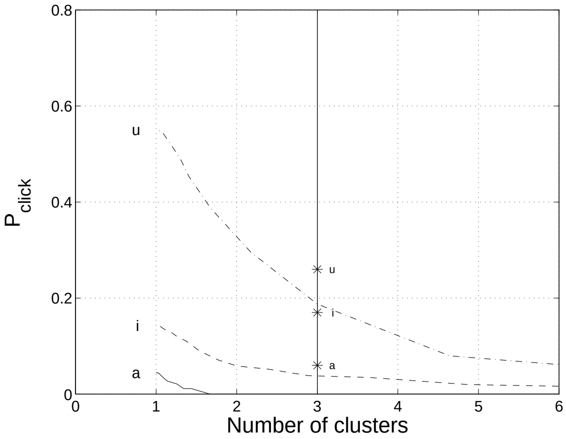


Figure 3.10: *Number of clusters versus P_{click}, the probability of a discontinuity arising, for /a/, /i/, and /u/. The lines represent the lower bound, i.e., the maximum improvement that might be obtained. The stars indicate the actual improvement obtained when using three clusters.*

that clustering does not reduce the amount of audible discontinuities leads to the conclusion that besides coarticulation there is always random variation in the pronunciation of the stimuli. This was also observed by Olive et al. (1998) who found F_2 variations in excess of 50 Hz for a vowel in repetitions of the exact same phrase as uttered by a highly professional speaker. Van Santen (1997) reports even larger F_1 and F_2 variations (up to 250 Hz) in the repeated pronunciation of /I/ in *six* and *million* by a professional speaker. This indicates the need to record several instances of a diphone and choose the one that is optimal for the database.

Third, the bottom panel in Figure 3.11 shows that when the diphones are recorded in asymmetrical nonsense words, the formant trajectories are no longer stable, but change gradually from start to finish. In that case, it may make sense to optimise the cutting point of the diphone boundary as proposed by Conkie and Isard (1997).

3.5 Conclusion

This chapter reported on the occurrence of audible discontinuities in diphone synthesis caused by spectral mismatch at the diphone boundaries.

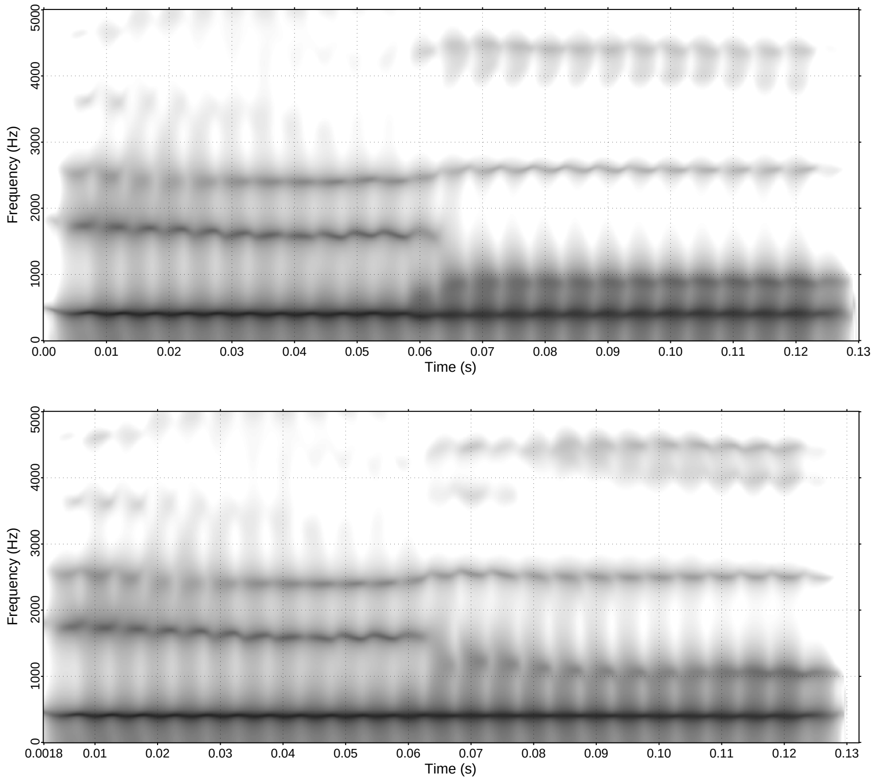


Figure 3.11: *Improvement in the concatenation of /du/ and /uk/ using a different diphone for /uk/. Instead of an abrupt jump in F_2 as is visible in the top panel, a gradual transition is observable in the bottom panel.*

A perceptual experiment was conducted to investigate the extent to which this phenomenon occurs. The results revealed that there are considerable differences between the vowels under investigation. The /a/ showed the least amount of perceived discontinuities, followed by the /i/. The /A/ and /I/ showed more perceived discontinuities, but the largest percentage of perceived discontinuities was found in the /u/. Research by Van Son (1988) and Van Bergem (1995) has shown that coarticulation has a centralizing effect. Maybe that is the reason why /u/ is more affected in terms of F_2 and /A/ more in terms of F_1. The scores obtained in the perceptual experiment were correlated with several spectral distance measures to find an objective measure to predict the occurrence of audible discontinuities. The correlation was performed using Receiver Operator Characteristic curves. The Kullback-Leibler distance, coming from statistics, was shown to be most adequate for the task.

In the second part of this chapter, a possible solution was presented. Context-sensitive diphones were added to the database. In order to reduce the number of additional diphones, the KL distance was used to cluster consonantal contexts that have the same spectral effects on the neighbouring vowels. A second perceptual experiment was conducted to evaluate the improvement obtained with this addition to the database. A significant improvement was obtained for /u/ and /i/ both in terms of the objective KL distance and the subjective scores. For /a/ there was only a subjective improvement, but objectively, in terms of KL distance, the improvement was not significant. This is not a problem, however, as the number of discontinuities in /a/ was already low to begin with.

The approach can be further extended to other phonemes prone to coarticulation effects, such as closed back vowels like /o/, /O/, /y/, and semivowels like /j/, /l/ and /w/.

The KL distance can also be applied to find the optimal cutting point in diphones or other concatenative units.

Chapter 4

Modelling segmental duration

4.1 Introduction

The quality of synthetic speech not only depends on the concatenative units themselves, but also on the prosodic modelling that is applied after concatenation. One aspect of prosodic modelling concerns the segmental durations. In speech perception, variation in segmental duration serves as a cue to the identity of a speech sound and helps to segment a continuous flow of sounds into words and phrases. In natural speech production, segmental duration is systematically varied depending on the phonetic context. Contextual factors influencing segmental duration operate on different phonological levels, from the phoneme and syllable level to the phrase level. If a duration model can take all important contextual factors into account, the resulting speech will not only sound more natural, but will probably also be more intelligible.

This section reports on the development of a new duration model for Dutch, which replaces the traditional sequential rule system in Calipso. The old system gave unsatisfactory results, probably because interactions among factors were not sufficiently modelled and some important higher-level prosodic and positional factors were not taken into account. Moreover, this system was not specifically modelled after the speaker of the diphones.

One general problem in constructing a duration model is the issue of *interacting factors*. When two factors interact, it means that the effect of one factor is amplified or attenuated by the other factor. An important characteristic of factors affecting duration prediction is that the effect of one factor does not reverse the effect of another factor, meaning that if one factor causes lengthening then another factor may influence the amount of lengthening, but it will not cause the first factor to induce shortening. This phenomenon is called *directional invariance* (Van Santen, 1992a).

A more specific problem in data-based systems is *data sparsity*. For all factors relevant for duration prediction levels or *features* can be identified, e.g., *unstressed* is a feature of the stress factor. For each segment whose duration we want to predict, a feature vector has to be computed describing the context in which the segment occurs. How this is done is explained in Section 4.3.2. Data sparsity refers to the problem that the feature space, consisting of all theoretically possible feature combinations, can be astronomically large. The linguistic space, a subset of feature vectors actually occurring in a particular language, only occupies a small fraction of the total feature space, due to phonotactic and other constraints. But even this fraction is still huge. The problem is that any large corpus used for analysis will cover only a small and unevenly distributed proportion of the feature space. Complete coverage is thus practically impossible. Because the number of very rare feature vectors is very large, it is not sufficient to simply model only the most common features. The statistical analysis methods used therefore have to be able to deal well with missing data in order to generalise to feature vectors not previously encountered.

In Section 4.2, different approaches to duration modelling will be introduced with their pros and cons, making a distinction between knowledge-based systems, data-based systems and systems that are based on knowledge as well as data. Section 4.3 discusses the steps involved in developing a new duration model. In Section 4.4, the resulting duration model for Dutch will be discussed together with the main effects found. Finally, in Section 4.5 an evaluation of the new against the old model is presented.

4.2 Possible approaches

4.2.1 Knowledge-based systems

The first type of duration model is a sequential rule system as applied by Klatt (1987) in the MITalk system. His model assumes that a) each phonetic segment type has an inherent duration specified as one of its distinctive features, b) each rule results in a percentage increase or decrease in the duration of the segment, but c) the segment cannot be compressed shorter than a certain minimum duration. This model is summarised in Equation 4.1.

$$DUR = MINDUR + \frac{(INHDUR - MINDUR)\ PERC}{100} \tag{4.1}$$

INHDUR is the inherent duration of a segment in milliseconds. MINDUR is the minimal duration of the segment if stressed. PERC is the percentage shortening or lengthening determined by applying a series of rules in sequence. These rules pertain among other things to clause- or phrase-final lengthening, non-word-final shortening, polysyllabic shortening, lengthening for emphasis, postvocalic context of vowels and shortening in consonant clusters. The problem with this type of approach is that the parameter values are based on small-scale studies in which the factors pertaining to a particular rule are varied while most others are kept constant. Since there may be interactions between contextual factors, the obtained parameter values are likely to be inaccurate for contexts where the factors held constant have different levels. For instance, in Dutch an interaction was found between stress and accent (Eefting, 1991; Cambier Langeveld, 2000). Accentuation affects the duration of the whole word, whereas stress only works on the syllable level. The interaction consists in the fact that accentuation affects stressed syllables more than unstressed syllables. In some investigations, the segments under analysis always occur in post-focal position. This means that an incomplete picture is presented of the behaviour of these segments. Using these results in duration prediction may result in underestimation of segment durations in unstressed syllables of accented words.

Many researchers have benefited from the work of Klatt. The rule-based system that was originally implemented in Calipso was also based on this framework. Since then there have been other knowledge-based approaches, that attempt to capture the rhythmic properties of a language by using a different level of description such as a hierarchical prosodic tree (Dirksen and Coleman, 1996; Local and Ogden, 1996). Dirksen and Coleman (1996) assign a metrical structure to the phonetic string (weak vs. strong nodes) (see Figure 4.1 for an example). A syllable is divided into a weak onset node and a strong rhyme node. The rhyme is further divided into a strong nucleus and a weak coda. The onset consists of zero to three consonants which are in a left-branching weak-strong order. This division may be language-specific. A total duration of a constituent is assigned to the head of the constituent. The segments within an onset constituent do not have inherent durations but these follow from statements of non-overlap between sister constituents within an onset. Coda constituents do have inherent durations as well as statements about non-overlap between sister constituents within the coda. The CD-ROM provided with the book of Van Santen, Sproat, Olive, and Hirschberg (1996) gives some examples generated by Dirksen and Coleman (1996).

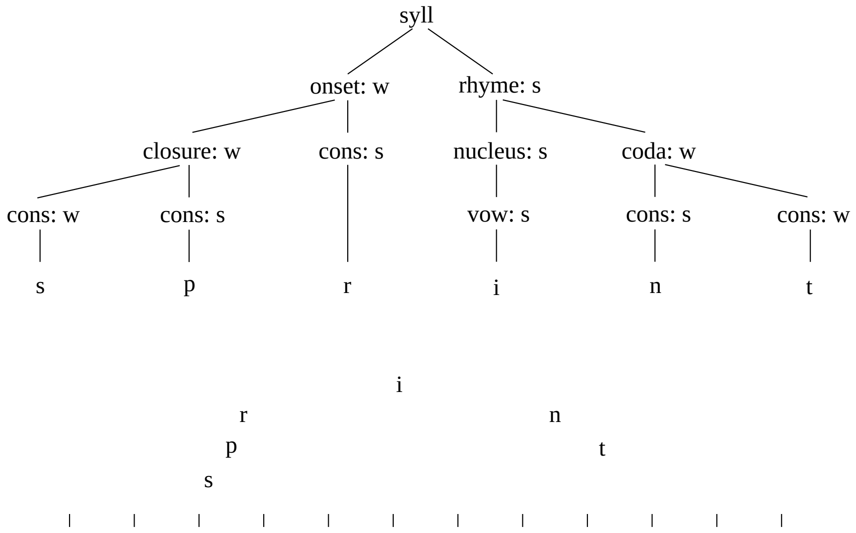


Figure 4.1: *Example of a strong/weak branching tree and durational composition of the word* sprint *according to the metrical approach proposed by Dirksen and Coleman (1996).*

4.2.2 Data-based systems

When large speech corpora and the computational means for analysing these corpora became available, new statistical approaches were proposed.

Riley (1992) proposed the use of Classification and Regression Trees (CART) to model duration. The principle of CART is to construct a binary branching tree from which segmental durations can be looked up. The branches in the tree are determined by contextual factors such as manner of articulation of the target segment, manner and place of articulation of the preceding and the following segments, lexical stress and the number of segments up to the end of the word. The grouping of contexts to the left or right branch is determined by minimising the variance of the estimated error. The problem with this type of approach is that it is impossible to exhaustively cover all possible feature vectors, even with huge amounts of training data. It is not possible to generalise to unseen cases, other than by pooling some seen cases that come close to the desired unseen case. For this reason, this method doesn't cope well with data sparsity. Overtraining is also a serious problem, which happens when the distinctions in the tree are so fine that they fit the training corpus well, but not any other corpus. This can be overcome by pruning the tree such that more general distinctions remain. Deciding on the amount of pruning necessary to obtain good results is not trivial, however.

Another approach uses Neural Networks to predict durations (Campbell, 1992). The syllable is taken as the basic unit of duration prediction, which can be calculated without prior knowledge of phonetic constraints. A neural net is trained taking into account the number of phonemes in the syllable, the nature of the syllabic peak (e.g., reduced, lax, tense vowel), the position in the tone group, the type of foot, stress and word class. Segmental durations are calculated as follows. For each phoneme class, there is a distribution of log durations that resembles a normal distribution and that is characterised by a mean duration and a standard deviation. One factor k is computed that is applied to each segment in the syllable in terms of standard deviations around its mean to produce an optimal fit to the overall syllable duration. This relies on the assumption that all segments in a syllable are lengthened or shortened equally in terms of standard deviation. This does not necessarily have to be the case. This approach does not take the segmental make-up of the syllable into account, which can contribute considerably to the overall syllable duration. For instance, in Dutch there is a difference in duration between *nar (joker)* and *nier (kidney)*. Both /A/ and /i/ are in principal short vowels, but the /i/ undergoes considerable lengthening when followed by /r/, whereas the /A/ does not (Moulton, 1962). This cannot be captured in this syllabic model. Like CART, this approach can suffer from overtraining and is unable to come up with good predictions when the data is sparse.

4.2.3 Knowledge-based data systems

The solution proposed by Van Santen (1992b) is the application of a broad class of mathematical models called *sums-of-products models*. This approach takes advantage of the fact that most interactions are directionally invariant, which allows describing these interactions with equations consisting of sums and products. Phonemes that are affected similarly by the various factors are grouped into subclasses. The decisions concerning this grouping are based on exploratory data analysis and phonetic/phonological literature. For each subclass of segments, a separate sums-of-products model is trained. Exploratory data analysis is an essential step in order to appropriately decide which factors are important and how many levels on a factor should be distinguished. For each subclass this can lead to different results. Thus, a great deal of phonetic and phonological knowledge can be incorporated in the model.

A sums-of-products model (Equation 4.2) gives the duration for a phoneme/context combination, as described by the feature vector $\vec{f}$. K is a set of indices, each corresponding to a product term. I_i is the set of indices of factors occurring in the i-th product term. In other words, a sums-of-products model is an equation whose parameters correspond to factor levels, which are combined by taking products of sub-groups of parameters and then adding the products. Two examples of sums-of-products models are the additive model (where $K = \{1, \ldots, N\}$, and $I_i = \{i\}$) and the multiplicative model (where $K = \{1\}$ and $I_1 = \{1, \ldots, N\}$).

$$DUR(\vec{f}) = \sum_{i \epsilon K} \prod_{j \epsilon I_i} S_{i,j}(f_j) \tag{4.2}$$

Evidence in favour of this approach is provided by Maghbouleh (1996), who compared the use of the sums-of-products approach to the use of CART. Two important findings were reported. First, while both methods perform equally well on a test set that comes from the same corpus as the training set, the sums-of-products approach needs only one-tenth of the data that CART needs to come to its optimal performance. Second, the sums-of-products approach performs better on new texts. Even when CART was restricted to use the same terms as in the sums-of-products model, its performance on new texts did not improve. Apparently, CART is unable to capture interactions as accurately as the sums-of-products approach. The sums-of-products approach has successfully been applied to many languages, including American English, French, German, Italian, Spanish, Japanese and Mandarin Chinese (Sproat, 1998).

4.3 The sums-of-products approach

For the reasons mentioned in the previous paragraph, the sums-of-products approach was used to develop a new duration model for IPO's speech synthesis system Calipso. Additional reasons for using this approach are that it requires a fairly small amount of training data to construct a reliable model, the resulting model is generalisable to many text types, and it allows a thorough analysis of the durational characteristics of a speaker. The duration model is created by performing a number of steps, as they are listed below. Each of these steps will be explained in more detail later on.

1. Obtain a large text corpus (Section 4.3.1).
2. Compute from this text all information relevant for duration prediction, encode it in feature vectors (Section 4.3.2).
3. Select a subset of the corpus using a greedy algorithm (Section 4.3.3).
4. Record the sentences of the subset and segment the recordings (Section 4.3.4).
5. Perform statistical analyses to construct a duration model consisting of a) a category tree in which all segments are grouped into subclasses that are affected similarly by contextual factors, (Section 4.3.5) and b) a set of sums-of-products models for each of the leaves in the tree (Section 4.3.6).

4.3.1 Corpus

The source of our database is a large collection of news transcriptions provided by the "Instituut voor Nederlandse Lexicologie" (INL)[1]. The collection consists of approximately 27,000 sentences from two types of news: the 'NOS Journaal' (public broadcast news) and 'Jeugdjournaal' (a special broadcast news edition for children). Since the sentences are literal transcriptions of spoken news reports, they are better suited to our needs than sentences taken from newspapers. The children's news has the additional advantage that sentences are shorter and have a simpler structure than

[1] We acknowledge the Institute for Dutch Lexicology for providing us with the text database.

standard news editions. These sentences were phonetically transcribed using GRAFON, a grapheme-to-phoneme converter developed at the University of Nijmegen, department of Language and Speech (Kerkhoff, Rietveld, Gussenhoven, and Elich, 1995). The database ultimately consisted of 1,549,946 segments: 480,796 vowels and 1,069,150 consonants.

4.3.2 Feature vectors

Each segment in the text is annotated with a list of contextual features that are relevant for duration prediction. The following factors were included in the annotation based on literature on duration in Dutch (e.g., Nooteboom (1972); Eefting (1991)) and other languages (e.g., Van Santen (1992b); Van Santen (1994)). The number of factor levels per factor is indicated between brackets. The meaning of those levels will be explained after the listing of the factors.

1. Identity of the current segment (53)
2. Phoneme class of the preceding segment (27)
3. Phoneme class of the following segment (27)
4. Stress (3)
5. Accent (3)
6. Word class (3)
7. Word frequency (3)
8. Left position (LPOS) (16)
    - Left position of the segment in the syllable (2)
    - Left position of the syllable in the word (2)
    - Left position of the word in the phrase (2)
    - Left position of the phrase in the sentence (2)
9. Right position (RPOS) (16)
    - Right position of the segment in the syllable (2)
    - Right position of the syllable in the word (2)

- Right position of the word in the phrase (2)
- Right position of the phrase in the sentence (2)

Factor 1, *Identity of the current segment*, has 53 levels that include the 44 phonemes that are indicated in Table A.1. These phonemes consist of 19 vowels (3 of which are diphthongs, and 3 of which are used only in loan words), 25 consonant segments and the glottal stop. The 7 stops are split up into two parts: a closure and a burst phase, so that we get a total of 51 levels. An additional symbol /R/ was introduced to distinguish the velar vowel-like /R/ from the fricative-like /r/. The last symbol /sil/ is for silence.

Factors 2 and 3, *Phoneme class of the preceding/following segment*, have 27 levels. They are classified with basic phonetic labels expressing manner of articulation and voicing, such as unvoiced stop, nasal or liquid[2]. Silence is also a possible value. Place of articulation was not encoded since previous research proved it to be hardly relevant (Crystal and House, 1988). Furthermore, if a preceding or following consonant is part of a consonant cluster, the size of the cluster is appended to the phonetic class. This rule applies within a syllable, so two consonants separated by a syllable boundary are not considered as being part of a cluster. The syllable boundaries have been determined automatically by the grapheme-to-phoneme converter.

Factor 4, *Stress*, has three levels, where 0 means the syllable has no lexical stress, 1 means primary stress and 2 means secondary stress. Factor 5, *Accent*, also has three levels, 0, 1, and 2, for unaccented, accented and cliticised words (e.g., *z'n (his)*, and *'m (him)*). This factor deals with pitch accent and operates on the word level. Factor 6, *Word class*, makes a distinction between function words, content words with less than four syllables and content words with more than four syllables. Factor 7, *Word frequency*, distinguishes between low-, medium-, and high-frequency words. Factor 8, *Left position*, consists of a combination of 0 and 1 values to encode position, e.g. *1111* means that the phoneme is the initial phoneme in the syllable, the syllable is initial in the word, the word is initial in the phrase and the phrase is the initial in the sentence. In the same manner, *0000* means that it is non-initial on all levels. Factor 9, *Right position*, is similar to factor 8, except that here it is checked whether the position is final vs. non-final.

[2] These will be abbreviated to ustop, nas, liq.

4.3.3 Selecting a subset

Multiplying all factor levels, we could in theory get a total of 770,943,744 unique factor combinations. The current corpus contains 99,206 distinct factor combinations. This is less than 0.01% of the number of theoretical possibilities. Gaps in the factor space are caused for a large part by phonotactic constraints of the language. To some extent accidental gaps occur. No matter how large a corpus, there will always be legal factor combinations that are not included. This illustrates exactly the data sparsity problem mentioned earlier. In unrestricted text-to-speech, the chance of a rare feature vector occurring is quite big. Therefore, the duration model has to be able to generalise to factor combinations never seen before. Nevertheless, it is important to find the smallest subset that covers the same unit types as the entire corpus. This can be done by applying a greedy algorithm (Van Santen and Buchsbaum, 1997). In our case, sub-vectorisation was applied so that complete coverage was achieved for sub-vectors (Van Santen and Sproat, 1998). This means that from each feature vector, a set of sub-vectors is selected that corresponds to precisely those components that are expected to interact. The sub-vectors used were {1}, {2}, {1,3}, {1,4,5}, {6,7}, {1,8}, {1,9}. Then the greedy algorithm is applied to a list in which each sentence is represented as a set of sub-vectors instead of full vectors. In this way, parameters can be estimated on far fewer sentences than are required to cover the entire feature space. For the current data set, the 4069 distinct sub-vectors occurring in the database of 27,000 sentences could be covered completely in just 297 sentences, containing 16,775 segments (4968 vowels and 11,807 consonants). Some cells consist of just one observation, because they contain a rare segment/context combination. However, after grouping the segments into subclasses, each cell contains at least five occurrences.

4.3.4 Recording and segmenting

After manual correction of the phonetic transcriptions, these 297 sentences were recorded using the same semi-professional female speaker that was used for the diphone database and the phrase concatenation recordings discussed in Chapter 2. The recordings were made in a sound-treated room on a DAT-tape with a 48 kHz sampling frequency. The speech was down-sampled to 16 kHz. It was then segmented by hand using Waves+ (Entropic inc.) on a Silicon Graphics workstation. At the time there was no automatic segmentation software available for Dutch that could identify segment boundaries with such small deviations as to make them useful for duration prediction.

Both the waveform and the spectrogram were used to determine the segment boundaries, and the placement of the boundaries was confirmed by listening. For stops, the closure and burst part were segmented separately. In utterance- and phrase-initial unvoiced stops, it was difficult to determine the start of the closure, so it was put at a reasonable distance before the release of the burst. These closures were eventually left out of the analyses, because they displayed a greater variability than closures at other positions. When a stop followed a nasal, the closure phase proved very short and was sometimes difficult to detect. In most cases, a reduced energy in the spectrogram signalled the position of the closure. In case it was not visible at all, it was assigned a zero duration.

When listening to the recordings, it was decided to assign different labels to different instances of /r/. The label *r* was reserved for the fricative /r/. The label *R* was used to mark the velar vowel-like /r/ that occurs in postvocalic position in some dialects of Dutch. It resembles the American /r/. This sound was more difficult to segment. The boundary point was chosen to be the point where a sharp decrease in F_3 could be detected.

In case of degemination, where two identical consonants occur next to each other on either side of a word boundary, the phoneme boundary between the two was not distinguishable and was put halfway between the start and end label.

Vowel-vowel transitions were most difficult to segment. Usually, two vowels in Dutch are separated by a glottal stop, but it is also possible to connect them with a /j/- or /w/-like transition. When this was the case, the midpoint of the transition was chosen as a cutting point. Between two consonants in a coda, such as /L-m/ or /r-k/, a very short schwa was sometimes inserted by our speaker. We decided to mark this as *&*. This only occurred 14 times in our corpus, insufficient to come up with a reliable model. This label was discarded in the analysis.

4.3.5 Constructing the category tree

The first step in constructing a duration model using the sums-of-products approach is to divide the feature space into subclasses (e.g. 'stressed' vs.

'unstressed'), such that within each subclass the cases are similarly affected by the factors. This is different from CART, where cases with similar durations are grouped together. This process takes shape by using knowledge about standard phonetic and phonological distinctions and effects reported on in the literature. The distinctions are visualised in a category tree (Figure 4.2).

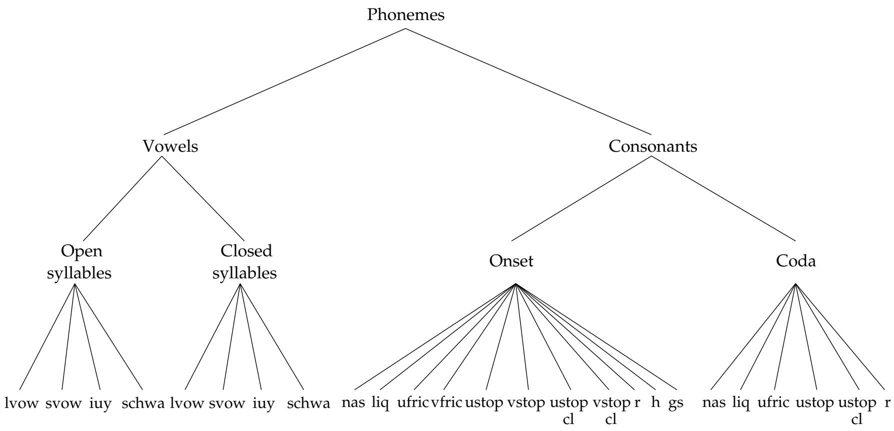


Figure 4.2: *Category tree of the Dutch duration system; lvow/svow = long/short vowels, iuy = /i/, /u/ and /y/, nas = nasals, liq = liquids/glides, ufric/ustop = unvoiced fricatives/stop bursts , vfric/vstop = voiced fricatives/stop bursts, ustop-cl/vstop-cl = unvoiced/voiced stop closures, gs = glottal stop.*

Vowels vs. consonants. This is a rather obvious distinction based on well-established phonetic and phonological knowledge. For instance, lengthening or shortening as a result of a change in speaking rate has a different effect on vowels than on consonants. Consonants are compressed more than vowels (Covell, Withgott, and Slaney, 1998).

Vocalic distinctions. Rietveld and Frauenfelder (1987) report that vowels in Dutch are longer in open than in closed syllables, i.e., that a vowel is shorter when followed by at least one consonant in the same syllable. But more importantly, the effect of the following consonant may be different dependent on whether it occurs in the coda of the same syllable, or in the onset of the next syllable. Therefore, it was decided to split these cases and create two separate branches in the category tree for vowels in open vs. closed syllables. Within each of these branches, we divide the vowels into long vowels, short vowels, the vowels /i/, /u/, and /y/ (in short: iuy), and schwa. The category of long vowels consists of /e, a, o, 2/, the diphthongs /Ei, Au, 9y/ and the loan vowels /E:, O:, 9:/, which only occur in loan words (see Table A.1 for examples). The diphthongs and the loan vowels are generally longer in duration than the other vowels, but analysis has shown them to behave similarly. The short vowels are /I, A, E, O, Y/. The short vowels /i/, /u/, and /y/ are classified separately. Their durations are similar to those of short vowels, except before /r/ where considerable lengthening is observed (Nooteboom, 1972). Distributionally, they behave as long vowels. Where short vowels can be followed by a maximum of two consonants in the coda, long vowels can be followed by maximally one (e.g., /lAmp/ vs. */lamp/ vs. */limp/). Moulton (1962) hypothesises that /i/, /u/, and /y/ probably started out as long vowels, but in the evolution of Dutch changed to short vowels, while retaining their long vowel characteristic before /r/. The schwa /@/ is treated as a separate category, because it only occurs in unstressed syllables.

Consonantal distinctions. The first distinction to be made is that between onsets and codas. The key reason to treat them separately is that the same factor may have different effects (e.g., stress lengthens consonants in onsets more than in codas) and the phonotactics of consonants in onsets and codas are totally different, producing phonemic context factors that are hard to reconcile. The consonants are grouped according to manner of articulation and voicing. The categories are unvoiced stop closures, /pcl, tcl, kcl, ccl/ (closures for /p/, /t/, /k/, and /c/), voiced stop closures /bcl, dcl, gcl/, unvoiced stop bursts /p, t, k/, voiced stop bursts /b, d, g/, unvoiced fricatives /f, s, x, S, c/, voiced fricatives /v, z, Z/, nasals /m, n, N, J/, liquids & glides /l, j, w/, and the *r*, *h*, and glottal stop. The /r/ is not included with the liquids and glides, because it is suspected to behave differently. In Dutch, the number of categories in codas is more limited than in onsets because syllable-final devoicing applies. The /h/ and /c/ do not occur in the coda either and the dark /L/ occurs as an allophonic variant of /l/.

The resulting category tree differs to some extent from the American English tree used by Van Santen (1994). There, consonants are divided into intervocalic consonants and consonants in clusters, which are further divided into onsets and codas. In our study, it was decided to simplify matters by not considering intervocalic consonants separately from the other consonants. In codas, Van Santen (1992a) makes a distinction made between phrase-medial and phrase-final codas to accommodate interactions between post-vocalic consonant class and phrasal position. This interaction is specific to the English language and is thus not included in our tree. In the Dutch duration model, a further distinction is made between vowels in open and in closed syllables.

4.3.6 Training the sums-of-products models

For each leaf in the category tree, a sums-of-products model has to be trained that will predict the duration of the segments in that subclass in all possible contexts. The current analysis was restricted to the use of pure multiplicative models following Möbius and Van Santen (1996), because these perform better than strictly additive models, and require less training data than any other combination of sums and products. The model is summarised in Equation 4.3, where $Dur_{i(f2,\ldots,fn)}$ is the predicted duration of a given vowel i with factor levels $f2, \ldots, fn$ for factors F2, $\ldots, Fn$ respectively. $BaseDur_i$ is the basic duration of the segments, and F2$_{f2}, \ldots, Fn_{fn}$ are the coefficients of the other factor levels.

$$Dur_{i(f2,\ldots,fn)} = BaseDur_i \times F2_{f2} \times F3_{f3} \times \cdots \times Fn_{fn} \tag{4.3}$$

The calculations take place in the logarithmic domain instead of the linear domain (see Equation 4.4). This reduces the skewness of the duration distribution and allows a more accurate prediction of small durations. Multiplication in the linear domain is equivalent to addition in the logarithmic domain.

$$log(Dur_{i(f2,\ldots,fn)}) = log(BaseDur_i) + log(F2_{f2}) + log(F3_{f3}) + \cdots + log(Fn_{fn}) \tag{4.4}$$

Training the multiplicative models is an iterative process. It involves exploratory data analysis to determine which levels have to be distinguished on a given factor for each category and which factors interact. By discarding irrelevant factors and reducing the number of levels on a factor, the number of gaps in the database will decrease and the number of observations per factor level will increase. If the correct levels are collapsed, the predictions will be more accurate. Interacting factors can be joined to incorporate them in the model appropriately. These decisions are based on a number of information sources: 1) phonological/phonetic knowledge, 2) the means that are computed by each factor to determine the size of the effect of each factor levels, 3) the two-way means that are computed to determine whether any two factors interact, 4) and the outcome of the multiplicative model, in terms of prediction error.

Most durational studies use raw means for analysing data, which is fine as long as the experiments are carefully controlled and the factors involved are balanced. In a natural speech database, the frequency distribution is unbalanced and thus the raw means can be quite misleading. The result may be biased by a segment occurring more often in some environments than in others. To overcome this problem, analyses are executed using *corrected means* (Van Santen and Sproat, 1998), which account for the fact that the corpus is not balanced. Therefore, results are similar to the ones that would have been obtained if a balanced corpus was used. It works as follows. If F_1 is the factor of interest, then a new factor $F_{remainder}$ can be defined, which is the product of all remaining factors, $F_2, \ldots, F_n$. A 2-factor additive or multiplicative model can be applied to this new factorisation, and the parameters Dur_i estimated. These parameter estimates can be interpreted as means of the levels on the factor of interest that *have been corrected for the effects of the remaining factors*.

Two-way corrected means are computed to investigate whether two factors interact. When a clear interaction is found, the two factors will be combined into one. Then, a multiplicative model can be trained. All model fitting is performed with least-squares estimation in the log duration domain (Van Santen and Sproat, 1998). The resulting model can be assessed by considering the set of predicted durations (in terms of corrected means) for each factor level, the correlation between the predicted and observed duration, and the prediction error. Factor levels with similar means can be collapsed, factors with only a minor effect can be discarded, and interacting factors can be combined into one. Every time a change is made to the model, the analysis is repeated until a satisfactory multiplicative model is obtained.

4.4 A duration model for Dutch

4.4.1 Fitted duration

The performance of the entire duration model is established by comparing the observed durations in the 297 sentences of our speech corpus with the predicted values. Table 4.1 displays the performance of the new Dutch duration model in comparison to the German and French duration models constructed with the same sums-of-products approach (Möbius and Van Santen, 1996; Tzoukermann and Soumoy, 1995). The number of obser-

	Dutch	German	French
Observations	12,948	24,240	7,143
Number of Sub-classes	25	30	16
Number of parameters	404	674	782
Correlation	0.760	0.896	0.847
RMSE (ms)	27	19	25
Obs. mean (ms)	73	60	74
Obs. std. (ms)	34	43	47

Table 4.1: *Results of model parameter estimation for the Dutch sums-of-products model compared to the German and French sums-of-products models.*

vations in the Dutch database is about half the size of the German database and almost double the size of the French. The correlation between predicted and observed duration is less for the Dutch duration model than for the French and German model, but the RMSE (Root-Mean Squared Error) between predicted and observed duration is almost as good as for the French model. The RMSE is a measure indicating the precision of prediction that is commonly used in duration prediction (see Equation 4.5).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(obs\_dur_i - pred\_dur_i)^2}{n}} \tag{4.5}$$

In this case, it is to the regression line what the standard deviation is to the mean. A fraction of 68% of the data points lie within one RMSE from the regression line, which in this case is 27 ms. 95% of the data points lie within two RMSE from the regression line. Figure 4.3 plots the correlation between observed and predicted duration values. The regression line and the two lines that define the area with data points within one RMSE are also indicated. For each factor it was determined what the contribution was to the overall prediction. Two factors did not have an effect, those for word frequency and word class. Thus, it did not matter whether the word was a function word, a content word with less than four syllables or a content word with more than four syllables (including most compound nouns). It also did not make a difference whether the word had a low, medium or high frequency. These two factors were therefore discarded. For the stress factor, no significant difference was found between primary and secondary stress, so they were combined. For the accent factor, the levels cliticised and unaccented were not significantly different, so they too were combined.

An interaction was found between stress and accent, confirming findings

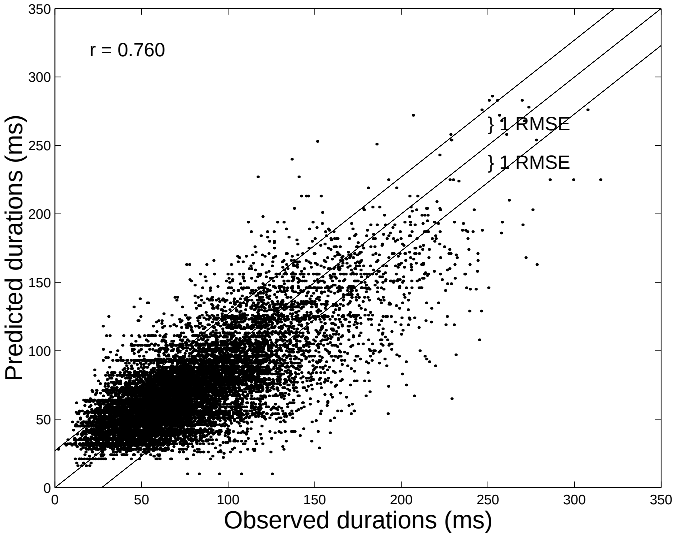


Figure 4.3: *Correlation between observed durations and durations computed with the sums-of-products duration model (n = 12948). 68% of the observations lie within one RMSE from the regression line. 95% of the observations lie within two RMSE from the regression line.*

by Eefting (1991). Figure 4.4 illustrates this with a clear example found in our database, pertaining to liquids in codas. One can see that the duration of liquids in codas is more affected by accentuation in a stressed syllable than in an unstressed syllable. The relative difference between accented and unaccented liquids in stressed syllables is 25.5%, whereas in unstressed syllables this difference is 6.3%. The interaction was modelled in the multiplicative models by joining the two factors into one factor with four levels: 00 stands for unstressed unaccented, 01 for unstressed accented, 10 for stressed unaccented, and 11 for stressed accented. For American English (Van Santen, 1992b), these factors were also jointly analysed.

In the next sections, the durational characteristics of our speaker's speech will be discussed by looking at how the factors affect the durations of segments in different categories. All results are presented in terms of corrected means in milliseconds. Whenever they are presented in graphical form, they are plotted on a logarithmic scale, because it is generally believed that this better resembles the perceptual behaviour of listeners.

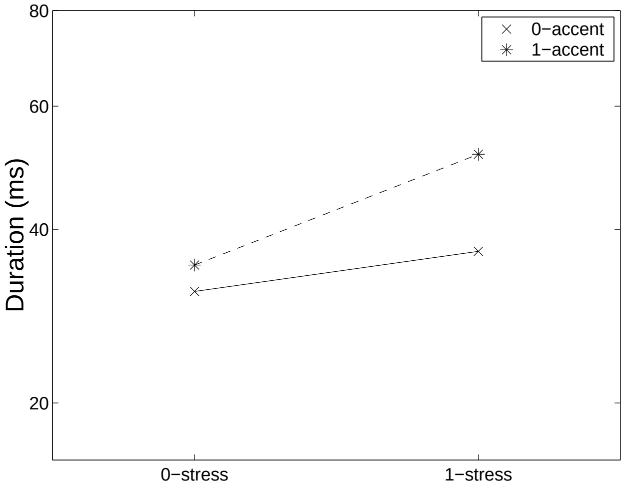


Figure 4.4: *The effect of stress and accent on the duration of liquids in codas.*

Van Santen (1992a) explains this by saying that "a 50 ms discrepancy between 65 and 15 ms is perceptually far more salient than a 50 ms discrepancy between 200 and 150 ms". The complete listing of corrected means for all subclasses and their relevant factor levels is given in Appendix B.

4.4.2 Vowels

Within the class of vowels a distinction is made between vowels occurring in open and in closed syllables. Figure 4.5 illustrates the differences between the two classes in terms of corrected means. The difference is especially large in the case of long vowels and diphthongs. For the other vowel categories, the difference is negligible. Rietveld and Frauenfelder (1987) also conducted a study into this phenomenon, but the scope of their study was restricted to the long vowels /e/ and /o/. The reason for maintaining the distinction between open and closed syllables even for those categories that do not differ in their corrected means, is that the effect of the following segment is probably different dependent on the presence or absence of a syllable boundary.

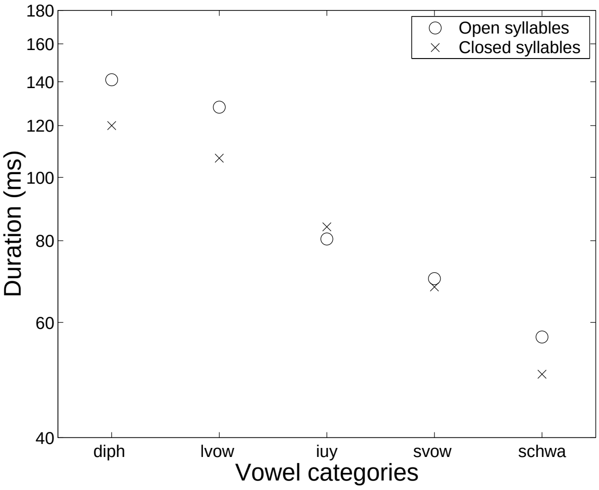


Figure 4.5: *Corrected means of vowels in open and closed syllables.*

Segment class	Previous	Next	Stress/ Accent	LPOS	RPOS
lvow	1.15	1.22	1.53	1.11	1.71
svow	1.13	1.59	1.44	-	-
iuy	1.69	2.11	1.41	1.27	2.69
schwa	1.37	1.39	1.17	1.19	2.57

Table 4.2: *Range ratios, indicating the proportional difference between the shortest and longest values for all levels on a factor, for vowels in open syllables.*

Vowels in open syllables

Table 4.2 lists the range ratios for all factors. The *range ratio* is obtained by dividing the largest value by the smallest value on the levels of a factor, to give an indication of the amount of lengthening/shortening a certain factor induces. (An overview of the corrected means per factor can be found in Tables B.1 to B.4.) A few observations can be made here. For instance, for iuy all factors have a large effect on the duration.

The *previous segment* has a minor effect on the vowels. A vowel is generally longer after a sonorant than after an obstruent and often longer after a voiced than after a voiceless obstruent. The difference between voiced obstruents and sonorants is quite small, but consistent, which is why we maintained the distinction. For short vowels there is no distinctive difference between the voiced and the voiceless context. The schwa shows deviant behaviour, because it is shortest after a sonorant and longest after an unvoiced obstruent.

The *next segment* has a considerable effect on the vowel duration. This is visualised in Figure 4.6. The most obvious finding is that of lengthening before /r/. This is most extreme for iuy, but is also considerable for other vowel categories. The phenomenon has been investigated before, among others by Nooteboom (1972). In his study, the vowels were produced in the context *pVrt*, so the vowel always occurred in a closed syllable. In open syllables, the following consonant is always in the onset of the next syllable, and it is all the more remarkable that the effect is so large (the vowels /i/, /u/, and /y/ are twice as long before /r/ as before another sonorant).

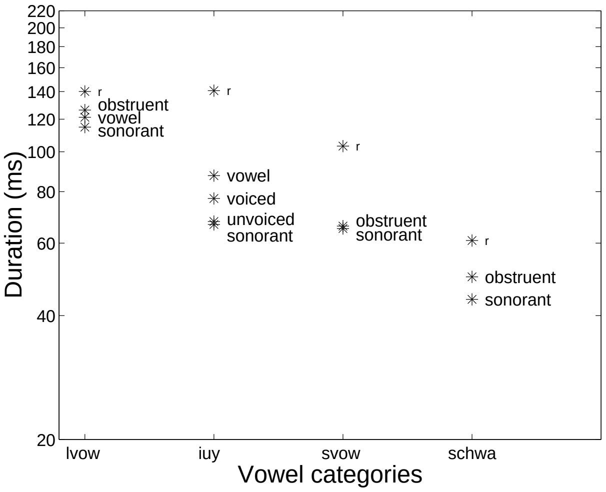


Figure 4.6: *Effect of next segment on vowels in open syllables.*

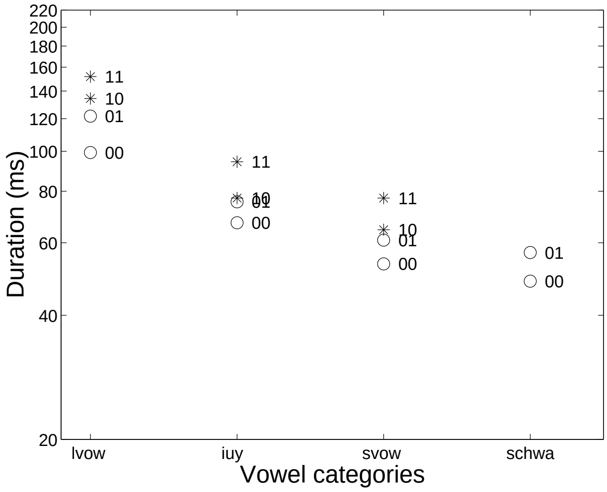


Figure 4.7: *Effect of stress and accent on vowels in open syllables; 00 = unstressed/unaccented, 01 = unstressed/accented, 10 = stressed/unaccented, 11 = stressed/accented.*

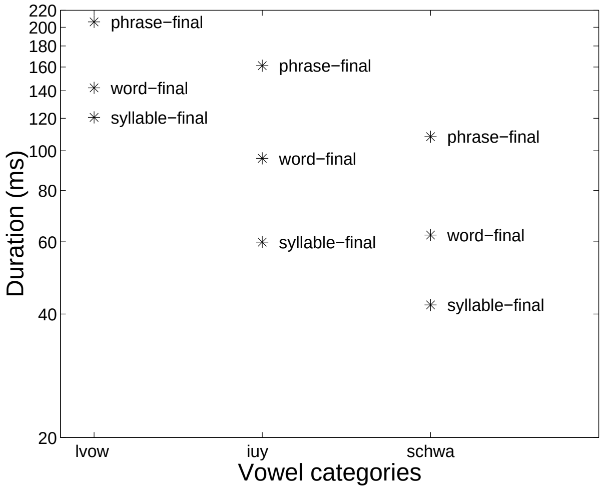


Figure 4.8: *Effect of right position on vowels in open syllables.*

When two vowels follow each other in Dutch, there is an optional but commonly applied rule to insert a glottal stop between the two. In the few cases where this does not occur, we see that the vowels are lengthened. This is caused by a glide-like change between the vowels. Since the segment boundary is put in the middle of the glide, the lengthening occurs in both the first and the second vowel.

The effect of *stress and accent* is very consistent in all vowel conditions (see Figure 4.7). Vowels are always shorter in unstressed, unaccented position than in unstressed, accented position and also always shorter in stressed, unaccented position than in stressed, accented position. This nicely illustrates the concept of directional invariance. It is also apparent that in each of the four conditions, even in unstressed unaccented condition, the long vowels are always longer than iuy and the short vowels. The effect of stress and accent is largest in long vowels, having a range ratio of 1.5. The schwa is only allowed in unstressed syllables. Even there, one can see significant lengthening due to the presence of a pitch accent on the word. It was not examined whether there was a difference between the duration of schwa in pre-focal or post-focal position.

The effect of *left position* is small. A distinction is maintained between word-initial and non-initial vowels, where the vowels are longer in the first context. This is in agreement with findings by Nooteboom (1972), who in unaccented three-syllable nonsense words also found the vowel in the first syllable to be longer than the one in the second syllable. This implies that the speaker signals word boundaries by durational lengthening at both ends of the word. The difference between vowels in word-initial and in non-initial syllables averages 8 ms.

The effect of *right position* is very large, which makes sense considering that the vowel in an open syllable always constitutes the final segment in the syllable (see Figure 4.8). The number of levels is reduced to three: syllable-final, word-final, and phrase-final. Short vowels do not occur in word-final position in open syllables. They have to be followed by at least one consonant in the coda. For the other vowels, the most extreme lengthening occurs on phrase-final syllables, but the difference between syllable-final and word-final syllables is also considerable.

Segment class	Previous	Next	Stress/ Accent	LPOS	RPOS
lvow	1.11	1.54	1.46	1.05	1.48
svow	1.36	1.31	1.35	-	1.52
iuy	1.03	2.07	1.43	1.11	1.37
schwa	1.64	1.39	1.10	1.48	1.63

Table 4.3: *Range ratios, indicating the proportional difference between the shortest and longest values for all factors, for vowels in open syllables.*

Vowels in closed syllables

The range ratios for the vowel classes in the closed syllables are given in Table 4.3. (An overview of the corrected means per factor can be found in Tables B.5 to B.8.) The effect of the *previous segment* is quite large for short vowels and schwa but is smaller for long vowels. The distinctions made in decreasing order are vowel, sonorant and obstruent. The distinction between voiced and voiceless obstruents was not relevant.

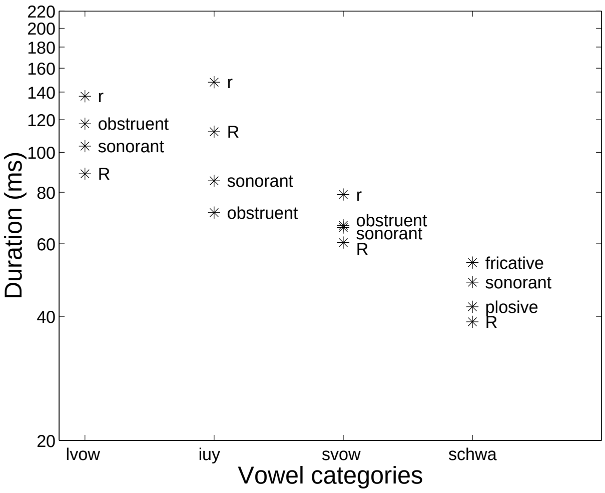


Figure 4.9: *Effect of next segment on vowels in closed syllables.*

The effect of the *next segment* is again dramatic (see Figure 4.9). The influence of lengthening due to /r/ is even more pronounced for our speaker than was found by Nooteboom (1972). Moreover, it is not restricted to iuy,

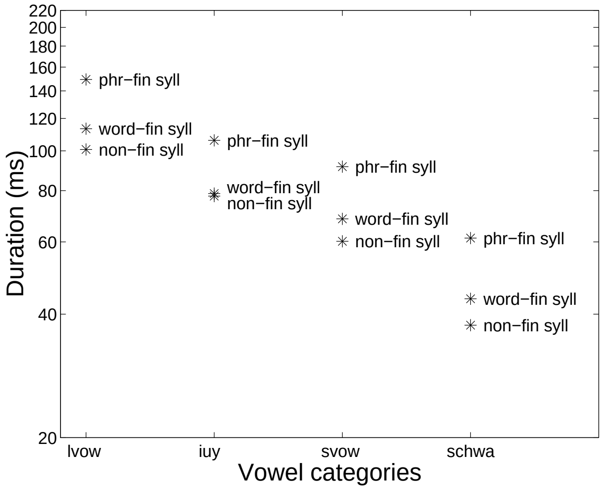


Figure 4.10: *Effect of right position on vowels in closed syllables.*

but also other vowels are affected, albeit to a smaller extent.

An interesting feature of Dutch is that in the coda two different allophones of /r/ are allowed, either the fricative-like /r/ or the velar /R/ which is similar to the American English /r/. The choice for either of these is largely determined by dialectical background of the speaker. In Nooteboom's research the speaker only used /R/. Our speaker sometimes used the /r/ and sometimes the /R/, which is very useful for comparison purposes. A consistent difference can be found between the two contexts. The vowels in the subclass iuy are lengthened before /R/, but even more so before /r/. For the other vowel categories, shortening is observed before /R/ and lengthening before /r/. A possible explanation is the fact that some centralisation of iuy takes place before /R/ and /r/. It could be that for the other vowels this phenomenon only occurs before /r/.

The effect of *stress and accent* on vowels in closed syllables is similar to that in open syllables, except that the durations of the long vowels are somewhat shorter.

Like in open syllables, the effect of *left position* is small, but the same consistent behaviour can be seen. Vowels in word-initial syllables are slightly longer than in non-initial syllables.

The effect of *right position* persists even when the vowel is followed by a coda comprising one or more consonants. The number of levels is reduced to three: non-final syllables, word-final syllables, and phrase-final syllables. Because at least one consonant follows the vowel in the same syllable, a vowel can never be the final segment in the syllable. The effect is smaller in size, though (see Figure 4.10). The picture is clear. Vowels are lengthened when in a phrase-final syllable. Additionally, they are longer in word-final than in non-final syllables.

4.4.3 Consonants

Figure 4.11 displays the corrected means for all consonants in onset and coda position. The consonants are grouped by category. Obstruents are longer than nasals, which in turn are longer than glides and liquids. Voiceless obstruents are longer than voiced ones. These findings are in line with Waals (1999), who found an increasing duration as sonority decreases.

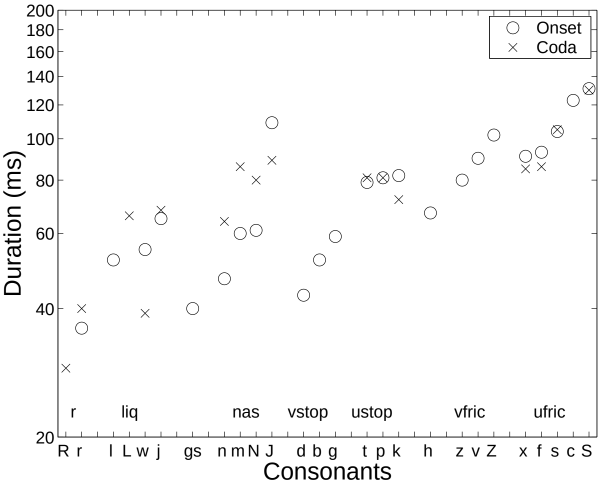


Figure 4.11: *Corrected means of consonants in onsets and codas.*

Consonants in codas were found to be longer in duration than in onsets by Crystal and House (1988) and Waals (1999). For our speaker, this difference is most visible in nasals. Other categories seem to have similar durations in both onsets and codas.

Onsets

Segment class	Previous	Next	Stress/ Accent	LPOS	RPOS
nas	1.94	1.08	1.66	1.37	1.04
liq	1.63	1.22	1.33	1.22	-
r	2.56	-	1.09	-	-
ufric	1.66	1.13	1.36	1.16	1.16
vfric	3.83	-	1.45	1.21	1.08
ustop	1.42	1.52	1.06	-	1.12
ustop-cl	2.23	1.47	1.43	1.39	-
vstop	1.28	1.30	1.05	-	-
vstop-cl	2.59	1.25	1.61	1.33	-
h	1.35	1.76	1.11	-	-
gs	1.61	-	2.70	-	1.22

Table 4.4: *Range ratios, indicating the proportional difference between the shortest and longest values for all factors, for vowels in open syllables.*

Let us first focus on consonants in onsets to see what phenomena we can observe. (An overview of the corrected means per factor can be found in Tables B.9 to B.19.) In Table 4.4 we can see that the previous segment has a considerable effect. For some subclasses, the next segment is also important. Stress and accent have a large effect on most categories.

It is more difficult to generalise the effects of the previous and next factor to all consonant categories. There is much variation in them due to the different distributional properties of the consonants. According to Dutch phonology, the onset can consist of a minimum of zero to a maximum of three consonants. When there is just one consonant, all consonants can occur except /R/, /N/ and /L/, which occur only in codas. Bi-consonantal onsets can consist either of an /s/ followed by an obstruent or sonorant, or an obstruent followed by a liquid (/r/ or /l/). /zw/ is the only voiced obstruent-liquid pair. Tri-consonantal clusters occur only word-initially. They always start with /s/, which is considered to be extra-syllabic, followed by a voiceless obstruent and a liquid.

When sonorants (nasals, liquids and glides) occur on their own in the onset, their duration is affected by the segment class of the last segment in the preceding syllable. Sonorants are shortest when preceded by an obstruent, which in a Dutch coda is always unvoiced, and longer after vowels or sonorants. As an example, the corrected means for nasals are 51 ms after an obstruent, 74 ms after a vowel and 82 ms after a sonorant. When the obstruent is part of the same consonant cluster, the sonorant is shortened even more. A nasal is then 42 ms, indicating a 17% shortening compared to the obstruent-sonorant pair that is separated by a syllable boundary. No difference is observed between the duration of /l/ in bi- and tri-consonantal clusters. In addition, for /l/ in a bi-consonantal cluster, the voiced-unvoiced property of the preceding obstruent is irrelevant. For all sonorants, there is a small but consistent effect of the quantity of the next vowel. They are slightly longer (5-8 ms) before a long vowel than before a short vowel.

The consonant /r/ is often counted among the liquids but our results indicate different behaviour with respect to previous and next context. It has the same distributional properties as /l/, but voicing of the preceding consonant in the cluster matters and shortening is greater in tri-consonantal clusters than in bi-consonantal clusters. The duration of /r/ is 46 ms when preceded by a voiced obstruent in the same syllable, 38 ms when preceded by an unvoiced obstruent and 29 ms when preceded by an /s/ plus an unvoiced obstruent. Also, it is the only consonant that displays an effect of intervocalic shortening. The quantity of the next vowel (short or long vowel) is irrelevant.

The unvoiced fricatives /S/ and /c/ occur only as singletons. For all fricatives occurring as singletons, their duration is shorter when preceded by an obstruent than when preceded by a vowel or sonorant. Bi-consonantal onsets can either consist of obstruent-obstruent combinations, for instance /sf/, /sx/, and /ts/, or of obstruent-sonorant combinations, for instance /xr/ and /fr/. In the first case, shortening of the fricative lies around 28%. In the latter case, we observe a small lengthening effect. For instance, /x/ is 8 ms longer when followed by a sonorant than by a vowel.

In unvoiced stops, the closures are more affected by their segmental context than the bursts. The bursts are shortest when the stops occur as second consonant in a bi- or tri-consonantal cluster and the same goes for the closures. When they are the first consonant in a cluster, the duration is not shortened. Only the identity of the preceding segment plays a role. After nasals, closures are shorter than after unvoiced obstruents. They are longest after other sonorants. When followed by a vowel, the bursts are slightly shorter when the vowel is a schwa.

Voiced fricatives can either occur as singletons or as the first consonant in a bi-consonantal cluster followed by /l/ or /r/. The duration does not change in these two conditions. When preceded by a sonorant in the previous syllable, it is longer than when preceded by an obstruent. In eight cases, a voiced fricative occurred in a consonant cluster preceded by a voiced obstruent. These were all loan words from English, e.g., *George /dZORtS/*. In these cases, the /Z/ was extremely shortened.

The bursts in voiced stops were hardly effected by the context. The closures were shortest when preceded by a nasal and longest when preceded by a vowel. When followed by a schwa, it was also shortened somewhat in comparison to other contexts.

When the first phoneme in the onset has the same phoneme identity as the last phoneme in the previous word, degemination occurs. This is a phonological phenomenon that deletes one of the consonants. A study by Martens and Quené (1994) shows that complete deletion does not occur. The duration is shorter than the sum of the two consonants normally would be, but they are longer than in the case of a single consonant onset. Speech rate affects the amount of shortening. Our data shows the duration to be around 75% of the total duration of the two consonants rather than 50%. In the case of stops, one stop is left which is about 15% longer than when it would have occurred as a singleton.

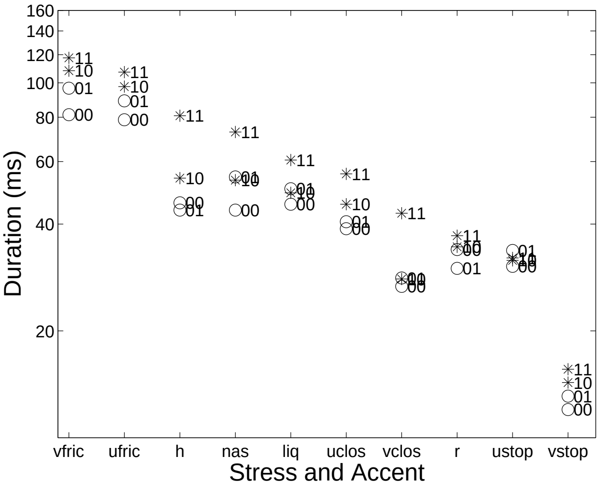


Figure 4.12: *Effect of stress and accent on consonants in onsets; 00 = unstressed/unaccented, 01 = unstressed/accented, 10 = stressed/unaccented, 11 = stressed/accented.*

The effect of *stress and accent* varies greatly among categories (see Figure 4.12). The largest effects can be observed in /h/, fricatives, and stop closures. On stop bursts, this factor has only a small effect.

The effect of *left position* is small but consistent across all categories. Consonants are slightly longer when in a word-initial syllable than in a non-initial syllable. The effect of *right position* is equally small. Here, consonants are slightly longer in word-final syllables than in non-final syllables. Only in the case of unvoiced stops is there no difference between non-final and word-final, but a small effect in phrase-final syllables can be observed, which implies that the entire final syllable in the phrase is affected to some extent. The setup of this study does not allow separate analysis of penultimate syllables, so we do not know for sure whether the effect persists even there. But considering the small effect on consonant onsets in the final syllable, an effect in the penultimate syllable is not very likely.

Codas

In Dutch codas, all obstruents are devoiced. Furthermore, the /h/ and glottal stop do not occur. This reduces the categories to six. Bi-consonantal codas consist either of two obstruents (e.g., /pt/, /ts/, /ft/), two sonorants (e.g., /lm/, /rn/), or a sonorant and an obstruent (e.g., /lp/, /nt/, /ft/). In tri-consonantal codas the last consonant is always a dental obstruent (/t/ or /s/). The first two consonants can be two obstruents or a sonorant followed by an obstruent.

Segment	Previous	Next	Stress/	LPOS	RPOS
class			Accent		
nas	1.39	1.10	1.99	1.60	3.10
liq	-	1.75	1.38	1.20	2.18
r	-	1.78	1.92	-	2.21
ufric	1.27	1.42	1.46	-	2.49
ustop	1.37	1.75	1.39	-	5.29
ustop-cl	2.30	1.45	1.26	-	1.37

Table 4.5: *Range ratios, indicating the proportional difference between the shortest and longest values for all factors, for vowels in open syllables.*

(An overview of the corrected means per factor can be found in Tables B.20 to B.25.) The range ratios listed in Table 4.5 show that the right position is the factor with the largest effect on codas. The effect of the previous segment was only relevant for unvoiced obstruents and nasals. For the other categories, it was discarded. The next factor also had a considerable influence.

Of the unvoiced fricatives it is usually the /s/ occurring as the second element in a bi- or tri-consonantal cluster. It is shorter after a consonant than after a vowel, and it also makes a difference whether the preceding consonant is an unvoiced obstruent or a sonorant. The following segment has a lengthening effect when it is a consonant and a shortening effect when it is a vowel. The effect of left position was only relevant for nasals and liquids. For other subclasses it was discarded.

The bursts in unvoiced stops are not affected very much by their segmental context. They are shortest when followed by a voiced stop in the following syllable. Other than that, it is hardly relevant whether it is part of a larger cluster or not. The closures however, are much more affected. In Table B.25 we can see than when the unvoiced stop closure is in a cluster preceded by a nasal the duration is 20 ms, when preceded by an unvoiced stop it is 29 ms, and when preceded by a vowel or sonorant it is 45 ms. A stop can also be the first consonant in a cluster followed by /s/. In that case, the closure is slightly shorter than when it marks the end of the syllable.

For liquids and glides, the difference in duration between occurrence as a singleton or in a bi-consonantal cluster is only small. This is in line with findings by Waals (1999). /L/ is shortest when followed by a schwa. For /r/, only the following segment matters. The /r/ is shortest when followed by a vowel. This is related to the intervocalic shortening observed in the onset /r/. When it is followed by an unvoiced obstruent in a bi-consonantal cluster, it is about 10 ms shorter than when followed by a syllable boundary.

There are only nine cases where a nasal is preceded by another sonorant in a coda. They are not shorter than when preceded by a vowel. When a nasal is followed by a voiced obstruent in the onset of the next syllable, it is longer than when it is followed by any other category. In cases where the nasal is followed by an unvoiced obstruent in a bi-consonantal cluster, shortening takes place, although for /n/ this effect is minor. For /m/ the difference is 30%.

Figure 4.13 shows similar behaviour to the effect of stress and accent on onset consonants. Nevertheless, the range ratios vary greatly between classes. This also shows that the amount of interaction between stress and accent is different for various subclasses.

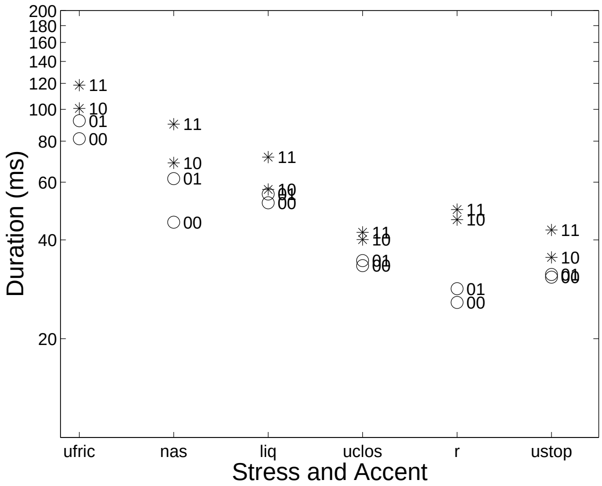


Figure 4.13: *Effect of stress and accent on consonants in codas; 00 = unstressed/unaccented, 01 = unstressed/accented, 10 = stressed/unaccented, 11 = stressed/accented.*

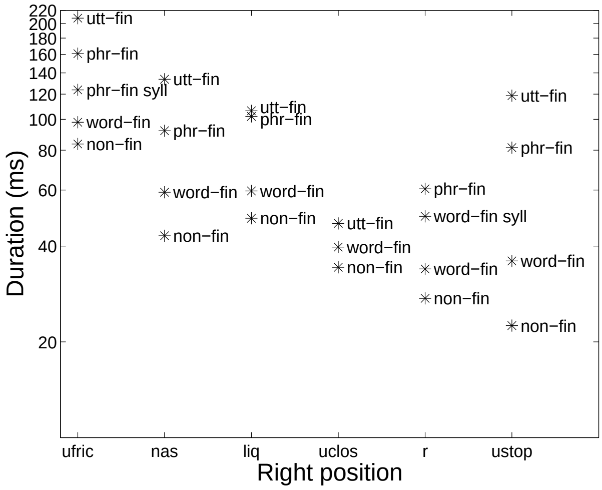


Figure 4.14: *Effect of right position on consonants in codas.*

Figure 4.14 shows that the effect of right position on consonants in codas is considerable. It has often been reported that lengthening occurs at the end of a phrase or sentence. Our results also show word-final consonants to be longer than non-final ones. The greatest effect is achieved on unvoiced stop bursts. The unvoiced stop closures however, are hardly affected.

4.4.4 Conclusion

This concludes the discussion of the duration model developed for Dutch using the sums-of-products approach. We have seen that with a relatively small corpus of 297 sentences (16,775 segments) a duration model can be constructed with a RMSE of 27 ms. The use of phonetic and phonological knowledge in the construction of the category tree and the data analysis, which was designed specifically for this purpose, results in a model that reflects the most important effects and interactions found in factors influencing duration prediction.

It could be observed that directional invariance holds for most categories, justifying the approach taken to generalise to unseen cases in the sums-of-products model. Even though the results only reflect data from one particular speaker, some effects reported on in literature on Dutch duration could be confirmed. First, the large effect of /r/ on the vowels /i/, /u/, and /y/ reported on by Moulton (1962) and Nooteboom (1972) was confirmed. Because the data used in our study is closer to natural speech than in their small-scale studies, it can even be refined. It was found that considerable lengthening takes place regardless of whether the /r/ occurred in the coda of the same syllable or in the onset of the next syllable, although the lengthening effect was not so great in the latter case. Furthermore, the effect was larger when iuy were followed by a fricative /r/ than by a velar /R/.

Second, the interaction between stress and accent as found by Eefting (1991) was confirmed. For this reason, these two factors were combined into one factor.

Third, it was found that lengthening of consonants and vowels occurs both in word-initial and in word-final syllables. This is probably an important cue for listeners to segment a stream of sounds into words.

Two additional observations have been made with respect to the data that are not included in the model. The first observation is that for all consonants, the coronal consonants usually show a larger effect than labial and post-coronal consonants. Van Son and Van Santen (1997) found the same result for American English speech data. We decided not to include this parameter in our model, as the training corpus would need to be extended considerably and only a small improvement was to be expected.

The second observation is an interaction between stress and accent and right position. The effect of right position, i.e., lengthening in phrase-final syllables, appears to be attenuated by the effect of stress and accent. This is in line with findings by Cambier Langeveld (2000) on Dutch duration data. Cambier Langeveld (2000) compared the Dutch data to English data, but found no such interaction there. This interaction was not included in our model, again because a considerable extension of the training corpus would be necessary.

4.5 Evaluation

In this section, the performance of the new sums-of-products duration model (NDM) and the old rule-based duration model (ODM) will be compared. Section 4.5.1 discusses a perceptual experiment in which sentences generated with the diphone synthesis system using either the NDM or the ODM were presented to a group of listeners in a pairwise comparison. Section 4.5.2 discusses a quantitative comparison between the NDM and the ODM.

4.5.1 Subjective evaluation

In order to find out which duration model listeners prefer, a paired comparison listening experiment was performed.

Material

The material for the experiment consisted of 20 sentences taken from news reports on Teletext. Each sentence had a duration of between 6 and 8 seconds, i.e., long enough to get a good impression of its durational structure. The number of phonemes per sentence ranged from 58 to 93. Each sentence was generated twice with the same diphone synthesis system, once using the old duration model and once using the new model. The durations were not normalised for differences in speaking rate, since this difference was less than 5%. Moreover, if changes in speaking rate are achieved by linear lengthening or shortening, they might have an effect on the accuracy of the predicted durations and thus on the output quality. Prior to the start of the experiment, an example pair was played to introduce the task at hand. See Appendix C for a description of the sentences used. They can be listened to on the accompanying web site.

Procedure

The experiment was presented over the internet, using a standard Webbrowser. The participants needed a PC or SGI station with audio output, a speaker and a headset. They had to press a button to listen to a pair of sentences and had to express their preferences on a five-point scale which correspond to the following statements:

- -2: I prefer the first sentence strongly over the second sentence.
- -1: I prefer the first sentence slightly over the second sentence.
- 0: I do not have a preference for either the first or the second sentence.
- 1: I prefer the second sentence slightly over the first sentence.
- 2: I prefer the second sentence strongly over the first sentence.

The sentences were presented in random order. In 10 out of 20 cases, the first sentence was generated with the ODM and the second sentence was generated with the NDM (order i, j), and in the 10 other cases it was the other way around (order j, i). The presentation of these different orders was also randomised for each participant.

Subjects

Twenty participants with a background in phonetics took part in the experiment. They had no reported hearing problems and were not paid for their participation.

Results

The scores were processed such that negative scores indicated a preference for the old model and positive scores indicated a preference for the new model. With 20 participants rating 20 sentence pairs, a total of 400 observations was obtained. Figure 4.15 displays the distribution of scores in the experiment. The mean score was 0.1125, which reveals a slight preference for the new model over the old one. The standard deviation was 1.17. A one-sample t-test was conducted (t(399) = 1.920, p = 0.056), which revealed that the difference of the mean from zero was not significant. The 95% confidence interval for the mean ranged from -0.003 to 0.23.

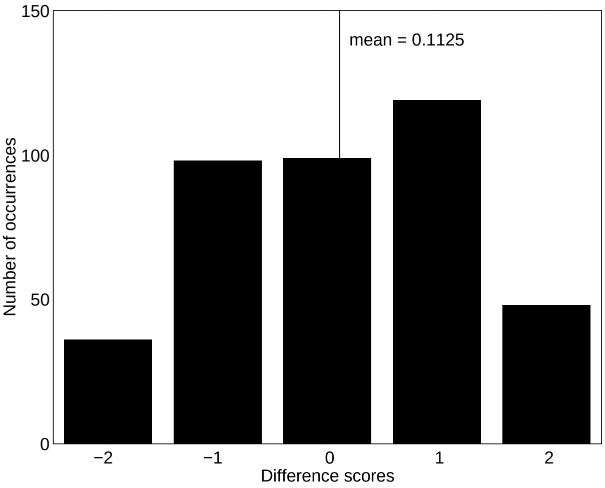


Figure 4.15: *Results from the preferential choice experiment.*

A Wilcoxon Matched-Pairs Signed Ranks Test (Sheskin, 1996) was performed. It is more sensitive than the sign test in that it considers the size of the differences in positive and negative direction and tests whether they have the same probability. Zero scores are ignored. The result (W- 20024, W+ 25427, N=301, p <= 0.06) shows that there are more and larger positive scores than negative ones, but the difference is not significant. So one has to conclude that the NDM is not judged as being significantly better than the ODM. Next to this overall impression, the results per sentence were also considered to detect any differences between them. Per sentence, a Wilcoxon Matched-Pairs Signed Ranks Test was performed. It proves that the NDM is judged as being significantly better in 5 cases and the ODM is judged as being significantly better in 3 cases (For details see Table 4.6). In the other 12 cases, the difference is not significant.

NDM preferred				
sentence 2	W- 12	W+ 66	N = 15	$p < 0.03$
sentence 4	W- 15	W+ 156	N = 18	$p < 0.001$
sentence 6	W- 4.5	W+ 148.4	N = 17	$p < 0.001$
sentence 8	W- 17.5	W+ 73.5	N = 13	$p < 0.05$
sentence 11	W- 4	W+ 51	N = 10	$p < 0.02$
sentence 5	W- 32	W+ 104	N = 16	n.s.
sentence 9	W- 34	W+ 102	N = 16	n.s.
sentence 10	W- 41	W+ 95	N = 16	n.s.
sentence 13	W- 28.5	W+ 37.5	N = 11	n.s.
sentence 15	W- 57.5	W+ 132.5	N = 19	n.s.
ODM preferred				
sentence 1	W- 97.5	W+ 22.5	N = 15	$p < 0.03$
sentence 16	W- 116.5	W+ 19.5	N = 16	$p < 0.01$
sentence 19	W- 145.5	W+ 44.5	N = 19	$p < 0.04$
sentence 3	W- 80	W+ 73	N = 17	n.s.
sentence 7	W- 105.5	W+ 47.5	N = 18	n.s.
sentence 12	W- 33	W+ 22	N = 10	n.s.
sentence 14	W- 69	W+ 51	N = 15	n.s.
sentence 17	W- 52.5	W+ 52.5	N = 14	n.s.
sentence 18	W- 114.5	W+ 38.5	N = 17	n.s.
sentence 20	W- 61	W+ 30	N = 13	n.s.

Table 4.6: *Wilcoxon Matched-Pairs Signed Ranks Test for each of the sentences in the evaluation experiment.*

4.5.2 Quantitative comparison between observed and predicted durations

In order to make some statement about the objective performance of the ODM and NDM, correlations and RMS errors were computed for both models relative to the observed durations in the training corpus. After computing these measures for the entire corpus, they were computed separately for each of the phonetic subclasses. Table 4.7 shows the data obtained. In the field of automatic speech recognition and pattern recognition, it is common practice to evaluate the performance on an independent test set. In this study, in order to get a good view of the performance of both duration models, a large test set would be required, covering all subclasses and factors sufficiently. It would require a lot of effort to record and segment this test set, which was not available at the time. However, Maghbouleh (1996) has shown that RMSE and correlations are approximately equal on training and testing data for the sums-of-products approach. Therefore, it is expected that the results presented in Table 4.7 will be similar to the situation where an independent test set is used.

			NDM		ODM	
Class		N	Corr	RMSE (ms)	Corr	RMSE (ms)
	all	12918	0.76	26.96	0.62	31.88
Nucleus	lvow	739	0.70	33.30	0.47	42.13
(open)	svow	214	0.33	21.17	0.27	20.85
	iuy	358	0.68	27.82	0.53	33.98
	schwa	692	0.40	24.66	0.41	23.49
Nucleus	lvow	601	0.61	36.27	0.49	41.89
(closed)	svow	1424	0.56	19.28	0.49	20.19
	iuy	164	0.75	26.14	0.64	28.96
	schwa	739	0.20	30.16	0.29	32.96
onset	ustop	797	0.48	24.17	0.43	31.72
	ufric	685	0.53	26.20	0.26	31.01
	vstop	852	0.53	24.36	0.41	25.63
	vfric	583	0.52	25.39	0.15	39.17
	nas	493	0.65	21.64	0.38	26.80
	liq	556	0.55	19.31	0.34	23.77
	r	412	0.41	23.42	0.30	33.00
	h	148	0.38	28.76	0.35	27.03
coda	ustop	848	0.75	37.56	0.59	43.42
	ufric	637	0.77	30.07	0.58	38.85
	nas	1116	0.60	27.17	0.42	29.24
	liq	247	0.66	21.34	0.25	27.27
	r	513	0.54	19.34	0.45	29.53

Table 4.7: *Objective evaluation of NDM and ODM in terms of Pearson's r correlation and RMSE relative to the observed durations.*

Table 4.7 shows that the performance of the NDM on the entire corpus is better than that of the ODM. The correlation is higher (0.76 vs. 0.62) and the RMSE is lower (27 vs. 32 ms). A paired-samples t-test which was performed comparing the correlations of the subclasses for the new and old model, shows that the correlations in the new model are significantly higher than in the old one (t(20) = 5.637, $p < 0.001$). The mean difference was 0.15 with a 95% confidence interval running from 0.09 to 0.20. The difference in RMSE is also significant (t(20) = -5.599, $p < 0.001$). On average the RMSE is 4.9 ms shorter with a 95% confidence interval between 3.09 and 6.76. The largest difference can be observed in the subclass containing voiced fricatives in onsets, where the difference in RMSE is 14 ms. The schwa is difficult to predict with high accuracy as the low correlation suggests. The short vowels are predicted equally well by the new and the old duration model.

4.6 Discussion

4.6.1 Explanation of findings

In the previous section, it was observed that although the objective evaluation shows an improvement of the new duration model over the old one, the subjective evaluation does not convincingly support that finding. There could be several reasons for this. First, the objective differences between the NDM and ODM are often very small. It could be the case that they are below threshold. Second, it is observed that the occurrence of one or more errors in the prediction has an effect on the perceived quality of the entire sentence. Third, it could be that the poor quality of the synthesis itself causes differences in duration to go unnoticed.

In informal listening we have observed that the new duration model usually has a better rhythmic build-up than the old model. Analysis of the twenty sentences used in the experiment reveals the differences in prediction quite well.

1. Diphthongs are usually longer in the ODM than in the NDM. This was known to be a problem in the ODM. RMSE values obtained in the comparison of both duration models with the observed durations in the corpus show that the new model indeed has a smaller prediction error for diphthongs than the old model (29.4 versus 44.3 ms).

2. Voiced fricatives are often much longer in the new model than in the old model. Because the objective evaluation presented in Table 4.7 reveals that the voiced fricatives are predicted much more accurately by the new model than by the old one (RMSE of 25.4 vs. 39.2 ms), we feel safe in saying that in this respect the NDM is superior.

3. Lengthening of the vowels /i/, /u/, and /y/ before /r/ is often

much greater in the new model than in the old one. The RMSE values confirm that the new model is more accurate (37.9 versus 49.9 ms).

4. Phrase- and sentence-final lengthening is sometimes much larger for the NDM than for the ODM.

This chapter has shown that a duration model using the sums-of-products approach can be produced in a very straightforward manner, resulting in a quality that is as good as the old duration model, and in some cases even better. There are still a few things left for future research. First, it would be worthwhile to record the twenty sentences of the subjective evaluation experiment using the same female speaker, and compare the observed durations to those predicted by the ODM and the NDM. The availability of these sentences in natural speech also makes it possible to transplant the predicted durations onto these sentences, such that the poor quality of the synthesis is factored out. The perceptual experiment can then be repeated. Second, the NDM can be further improved by adding more training data.

In a study by Bellegarda (1998), an improved duration modelling with the sums-of-products approach was proposed using a root-sinusoidal transformation on the raw durations instead of a logarithmic one. In his view, short durations are underestimated and long durations are overestimated in the sums-of-products approach. The transformation is carried out using Equation 4.6. There x represents the raw duration, A is the minimum observed duration in the training corpus and B is the maximum observed duration.

$$F(x) = \sin\left\{\frac{\pi}{2}\left(\frac{x-A}{B-A}\right)^{0.8}\right\}^{2} \tag{4.6}$$

Applying this transformation to the durations obtained with the NDM did not give a significant improvement. The new correlation with the observed duration is 0.761 and the RMSE is 30.91, which is worse than what was obtained with the standard sums-of-products model. Apparently, it is not the case that short durations are underestimated and long durations are overestimated (as can also be deducted from Figure 4.3), or else the improvement would have been clear.

4.6.2 Finding a meaningful predictor for perceived quality

Another important issue is that it is not certain whether the RMSE is an adequate predictor of perceived quality. It does not take into account the fact that errors in prediction may be less or more acceptable dependent on the phoneme class (e.g. vowels vs. consonants), the stress or accent status, or the position of the syllable in the word, etc. In a study by Kato, Tsuzaki, and Sagisaka (1998), it was investigated what the listeners' acceptability for temporal modification was of single vowel segments in isolated words. It was shown that the acceptable range of modification depended on the position of the vowel in the word, such that the range was narrower for vowels in the first moraic position versus the third moraic position. It was also narrower for the vowel /a/ than for /i/ and similarly narrower for vowels followed by unvoiced consonants than voiced consonants. This suggests that the RMSE may not be a suitable predictor of acceptability because no weighting of factors takes place.

Córdoba et al. (1999) proposed the use of a relative RMSE instead of an absolute RMSE, in which the average duration $\overline{t}$ and the optimum durations t_i are taken into account.

$$RelativeRMSE = \frac{RMSE}{\sqrt{(t_i - \overline{t})^2}} \tag{4.7}$$

However, Kato et al. (1998) also found in their study that the base duration of the vowels did not affect the acceptable range of modification, despite the fact that the durations varied widely (between 35 and 145 ms). This implies that a relative RMSE will not be a better predictor than the absolute RMSE. It must be marked however, that the results presented by Kato et al. (1998) have been obtained in a small-scale study using isolated words and it is not obvious how this generalises to natural running speech as used in our own database.

Chapter 5

Summary and conclusion

5.1 Summary of findings

This thesis aimed at finding segmental and prosodic improvements in speech generation. Two types of speech generation were investigated: diphone synthesis and phrase concatenation. The research was carried out in the context of a spoken dialogue system called OVIS, which gives train timetable information over the telephone. In this type of applications, the approach to speech generation is often very simplistic. The speech output is achieved by a straightforward concatenation of words and phrases. They literally correspond to the text that is to be spoken. This approach has a major drawback in that it lacks variability in accentuation and the marking of phrase boundaries, which is essential for creating natural speech. Proper accentuation is necessary to highlight the information structure of the messages by accenting new or contrastive information and deaccenting given information, and prosodic boundary marking helps to highlight the linguistic structure of the messages by dividing them into smaller pieces that can be comprehended more easily. The fact that the words and phrases to be concatenated are often spoken in isolation results in mismatches in pitch, loudness and tempo which makes the speech sound less fluent.

In OVIS, the natural language generation module provides the prosodic information needed to introduce the required variability. It generates the location of accents and phrase boundaries on the basis of syntactic, semantic and pragmatic information. These locations can be predicted with more accuracy via a data-to-speech system than when a text-to-speech system computes them from unknown text.

The speech generation module had to be extended, so that it could accommodate the prosodic variability introduced by the natural language generation module. To this end, an advanced phrase concatenation technique was developed that is superior to a straightforward sequencing of words and phrases. Different versions of words were recorded dependent on whether they occur in an accented or unaccented condition and on whether they occur before a phrase boundary of some depth. All words and phrases were recorded in the proper context, to automatically elicit the appropriate versions. That and a careful recording resulted in speech output that sounded almost as fluent as natural speech and that minimised the occurrence of mismatch in pitch, loudness and tempo.

The drawback of this phrase concatenation approach is that a considerable amount of phonetic knowledge is required to construct a corpus that covers the sentences the natural language generation module creates with their prosodic variations. For every new application, this has to be figured out again. Moreover, this approach to speech generation only works for applications that suffice with a medium-sized stable vocabulary. There is a maximum to the amount of speech a speaker can utter with adequate constancy and regular updates of the recordings will soon become impractical. When the need for flexibility increases, speech synthesis is the only alternative that is left. Although the most popular implementation, i.e., diphone synthesis, is fairly intelligible, listeners still judge the overall quality too low for it to be used in commercial applications.

Chapter 3 and 4 concentrated on two problems affecting the quality of diphone synthesis. A significant problem that interferes with the segmental quality is the occurrence of audible discontinuities at diphone boundaries. In Chapter 3, it was shown that the addition of context-sensitive diphones to the standard diphone database reduces the number of audible discontinuities. These discontinuities are mainly caused by spectral mismatch at the diphone boundaries. A listening experiment was conducted to investigate the extent to which discontinuities are perceived. The results revealed considerable differences between the vowels under investigation. The /a/ showed the smallest percentage of perceived discontinuities, followed by the /i/. The /A/ and /I/ showed a higher percentage, but the largest percentage of perceived discontinuities was found in the /u/. The alveolars account for most of the coarticulation, probably because their F_2 is much higher than that for /u/. The scores obtained in the listening experiment were correlated with several spectral distance measures to find an objective measure that predicts the occurrence of audible discontinuities. The correlation was performed using Receiver Operator Characteristics. The Kullback-Leibler (KL) distance, coming from statistics, was shown to be most adequate for the task.

Context-sensitive diphones were added to the diphone database. In order to limit the number of additional diphones, the KL distance was used to cluster consonantal contexts with similar spectral effects on the neighbouring vowels. A second listening experiment was conducted for the vowels /a/, /i/, and /u/, to evaluate the improvement obtained with this addition to the database. The results show that the number of audible discontinuities has significantly decreased for all vowels. The KL distance has significantly decreased for /i/ and /u/, but not for /a/. Future research in this area could be directed at extending the diphone database even further, including context-sensitive diphones for other phonemes that are heavily affected by coarticulation, for instance /o/, /O/ and /y/ and semivowels like /l/ and /w/. An additional improvement might be obtained by recording several repetitions of each diphone to be able to select the one that best fits a particular context.

A second problem that affects the prosodic quality is the inadequate prediction of segmental durations. Therefore, a new duration model was developed using the sums-of-products approach of Van Santen (1992a). Chapter 4 reports on the development of this new model. With a relatively small corpus of 297 sentences, containing 16,775 segments, a duration model could be constructed, which compares well with similar models constructed for French and German. The use of phonetic and phonological knowledge in the construction of the category tree and the data analysis, which was designed specifically for this purpose, results in a model that incorporates the most important effects and interactions found in factors influencing duration prediction.

The performance of the old and the new duration models was evaluated by comparing them to the actual segmental durations in the 297-sentence training corpus. The new model was shown to be a significant improvement over the old one. A perceptual pairwise comparison experiment using sentences generated with the old and new model, did not provide conclusive evidence for the improvement. There could be additional reasons for this. First, the objective differences between the new duration model and the old one are often very small. It could be the case that they are below threshold. Second, it was observed that the occurrence of one or more errors in the prediction has an effect on the perceived quality of the entire sentence.

5.2 Conclusion and future research

The research presented in this thesis has shown that with basic phonetic research and detailed perceptual experiments, some major problems in diphone synthesis and phrase concatenation could be brought closer to a solution. Unfortunately, the inherent disadvantages of these two approaches still remain. Although phrase concatenation can now produce near-natural speech, it is still inflexible in that units have to be carefully selected, recorded and excised. Diphone synthesis still produces unnatural results. Probably the most important problem in diphone synthesis is the deterioration of the speech quality due to signal modifications. This can be remedied in two ways, either by coming up with a new signal representation that leaves the original speech quality intact or by avoiding signal modification as much as possible.

As a possible solution for both speech generation techniques, an extension to diphone synthesis often referred to as *unit selection* has become quite popular recently (Campbell and Black, 1997; Black and Taylor, 1997; Balestri, Pacchiotti, Quazza, Salza, and Sandri, 1999). It differs from conventional diphone synthesis systems in a number of ways. First, the corpus can often be many times larger than the standard diphone database, allowing for several instances of units that vary in context, in terms of e.g., surrounding phonemes, stress, accentuation, and position in the word or phrase. Second, the size of the unit can be non-uniform. By segmenting the database into half phones, units ranging from phones and diphones to syllables, words and even phrases can be selected. And third, the units need not be extracted off-line, but can be selected on-line on the basis of automatic selection criteria that compute the *concatenation cost*, i.e., how well will this unit concatenate with the previously selected unit, and the *target cost*, i.e., how well does this unit fit the computed prosodic characteristics.

When the database is tuned to a particular application, the vocabulary coverage increases, which makes the approach similar to the phrase concatenation approach described in Chapter 2 of this thesis. The only difference is that the units are not pre-excised. A first attempt is reported by Stöber et al. (1999), who used it in the Verbmobil project, an application which offers translation assistance in a dialogue situation.

The general expectation of this unit selection approach is that less audible discontinuities will occur because the units are selected from the best suitable context and because the number of concatenative boundaries is usually lower. Moreover, because the units are taken from the appropriate prosodic and acoustic context their prosodic characteristics will better fit the previously selected units, thus making prosodic modifications unnecessary in many cases.

The success of unit selection crucially depends on the knowledge that was gained with diphone synthesis and phrase concatenation in the past. The construction of the corpus is the most important step in the development of such a synthesis system. Similar to the duration corpus presented in Chapter 4, it has to cover a large feature space taking into account preceding and following phonemes, stress, accentuation and positional factors. In many cases complete coverage will not be possible, thus requiring diphones as a backup. Computation of F_0 and duration is still necessary to compare units to target values and to make slight modifications whenever needed. A suitable spectral distance measure is essential to compute the concatenation cost in order to avoid audible discontinuities at the unit boundaries. Even then, discontinuities can occur due to differences in loudness or voice quality.

Whether unit selection is the solution to generate natural sounding speech, remains to be seen. This question can only be answered by continuing basic phonetic, prosodic and perceptual research in the area of speech synthesis.

Bibliography

Aust, H., M. Oerder, F. Seide, and V. Steinbiss (1995). The Philips automatic train timetable information system. *Speech Communication 17*, 249–262.

Balestri, M., A. Pacchiotti, S. Quazza, P. Salza, and S. Sandri (1999). Choose the best to modify the least: A new generation concatenative synthesis system. In: *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH'99), Budapest, Hungary*, Volume 5, pp. 2291–2294.

Bellegarda, J. (1998). Improved duration modeling of English phonemes using a root sinusoidal transformation. In: *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'98), Sydney, Australia*, Volume 2, pp. 21–24.

Black, A. and P. Taylor (1997). Automatically clustering similar units for unit selection in speech synthesis. In: *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH'97), Rhodes, Greece*, Volume 2, pp. 601–604.

Boersma, P. and D. Weenink (1996). Praat, a system for doing Phonetics by Computer. Report 132, Institute of Phonetic Sciences of the University of Amsterdam, Amsterdam. http://www.praat.org/.

Bonnema, R., R. Bod, and R. Scha (1997). A DOP model for semantic interpretation. In: *Proceedings of the ACL/EACL 1997, Madrid, Spain*, pp. 159–167.

Cambier Langeveld, T. (2000). *Temporal marking of accents and boundaries*. Ph. D. thesis, University of Amsterdam.

Campbell, N. and A. Black (1997). Prosody and the selection of source units for concatenative synthesis. In: J. Van Santen, R. Sproat,

J. Olive, and J. Hirschberg (Eds.), *Progress in speech synthesis*, pp. 279–292. New York: Springer-Verlag.

Campbell, W. (1992). Syllable-based segmental durations. In: G. Bailly, C. Benoît, and T. Sawallis (Eds.), *Talking machines: Theories, models and designs*, pp. 43–60. Amsterdam: Elsevier.

Carvalho, P., L. Oliveira, I. Trancoso, and M. Viana (1998). Concatenative speech synthesis for European Portuguese. In: *Proceedings of the 3rd ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan Caves, Australia*, pp. 159–163.

Chappell, D. and J. Hansen (1998). Spectral smoothing for concatenative speech synthesis. In: *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'98), Sydney, Australia*, Volume 5, pp. 1935–1938.

Charpentier, F. and E. Moulines (1989). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. In: *Proceedings of the 1st European Conference on Speech Communication and Technology (EUROSPEECH'89), Paris, France*, Volume 2, pp. 13–19.

Collier, R. and A. Houtsma (1995). Hearing and Speech: Developments. In: *IPO Annual Progress Report*, Volume 30, Eindhoven, pp. 20.

Collier, R. and J. 't Hart (1981). *Cursus Nederlandse Intonatie*. Leuven: Acco.

Conkie, A. and S. Isard (1997). Optimal coupling of diphones. In: J. Van Santen, R. Sproat, J. Olive, and J. Hirschberg (Eds.), *Progress in speech synthesis*, pp. 293–304. New York: Springer-Verlag.

Córdoba, R., J. Vallejo, J. Montero, J. Gutierrez-Arriola, M. López, and J. Pardo (1999). Automatic modeling of duration in a Spanish text-to-speech system using neural networks. In: *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH'99), Budapest, Hungary*, Volume 4, pp. 1619–1622.

Covell, M., M. Withgott, and M. Slaney (1998). MACH1 for nonuniform time-scale modification of speech: Theory, technique, and comparisons. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'98), Seattle, Washington*, pp. 349–352.

Crystal, T. and A. House (1988). Segmental durations in connected speech signals: Current results. *Journal of the Acoustical Society of America 83*(1), 1553–1573.

De Pijper, J. (1996). High quality message-to-speech generation in a practical application. In: J. van Santen, R. Sproat, J. Olive, and J. Hirschberg (Eds.), *Progress in speech synthesis*, pp. 575–586. New York: Springer-Verlag.

Dirksen, A. and J. Coleman (1996). All prosodic speech synthesis. In: J. Van Santen, R. Sproat, J. Olive, and J. Hirschberg (Eds.), *Progress in speech synthesis*, pp. 91–108. New York: Springer-Verlag.

Dixon, R. and H. Maxey (1968). Terminal analog synthesis of continuous speech using the diphone method of segment assembly. *IEEE Transactions on Audio and Electro-acoustics AU-16*, 40–50.

Dutoit, T. (1997). *An introduction to text-to-speech synthesis*. Dordrecht: Kluwer Academic Press.

Eefting, W. (1991). The effect of ”information value” and ”accentuation” on the duration of dutch words, syllables and segments. *Journal of the Acoustical Society of America 89*(1), 412–424.

Gibbon, D., R. Moore, and R. Winski (Eds.) (1997). *Handbook of standards and resources for spoken language systems*. Berlin: Mouton de Gruyter.

Gigi, E. and L. Vogten (1997). A mixed-excitation vocoder based on exact analysis of harmonic components. *IPO Annual Progress Report 32*, 105–110.

Gray, A. and J. Markel (1976). Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech and Signal Processing 24*(5), 380–391.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America 87*(4), 1738–1752.

Hermansky, H. and J. Junqua (1988). Optimization of perceptually-based ASR front-end. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’88)*, pp. 219–222.

Hunt, M. (1995). Signal representation. In: R. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, and V. Zue (Eds.), *Survey of the State of the Art in Human Language Technology*, pp. 11–16. Cambridge, UK: Cambridge University Press.

Itakura, F. (1975). Minimum prediction residual applied to speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing 23*(1), 67–72.

Kato, H., M. Tsuzaki, and Y. Sagisaka (1998). Acceptability for temporal modifications of single vowel segments in isolated words. *Journal of the Acoustical Society of America 104*(1), 540–549.

Kerkhoff, J., T. Rietveld, C. Gussenhoven, and L. Elich (1995). NIROS: The Nijmegen Interactive Rule Oriented Speech-synthesis system, an overview. Internal report, Dept. of Language and Speech, University of Nijmegen.

Klabbers, E., J. Odijk, J. De Pijper, and M. Theune (1996). GoalGetter: From teletext to speech. In: *IPO Annual Progress Report*, Volume 31, Eindhoven, the Netherlands, pp. 66–75.

Klabbers, E. and R. Veldhuis (1998). On the reduction of concatenation artefacts in diphone synthesis. In: *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'98), Sydney, Australia*, Volume 5, pp. 1983–1986.

Klatt, D. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America 82*(3), 737–783.

Krishnan, S. and P. Rao (1996). A comparative study of explicit frequency and conventional signal representations for speech recognition. *Digital Signal Processing 6*, 249–284.

Kullback, S. and R. Leibler (1951). On information and sufficiency. *Annals of Mathematical Statistics 22*, 79–86.

Local, J. and R. Ogden (1996). A model of timing for non-segmental phonological structure. In: J. Van Santen, R. Sproat, J. Olive, and J. Hirschberg (Eds.), *Progress in speech synthesis*, pp. 109–119. New York: Springer-Verlag.

Luce, R. and C. Krumhansl (1988). Measurement, scaling and psychophysics. In: S. Stevens (Ed.), *Handbook of experimental psychology*, Chapter 1, pp. 3–73. New York: Wiley.

Macon, M., A. Cronk, and J. Wouters (1998). Generalization and discrimination in tree-structured unit selection. In: *Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan Caves, Australia*, pp. 195–200.

Maghbouleh, A. (1996). An empirical comparison of automatic decision tree and linear regression models for vowel durations. In: *Proceedings of the second meeting of the ACL Special Interest Group in Computational Phonology, Santa Cruz*.

Makhoul, J. and L. Cosell (1976). LPCW: An LPC vocoder with linear predictive spectral warping. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'76), Philadelphia, PA*, pp. 466–469.

Markel, J. and A. Gray (1976). *Linear prediction of speech*. Berlin, Germany: Springer-Verlag.

Martens, L. and H. Quené (1994). Degemination of Dutch fricatives in three different speech rates. In: R. Bok-Bennema and C. Cremers (Eds.), *Linguistics in the Netherlands*, pp. 119–126. Amsterdam: John Benjamins, AVT Publications.

Möbius, B. and J. Van Santen (1996). Modeling segmental duration in German text-to-speech synthesis. In: *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP'96), Philadelphia, PA*, pp. 2395–2398.

Moore, B., B. Glasberg, and T. Bear (1997). A model for the prediction of thresholds, loudness and partial loudness. *Journal of the Audio Engineering Society 45*(4), 224–239.

Moulines, E. and F. Charpentier (1990). Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication 9*(5-6), 453–467.

Moulton, W. G. (1962). The vowels of Dutch: phonetic and distributional classes. *Lingua 11*, 294–312.

Nederhof, M., G. Bouma, R. Koeling, and G. Van Noord (1997). Grammatical analysis in the OVIS spoken dialogue system. In: *Proceedings of the ACL/EACL Workshop on Spoken Dialogue Systems, Madrid, Spain*.

Nooteboom, S. G. (1972). *Production and perception of vowel duration: A study of the durational properties of vowels in Dutch*. Ph. D. thesis, Rijksuniversiteit Utrecht.

Olive, J., J. Van Santen, B. Möbius, and C. Shih (1998). Synthesis. In: R. Sproat (Ed.), *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, pp. 192–228. Boston: Kluwer Academic Publishers.

Pols, L. (1994). Speech technology systems: Performance and evaluation. In: R. Asher (Ed.), *The Encyclopedia of Language and Linguistics*, pp. 4289–4296. Oxford: Pergamon Press.

Rabiner, L. and B. Juang (1993). *Fundamentals of speech recognition*. New Jersey: Englewood Cliffs.

Rietveld, A. C. M. and U. H. Frauenfelder (1987). The effect of syllable structure on vowel duration. In: *Proceedings of the International Conference on Phonetic Science (ICPhS'87), Tallinn, USSR*, pp. 28–31.

Rietveld, T., J. Kerkhoff, M. Emons, E. Meijer, A. Sanderman, and A. Sluijter (1997). Evaluation of speech synthesis systems for Dutch in telecommunication applications in GSM and PSTN networks. In: *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH'97), Rhodes, Greece*, Volume 2, pp. 577–580.

Riley, M. (1992). Tree-based modeling for speech synthesis. In: G. Bailly, C. Benoît, and T. Sawallis (Eds.), *Talking machines: Theories, models and designs*, pp. 265–273. Amsterdam: Elsevier.

Sheskin, D. (1996). *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, Fla: CRC Press LLC.

Sproat, R. (Ed.) (1998). *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*. Dordrecht: Kluwer Academic Publishers.

Steeneken, H. (1992). Quality evaluation of speech processing systems. In: N. Ince (Ed.), *Digital speech coding: Speech coding, synthesis and recognition*, pp. 127–160. Norwell, USA: Kluwer.

Stöber, K., T. Portele, P. Wagner, and W. Hess (1999). Synthesis by word concatenation. In: *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH'99), Budapest, Hungary*, Volume 2, pp. 619–622.

Strik, H., A. Russel, H. Van den Heuvel, C. Cucchiarini, and L. Boves (1996). Localizing an automatic inquiry system for public transport information. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP'96), Philadelphia, PA*, pp. 853–856.

Stylianou, Y., T. Dutoit, and J. Schroeter (1997). Diphone concatenation using a harmonic plus noise model of speech. In: *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH'97), Rhodes, Greece*, Volume 2, pp. 613–615.

't Hart, J., R. Collier, and A. Cohen (1990). *A perceptual study of intonation: An experimental phonetic approach to speech melody*. Cambridge: Cambridge University Press.

Terken, J. (1996). Language and speech technology: Developments. In: *IPO Annual Progress Report*, Volume 31, Eindhoven, the Netherlands, pp. 64.

Theune, M. (2000). *From Data to Speech*. Ph. D. thesis, Eindhoven University of Technology. To appear.

Theune, M., E. Klabbers, J. Odijk, and J. De Pijper (1997). Computing prosodic properties in a data-to-speech system. In: *Proceedings of the ACL/EACL Workshop on Concept-to-Speech Generation Systems, Madrid, Spain*, pp. 39–46.

Tzoukermann, E. and O. Soumoy (1995). Segmental duration in French text-to-speech synthesis. In: *Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH'95), Madrid, Spain*, pp. 607–611.

Van Bergem, D. (1995). *Acoustic and lexical vowel reduction*. Ph. D. thesis, IFOTT, University of Amsterdam.

Van Coile, B. (1989). *Tekst-naar-spraak omzetting: een taalkundig, fonetisch en akoestisch probleem*. Ph. D. thesis, Rijksuniversiteit Gent.

Van den Heuvel, H., B. Cranen, and T. Rietveld (1996). Speaker variability in the coarticulation of /a,i,u/. *Speech Communication 18*, 113–130.

Van Dinther, R., P. Rao, R. Veldhuis, and A. Kohlrausch (1999). A measure for predicting audibility discriminatoin thresholds. In: *to appear in IPO Annual Progress Report*, Volume 34, Eindhoven, the Netherlands.

Van Santen, J. (1992a). Contextual effects on vowel duration. *Speech Communication 11*, 513–546.

Van Santen, J. (1992b). Deriving text-to-speech durations from natural speech. In: G. Bailly, C. Benoît, and T. Sawallis (Eds.), *Talking machines: Theories, models and designs*, pp. 275–285. Amsterdam, the Netherlands: Elsevier.

Van Santen, J. (1994). Assignment of segmental duration in text-to-speech synthesis. *Computer, Speech and Language 8*, 95–128.

Van Santen, J. (1997). Prosodic modeling in text-to-speech synthesis. In: *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH'97), Rhodes, Greece*, pp. KN19–28.

Van Santen, J. and A. Buchsbaum (1997). Methods for optimal text selection. In: *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH'97), Rhodes, Greece*, Volume 2, pp. 553–556.

Van Santen, J., L. Pols, M. Abe, D. Kahn, E. Keller, and J. Vonwiller (1998). Report on the Third ESCA TTS Workshop Evaluation. In: *Pro-*

ceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan Caves, Australia, pp. 329–332.

Van Santen, J. and R. Sproat (1998). Methods and tools. In: R. Sproat (Ed.), *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, pp. 7–30. Boston: Kluwer Academic Publishers.

Van Santen, J., R. Sproat, J. Olive, and J. Hirschberg (Eds.) (1996). *Progress in Speech Synthesis*. New York: Springer Verlag.

Van Son, R. (1988). *Spectro-temporal features of vowel segments*. Ph. D. thesis, IFOTT, University of Amsterdam.

Van Son, R. and J. Van Santen (1997). Strong interaction between factors influencing consonant duration. In: *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH'97), Rhodes, Greece*, pp. 319–322.

Veldhuijzen van Zanten, G. (1998). Adaptive mixed-initiative dialogue modelling. In: *Proceedings of IVTTA'98: IEEE 4th Workshop on Interactive Voice Technology for Telecommunications Applications, Turin, Italy*, pp. 65–70.

Veldhuis, R. and M. Breeuwer (1993). *An introduction to source coding*. UK: Prentice Hall International Ltd.

Waals, J. (1999). *An experimental view on the Dutch syllable*. Ph. D. thesis, Utrecht University.

Waterworth, J. (1983). Effect of intonation form and pause durations of automatic telephone number announcements on subjective preference and memory performance. *Applied Ergonomics 14*(1), 39–42.

Wouters, J. and M. Macon (1998). A perceptual evaluation of distance measures for concatenative speech synthesis. In: *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'98), Sydney, Australia*, Volume 6, pp. 2747–2750.

Appendix A

Listing of the phoneme inventory

This appendix contains a listing of the phoneme inventory that was used in our synthesis system. The notation is derived from SAMPA. It is different in that long vowels are represented with lower case letters, not followed by a colon as in the official SAMPA notation. Furthermore, the symbols /c/, representing an assimilation of /t/ and /j/, and /J/, representing an assimilation of /n/ and /j/, are added.

Consonants			Vowels		
p	pak	pAk	a	naam	nam
t	tak	tAk	e	veer	ver
k	kap	kAp	2	deur	d2r
b	bak	bAk	o	door	dor
d	dak	dAk	Ei	fijn	fEin
g	goal	goL	Au	goud	xAut
f	fel	fEL	9y	huis	h9ys
s	sein	sEin	E:	crème	krE:m
x	toch	tOx	O:	roze	rO:z@
c	katje	kAc@	9:	freule	fr9:l@
S	show	So	i	vier	vir
v	vel	vEL	u	voer	vur
z	zijn	zEin	y	vuur	vyr
Z	bagage	bAxaZ@	A	pad	pAt
h	hand	hAnt	E	pet	pEt
m	met	mEt	O	pot	pOt
n	net	nEt	I	pit	pIt
N	bang	bAN	Y	put	pYt
J	oranje	orAJ@	@	bange	bAN@
l	land	lAnt			
L	bal	bAL			
r	rand	rAnt			
j	ja	ja			
w	wit	wIt			
?	aap	?ap			

Table A.1: *The consonants and vowels of Dutch in SAMPA-like notation. Their meaning is illustrated with an example, presented in orthographic and phonetic transcription.*

Appendix B

Duration factors

This appendix lists the results found in the duration analysis that is reported on in Chapter 4. For each category in the category tree, the relevant factors and factor levels are listed. The number of observations is given to give an indication of the coverage of the corpus. The effect of a factor level is indicated by the corrected means duration (in milliseconds).

Factors	Factor levels	Number of observations	Corrected means (ms)
Previous	unvoiced obstruent	190	119.9
	sonorant	362	130.2
	voiced obstruent	187	137.5
Next	sonorant	139	114.8
	vowel	35	121.3
	obstruent	505	126.3
	/r/	60	140.1
Stress/Accent	00	85	121.7
	01	226	99.3
	10	88	134.3
	11	340	151.8
Left position	word-initial syllable	412	130.6
	other	297	134.4
	syllable-initial	30	144.6
Right position	syllable-final	581	120.6
	word-final	113	142.3
	phrase-final	45	206.1

Table B.1: *Factor effects on long vowels in open syllables.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Previous	obstruent	121	64.4
	sonorant	93	72.6
Next	obstruent	105	65.1
	sonorant	101	66.0
	/r/	8	103.2
Stress/Accent	00	7	53.3
	01	84	60.9
	10	31	64.5
	11	92	77.0

Table B.2: *Factor effects on short vowels in open syllables.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Previous	unvoiced obstruent	112	58.6
	vowel	8	98.8
	sonorant	153	85.9
	voiced obstruent	85	79.8
Next	sonorant	65	66.6
	unvoiced obstruent	144	67.7
	voiced obstruent	89	77.0
	vowel	39	87.5
	/r/	21	140.7
Stress/Accent	00	58	67.1
	01	199	75.4
	10	10	77.0
	11	91	94.3
Left position	non-initial	205	70.4
	word-initial syllable	128	78.2
	phrase-initial syllable	15	161.2
Right position	syllable-final	249	59.9
	word-final	94	95.8
	phrase-final	15	161.2

Table B.3: *Factor effects on /i, u, y/ in open syllables.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Previous	sonorant	110	47.1
	voiced obstruent	342	54.2
	unvoiced obstruent	243	64.6
Next	sonorant	148	43.8
	obstruent	491	49.7
	/r/	56	60.9
Stress/Accent	00	344	48.4
	01	351	56.8
Left position	non-initial	329	58.5
	syllable-initial	11	69.5
	word-initial syllable	355	66.5
Right position	syllable-final	252	42.1
	word-final	418	62.3
	phrase-final	25	108.2

Table B.4: *Factor effects on schwa in open syllables.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Previous	obstruent	401	104.3
	sonorant	200	115.6
Next	/R/	144	88.7
	sonorant	192	103.5
	obstruent	206	117.3
	/r/	59	136.8
Stress/Accent	00	239	93.8
	01	69	109.0
	10	48	109.2
	11	245	137.1
Left position	word-initial syllable	434	110.8
	non-initial	167	116.5
Right position	non-final	147	100.8
	word-final syllable	147	100.8
	phrase-final syllable	88	149.3

Table B.5: *Factor effects on long vowels in closed syllables.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Previous	obstruent	1000	62.8
	sonorant	404	72.9
	vowel	21	85.5
Next	/R/	77	60.4
	sonorant	735	65.7
	unvoiced obstruent	534	66.5
	/r/	79	79.0
Stress/Accent	00	612	57.4
	01	312	64.8
	10	85	71.8
	11	416	77.7
Right position	non-final	488	60.2
	word-final syllable	786	68.3
	phrase-final syllable	151	91.5

Table B.6: *Factor effects on short vowels in closed syllables.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Next	unvoiced obstruent	86	71.4
	sonorant	48	85.3
	/R/	19	112.2
	/r/	11	147.9
Stress/Accent	00	30	65.3
	01	41	69.9
	10	8	80.6
	11	85	93.5
Left position	word-initial syllable	96	80.4
	non-initial	68	89.0
Right position	non-final	42	77.6
	word-final syllable	89	78.7
	phrase-final syllable	33	106.1

Table B.7: *Factor effects on /i, u, y/ in closed syllables.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Previous	other	522	43.0
	sonorant	207	55.9
	vowel	19	70.5
Next	/R/	82	38.8
	stop	102	42.2
	sonorant	510	48.4
	fricative	54	54.0
Stress/Accent	00	329	44.9
	01	419	49.4
Left position	non-initial	543	40.6
	word-initial syllable	172	51.5
	phrase-initial syllable	33	60.2
Right position	non-final	147	37.6
	word-final syllable	434	43.6
	phrase-final syllable	167	61.3

Table B.8: *Factor effects on schwa in closed syllables.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Previous	obstruent2	9	42.0
	same	62	47.2
	obstruent	106	50.9
	vowel	193	74.1
	sonorant	87	81.6
Next	short vowel	221	55.2
	long vowel	236	59.8
Stress/Accent	00	102	43.8
	01	179	54.3
	10	29	53.2
	11	147	72.7
Left position	non-initial	236	49.3
	word-initial	221	67.7
Right position	non-final	208	56.6
	word-final syllable	249	58.8

Table B.9: *Factor effects on nasals in onsets.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Previous	obstruent2	104	37.6
	obstruent	88	51.3
	vowel	214	60.3
	sonorant	109	61.2
Next	schwa	78	43.5
	short vowel	191	52.0
	long vowel	236	53.2
Stress/Accent	00	99	45.5
	01	167	50.3
	10	42	49.0
	11	207	60.6
Left position	non-initial	263	47.2
	word-initial	252	57.6

Table B.10: *Factor effects on liquids and glides in onsets.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Previous	vowel	145	23.3
	unvoiced obstruent3	36	28.9
	unvoiced obstruent2	105	37.6
	voiced obstruent2	83	45.6
	other	40	59.7
Stress/Accent	00	33	33.9
	01	187	30.0
	10	31	34.5
	11	160	37.1

Table B.11: *Factor effects on /r/ in onsets.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Previous	obstruent2	44	64.0
	same	11	68.1
	obstruent	143	85.2
	vowel	305	102.1
	sonorant	168	106.1
Next	obstruent2	126	94.2
	vowel	480	98.8
	sonorant2	65	106.5
Stress/Accent	00	118	78.7
	01	339	89.0
	10	34	97.6
	11	180	107.2
Left position	non-initial	403	87.9
	word-initial	268	102.0
Right position	non-final	359	87.9
	word-final syllable	313	102.0

Table B.12: *Factor effects on unvoiced fricatives in onsets.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Previous	unv. fricative2	108	25.2
	unvoiced fricative	116	28.5
	same	12	33.1
	unvoiced stop	59	33.5
	sonorant	472	35.8
Next	consonant3	33	24.9
	consonant2	83	31.2
	vowel	620	31.3
	sonorant2	42	37.7
Stress/Accent	00	118	30.4
	01	339	33.7
	10	34	31.6
	11	180	32.1
Right position	non-final	272	32.0
	phrase-final syllable	136	35.8

Table B.13: *Factor effects on unvoiced stops in onsets.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Previous	unvoiced obstruent2	108	26.2
	nasal	116	36.2
	unvoiced obstruent	176	39.6
	sonorant	383	58.3
Next	unvoiced fricative2	19	31.4
	schwa	232	39.4
	sonorant2	138	41.5
	vowel	394	46.1
Stress/Accent	00	128	38.8
	01	390	40.6
	10	46	45.5
	11	219	55.4
Left position	non-initial	513	38.8
	word-initial	270	54.0

Table B.14: *Factor effects on unvoiced stop closures in onsets.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Previous	voiced stop2	8	36.5
	unvoiced obstruent	123	85.3
	schwa	86	92.5
	sonorant	305	139.4
Stress/Accent	00	199	81.3
	01	153	96.6
	10	22	108.2
	11	148	117.6
Left position	non-initial	201	88.2
	word-initial	321	106.8
Right position	non-final	221	94.4
	word-final syllable	301	101.9

Table B.15: *Factor effects on voiced fricatives in onsets.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Previous	sonorant	601	12.6
	unvoiced obstruent	154	16.1
Stress/Accent	00	310	12.0
	01	268	13.1
	10	36	14.3
	11	141	15.6
Left position	non-initial	346	12.9
	word-initial	409	13.6

Table B.16: *Factor effects on voiced stops in onsets.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Previous	nasal	228	15.2
	other	291	29.7
	vowel	236	39.4
Next	schwa	370	26.4
	consonant2	67	33.1
	vowel	318	33.0
Stress/Accent	00	310	26.7
	01	268	28.2
	10	36	28.0
	11	141	42.9
Left position	non-initial	346	22.9
	word-initial	409	30.5

Table B.17: *Factor effects on voiced stop closures in onsets.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Previous	unvoiced fricative	12	46.7
	vowel	45	54.2
	unvoiced stop	32	55.9
	sonorant	42	63.2
Stress/Accent	00	49	45.9
	01	24	43.8
	10	12	53.9
	11	46	80.8
Left position	non-initial	32	53.0
	word-initial	99	58.9

Table B.18: *Factor effects on /h/ in onsets.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Previous	unvoiced obstruent	186	28.7
	long vowel	43	36.3
	sonorant	154	40.4
	schwa	71	46.1
Stress/Accent	00	241	23.5
	01	54	32.9
	10	24	41.8
	11	135	63.5
Right position	non-final	189	32.8
	word-final syllable	265	39.9

Table B.19: *Factor effects on glottal stops in onsets.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Previous	vowel	1161	57.4
	sonorant2	9	79.9
Next	unvoiced obstruent2	144	51.5
	vowel	49	55.1
	other	632	56.8
	voiced obstruent	345	67.8
Stress/Accent	00	557	45.3
	01	369	61.5
	10	35	68.7
	11	209	90.1
Left position	non-initial	541	48.8
	word-initial syllable	629	78.2
Right position	non-final	282	43.1
	word-final	680	59.0
	phrase-final	97	92.1
	utterance-final	111	133.7

Table B.20: *Factor effects on nasals in codas.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Next	schwa	21	36.3
	consonant2	55	54.0
	other	174	63.6
Stress/Accent	00	56	51.9
	01	83	55.2
	10	18	57.0
	11	93	71.5
Left position	non-initial	109	54.2
	word-initial syllable	141	65.3
Right position	non-final	132	48.9
	word-final	82	59.6
	phrase-final	20	102.2
	utterance-final	16	106.4

Table B.21: *Factor effects on liquids and glides in codas.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Next	vowel	26	22.5
	obstruent2	109	27.6
	other	412	34.1
Stress/Accent	00	186	25.8
	01	182	28.4
	10	38	46.1
	11	143	49.5
Right position	non-final	221	27.3
	word-final	197	33.9
	word-final syllable	99	49.6
	phrase-final	32	60.5

Table B.22: *Factor effects on /r/ in codas.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Previous	unvoiced obstruent2	76	86.9
	sonorant2	113	101.3
	vowel	450	110.2
Next	same	11	77.7
	vowel	23	74.3
	obstruent	540	91.9
	sonorant	65	105.3
Stress/Accent	00	164	81.4
	01	193	92.3
	10	34	100.7
	11	248	118.5
Right position	non-final	185	83.7
	word-final	340	97.9
	phrase-final syllable	30	123.8
	phrase-final	44	160.8
	utterance-final	40	208.0

Table B.23: *Factor effects on unvoiced fricatives in codas.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Previous	unvoiced obstruent2	118	30.2
	vowel	535	38.1
	sonorant2	191	41.4
Next	voiced stop	86	19.7
	obstruent2	87	20.3
	vowel	43	26.6
	other	628	34.4
Stress/Accent	00	363	30.8
	01	161	31.4
	10	51	35.4
	11	269	42.9
Right position	non-final	153	22.5
	word-final	570	35.9
	phrase-final	52	81.5
	utterance-final	69	118.7

Table B.24: *Factor effects on unvoiced stops in codas.*

Factors	Factor levels	Number of observations	Corrected means (ms)
Previous	nasal2	107	19.6
	unvoiced obstruent2	120	29.4
	sonorant	628	45.0
Next	unvoiced obstruent2	87	32.5
	other	559	37.0
	nasal	49	42.3
	unvoiced obstruent	148	46.3
	same	12	47.0
Stress/Accent	00	369	33.4
	01	163	34.6
	10	51	40.1
	11	272	42.2
Right position	non-final	152	34.3
	word-final	623	39.7
	utterance-final	80	47.1

Table B.25: *Factor effects on unvoiced stop closures in codas.*

Appendix C

Audio examples

This appendix gives a description of the audio examples that are contained on the following web site:

http://www.ipo.tue.nl/homepages/eklabber/audio.html

The numbers associated with each of the speech examples correspond to the filename of the examples on the site. All examples are stored in .wav format. Section C.1 presents examples of IPO's phrase concatenation approach (for Dutch and German) in comparison with IPO's diphone synthesis and two instances of conventional phrase concatenation. These examples serve as illustrations for Chapter 2. Section C.2 presents some examples of audible discontinuities, which are discussed in Chapter 3. Section C.3 presents twenty sentences that were used in the duration evaluation experiment of Chapter 4. They have been taken from the 'Nieuws voor Doven en Slechthorenden', a teletext service providing concise news reports to hearing impaired people.

C.1 IPO’s diphone synthesis and phrase concatenation

This section contains recordings of IPO’s diphone synthesis and phrase concatenation technique for three applications: GoalGetter, OVIS and a German train timetable system. GoalGetter is a data-to-speech system that generates spoken soccer reports on the basis of tabular information from Teletext.

Example 01: IPO’s phrase concatenation Example 02: IPO’s diphone synthesis	De wedstrijd tussen PSV en Ajax eindigde in een - drie. 25.000 toeschouwers bezochten het Philipsstadion. Ajax nam na 5 minuten de leiding door een treffer van Kluivert. Dertien minuten later liet de aanvaller zijn tweede doelpunt aantekenen. De verdediger Blind verzilverde in de 83e minuut een strafschop voor Ajax. Vlak voor het eindsignaal bepaalde Nilis van PSV de eindstand op een drie. Scheidsrechter van Dijk leidde het duel. Valckx van PSV kreeg een gele kaart.

This section contains three recordings of a sample train connection taken from OVIS using three speech generation methods. The first is IPO's phrase concatenation method, the second is conventional phrase concatenation and the third is diphone synthesis. The conventional concatenation method differs from IPO's phrase concatenation in that the concatenative units have been recorded in isolation, and all words and phrases have been recorded in one prosodically neutral version only. The recordings were used in the evaluation experiment of Chapter 2.

Example 03: IPO's phrase concatenation.	Ik heb de volgende verbinding gevonden. Met de sneltrein, vertrek vanuit Groningen, om elf uur vierenvijftig, aankomst in Oudenbosch, om tien uur tweeenveertig. Daar verder met de stoptrein, vertrek om vier uur vierendertig, aankomst in Leidschendam-Voorburg, om een uur tweeentwintig. Wilt u nog een andere verbinding weten?
Example 04: conventional phrase concatenation.	Ik heb de volgende verbinding gevonden. Met de sneltrein, vertrek vanuit Enkhuizen, om elf uur eenenvijftig, aankomst in Oosterbeek, om drieentwintig uur twee. Daar verder met de stoptrein, vertrek om dertien uur zeventien, aankomst in Stavoren, om drie uur achtentwintig. Wilt u nog een andere verbinding weten?
Example 05: IPO's diphone synthesis.	Ik heb de volgende verbinding gevonden. Met de sneltrein, vertrek vanuit Harlingen, om zeventien uur twintig, aankomst in Pijnacker, om eenentwintig uur drie. Daar verder met de stoptrein, vertrek om twee uur drieenveertig, aankomst in Roosendaal, om dertien uur achttien. Wilt u nog een andere verbinding weten?

This section contains two examples of a German train connection. The conventional concatenation method also uses units that have been recorded in one prosodically neutral version and in isolation.

Example 06: Conventional phrase concatenation	Mit Euronight 6789 Abfahrt in Rotterdam um 13 Uhr 08 Ankünft in Timmendorfer Strand um 14 Uhr 25. Dort weiter mit Intercity 288
Example 07: IPO's phrase concatenation	Abfahrt um 08 Uhr 20 Ankünft in Zwillingen um 05 Uhr 08.

C.2 Audible discontinuities

This section contains two examples of vowel segments of the /a/, /i/, and /u/, which were constructed using diphone synthesis. One has a smooth transition at the diphone boundary in the middle of the vowel, the other has a discontinuous transition. These stimuli were used as examples in the training session of the first listening experiment presented in Chapter 3.

Smooth	Discontinuous	Vowel
Example 08	Example 09	/a/
Example 10	Example 11	/i/
Example 12	Example 13	/u/

Additionally, this section contains two examples of the word /duk/, one generated with the original diphones and the other generated with an additional /uk/ diphone from a context-sensitive cluster. Their spectra can be seen in Figure 3.11.

Example 14:	/duk/ generated with original database
Example 15:	/duk/ generated with new database

C.3 Duration evaluation

This section contains twenty sentences that were used in the duration experiment of Chapter 4. Each sentence was generated with diphone synthesis using either the new duration model or the old duration model.

New	Old	Sentence
Example 16	Example 17	Er wordt rekening mee gehouden dat het dodental oploopt tot ruim veertig duizend
Example 18	Example 19	De VS heeft verklaard dat de NAVO de aanvallen voortzet, ondanks de vrijlating van drie krijgsgevangenen
Example 20	Example 21	Staatssecretaris Cohen zegde vorige week toe dat Nederland zo'n 2000 ontheemden zal opnemen
Example 22	Example 23	De vierde groep vluchtelingen uit Kosovo is maandag aangekomen op de vliegbasis Eindhoven
Example 24	Example 25	Ondanks een vrij rustig verloop zijn bij de oud- en nieuw-viering toch weer slachtoffers gevallen
Example 26	Example 27	De regeringscrisis in Israel neemt verder toe, nu ook minister Neeman van Financien met aftreden dreigt
Example 28	Example 29	De storm die dit weekeinde over het noordwesten van Europa raasde, heeft aan 14 mensen het leven gekost
Example 30	Example 31	Korpschef Veenstra trad gisteren af, na de bekendmaking van een voor hem negatief onderzoeks-rapport
Example 32	Example 33	Bijna alle werknemers gaan er dit jaar netto toch op vooruit, blijkt uit een onderzoek van de NOS

New	Old	Sentence
Example 34	Example 35	Oud-korpschef Brand van de politie Haaglanden is gevraagd om tijdelijk de politie in Groningen te leiden
Example 36	Example 37	Brand, die in Den Haag vervroegd met pensioen ging, is door burgemeester Ouwerkerk van Groningen benaderd
Example 38	Example 39	In België zijn bij een overval op een geldtransport de chauffeur en een bewaker doodgeschoten
Example 40	Example 41	Het is niet bekend of de rovers de transportwagen wisten te openen en hoe groot de eventuele buit is
Example 42	Example 43	Leraren in het voortgezet onderwijs leggen op 29 januari het werk neer uit onvrede over de grote werkdruk
Example 44	Example 45	Koningin Beatrix ontvangt vandaag de fractieleiders van de kleinere partijen inzake de kabinetscrisis
Example 46	Example 47	Vermoedelijk wijst de koningin daarna een formateur aan die moet proberen de breuk te lijmen
Example 48	Example 49	Minister de Vries van Sociale Zaken stelt een onderzoek in naar fraude met Europees subsidiegeld
Example 50	Example 51	Voor werkgelegenheidsprojecten werden te veel deelnemers opgegeven waardoor meer subsidie werd geïnd
Example 52	Example 53	In België mogen vetrijke vleeswaren niet meer verkocht worden vanwege het gevaar voor dioxine-besmetting
Example 54	Example 55	Korthals had de Kamer maandag al achter gesloten deuren geinformeerd over de cocaine zaak

Summary

This thesis aims at finding segmental and prosodic improvements to speech generation.

In Chapter 2 two types of speech generation are discussed: diphone synthesis and phrase concatenation. The research is carried out in the context of a spoken dialogue system called OVIS, which gives train timetable information over the telephone. In this type of applications, the approach to speech generation is often very simplistic. The speech output is achieved by a straightforward concatenation of words and phrases that literally correspond to the text that is to be spoken. This approach has a major drawback in that it lacks variability in accentuation and the marking of phrase boundaries, which is essential for creating natural speech. The fact that the words and phrases to be concatenated are often spoken in isolation results in mismatches in pitch, loudness and tempo which makes the speech sound disfluent.

The speech generation module is extended, to accommodate the prosodic variability introduced by the natural language generation module. Different versions of words have been recorded dependent on whether they occur in an accented or unaccented condition and on whether they occur before a phrase boundary of some depth. All words and phrases have been recorded in the proper context, to automatically elicit the appropriate versions. That and a careful recording results in speech output that sounds almost as fluent as natural speech and that minimises the occurrence of mismatch in pitch, loudness and tempo.

The drawback of this phrase concatenation approach is that a considerable amount of phonetic knowledge is required to construct an application-dependent corpus that covers the sentences the natural language gener-

ation module creates with their prosodic variations. Moreover, this approach to speech generation only works for applications that suffice with a medium-sized stable vocabulary. When the need for flexibility increases, speech synthesis is the only alternative that is left. Although the most popular implementation, i.e., diphone synthesis, is fairly intelligible, listeners still judge the overall quality too low for it to be used in commercial applications.

Chapter 3 concentrates on a significant problem that interferes with the segmental quality, namely the occurrence of audible discontinuities at diphone boundaries. These discontinuities are mainly caused by spectral mismatch at the diphone boundaries and are most obvious in vowels. A listening experiment has been conducted with five vowels /a/, /i/, /A/, /I/ and /u/, to investigate the extent to which discontinuities are perceived. The results reveal considerable differences between the vowels under investigation. The /u/ has the highest percentage of perceived discontinuities, whereas the /a/ has the lowest percentage. The scores obtained in the listening experiment have been correlated with several spectral distance measures to find an objective measure that predicts the occurrence of audible discontinuities. The correlation is performed using Receiver Operator Characteristics. The Kullback-Leibler (KL) distance was shown to be most adequate for the task.

To reduce the number of audible discontinuities context-sensitive diphones have been added to the diphone database. In order to limit the number of additional diphones, the KL distance is used to cluster consonantal contexts with similar spectral effects on the neighbouring vowels. A second listening experiment has been conducted for the vowels /a/, /i/, and /u/, to evaluate the improvement obtained with this addition to the database. The results show that the number of audible discontinuities has significantly decreased for all vowels. However, the objective KL distance has significantly decreased only for /i/ and /u/, but not for /a/. But the KL distances were already lower in the /a/ to begin with.

Chapter 4 concentrates on a problem that affects the prosodic quality of speech synthesis, namely the inadequate prediction of segmental durations. A new duration model has been developed using the sums-of-products approach of Van Santen (1992a). With a relatively small corpus of 297 sentences, containing 16,775 segments, a duration model could be constructed that compares well with German and French duration models that have been constructed using the same sums-of-products approach. The use of phonetic and phonological knowledge in the construction of the category tree and the data analysis, which was designed specifically for this purpose, result in a model that incorporates the most important effects and interactions found to influence duration prediction.

The performance of the old and the new duration model has been evaluated quantitatively by comparing the computed to the actual segmental durations in the 297-sentence training corpus. The new model was shown to be a significant improvement over the old one. A perceptual pairwise comparison experiment using sentences generated with the old and new model, did not provide conclusive subjective evidence for the improvement.

Samenvatting

Dit proefschrift heeft als doel om segmentele en prosodische verbeteringen in spraakgeneratie tot stand te brengen.

Hoofdstuk 2 bespreekt twee soorten spraakgeneratie: difoonsynthese en frasenconcatenatie. Het onderzoek is uitgevoerd in de context van een gesproken dialoogsysteem OVIS genaamd, dat telefonisch informatie over treinverbindingen geeft. In dit type van applicaties is de spraakgeneratie aanpak vaak erg simplistisch. De spraakuitvoer wordt bereikt via het rechttoe-rechtaan aan elkaar plakken van opgenomen woorden en frasen die letterlijk corresponderen met de tekst die uitgesproken dient te worden. Het voornaamste nadeel van deze aanpak is dat het variabiliteit mist in accentuering en markering van frasegrenzen, die beiden essentieel zijn voor de creatie van natuurlijke spraak. Het feit dat de woorden en frasen die uitgesproken moeten worden vaak in isolatie worden opgenomen, resulteert in abrupte verschillen in toonhoogte, luidheid en tempo, waardoor de spraak niet vloeiend klinkt.

De spraakgeneratie module is uitgebreid, om de prosodische variatie te kunnen uiten die door de taalgeneratie module geïntroduceerd wordt. Verschillende versies van woorden zijn opgenomen afhankelijk van of ze in een geaccentueerde of een ongeaccentueerde conditie voorkomen en of ze voor een frasegrens van enige diepte voorkomen. Alle woorden en frasen zijn in de juiste context opgenomen, om automatisch de juiste versies te laten uitspreken. Dit in combinatie met een zorgvuldige manier van opnemen heeft geleid to spraakuitvoer die bijna net zo vloeiend klinkt als natuurlijke spraak en waarbij het voorkomen van abrupte verschillen in toonhoogte, luidheid en tempo is geminimaliseerd.

Het nadeel van deze frasenconcatenatie methode is dat aardig wat fonetische kennis vereist is om een applicatie-specifiek corpus te construeren dat de zinnen dekt die de taalgeneratie module genereert met hun prosodische variatie. Bovendien werkt deze spraakgeneratie aanpak alleen voor applicaties met een niet te groot, stabiel vocabulaire. Wanneer de behoefte aan flexibiliteit toeneemt, is spraaksynthese het enige alternatief. Hoewel de meest populaire implementatie, d.w.z., difoonsynthese, redelijk goed verstaanbaar is, vinden luisteraars de globale kwaliteit nog te laag om het gebruik daarvan te accepteren in commerciële applicaties.

Hoofdstuk 3 concentreert zich op een belangrijk probleem dat de segmentele kwaliteit aantast, namelijk het probleem van hoorbare discontinuïteiten op difoongrenzen. Deze discontinuïteiten worden voornamelijk veroorzaakt door spectrale verschillen op de difoongrenzen en zijn het duidelijkst in klinkers. Er is een luisterexperiment uitgevoerd met de klinkers /a/, /i/, /A/, /I/ en /u/, om te onderzoeken in welke mate deze discontinuïteiten worden waargenomen. De resultaten laten zien dat er flinke verschillen tussen de verschillende klinkers zijn. De /u/ heeft het grootste percentage waargenomen discontinuïteiten, terwijl de /a/ het laagste percentage heeft. De scores die in het luisterexperiment zijn verkregen werden gecorreleerd met verschillende spectrale afstandsmaten. De correlatie is via Receiver Operator Characteristics tot stand gebracht. De Kullback-Leibler (KL) afstand kwam als beste uit de bus.

Om het aantal hoorbare discontinuïteiten te verminderen zijn context-afhankelijke difonen aan de difoondatabase toegevoegd. Om het aantal extra difonen te beperken, is de KL afstand gebruikt om medeklinkers te clusteren die gelijke spectrale effecten op de omringende klinkers hebben. Er is een tweede luisterexperiment uitgevoerd voor de klinkers /a/, /i/, en /u/, om de verbetering die deze extra difonen met zich meebrengen te evalueren. De resultaten laten zien dat het aantal hoorbare discontinuïteiten voor alle drie de klinkers significant is verminderd. Als we kijken naar de vermindering van de KL afstand is die alleen significant voor de /i/ en de /u/. Daarbij moeten we zeggen dat de /a/ al veel minder discontinuïteiten bevatte om mee te beginnen.

Hoofdstuk 4 concentreert zich op een tweede probleem dat de prosodische kwaliteit van spraaksynthese aantast, namelijk de inadequate voorspelling van segmentele duren. Daarom is er een nieuw duurmodel ontwikkeld met gebruikmaking van de sums-of-products methode van Van Santen (1992a). Met een relatief klein corpus van 297 zinnen, bestaande uit 16,775 segmenten, kon een duurmodel geconstrueerd worden dat zich goed meet met duurmodellen voor het Frans en Duits die met dezelfde sums-of-products aanpak geconstrueerd zijn. Het gebruik van fonetische en fonologische kennis in de constructie van de boom met categorieën en de data analyse, die specifiek voor dit doel is ontwikkeld, resulteert in een model dat de meest belangrijke effecten en interacties bevat van factoren die de duurvoorspellingen beïnvloeden.

De werking van het oude en het nieuwe duurmodel is kwantitatief geëvalueerd in een vergelijking met de segmentele duren in het trainingcorpus van 297 zinnen. Het nieuwe model was significant beter dan het oude. Een perceptief paarsgewijs vergelijkingsexperiment, dat zinnen gebruikte die zowel met het oude als het nieuwe model gegenereerd waren, gaf geen uitsluitsel over de perceptieve verbetering.

Biography

Esther Klabbers was born in Nijmegen on 11 July 1973. She attended the Nijmeegse Scholengemeenschap Groenewoud in Nijmegen from 1985 to 1991, where she obtained her VWO diploma in 1991. From 1991 to 1992 she studied English Language and Literature at the University of Nijmegen. From 1992 to 1995 she studied Language and Computer Science at the University of Nijmegen. She obtained her Master's degree in 1995 in Language and Computer Science, with a specialisation in speech technology. From 1996 until 1999 she was employed as a PhD student by NWO (Netherlands Organisation for Scientific Research) for the 'Language and Speech Technology' Priority Programme. During this time she was stationed at IPO, Center for User-System Interaction at the Eindhoven University of Technology.