KTH Computer Science
and Communication

Modelling Phone-Level Pronunciation in Discourse Context

Per-Anders Jande

Doctoral Thesis
Stockholm, Sweden, 2006

GSLT Graduate School of Language Technology
Faculty of Arts, Göteborg University
405 30 Göteborg, Sweden

KTH Computer Science and Communication
Department of Speech, Music and Hearing
100 44 Stockholm, Sweden

TRITA-CSC-A 2006:25
ISSN 1653-5723
ISRN KTH/CSC/A–06/25–SE
ISBN 91-7178-490-X

Akademisk avhandling som med tillstånd av Kungliga Tekniska Högskolan framlägges till offentlig granskning för avläggande av filosofie doktorsexamen måndagen den 11 december klockan 13.00 i sal F3, Kungliga Tekniska Högskolan, Lindstedtsvägen 26, Stockholm.

Typeset by the author using LaTeX
Printed by Universitetsservice US AB

Abstract

Analytic knowledge about the systematic variation in a language has an important place in the description of the language. Such knowledge is interesting e.g. in the language teaching domain, as a background for various types of linguistic studies, and in the development of more dynamic speech technology applications. In previous studies, the effects of single variables or relatively small groups of related variables on the pronunciation of words have been studied separately. The work described in this thesis takes a holistic perspective on pronunciation variation and focuses on a method for creating general descriptions of phone-level pronunciation in discourse context. The discourse context is defined by a large set of linguistic attributes ranging from high-level variables such as speaking style, down to the articulatory feature level. Models of phone-level pronunciation in the context of a discourse have been created for the central standard Swedish language variety. The models are represented in the form of decision trees, which are readable for both machines and humans. A data-driven approach was taken for the pronunciation modelling task, and the work involved the annotation of recorded speech with linguistic and related information. The decision tree models were induced from the annotation. An important part of the work on pronunciation modelling was also the development of a pronunciation lexicon for Swedish. In a cross-validation experiment, several sets of pronunciation models were created with access to different parts of the attributes in the annotation. The prediction accuracy of pronunciation models could be improved by 42.2% by making information from layers above the phoneme level accessible during model training. Optimal models were obtained when attributes from all layers of annotation were used. The goal for the models was to produce pronunciation representations representative for the language variety and not necessarily for the individual speakers, on whose speech the models were trained. In the cross-validation experiment, model-produced phone strings were compared to key phonetic transcripts of actual speech, and the phone error rate was defined as the share of discrepancies between the respective phone strings. Thus, the phone error rate is the sum of actual errors and discrepancies resulting from desired adaptations from a speaker-specific pronunciation to a pronunciation reflecting general traits of the language variety. The optimal models gave an average phone error rate of 8.2%.

Keywords

Pronunciation modelling · Pronunciation variation · Discourse-context · Phone-level variation · Central standard Swedish · Spoken language annotation · Data-driven methods · Machine learning · Decision trees · Pronunciation lexicon development · Machine-readable lexicon · Phonology · Discourse · Lexicon

Sammanfattning

Analytisk kunskap om den systematiska variationen i ett språk har en viktig plats i beskrivningen av språket. Sådan information är intressant t.ex. inom språkundervisningsområdet, som bakgrund till olika typer av lingvistiska studier och i utvecklandet av mer dynamiska takteknologitillämpningar. I tidigare studier har effekterna av enstaka variabler eller relativt små grupper av relaterade variabler på uttalet av ord undersökts separat. Arbetet som beskrivs i denna avhandling tar ett helhetsperspektiv på uttalsvariation och fokuserar på en metod för att skapa generella beskrivningar av fonnivåuttal i diskurskontext. Diskurskontexten definieras av en stor mängd lingvistiska attribut som sträcker sig från högnivåvariabler som talstil ner till artikulatoriska särdragsnivån. Modeller av fonnivåuttal i kontexten av en diskurs har skapats för språkvarieteten central standardsvenska. Modellerna är representerade i formen av beslutsträd, vilka är läsbara för både maskiner och människor. Ett datadrivet angreppssätt antogs för uttalsmodelleringsuppgiften och arbetet innefattade uppmärkning av inspelat tal med lingvistisk och relaterad information. Beslutsträdsmodellerna inducerades från uppmärkningen. En viktig del av uttalsmodelleringsarbetet var också utvecklandet av ett uttalslexikon för svenska. I ett korsvalideringsexperiment skapades ett flertal uttalsmodeller med tillgång till olika delar av attributen i uppmärkningen. Precisionen i uttalsmodellers förutsägelser kunde förbättras med 42,2% genom att göra information från lager ovanför fonemnivån tillgängliga vid träningen av modeller. Optimala modeller uppnåddes när attribut från alla uppmärkningslager användes. Målet för modellerna var att producera uttalsrepresentationer som är representativa för språkvarieteten och inte nödvändigtvis för de enskilda talare på vilkas tal modellerna tränats. I korsvalideringsexperimentet jämfördes modellgenererade fonsträngar med facittranskriptioner av faktiskt tal och fonfelfrekvensen definierades som andelen avvikelser mellan de respektive fonsträngarna. Fonfelfrekvensen är således summan av faktiska fel och avvikelser som uppkommit genom önskade anpassningar från ett talarspecifikt uttal till ett uttal som speglar generella drag hos språkvarieteten. De optimala modellerna gav en genomsnittlig fonfelfrekvens på 8,2%.

Acknowledgements

I would like to thank my supervisor Rolf Carlson for his ideas and valuable comments and suggestions during my thesis work. I would also like to thank Beáta Megyesi for introducing me to the Department of Speech, Music and Hearing at KTH.

Thanks to Beáta Megyesi, Sara Rydin, Håkan Melin and Botond Pakucs for helping out with Linux-oriented questions. Thanks to Jens Edlund for perl advice and help with databases and sql. Thanks to Jonas Beskow for advice on tcl/tk and help with the KTH text-to-speech system. Thanks to Rolf Carlson for help with Rulsys. Thanks to Botond Pakucs for php advice. Thanks to Giampiero Salvi for advice and discussions on LaTeX. Thanks to Leif Grönquist for tools, information and tips related to the Göteborg Spoken Language Corpus.

Thanks to Rolf Carlson, Sheri Hunnicutt, Kjell Gustafson and David House for proof reading and commenting the thesis and related papers. Thanks to Kjell Gustafson for discussions on CentLex matters and many other things and for good co-operation in the CentLex project.

Thanks also to everyone who have made tools and resources used for the work described in this thesis available. Special thanks Robert Bannert and Peter Czigler for their VaKoS database, to the Grog project participants for the Radio speech data and annotation, to the department of Linguistics at Göteborg University for access to the Göteborg Spoken Language Corpus, to Kåre Sjölander for his excellent free software and for access to his aligner and phoneme models, and to Beáta Megyesi for part-of-speech tagging and parsing software.

Thanks to *Grötgänget*, Sheri Hunnicutt, Karl-Erik Spens, Inger Karlsson, Kjell Elenius, Mats Blomberg, Preben Wik and others for many culinary experiences and interesting discussions over lunch.

The research reported in this thesis was carried out at the Centre for Speech Technology (CTT), a competence centre at KTH, supported by Vinnova (the Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organisations. The research was supported by the Swedish National Graduate School of Language Technology (GSLT).

Per-Anders Jande's Publications on Pronunciation Modelling

- Per-Anders Jande (2003). Evaluating Rules for Phonological Reduction in Swedish. In *Proceedings of Fonetik*, pp. 149–152, Lövånger, Sweden June 2–4 2003.

- Per-Anders Jande (2003). Phonological Reduction in Swedish. In *Proceedings of Proceedings of the International Congress of Phonetic Sciences (ICPhS)* pp. 2557–2560, Barcelona, Catalonia, August 3–9 2003.

- Per-Anders Jande (2004). Pronunciation variation modelling using decision tree induction from multiple linguistic parameters. In *Proceedings of Fonetik* pp. 12–15, Stockholm, Sweden, May 26–28 2004.

- Per-Anders Jande (2005). Annotating Speech Data for Pronunciation Variation Modelling. In *Proceedings of Fonetik* pp. 25–28, Göteborg, Sweden, May 25–27 2005.

- Per-Anders Jande (2005). Inducing Decision Tree Pronunciation Variation Models from Annotated Speech Data. In *Proceedings of Interspeech* pp. 1945–1948, Lisbon, Portugal, September 4–8 2005.

- Per-Anders Jande (2006). Integrating Linguistic Information from Multiple Sources in Lexicon Development and Spoken Language Annotation. In *Proceedings of the LREC workshop on merging and layering linguistic information* pp. 1–8, Genoa, Italy, may 23 2006.

- Per-Anders Jande (2006). Modelling Pronunciation in Discourse Context. In *Proceedings of Fonetik* pp. 69–72, Lund, Sweden, June 7–9 2006.

- Per-Anders Jande (Submitted). Spoken Language Annotation and Data-Driven Modelling of Phone-Level Pronunciation in Discourse Context. Submitted to *Speech Communication* in May 2006.

Contents

Chapter 1

Introduction

Language can take many different forms and vary along many dimensions. Different speakers will use a language in somewhat different ways, even if they are speakers of what is generally conceived of as the *same* language. A language changes over time and this means that language users of different ages will differ more or less in their language performance. Sub-communities of language users based on e.g. geographical or social factors may also differ in how they use the language. There is also variation in language performance between individual language users, *idiomatic* variation.

In addition to this variation between language users and groups of language users, there is also variation in performance within a language group and within an individual speaker. For example, there is considerable variation in the spoken language performance of an individual speaker depending on the speaking situation. The choice of words as well as the pronunciation of words may change with the situation. Word pronunciation depends heavily on i.a. speaking style (cf. e.g. Ostendorf et al., 1996; Van Bael et al., 2004) and speech rate (cf. e.g. Fosler-Lussier and Morgan, 1999; Zheng et al., 2000).

The pronunciation of a certain word also depends on its local context, such as adjacent phonemes and the predictability of the word in its context. The predictability of a word can be estimated from variables such as global word frequency and n-gram probabilities (cf. e.g. Fosler-Lussier and Morgan, 1999; Jurafsky et al., 2001) and is also dependent on e.g. the new/given status of the word (Horne et al., 1994; Jurafsky et al., 1997).

Pronunciation variation is an exciting and important area of research with many relevant applications in the speech technology area and linguistic research. More knowledge about systematic variation in a language is an important part of the description of the language and interesting for e.g. language teaching and contrastive linguistic and phonological studies.

Knowledge about systematic pronunciation variation is also a necessity for developing more dynamic and human-like synthetic speech for e.g. human-computer dialogue systems and for interpreting human speech in automatic speech recognition and understanding systems.

The variation in pronunciation is manifested on many levels. There is variation in prosodic features, such as speech rate, intonation, rhythm and accent. There is also variation in the phone-level realisation of words and in the fine-phonetic realisation of speech segments.

Although there is a certain degree of individual (idiomatic) and random variation in the pronunciation of words in context, the variation due to context factors is largely systematic within a restricted, relatively homogeneous group of language users. This agreement on systematic variation strategies can be seen as a property of the language variety (e.g. dialect) spoken by a specific group. A language variety thus has its rules and frames for variation and this systematicity is a prerequisite for the language to ensure successful communication.

In their discussion on the concept of *distinctive features*, (Jakobson et al., 1963, section 1.3) stated that “for the study of speech sounds on any level whatsoever their linguistic function is decisive” and it is obvious that the phonetic realisation of words varies depending on their linguistic-semantic prominence, following the principle of “sufficient discriminability” (Lindblom, 1990) in the particular context.

Earlier studies of pronunciation variation have mostly either been aimed at neutralising the variation or at explaining the variation. Studies aimed at neutralising variation, prototypically in an automatic speech recognision system, deal with pronunciation variation in general, but mostly give no explanatory model of the variation. In contrast, studies aimed at explaining variation are concerned with describing the causal relations between variables governing pronunciation variation and the types of variation occurring in speech. However, these studies often deal only with a specific type of variation or a specific source of variation, investigating the effects of a single variable or a small group of related variables on pronunciation separately.

The focus of this thesis is a method for modelling the pronunciation of words in discourse context, with *discourse context* including the speaking situation characterised by speaking style, discourse type etc. and linguistic and related context on different levels down to the articulatory feature level. The aim is thus to create an explanatory model of pronunciation in general allowing a large set of variables to co-operate in a transparent model.

More precisely, the aim is to model systematic discourse context-induced variation in phone-level pronunciation inherent in a language variety. The goal is thus to create models of the subset of variation in language performance which occurs within a restricted language group due to discourse context and the aspect of pronunciation variation modelled is pronunciation variation *on the phone level.* In creating such models, the aim is to find patterns common to the language variety being modelled while idiomatic variation specific to individual language users are avoided.

The target language variety is *central standard Swedish*, the standard variety of Swedish spoken in the Stockholm area. The methods can, however, easily be adapted for modelling other language varieties and languages.

The method used for pronunciation modelling is data-driven. Spoken language is annotated with various kinds of linguistic and related information and machine learning is used to create pronunciation models from the annotation. The phoneme is the central unit in the approach and the annotation is aimed at describing the discourse context of a phoneme from high-level linguistic variables such as speaking style, down to the articulatory feature level.

An important part of the work has been to develop an annotation scheme, so that data can be organized in a way that is theoretically and practically appropriate for the current purposes. Another important part of the work has been the development of methods for data annotation.

The work described in this thesis is partly driven by an interest in human language processing and the factors involved in how humans choose to alter their speech over different situations. The work is also partly driven by an interest in using knowledge about human language performance to improve speech technology systems, such as synthetic speech.

To accommodate both of these aspects of pronunciation in discourse context, the pronunciation models created have to meet two important specific requirements. The first requirement is that the models must be able to predict phone-level pronunciation in context with high accuracy. The second requirement is that the models can serve as linguistic descriptions of pronunciation variation. This second requirement calls for a method making it possible to create transparent models, revealing which variables are the most important for predicting pronunciation in context and how variables co-operate to make predictions.

The decision tree induction paradigm is not impeded by the fact that the data from which a model is to be induced is of disparate kinds, as it is in the annotation described in this thesis. For example, the annotation includes discrete variables, such as Part of Speech, continuous integer-valued variables, such as the position of a phoneme in a cluster and continuous real-valued variables such as mean phoneme duration. The decision tree paradigm also produces transparent models, which can easily be transformed into rules. Models produced by a decision tree inducer are thus able to meet the specified requirements and decision tree induction is the method selected for creating models.

1.1 Method Selection Rationale

Modelling pronunciation variation in discourse context is interesting for the description of a language variety. The influence of a large variety of variables on the pronunciation of words has been studied in previous research projects over the years. However, the variables have mostly been studied in isolation or in small sets. A detailed description of a discourse, including a large variety of linguistic and related variables, enables studies of the interplay between various information sources on e.g. phone-level pronunciation and allow data-driven creation of models for prediction of word pronunciation in context.

For creating models of pronunciation variation with the specific purpose to get the most natural-sounding speech synthesis, the best strategy may be to try to mimic the pronunciation variation behaviour of one single speaker and use this speaker to build the concatenation database (cf. e.g. Miller, 1998a; Bennett and Black, 2003, 2005).

However, if the aim is to describe a language variety from a pronunciation variation point of view or to develop a model that for other reasons is general to the language variety (e.g. if the model is to serve as a general resource for speech technology applications), looking at many speakers of the particular language variety for creating models is necessary. Statistics can then be used to single out common patterns from individual patterns. Further, if a pronunciation variation model is to be generally applicable in e.g. speech synthesis contexts, it cannot be specific for the pronunciation variation of a single speaker. Models of the type described in this thesis will be models general for a language variety and are thus induced from the annotated speech of several speakers.

1.2 A Canonical Pronunciation Lexicon

The point of origin for the pronunciation models is a maximally detailed, context independent *canonical* 'citation form' pronunciation description, which can be transformed into its context dependent counterpart given a description of the context. The canonical pronunciation description serving as the basis for the pronunciation modelling method presented in this thesis corresponds to a phonemic description of the type which can be found in a pronunciation lexicon.

For the method to be successful, it is important that the phonemic pronunciation descriptions are of high and consistent quality. The same demands on lexical resources are shared by most research in the field of speech technology, and lexica are also important in the development of speech technology applications.

For these reasons, a part of the pronunciation modelling research has been to develop a lexical resource called CentLex. This lexicon is based on lexical data resulting from a number of projects at the Department of Speech, Music and Hearing (TMH) and the Centre for Speech Technology (CTT) at KTH over the years. The resource has been expanded and edited and now constitutes a general resource which can be used e.g. for development of automatic speech recognition systems and speech synthesis systems and in various speech technology applications, and generally for research purposes. CentLex is centrally available for the department and for the partners of CTT and tools for co-operative continuous lexicon development have been created.

1.3 The Approach to Pronunciation Modelling

The approach to pronunciation variation modelling used for the work presented in this thesis can be described as consisting of nine basic steps, as listed below. Each of these steps is described in a separate chapter of the thesis.

- Surveying previous research in the area
- Developing a canonical pronunciation lexicon
- Evaluating the canonical pronunciation lexicon
- Developing an annotation scheme and methods for annotation
- Collecting and annotating speech corpora
- Creating pronunciation models from the annotation
- Evaluating the pronunciation models
- Phonologically analysing the pronunciation models
- Discussing how the models can be used in speech technology applications

1.4 Thesis Overview

The next chapter of this thesis (Chapter 2) gives a background to pronunciation variation modelling, including definitions of some basic concepts and terms, a review of previous research in the area and the presentation of an evaluation of a tentative rule system for phonological reduction.

Chapter 3 gives a description of the work with developing CentLex—the pronunciation lexicon that has served as the basis for the phoneme-level annotation used for pronunciation model induction. Further, the structure of the lexicon and tools created for continuous lexicon development are described in this chapter. CentLex has been built to be a general lexical resource and has been used in several contexts for research and development of speech technology applications. These more general aspects of the CentLex database are also discussed.

In Chapter 4, an evaluation of the coverage of CentLex over different text types and an evaluation of the quality of the pronunciation representations in CentLex are presented. Also, strategies for increasing the coverage and the accuracy of the lexicon are discussed in the chapter.

Chapter 5 describes the speech databases used for pronunciation modelling and the system used for annotating the data. The chapter further includes descriptions of methods used for segmenting the speech data into units in several layers and for obtaining some of the information included in the annotation.

Chapter 6 gives a detailed account of all information included in the annotation and the rationale behind including the information. Each layer of annotation is presented separately and the information variables associated with each layer and their possible values are listed.

Chapter 7 describes the use of decision tree induction for creating transparent models of pronunciation in discourse context. It is explained how training examples, which can be seen as context-dependent phonemes, are constructed by attaching attribute values derived from the annotation to phoneme-sized units.

Chapter 8 presents a tenfold cross validation evaluation of decision tree models induced from the annotated speech data. One of the aspects evaluated is how the phone error rate is affected by using different amounts of the speech data and different sets of attributes at model training. There is also a closer examination of the attributes and a presentation of which attributes are the best predictors of pronunciation variation.

Chapter 9 looks closer at the phoneme-to-phone conversions made by a final pronunciation model trained on all available data. The distributions of realisations for each phoneme, the shares of correct classifications and the rules employed by the model are discussed.

Chapter 10 discusses how pronunciation models can be used in speech synthesis systems to produce more dynamic synthetic speech. Chapter 11 gives a brief summary of the thesis and ends with some concluding remarks.

Chapter 2

Background

This chapter will give a background to pronunciation variation modelling, starting with a presentation of the theory on which the current research builds and definitions of some basic concepts and terms. This chapter will also include a review of previous research in the area and a summay of an initial listening experiment testing the hypothesis that a general reduction rule system can be used to increase the perceived naturalness of speech synthesis at high speech rates.

2.1 Basic Concepts

This section will include brief descriptions of some basic concepts associated with the field of pronunciation variation modelling and definitions of the central terms used in the thesis.

2.1.1 Pronunciation Variation Theory

The pronunciation of a word can vary on a continuous scale of phonetic detail, with a maximally detailed pronunciation of the word at the one end and *no realisation* at the other end. How a word is realised in a particular situation depends on many factors.

We speak in order to communicate, and thus on the most abstract level, we want to pronounce words in such a manner that communication is maximally facilitated. We must convey enough information for the intended message to get through, but we do not want to convey too much information, both for production-economical reasons and for conversational-pragmatic reasons (cf. e.g. Lindblom, 1990). The conversational-pragmatic reasons can be summarised by H. P. Grice's second conversational maxim of quantity: "Do not make your contribution more informative than is required" (Grice, 1975, 1989) (originally referring to *semantic* information, but equally true for phonetic information). We also want to be able to contrast the more important parts of the sentence from the background (e.g. contrast the parts of an utterance conveying new information to the parts conveying given information). Varying the degree of phonetic detail over the words and syllables of an utterance is the conventional way of doing this.

Lindblom (1990) discusses this variation in phonetic detail over different parts of an utterance and over different speaking situations. Lindblom describes the speech as varying on a hyper-hypo-speech dimension, where hyper-articulated speech has a great deal of phonetic detail, while hypo-articulated speech has less phonetic detail. A hyper-articulated form of a word will resemble the *canonical* form on the segmental-phonetic level, while a hypo-articulated form will be *reduced* in relation to the canonical form.

In this thesis, a *canonical* pronunciation representation will refer to the phonological representation of the pronunciation of a word, i.e. to a pronunciation representation of the type that can be found in a pronunciation lexicon. Pronunciation lexica normally describe maximally detailed pronunciation of words, i.e. how words are pronounced in a clear fashion in isolation.

A *context-dependent* pronunciation will, in this thesis, refer to a representation of an actual pronunciation (supplied by a human or an automatic transcription system) or to a form produced by a pronunciation model. The context-dependent pronunciation is described as a *realisation* of the canonical phonological form.

2.1.2 Phonemes and Phones

In this thesis, a *phoneme* is defined as an abstract unit in the canonical phonemic form of a word. The sequence of phoneme units used to describe the 'underlying' pronunciation of a word is chosen so that each phoneme unit has a physical counterpart in a maximally detailed pronunciation of a word. A *phone* is an abstract context-dependent realisation of the phoneme and a phonetic *segment* is the physical correspondent of a phone in an actual speech signal.

The phoneme and the phone units are used as a way of describing pronunciation variation based on general phonological theory, however without making any claims about mental counterparts of the units or about human speech processing. In this thesis, the definition of a *phonemic representation* is that it is the canonical pronunciation representation that can be found in a pronunciation lexicon. A *phonetic representation* is a representation of a stream of often overlapping or partly overlapping speech segments, converted into a sequence of categorical classes.

A phoneme is really an 'umbrella class' for a group of allophones serving the same function. Substituting an allophone for another allophone of the same phoneme class in a string of phones constituting a word can never change the string into another word (although it may change the string to a less typical or anomolous pronunciation of the same word). However, if the allophone is substituted for an allophone from a different phoneme class, the string is changed into either another word or a non-word. Both the phoneme and the allophone are abstract units, only applicable in canonical descriptions of pronunciation. In practice, speech segments may not be aurally distinguishable, although belonging to different categories.

The phone has a more pragmatic definition than the phoneme. A phone symbol represents speech sounds, segments, which share acoustic and functional properties. The phone set used for describing the pronunciation of spoken language can be extended or curtailed, depending on the intended use of the description. For the work presented in this thesis, a limited set of phone symbols are used and the same phone symbol may be used to describe the (non-canonical) realisation of more than one phoneme.

The phone symbol set used is the same as the set of phonemes/allophones used to describe canonical word pronunciations. The decision to use this very restricted phone set was made since an automatic transcription system was used for obtaining phone sequences during the annotation of spoken language. The automatic transcription system uses a set of acoustic monophone models, including one model per phoneme in the central standard Swedish phomeme inventory.

The use of this phone symbol set is convenient when a model of pronunciation variation is to be used in an existing diphone speech synthesis system, with exactly those diphones available. On the other hand, the degree of phonetic detail modelled is limited by the small phone set. However, it would be possible to extended the phone set to describe more fine-phonetic variation within the framework of the current pronunciation modelling method.

Since the symbol sets for phonemes and phones overlap, phonemes and phonemic representations are often written in slashes, /fəʊniːm/, and phones and phonetic representations written in square brackets, [fəʊn], to distinguish between the two types of representations.

In the annotation of speech data, the phoneme symbols used have been those in the Swedish Technical Alphabet, STA. In the text of this thesis, the phoneme and phone symbols from the International Phonetic Alphabet, IPA, are used in most cases. There is a one-to-one correspondence between the symbol sets, as illustrated in Table A.1 in Appendix A.

2.2 Earlier Work on Pronunciation Variation

As stated in the introduction to this thesis, in earlier work on pronunciation variation, two main types of studies, differing in their general aim, can be discerned. Simply put, there are studies aimed at explaining the variation in pronunciation and studies aimed at neutralising the variation. Studies aimed at explaining variation are concerned with describing the causal relations between variables governing pronunciation variation and the types of variation occurring in speech. These studies often deal only with a specific type of variation or a specific source of variation. The purpose is generally to extend the phonetic or phonological description of a language or a language variety. A practical application of detailed knowledge about the sources of different kinds of pronunciation variation is in the speech synthesis area.

Studies aimed at finding methods for neutralising pronunciation variation are mostly concerned with variation in general, but not necessarily with describing or explaining the sources of the variation. The purpose of these studies is mostly to improve automatic speech recognition (ASR) systems. The studies may explore source-dependent solutions for neutralising pronunciation variation. For example, dialect and speech rate induced variation, respectively, may be handled in different ways in an ASR system.

In the context of this thesis, the focus of interest is which variables have been shown to affect phone-level pronunciation. Some examples of studies of variables governing pronunciation variation are presented in sections 2.2.1 and 2.2.2 below. The studies may be focused on neutralising variation in ASR systems or be studies aimed at general language description. Sections 2.2.3 to 2.2.5 focus on different types of methods for modelling pronunciation variation.

Since speech synthesis is an area where pronunciation models of the type described in this thesis can be applied, two sections presenting some earlier work on pronunciation modelling in speech synthesis, sections 2.2.6 and 2.2.7, are also included in this chapter. Finally, since the target language used for the work presented in this thesis is (central standard) Swedish, some previous work on pronunciation variation modelling for Swedish is presented in Section 2.2.8.

2.2.1 Variables Governing Pronunciation Variation

Variables that have been found to influence the within-speaker phone-level realisation of words in context are foremost speech rate, word predictability (often estimated by global word frequency) and speaking style. For example Fosler-Lussier and Morgan (1998, 1999) investigated the relationships between word predictability, speech rate and pronunciation in spontaneous American English speech. It was shown that reduction on the phone level in relation to the maximally detailed pronunciation is greater for highly predictable words and when speech rate is high. Van Bael et al. (2004) investigated the application probability of a set of phonological reduction rules operating on canonical pronunciation representations for recorded Dutch speech of three different speaking styles. They found that significantly more reduction rules needed to be applied when transforming the canonical representation into the transcript of spontaneous speech than when transforming it into the transcript of read speech and public lectures.

Jurafsky et al. (2001) report a higher probability of reduced pronunciation of frequent function words in American English spontaneous speech, when the function words occur in pairs with high bigram or reverse bigram probabilities. They also report that higher unigram probabilities for content words imply a higher probability for elision of a word final /t/ or /d/. Ostendorf et al. (1996, 1997) explore the use of speaking style-dependent lexical expansion rules. They also deal with speaking style prediction from acoustic observations (e.g. speaking rate and relative energy) and from information status (e.g. new vs. given information and content words vs. function words) for American English in an ASR context. Finke and

Waibel (1997) use information about word frequency, information status, estimated speech rate and f_0 to select the most probable pronunciation variant from an ASR lexicon during automatic phonetic transcription.

Zheng et al. (2000) report using parallel, rate-specific acoustic models for automatic recognition of American English, which improved the recognition accuracy compared to a model with a collapsed-rate acoustic model. This study is an example of pronunciation variation modelling on a higher level. The acoustic models used were trained in the usual way, but on different data sets (slow and fast speech) and it was shown that separating different types of speech can improve recognition.

Hazen et al. (2002) report using finite state transducer representations of phonological pronunciation variation in an ASR setting. What is different about this model is that there are rules using stress and syllable position as context. This information is captured with syllable position-dependent labels in the phonetic representations. Another interesting language model is reported by Bates and Ostendorf (2002). This model uses hand-labelled prosodic attributes (based on f_0, energy, and duration values) as context in pronunciation variation rules.

Nakajima et al. (2001) report comparing phonetic transcripts of Japanese continuous conversational speech and read versions of the conversations to derive phonological pronunciation variation rules. Although the rules are used in an ASR setting, there are aspects of the method for deriving rules, and of the results, which may be interesting also in the context of creating explanatory models of pronunciation variation. These aspects are that Part of Speech is used as context for the reduction rules and that Part of Speech-specific as well as general reduction rules were detected.

2.2.2 Studies Involving Several Pronunciation Predictors

Some previous studies on pronunciation variation have simultaneously taken several context variables into account. For example, Bates and Ostendorf (2001) used syntax and discourse-related features e.g. to predict the context-dependent phone-level realisation of words in the Switchboard corpus[1] with decision trees. The features used were the log trigram score, Part of Speech tags for the word and for the left and right adjacent words, the position of the word in the utterance, the position of the word in relation to a sentence pivot point (the main verb), and dialogue act classification. They also used a number of phone and stress-related baseline features.

The idea behind the pivot point concept is that it can capture some discourse-level information, since words before the main verb of a sentence (or more general macro-syntactic unit) typically convey *given* information, while words after this pivot point typically convey *new* information (Jurafsky et al., 1997, 1998b).

Bates and Ostendorf (2001) also included intermediate predictors as attributes for their decision trees. The intermediate predictors included a prediction of

[1] American English spontaneous telephone speech.

phonetic distance (number of deviating articulatory features) and a prediction of transformation type (*no transformation*, *insertion*, *reduction* or *other*) and were determined by decision trees trained to make these decisions. The best result, a phone error rate (PER) of 19.4%, was obtained when all attributes were used. Compared to the baseline model using phoneme-level attributes only, the reduction of PER was about 10%.

Bates and Ostendorf (2002) added prosodic variables (duration-based, energy-based, and fundamental pitch-based measures) to the attribute set used in Bates and Ostendorf (2001). The duration-based variables included the absolute duration of utterances, words and phones, respectively. They also included word duration normalised by (divided by) the utterance duration and phone duration normalised by the word duration. The energy-based variables included the mean, maximum and minimum values over the utterance, over the word and over windows of 15 and 30 frames, respectively. The energy values were also normalised by dividing them with the average energy of the first 10 frames of the particular conversation. The f_0-based variables included the mean, maximum and minimum values over the utterance, over the word, and over 15 and 30 frame windows, respectively. Further, the slope values and the number of slope changes over the utterance and the word were included. Pitch values were normalised both by subtracting and by dividing the f_0 with a speaker baseline. The prosodic attributes, especially word duration and word energy values, gave slight improvements in PER. More details are given in Bates (2003).

The fact that a word conveys given information as opposed to new information may have consequences for the pronunciation of the word in its context. Simply put, since there is more top-down information available for a word associated with given information than for a word associated with new information, there may be less need for bottom-up information. Thus, a speaker can pronounce a word associated with given information with less phonetic detail without increasing the demands on the listener. Horne et al. (1993, 1994) describe an algorithm developed to keep track of the semantic identity of lexical instances in a text as well as of semantic relations between words, as to avoid putting focal stress on a word conveying given information in a speech synthesis setting (Horne and Filipsson, 1996; Bruce et al., 1996). The fact that given information is generally not phonetically accented may of course not only affect stress patterns, but also the segmental realisation of words.

Greenberg et al. (2002) report a study of the relation between stress accent and pronunciation variation in American English based on a subset of the Switchboard corpus. The study revealed that the heavier the stress, the less a syllable will deviate from the canonical realisation. Unstressed syllables thus deviated more from the canonical realisation than did stressed syllables. Nuclei and codas were the parts of the syllable that were most affected by stress. The codas were subjected mostly to segment elision (in relation to the canonical realisation) and nuclei were subjected mostly to substitution. Onsets of unstressed syllables were also affected, and subjected to both elision and substitution, however to a lesser degree than the succeeding parts of the syllable.

Gregory et al. (1999) report investigations of the effects of word predictability on the realisation of certain words from the Switchboard corpus. The words investigated were words who in their canonical pronunciation end in /t/ or /d/. The realisation factors investigated were three 'shortening phenomena': 1) /t/ or /d/ elision, 2) /t/ or /d/ tapping and 3) word duration shortening.

A base model was created, including a set of well-known predictors of 'reduced pronunciation': *speech rate*, left and right adjacent *phone type* (vowel or consonant), right adjacent *vowel quality* (full or reduced), *word length* (syllables) and *word type* (function word or content word). Also included in the base model was information about the identity of the final consonant of the word (/t/ or /d/).

The predictability measures investigated were 1) *prior probability*, defined as the relative word frequency in Switchboard, 2) *collocational probability*, including several measures based on bigram and trigram probabilities and *mutual information* (the bigram probability divided by the product of the individual word probabilities), and 3) *discourse probabilities*, including a count of *word repetition thus far in the conversation* and a measure of *semantic relatedness* calculated through Latent Semantic Analysis, LSA (Landauer and Dumais, 1997; Landauer et al., 1998).

The addition of information resulting from combining the set of predictability measures with the base model was measured using multiple regression analysis. Only the mutual information measure was shown to affect tapping. Elision was affected by word frequency, mutual information, reverse trigram probability and forward trigram probability. Durational shortening was shown to be sensitive to word frequency, mutual information and semantic relatedness.

In a similar manner to Gregory et al. (1999), Bell et al. (2003) investigated the effects of disfluencies, predictability, and utterance position on 'lengthening phenomena' in 8,000 occurrences the ten most frequent English function words collected from the Switchboard corpus. It was concluded that the words are more likely to be longer or occur in a phonetically fuller form when 1) neighbouring disfluencies are present, 2) the predictability of the word is low and 3) when the word stands in utterance initial or utterance final position.

Duez (1998) studied elision and assimilation phenomena in consonant sequences in French spontaneous speech, taking into account phoneme context, word type (function word or content word), syllable prominence and the position of the phoneme in the syllable, word and phrase, respectively. The study was conducted on a spontaneous speech database with phonetic transcripts provided by human transcribers. The manually obtained transcripts were compared to canonical phonemic representations. Among other things, Duez (1998) found that reduction phenomena seem to be highly dependent on syllable boundaries (phonemes adjacent to syllable boundaries were much more prone to having reduced realisations than phonemes not in the immediate vicinity of syllable boundaries), on the position in the syllable (codas were more prone to phonological reduction than onsets) and on syllable prominence (non-prominent syllables were more prone to reduction than prominent syllables).

Duez (2001) describes additional reduction phenomena and assimilatory effects in French speech and relates the effects to different context factors. Observations relating to *phonological features of phonemes* were that sonorants are more often deleted or substituted than obstruents and that voiced obstruents are more often deleted or substituted than voiceless obstruents. An observation relating to *word type* (function word vs. content word) was that function words largely show individual patterns. Finally, an observation relating to the *position in the phrase* was that phrase-final syllables often are more prominent and less subjected to reduction processes than non-final syllables.

Välimaa-Blum (1998) reports a similar study for Finnish spontaneous speech, finding e.g. syllable length, stress (prominence) and vowel harmony to be important factors for how words are reduced. Su and Basset (1998) looked at vowel and consonant reduction phenomena in both French and Taiwanese Mandarin. They found that vowel reduction was more common in French and consonant reduction was more common in Taiwanese Mandarin. Further, coda consonants were reduced more often than onset consonants in French. For Taiwanese Mandarin, it was the other way around. This suggests that the language specific factors of reduction are very important. Greenberg and Fosler-Lussier (2000) use the fact that pronunciation variation is very sensitive to syllabic, lexical and phrasal context to argue for a view on pronunciation variation that is less centred on articulatory/bio-mechanical factors and more dependent on higher level linguistic organisation.

There are many reports on fine-phonetic studies of vowel reduction due to context factors, cf. e.g. Moon and Lindblom (1994) and Engstrand (1988).

2.2.3 Methods for Modelling Pronunciation Variation

Strik and Cucchiarini (1998, 1999) and Cucchiarini and Strik (2003) give overviews of the literature dealing with pronunciation variation for automatic speech recognition purposes. In these overviews, the methods used in the reviewed literature are characterised in terms of 1) information sources (knowledge-based vs. data-driven methods), 2) the type of pronunciation variation modelled, 3) information representation and, 4) level of modelling. Strik and Cucchiarini distinguish between two types of pronunciation variation modelled for ASR systems: *within-word* and *cross-word* variation and conclude that within-word variation is the most frequently modelled type. Further, two types of information representation are distinguished: *enumeration* and *formalisation.* Enumerations are representations where all possible variations are listed, e.g. in a lexicon. Formalisations can be rule systems (either derived from data or constructed from linguistic knowledge). For data-driven approaches, the derived formalisations can also be in the form of e.g. artificial neural networks or phone confusion matrices.

In sections 2.2.4 and 2.2.5 below, some studies using different combinations of information sources, variation types and information representations are presented. These are all studies aimed at improving speech recognition. They may include rule application probabilities or probability of use for different word pronunciations.

However, they do not say anything about when (in what contexts) a certain rule should be applied or when a certain word pronunciation should be used.

2.2.4 Knowledge-Based Methods

Automatic speech recognition performance in relation to using a static, canonical pronunciation lexicon can be improved with relatively simple methods. For example, a small set of knowledge-based phonological reduction rules can be used to extend the lexicon with reduced forms and application probabilities can be associated with the different pronunciation representations. In this section, some studies using relatively simple methods for improving ASR performance are presented. It is interesting to see that these straight-forward methods give improvements in ASR performance. However, these studies do not consider *when* (in what contexts) a certain pronunciation variant should be used, and are thus only of limited interest for the modelling of pronunciation variation with the aim to describe a language variety or to increase the naturalness of synthetic speech.

Adda-Decker and Lamel (2000) evaluate different hand-constructed reduction rules for German by creating different speech recognition lexica and comparing the word error rate from continuous speech recognition resulting from using the different lexica. The rules delete [ə] vowels before [l], [n] and [m] in unstressed syllables. Only slight improvements were seen using the lexica expanded with these simple rules. A similar experiment was conducted by Kessens et al. (1999) for Dutch. However, in this case, using five phonological rules to expand their recognition lexicon was enough to show a significant improvement in recognition performance. Kessens et al. (1999) also modelled cross-word variation using two different methods: 1) adding more phonetic realisation variants to the lexicon and 2) adding new lexical items in the form of multi-words. The second cross-word pronunciation modelling method proved better than the first method. In combination, within-word modelling and cross-word modelling using the better method showed an improvement in word error rate of 8.8%. Wester et al. (1998b) report further improvements when also adding pronunciation variant probabilities to the model. Another study using cross-word (phrase level) variation modelling for improving Dutch ASR is reported by Ordelman et al. (1999b,a).

Tajchman et al. (1995a,b) and Seneff and Wang (2002); Chung et al. (2004); and Seneff and Wang (2005) report using knowledge-based rules and data-driven methods to derive their application probabilities. Koval et al. (2002) use knowledge-based phonological rules to create hierarchical networks representing the possible pronunciations of words.

2.2.5 Data-Driven Methods

Data-driven methods have been used for creating more complex ASR pronunciation models and this section presents some experiments using data-driven methods for pronunciation modelling.

Wester et al. (2000) used a data-driven approach to the deduction of phonological elision rules. First, an ASR lexicon with one canonical phonemic representation per orthographic word was automatically expanded with all possible variants of each pronunciation representation with one or more phonemes deleted, given the restriction that at least one phoneme per syllable should be left. This naïve expansion of course produced many variants that are never used in actual speech. However, in a forced recognition step, the recogniser was used to transcribe a spontaneous speech corpus and thus select the most likely phonetic realisation for each word given the observation sequence. Elision rules for phonemes in the context of one preceding and one succeeding phoneme were derived by comparing the phonemic representations in the original lexicon with the selected phonetic realisations. Rules which were applied less than 100 times in the material transcribed were excluded.

The rules also had to have a minimal relative rule application (number of times the rule was applied divided by the number of times it could have been applied) of 0.2 to be included in the final system. A new recognition lexicon was generated from the lexicon with one representation per word, using the derived rules. Further, the acoustic models were re-estimated using the selected phonetic realisations and the language models were re-estimated, so that different variants received different probabilities. Using the four most common elision rules showed the same recognition performance as using a four-rule knowledge-based rule system, although the data-derived rules only produced a quarter of the pronunciation representations produced by the knowledge-based rules. In a similar comparison of knowledge-based and data-derived phonological reduction rules, Kipp et al. (1997) use a data-driven approach that outperforms a knowledge-based approach for German continuous speech segmentation. Fukada et al. (1998) report using a method similar to that used by Wester et al. (2000) to expand a lexicon for Japanese ASR.

Byrne et al. (1998) and Riley et al. (1999) describe the use of decision trees to model pronunciation variation in English continuous speech. Using decision trees on-line in evaluation (i.e., allowing all phoneme-level variation of each word found in the training data) proved to decrease recognition performance. Limiting over-generalisations by only allowing variants that occurred sufficiently often in the training material, however, increased the performance slightly. Multi-word units added to the lexicon further increased the performance.

Pastor-i-Gadea and Casacuberta (2001) automatically train finite state automata using Spanish speech data. The automata describe the different phonetic realisation variants of words. Like Byrne et al. (1998), Pastor-i-Gadea and Casacuberta also use the stop criterion of a certain number of appearances in the training data for a rule to be incorporated in the system. Kessens et al. (2002) show that this rule selection criterion (i.e., absolute frequency of rule application) is the most suitable selection criterion for a Dutch ASR.

2.2.6 Pronunciation Modelling in Speech Synthesis

A strategy used for modelling phone-level pronunciation variation for speech synthesis purposes has been to model the pronunciation of a single speaker, typically the speaker whose voice is used in the concatenation database (cf. e.g. Miller, 1998a; Sundaram and Narayanan, 2002; Bennett and Black, 2003, 2005). Miller (1998b) used syntactic and prosodic annotation at several linguistic levels to create an artificial neural network model of the pronunciation of a single speaker. When the specific purpose is to get the most natural-sounding speech synthesis, modelling the pronunciation of a single speaker is a good strategy.

However, a model created for a single speaker will not be general for any group of speakers. If the aim is to describe the *language variety* from a pronunciation variation point of view, it is necessary to study the behaviour of many speakers of the particular language variety. Statistics can then be used to single out common patterns from individual patterns.

Werner et al. (2004a,b) use a stochastic pronunciation net induced from a speech corpus including many speakers (thus being a more general pronunciation model) and a word duration model. They first determine adequate word durations using the probability of a word in its context and then estimate the appropriate phone sequence given the specified durations, the transition probabilities from the word pronunciation nets, and word transition probabilities. Listening experiments showed that the efforts gave rise to more colloquial and natural-sounding speech.

Hawkins et al. (1998) and Ogden et al. (2000) use knowledge-based syntactic and phonological information in a speech synthesis system called ProSynth. The information is used to model segmental-phonetic and prosodic pronunciation and to select concatenation units or generate parameters for parametric synthesis. The information is organised in a hierarchical structure with the linguistically motivated levels *intonational phrase*, *accent group*, *foot*, *syllable*, *syllable constituent* (onset, rhyme), *rhyme constituent* (nucleus, coda) and *phoneme*. The units at each level have different types of information attached to them. For the syllable, the information attributes include *strength*, which takes the values *strong* and *weak*, and *weight*, taking the values *heavy* and *light*. For the phoneme, the attributes are phonological features.

2.2.7 Pronunciation Modelling in HMM Synthesis

A Markov chain is a weighted finite-state atomaton, i.e., a set of interconnected states where the probabilities of transitioning from each particular state to another are different. A Hidden Markov Model (HMM) used for phoneme recognition has states corresponding to parts of phonemes, represented by parameters derived from a speech signal. Such a model can be used to convert an input speech stream into a string of phonemes through converting the speech stream into a series of observations, i.e., a series of parameterised samples of the signal. When the signal has been converted into a series of observations, it can be calculated which is the most likely way for the model to generate the observation sequence, i.e., which is the most probable path trough the chain with weighted transition probabilities given the series of observations. When the most probable path is known, a sequence of phonemes can be derived by back-tracking through the model.

Yoshimura et al. (1999) describe HTS, a Hidden Markov Model synthesis system where speech is generated from HMMs and in which the spectrum, f_0, and duration are simultaneously modelled. Context-dependent phone HMMs (models for single phones using five states corresponding to different parts of the phone) are built taking a large set of contextual attributes into account. Separate HMMs are built for spectrum and pitch, and state duration is modelled by Gaussian distributions.

In the original system for Japanese speech, the contextual factors considered were information about 1) the *number of morae contained by the sentence*, 2) the *position of the breath group in the sentence*, 3) the *number of morae* in the preceding, current and succeeding *breath group*, 4) the *position of the current accentual phrase in the current breath group*, 5) the *number of morae* in the preceding, current and succeeding *accentual phrase*, 6) the *accent type* of the preceding, current and succeeding *accentual phrase*, 7) the *Part of Speech* of the preceding, current and succeeding word, 8) the *position of the current phoneme* in the current *accentual phrase*, and 9) the *identity* of the preceding, current and succeeding *phoneme.*

Tokuda et al. (2002) used HTS for implementing a synthesis voice for English. For the English version, the set of contextual factors was extended. In the implementation for English, contextual factors connected to five types of linguistic units were considered. Connected to the *phoneme* unit was information about 1) the *identity* of the preceding, current and succeeding *phoneme*, and 2) the *position of the current phoneme in the current syllable.*

Connected to the *syllable* unit was information about 1) the *number of phonemes* in the preceding, current, and succeeding syllable, 2) the *accent* of the preceding, current, and succeeding syllable, 3) the *stress* of the preceding, current, and succeeding syllable, 4) the *position of the current syllable in the current word*, 5) the number of preceding and succeeding *stressed syllables*, respectively, in the current phrase, 6) the number of preceding and succeeding *accented syllables*, respectively, in the current phrase, 7) the *number of syllables* from the previous to the next stressed syllable, 8) the *number of syllables* from the previous to the next accented syllable, and 9) the *identity of the vowel* in the current syllable.

Connected to the *word* unit was information about 1) the *Part of Speech* (automatically obtained 'guess') of the preceding, current and succeeding word, 2) the *number of syllables* in the preceding, current and succeeding word, 3) the *position of the current word in the current phrase*, 4) the number of preceding and succeeding *content words*, respectively, in the current phrase, and 5) the *number of words from the previous to the next content word.*

Connected to the *phrase* unit was information about 1) the *number of syllables* in the preceding, current and succeeding phrase, 2) the *position in a major phrase*, and 3) the ToBI *end tone of the current phrase.* Connected to the *utterance* unit was information about the *number of syllables in the current utterance.*

The HTS system has also been implemented for Swedish by Lundgren (2005). It was unfeasible to create separate models for all combinations of context factors for any of the implementations, since the amount of training data needed would be too extensive. Further, it may not even be possible to collect a speech database containing all combinations of factors. To overcome this problem, context factors were clustered.

The HTS approach does not explicitly model phone-level pronunciation. However, phone-level pronunciation is modelled implicitly in that phonetically more or less salient spectral properties and durationally longer or shorter segments (and more or less salient pitch characteristics) are selected depending on the specific context factors.

Prahallad et al. (2006) take pronunciation variation modelling in HMM synthesis a step further and suggest five-state HMM phoneme models allowing either all possible interconnections between states or all possible connections in the forward direction. Such phoneme models showed better log likelihood scores when applied to forced recognition (where the phoneme sequence is known, as in the speech synthesis case) of affected readings of short stories. The better scores indicated a better fit to the data than when standard HMM phone models, where all states must be passed in left-to-right order, were used. An additional context factor was also introduced in the duration models, *the number of times the current word had been mentioned in the discourse* (for content words).

2.2.8 Pronunciation Variation in Swedish

There is a considerable corpus of studies on the variation in pronunciation of Swedish words uttered in context from a phonetic perspective. These studies are mostly detailed phonetic studies, but some are interesting also from a more phonological perspective. For example, Lindblom (1963) investigated the process of reduction of Swedish vowels resulting from increasing speech rate. Öhman (1967) investigated the coarticulation properties of Swedish dental stops in vowel context (vowel-consonant-vowel syllables). Engstrand (1988) investigated the articulatory effects of stress and speaking rate in Swedish vowel-consonant-vowel syllables. Engstrand and Krull (1988) and Engstrand (1992) investigated phonetic variation in natural Swedish discourse.

Both segmental-phonetic studies, as those described above, and prosodic-phonetic studies have been reported. Some examples of studies aimed at prosodic aspects of pronunciation are Gårding (1967), in which juncture and syllabification in various styles of connected Swedish speech was investigated, and Horne et al. (1995), reporting investigations of final lengthening at prosodic boundaries in Swedish continuous speech.

In addition to the above mentioned studies oriented towards sub-phonemic and supra-phonemic of pronunciation variation in Swedish, a number of studies on pronunciation variation in Swedish on the phone-level have been carried out. Examples of such studies are Gårding (1974); Eliasson (1986); Bruce (1985, 1986) and Bannert and Czigler (1999).

Gårding (1974) presents a rule system for transforming canonical phonemic representations (describing a maximally detailed *careful* pronunciation of a word) of consonant clusters at word boundaries into representations corresponding to a fast speech pronunciation. In this study, lists of examples chosen to include all consonant clusters at word boundaries possible in standard Swedish (described by Sigurd 1965) were recorded at different speech rates and phonetic transcripts of these recordings were used to derive pronunciation rules.

Bannert and Czigler (1999) studied variations in consonant clusters using a larger corpus of recorded speech. The corpus included both read-aloud examples of the same type as the Gårding (1974) speech data and speech of a more spontaneous type. Among other things, Bannert and Czigler report the frequency of occurrence of every type of consonant cluster they find in the spontaneous speech corpus and the word boundary, compound boundary, and other morphological boundary context of each cluster. They also report the frequencies of all the different types of elision and assimilation processes (devoicing and fricativisation) they found in the corpus.

Eliasson (1986) presents some common types of phonological processes describing the differences in pronunciation between words spoken in isolation and words spoken in context, and between compound constituents spoken in isolation and spoken in the context of their compound word, respectively. The main focus is on retroflexation (postalveolarisation). In central standard Swedish and several other Scandinavian dialects, combining a unit ending with an /ɹ/ in its isolated form with a unit beginning with a dental consonant in its isolated form, gives rise to retroflexation of the dental consonant and /ɹ/-dropping. The process is recursive for all dental in direct succession, with some exceptions.

Bruce (1986) discusses omissions of vowels and syllables in everyday speech pronunciation as compared to canonical pronunciation. According to Bruce (1986), omission phenomena are governed primarily by the syllable-bound rhythmical organisation of spoken language. The author proposes a set of rules compiled through working with (listening, transcribing) different types of speech material. The rules were tested using a read-aloud sample text containing many words with possible elisions encountered in earlier work with speech data.

Many descriptions of Swedish from different phonology-related perspectives not focusing on pronunciation variation are, of course, also available. Since the work described in this thesis involves linguistic annotation of many different kinds, much of this work has been used as references. These descriptions of Swedish will be discussed in the following chapters in association with the description of annotation relating to their particular subjects.

2.3 A Tentative Rule System for Phonological Reduction

Inspired by the results from the above mentioned studies on pronunciation variation in Swedish, Jande (2003a,b) constructed a tentative rule system for transforming canonical phonemic representations of words into representations corresponding to a fast speech rate. The rule system was used to create synthetic speech stimuli used in an assessment experiment.

Nine words resulting in a wide range of rule applications were selected and each word was placed in a carrier sentence. The sentences were converted into phonological pronunciation representations using a canonical pronunciation lexicon. The *canonical* pronunciation representations were then processed by the rule system, resulting in a set of *reduced* pronunciation representations. Both the canonical and the reduced form of each sentence were synthesised using a diphone synthesiser. Three different speech rates were produced for each sentence variant, *low* (the system default rate), *medium* (1.3 times the default rate) and *high* (1.7 times the default rate).

The sentences were presented in pairs to a group of subjects, with each sentence of a pair having the same target word and the same speech rate, but differing with respect to reduction. The subject's assignment was to select the most natural-sounding sentence from each pair. The experiment showed that the reduced stimuli were generally perceived as more natural for the medium and high speech rates while there was no significant difference in perceived naturalness between the reduced and the canonical stimuli for the low speech rate. It was further shown that the preference bias in favour of the reduced pronunciations increased with increasing speech rate.

Some sentences broke the general pattern and were preferred by a majority of the subjects either in their reduced form or in their canonical form, irrespective of the speech rate. Post hoc word frequency estimations from a newspaper text corpus revealed that the words always being preferred in their reduced form were high frequency words, while words always preferred in their canonical form were low frequency words. Since word predictability has been shown to affect word pronunciation in several previous studies, the word frequency is a possible explanation for certain words breaking the pattern. However, since only a few words were studied, there is no way to be certain what made some stimuli break the pattern. Whatever the cause, the results still support the notion that other things than only phonological context and speech rate play a role for the pronunciation of a word.

Thus, to conclude, this first study showed that a general reduction rule system can be used to increase the perceived naturalness of speech synthesis at high speech rates. It also suggested that more context variables than phonological context must be included in a model of pronunciation in discourse context.

2.4 Summary

In this chapter, some basic concepts relating to pronunciation modelling have been defined and a background to the pronunciation modelling research area has been presented. An experiment using a tentative phonological rule system for adapting synthetic speech to higher speech rates was summarised. The experiment showed that reduction rules can be used to increase the perceived naturalness of speech synthesis at high speech rates. However, the results suggested that more context variables than phonological context must be included in a model of pronunciation in discourse context.

The next chapter will describe the creation and structure of the CentLex lexicon database, a machine-readable lexicon containing canonical pronunciation representations, which are used as the basis for the pronunciation modelling approach described in this thesis.

Chapter 3

Pronunciation Lexicon Development

The point of departure for the data-driven pronunciation modelling method described in this thesis is a set of context-independent pronunciation representations that correspond to phonemic descriptions of the type that can be found in a pronunciation lexicon. For the method to be successful, it is important that the phonemic pronunciation descriptions are of high and consistent quality. For this reason, a part of the pronunciation modelling research reported in this thesis has been aimed at developing a canonical pronunciation lexicon for Swedish. Considerable effort has been put into the development of this lexicon, which has been named CentLex.

A high quality pronunciation lexicon is essential for many areas of speech and speech technology research and for most speech technology applications. CentLex has been built to be a central lexicon database for the Department of Speech, Music and Hearing at KTH and the Centre for Speech Technology (CTT) and to meet the specific demands of the phone-level pronunciation modelling work which is the focus of this thesis as well as general demands from speech technology research and application development. CentLex is built as a relational database, and for the lexicon to function as a central resource, tools for facilitating access and continuous, cooperative editing of the lexicon database have been developed.

For the reader interested in lexicon development, the book *Lexicon Development for Speech and Language Processing* (Van Eynde and Gibbon, 2000) gives “a survey of methods and techniques for structuring, acquiring and maintaining lexical resources for speech and language processing” (in the words of the publisher). Another book, *The Structure of the Lexicon: Human versus Machine* (Handke, 1995) can also be used as a reference for work on machine-readable lexica.

3.1 CentLex: A Central Lexicon Database

The department of Speech, Music and Hearing (TMH) and the Centre for Speech Technology (CTT)[1] at KTH have been involved in a large variety of projects over the years and various applications have been developed at TMH and within CTT. Many of these projects and applications have entailed the development of lexical resources. To facilitate access to the lexical information available at TMH/CTT, the lexical data have been mapped onto a common format and brought together in a central lexical database, CentLex.

The main ideas behind the CentLex database are that all lexical data used in reasearch and application development is stored centrally, so that the data is immediately and easily accessible for all researchers at the department and for all partners involved in the Centre. Lexicon-related work performed in different projects can be easily integrated with the central lexical resource, and the results immediately available for all users. Standards for mapping between the CentLex format and several commonly used formats have been developed to facilitate information sharing.

3.2 Information Included in the Lexicon

CentLex is a full-form lexicon, with each entry minimally containing an orthographic word form and a grammatical analysis (Part of Speech and morphology). An entry can also have an arbitrary number of phonemic representations, ordered by their probability of use. Each phonemic representation may be enriched with information about the intended context of the representation (e.g. *reduced form* or *foreign language*). Information about the source language is added e.g. for proper names, since orthographically identical names may be pronounced differently depending on the native language environment of the person bearing the name. An entry also contains information about the probability of the particular grammatical analysis, given the orthographic word (estimated from a large automatically tagged text corpus).

3.2.1 Formats

Grammatical analyses (Part of Speech and morphology) in CentLex are in the format used in the suc corpus (Ejerhed et al., 1992). Pronunciation representations are stored in a special CentLex meta-format similar to the Swedish Technical Alphabet (STA), presented in Appendix A, tables A.1 and A.2. The CentLex meta-format differs from the STA format in that the retroflex consonants that are part of the central standard variety of Swedish, but not of all variants of Swedish, are not part of the meta-format.

[1] CTT is supported by Vinnova (the Swedish Agency for Innovation Systems), KTH, and participating Swedish companies and organisations.

An orthographic sequence of an <r> and a dental consonant is in most cases pronounced as a retroflex consonant in central standard Swedish. However, in southern standard Swedish, the same sequence would be pronounced as an /ʁ/ followed by a dental consonant. STA was developed for representing the central standard Swedish pronunciation, while the CentLex representations are common to the language varieties.

Thus, what would be pronounced as a retroflex in central standard Swedish is represented by an /ɹ/ followed by a dental consonant in CentLex. What would be pronounced as a sequence of retroflex consonants in central standard Swedish is represented by an /ɹ/ followed by a sequence of dental consonants.

When creating an actual pronunciation lexicon from the CentLex meta-format, the representation can be interpreted literally, as an /r/ allophone followed by a dental consonant or sequence of dental consonants, if a southern standard Swedish lexicon is desired. However, if central standard Swedish pronunciations are the desired output, the /ɹ/ followed by a sequence of dental consonants can easily be converted into a sequence of retroflex consonants. A special star symbol (*) can be introduced to symbolise that the conversion should not be performed at conversion from the CentLex meta-format to a representation matching a central standard Swedish pronunciation (this can be the case e.g. for names with 'foreign' origin).

The CentLex phoneme inventory includes a set of xenophone symbols, so that loan words and names of 'foreign' origin can receive representations closer to the original language pronunciation than what is possible with only the more restricted set of phonemes used more generally in Swedish. Table B.1 in Appendix B shows the xenophone symbols used in CentLex. For further information on xenophones in Swedish, cf. e.g. Eklund and Lindström (2001) and Lindström (2003), which report investigations on the use of xenophones in Swedish and the implications of xenophones for speech technology applications.

Mappings from the CentLex meta-format have been developed for a set of other pronunciation representation formats. Some examples are the Swedish IPA (Engstrand, 1999) format, the SUO format—the format of *Svenska Språknämndens Uttalsordbok*, the pronunciation dictionary of the Swedish Language Council Garlén (2003), and the original STA format (cf. Table A.1). Mappings to several application-specific representation formats have also been developed, e.g. to the formats used by the Infovox/Acapela Group synthesis voices *Ingmar* and *Emma*, by the L&H+ Swedish speech synthesis, by the Loquendo Swedish speech synthesiser, and by the Nuance Swedish automatic speech recogniser, respectively. Further a mapping to a STA-similar format used for ASR with HTK (the Hidden Markov Model Toolkit) has been developed.

If a lexicon is to be generated e.g. for a specific speech synthesis system, the meta-format is mapped to a form that the system can handle. This mostly involves mapping xenophone symbols to their nearest equivalent present in the specific diphone or unit selection database.

3.3 Available Lexical Resources

As mentioned, CentLex incorporates lexical resources developed at the Department of Speech, Music and Hearing at KTH and the Centre for Speech Technology over the years.

The lexical resources constituting the backbone of the CentLex database are the *KTH text-to-speech lexicon*, developed during many years and used e.g. in the KTH text-to-speech system (Carlson and Granström, 1976), the Swedish Onomastica proper name lexicon (Carlson et al., 1990; Gustafson, 1995a,b, 1996; Trancoso, 1995) and the Swedish Dragon lexicon (Reimers et al., 1995) and various word lists and text data available at the Department of Speech, Music and Hearing at KTH.

For creating the Dragon lexicon, the 120,000 most frequent words (orthographic forms) were selected from a 166 million word text corpus[2] with approximately 1.8 million unique word forms. The text corpus included mostly newspaper text, but also novels, law text, parliament protocols, public information, etc. The grapheme-to-phoneme rules of the KTH text-to-speech system were used to supply phonemic pronunciation representations for the words, and the representations were then manually checked and corrected when necessary. Some word units not deemed appropriate for the lexicon were discarded and the final lexicon contains 110,587 word units with manually checked pronunciation representations.

The Onomastica lexicon contains 184,760 names with pronunciation representations. From these names, 5,967 are place names, 10,459 are first names, 39,245 are street names, 122,524 are family names and 6,565 are titles.

For the creation of CentLex, the Stockholm-Umeå Corpus, suc (Ejerhed et al., 1992), was used as one of several information sources for selecting and ordering pronunciation representations for specific entries and for verification of grammatical analyses. The specifics of this process are described in Section 3.6.

3.4 Tools for Generating Lexical Information

A morphological analyser called twol, building on the free PCkimmo[3] system (SIL International, 1995; Antworth, 1990, 1995), was used for producing grammatical analyses and pronunciation representations for the words in the lexica and word lists mentioned above. The special thing about twol is that it produces phonological pronunciation representations from morph pronunciation representations (Magnuson et al., 1990). Twol has also been extended with proper name pronunciations (Gustafson, 1996).

[2] Punctuation marks are included in this word count. The punctuation marks amounted to 20,5 million tokens.

[3] The PCkimmo analyser is an implementation of Koskenniemi's two-level morphology (Koskenniemi, 1983; Karttunen, 1983).

The grapheme-to-phoneme rules of the KTH text-to-speech system (Carlson and Granström, 1976) have been utilised in the development of the lexical resources included in CentLex and have also been used in the further extension of CentLex to suggest pronunciation representations when no other pronunciation representation sources were available.

3.5 Analysis Format Conversion

The next section (Section 3.6) describes the initial integration of lexical data, forming the basis of the CentLex database. More data has been added at later stages, with methods specific for the data integrated with CentLex at the particular data additions. The twol system has, however, been used in most cases and thus the mapping, described in this section, between the twol grammatical analysis format and the suc format used in CentLex is of general interest for the lexicon development efforts described in this chapter.

Although the grammatical analysis formats in some cases are organised in different ways for twol and suc, respectively, it is in most cases possible to find a one-to-one conversion between the formats. However, it is not always possible to make an unambiguous mapping from twol to suc. Below, some mapping problems that arise are briefly discussed together with the solutions employed.

In general, there is a greater use of 'unspecified tags' in the twol grammatical analyses than is allowed for well-formed suc analyses. For example, twol analyses for nouns can contain 'unspecified tags' for the *gender*, *number* and *definiteness* morphological parameters. For well-formed suc noun analyses, there can be no 'unspecified tags'. The solution to this problem is to split the analyses, so that a twol analysis N NEU DEF,INDEF PL NOM with the 'unspecified' definiteness tag DEF,INDEF is converted into the two suc analyses NN NEU PLU IND NOM and NN NEU PLU DEF NOM[4]. A twol analysis containing more than one 'unspecified tag' not allowed in suc is split to create all possible well-formed suc analyses. Thus, a twol analysis containing two non-approved 'unspecified tag' instances creates four suc analyses and a twol analysis containing three non-approved 'unspecified tag' instances creates eight suc analyses.

twol has two separate Part of Speech (PoS) tags for adverbs, ADV and AD-A. The latter tag denotes adverbs modifying adjectives only. Since the suc format does not have a corresponding tag, the ADV and the AD-A tags are collapsed into the single suc adverb tag, AB. Suc has a special tag for interrogative/relative adverbs, HA. This tag has no corresponding tag in the twol tag system and to compensate for this, the word forms of twol-generated adverb analyses are looked up in the suc corpus. If the word has a *HA* analysis present in the suc corpus, the translation algorithm simply uses the suc tag.

In suc, the participle is treated as a Part of Speech, while it is treated as a subclass of verbs in twol. However, the participles have a special morphological

[4] The suc tag set is described in Table C.1 in Appendix C.

tag in TWOL and are thus easily identifiable. TWOL has a morphological parameter with tags differentiating active and so called *-s forms* of verbs (the -s form mostly signals a passive voice, but may also signal medial or reciprocal meaning or be an active form). This morphological parameter is not included for participles in the SUC system and is simply excluded for participles during TWOL-to-SUC conversion.

The TWOL analysis format does not include a special verb particle tag, while SUC does. To compensate for this, if there is a verb particle analysis available in SUC for a certain orthographic word form, a verb particle analysis is created for the word.

The SUC format includes many subgroups of pronouns having separate PoS tags, while the TWOL format includes only one pronoun PoS tag. The TWOL pronoun analyses can be divided into several categories depending on the morphological information. There is, however, not enough information in the TWOL tag strings to enable a division into as many categories as are present in the SUC format. Thus, SUC corpus lookup is employed to guide the conversion from TWOL to SUC format also in this case. The conversion of a TWOL pronoun analysis into the SUC format can mean changing the PoS from *pronoun* to *determiner* (the SUC PoS category *WH-determiner*, HD, is tagged as a pronoun in the TWOL format).

An unspecified subjective/objective form tag appears in SUC, while there is no correspondent in the TWOL input. The SUC corpus lookup can guide the conversion algorithm on when to merge TWOL analyses into one SUC analysis and when to change a subjective form tag or an objective form tag into an unspecified tag.

TWOL has no special PoS tag for possessive pronouns, while SUC does. However, TWOL has a morphological genitive tag for pronouns. All pronouns in the genitive in TWOL are translated to SUC possessives.

The third person possessive pronouns (in the SUC format) are tagged radically differently in TWOL and SUC, respectively. There is only one possible analysis for the group in SUC, while there are several possible analyses in the TWOL format. There are no special characteristics in the tag sequences to indicate a third person possessive pronoun analysis in the TWOL tags and none of the tags match in any case in TWOL and SUC. This seems to be due to differences in the way the words are interpreted. In these cases, since we are dealing with a small, closed set of words, the analyses of all words of the particular PoS classes are listed in the SUC manual (Ejerhed et al., 1992). The SUC analyses literally 'from the book' can thus be used in the translation (given the orthographic word form and the fact that the TWOL PoS is 'pronoun').

All interrogative/relative possessive pronouns (*vems*, *vilkens*, *vilkets*, *vilkas*, and *vars*, all translating into *'whose'*) are tagged HS DEF in SUC and in the same way as other pronouns in TWOL. All TWOL pronoun analyses for the closed set of words listed above are simply converted into the SUC HS DEF tag sequence.

Ordinal numbers are tagged as adjectives in the TWOL analyses, but have a separate RO POS analysis in SUC. However, instead of a comparison tag, they have an ordinal number indicator, <O>, and are thus easy to convert into SUC ordinal number analyses. All morphological parameters except case (the only morphological information occurring for ordinal numbers in the SUC format) are excluded during conversion.

3.6 Creating CentLex Entries

For the initial integration of data to form the basis of CentLex, the Onomastica names were all tagged as PM NOM (*proper name, nominative*). They also received a PM GEN (*proper name, genitive*) analysis, if ending with an *-s*, the Swedish genitive suffix. Each combination of orthographic name form and grammatical analysis was introduced into CentLex as an entry. The pronunciation representation attached to each entry was collected directly from the Onomastica lexicon, which included one pronunciation representation per orthographic word.

For the Dragon lexicon and other word lists with manually supplied or checked pronunciation representations, the entry creation procedure was not quite as simple. Here, the grammatical analysis of a word was not given and analyses were obtained by processing the words of the lexica through the TWOL system and subsequently converting the analyses into the SUC format as described in Section 3.5.

Since TWOL generates all analyses that are possible by combining morphological constituents from a lexicon, identical analyses can sometimes be created in several ways. This happens almost exclusively for compounds where different morphological constituents can be combined to form the same orthographical string, as shown in Example 3.1 (with analysis in SUC format). In some cases, a single stem word and a compound can have the same orthographic form, as shown in Example 3.2. A third possibility is that different numbers of constituents can be combined to form a single orthographic form, as shown in Example 3.3[5].

självägande PC PRS UTR/NEU SIN/PLU IND/DEF NOM
`själv|ägande` 'self-owning'
`själ|vägande` 'soul-weighing' ex. 3.1

centrum NN NEU SIN IND NOM
`centrum` 'centre'
`cent|rum` 'cent room' ex. 3.2

publikdragande PC PRS UTR/NEU SIN/PLU IND/DEF NOM
`publik|dragande` 'audience-attracting'
`pub|lik|dragande` 'pub corpse-dragging' ex. 3.3

In the TWOL output (with analyses converted into the SUC format), there can thus be several identical combinations of orthographic word form and analysis.

[5] The participle ending *-ande* has the same orthographic form as the word *ande 'spirit/genie'*, which opens up for noun compound analyses of many words ending with this string (including examples 3.1 and 3.3).

However, the information decided on to define a CentLex entry was the *orthographic word form* and the *grammatical analysis*, and it is not possible to distinguish between differently derived entries with these identifiers being identical. Thus, the decision was made to only create one CentLex entry per unique combination of orthographic word form and grammatical analysis (although this is not a strict requirement in CentLex).

The different ways of arriving at a specific analysis for a particular word may result in a set of different pronunciation representations, since these are created from the phonological forms of the morphological constituents. Thus, the decision to create only one CentLex entry per combination of word and analysis required a strategy for selecting and sorting the different twol-generated pronunciation representations that might occur for an entry.

The different sources of a particular analysis are, although mostly all theoretically possible, not equally probable. For example, in Example 3.2, the single stem analysis *'centre'*, is much more probable than the slightly anomalous (except, perhaps, in a very specific context) compound analysis *'cent room'*. The twol system actually produces a probability score for the combination of source and analysis given the word. This means that the same analysis from different sources can be scored differently in the same way as different analyses can be scored differently.

The score is dependent on e.g. the number of compound constituents, and analyses originating from few source constituents are weighted lower (the lower the score, the more probable the analysis) than analyses originating from more constituents. During the creation of CentLex entries, the score was exploited for ordering the pronunciation representations when merging identical combinations of word and analysis to form unique combinations, *tentative CentLex entries*. The scores, and the number of compound constituents directly, were used also to exclude the pronunciation representations from the least probable sources when there were multiple pronunciation representations for the same analysis.

There was only one hand-checked pronunciation representation for each orthographic word to be included in CentLex. This representation might not be the correct one for all analyses of the word, butthere was no information about which analysis/analyses of an orthographic word a hand-checked representation was intended to match. Twol produced both grammatical analyses and pronunciation representations specific to each analysis, but since the automatically generated pronunciation representations were known to be less accurate than the hand-checked representations, it was nevertheless of interest to exploit the high quality of the manually checked data. Thus, for the creation of the initial version of CentLex, a procedure had to be developed for optimally selecting and ordering the pronunciation representations for an entry.

When the orthographic words collected from different sources had been processed through the twol system, the grammatical analyses of the twol output had been converted into the suc format, and identical combinations of word and analysis had been merged into unique combinations (tentative CentLex entries) with pronunciation representations ordered according to the twol score for the original combination of word, analysis and source, the procedure for creating CentLex entries and attaching pronunciation representations to the entries could be employed. This procedure is described in Section 3.6.1 below.

3.6.1 Entry Creation Procedure

First, the question of whether an entry should be created from a particular TWOL-generated analysis was addressed[6]. Since not all analyses could be manually checked, the decision was taken to accept a tentative entry (combination of an orthographic word and a grammatical analysis) as a CentLex entry if the combination either occurred in the SUC corpus or in a large, automatically tagged newspaper corpus. For combinations occurring in neither corpus, decisions about whether the combination should be included in CentLex were made manually by the author (using some simple script tools to speed up the process).

Second, the questions about which pronunciation representations should be associated with an entry and in what order, were addressed. If there was a match between the hand-checked pronunciation representation for the orthographic word and one of the automatically generated pronunciation representations of the tentative CentLex entry, the hand-checked representation was used as the first representation in the CentLex entry. If there were more (non-matching) TWOL-generated pronunciation representations[7], these were also associated with the entry in the order in which they occurred in the tentative entry.

The fact that the same pronunciation representation occurred in both the hand-checked lexicon and the TWOL-generated lexicon made it highly probable that the hand-checked pronunciation representation was intended to match the particular analysis of the word and that this was the most commonly occurring pronunciation representation for the entry. However, other pronunciation representations may also be valid for the entry.

If there was *no match* between the hand-checked pronunciation representation and a TWOL-generated representation, auxiliary information had to be used to determine the pronunciation representation association. Three cases could be identified:

Case 1. *The current* TWOL *analysis was the only* TWOL *analysis for the word in question.* It is assumed that the hand-checked pronunciation representations are of higher quality than the automatically generated ones. In case 1, the hand-checked pronunciation representation was thus assumed to have a higher probability of being a correct representation for the entry than any of the TWOL-generated representations. Both the hand-checked and the TWOL-generated pronunciation

[6] Twol generates all analyses possible from combining morphological constituents in such a way that the resulting orthographic form matches a particular orthographic word and that general rules for word construction are satisfied. Twol analyses may thus be erroneous or anomalous, cf. examples 3.1 to 3.3.

[7] For most tentative entries, there was only one TWOL-generated pronunciation representation to match the hand-checked representation against.

representations were included for the entry. The hand-checked representation was chosen to be the first representation and the TWOL-generated representations were added in the order in which they occured in the tentative entry (if there were more than one TWOL-generated pronunciation representation). Example 3.4 illustrates case 1 (x and y denote pronunciation representations, $x \neq y$).

Current TWOL analysis pronunciation:	y	
Hand-checked pronunciation:	x	ex. 3.4

Case 2. *The current* TWOL *analysis was not the only one for the word and there was a pronunciation representation for another* TWOL*-generated analysis that matched the hand-checked representation.* In this case, the probability of the hand-checked representation being the correct one for the current analysis is low. Thus, the TWOL-generated pronunciation (or pronunciations, in the order in which they occurred in the tentative entry) was (or were) used for the entry, and the hand-checked representation was not included. Example 3.5 illustrates case 2 (x and y denote pronunciation representations, $x \neq y$).

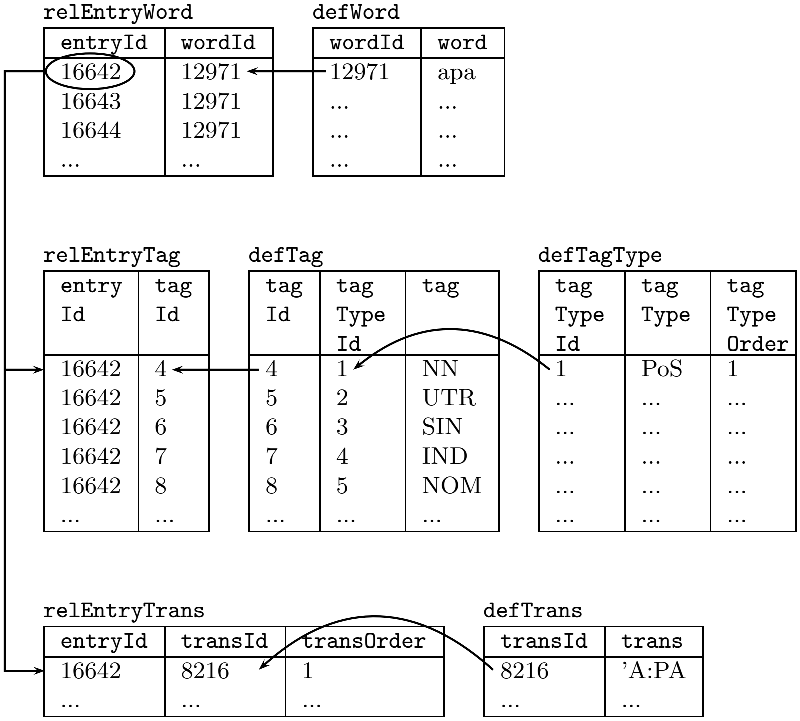
Current TWOL analysis pronunciation:	y	
Other TWOL analysis pronunciation:	x	
Hand-checked pronunciation:	x	ex. 3.5

Case 3. *The current* TWOL *analysis was not the only* TWOL *analysis for the word in question, but the hand-checked pronunciation representation did not match a pronunciation representation of any* TWOL *analysis.* The fact that the hand-checked representation did not match any TWOL-generated pronunciation representation was likely due to the fact that the representation had been edited by hand. It was, however, not known whether the hand-checked representation was meant to match the current analysis or not. Since there was no indication of whether the hand-checked representation was the correct one or not, the best bet was to place the hand-checked representation first (since this is likely the most common pronunciation for the word form) and to also include the TWOL-generated representation or representations. Example 3.6 illustrates case 3 (x and y denote pronunciation representations, $x \neq y$).

Current TWOL analysis pronunciation:	y	
Other TWOL analysis pronunciation:	y	
Hand-checked pronunciation:	x	ex. 3.6

3.7 Database Structure

The lexical information is stored in an SQL database with separate tables for different types of information. Each lexicon entry has a unique index and this index is used to relate the different types of information to the entry. In Figure 3.1, an



relEntryWord

entryId	wordId
16642	12971
16643	12971
16644	12971
...	...

defWord

wordId	word
12971	apa
...	...
...	...
...	...

relEntryTag

entry Id	tag Id
16642	4
16642	5
16642	6
16642	7
16642	8
...	...

defTag

tag Id	tag Type Id	tag
4	1	NN
5	2	UTR
6	3	SIN
7	4	IND
8	5	NOM
...	...	...

defTagType

tag Type Id	tag Type	tag Type Order
1	PoS	1
...	...	...
...	...	...
...	...	...
...	...	...
...	...	...

relEntryTrans

entryId	transId	transOrder
16642	8216	1
...	...	...

defTrans

transId	trans
8216	'A:PA
...	...

Figure 3.1: *Example showing the structure of the* CentLex *database.*

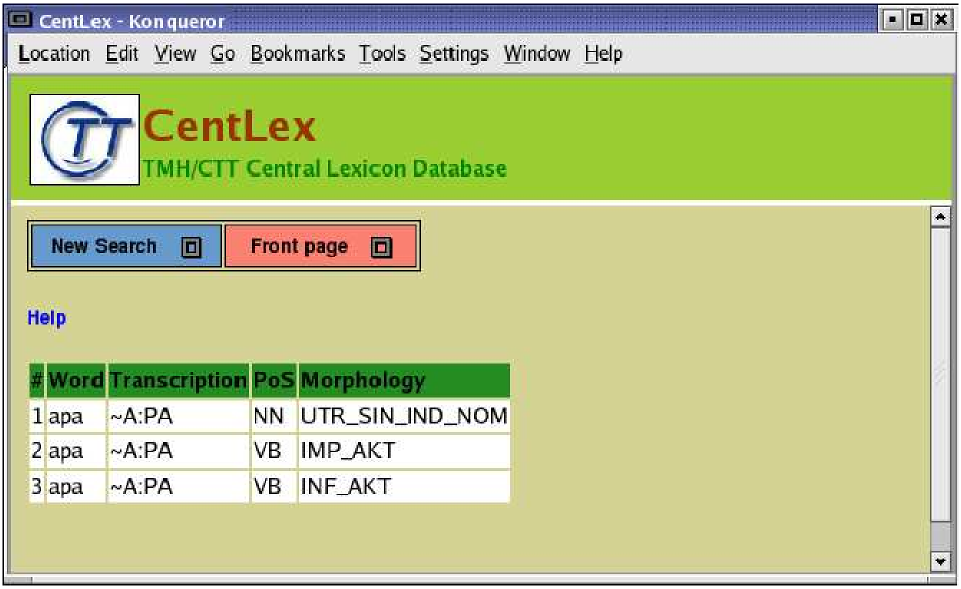
example of how information is related in the database is shown. The example does not include all information available in CentLex. The information of an entry is related to the entry as shown in the first row of tables in Figure 3.1. In the example, each unique orthographic word form receives a unique index in the `defWord` table, and the word is related to an entry index via this word index in table `relEntryWord`. Information about the probability of the entry given the orthographic word is related to the entry in a similar way in a table called `relEntryProb`.

In the second row of tables in Figure 3.1, it is shown how tags are related to an entry. Each tag is of a specific type, either it is a Part of Speech tag or it is a morphological tag of a certain type. The tag types are defined in the table `defTagType`. Each type has an order, the order in which it occurs in a suc tag string. It is possible to add tags of different types (ordered or unordered) to the table, although presently only Part of Speech tags and morphological tags occur.

Even though a pronunciation representation can be related to more than one entry, the order of pronunciation representations is unique for a specific entry, as shown in the last row of tables in Figure 3.1. Language tags and comments are stored in a table similar to the table `defTransOrder` table and related to the combination of entry index (`entryId`) and pronunciation representation index (`transId`) in a table similar to the `relTrans` table. However, while there can be only one `transOrder` per combination of `entryId` and `transId`, there can be an arbitrary number of language tags and comments related to a single pronunciation representation for a specific entry.

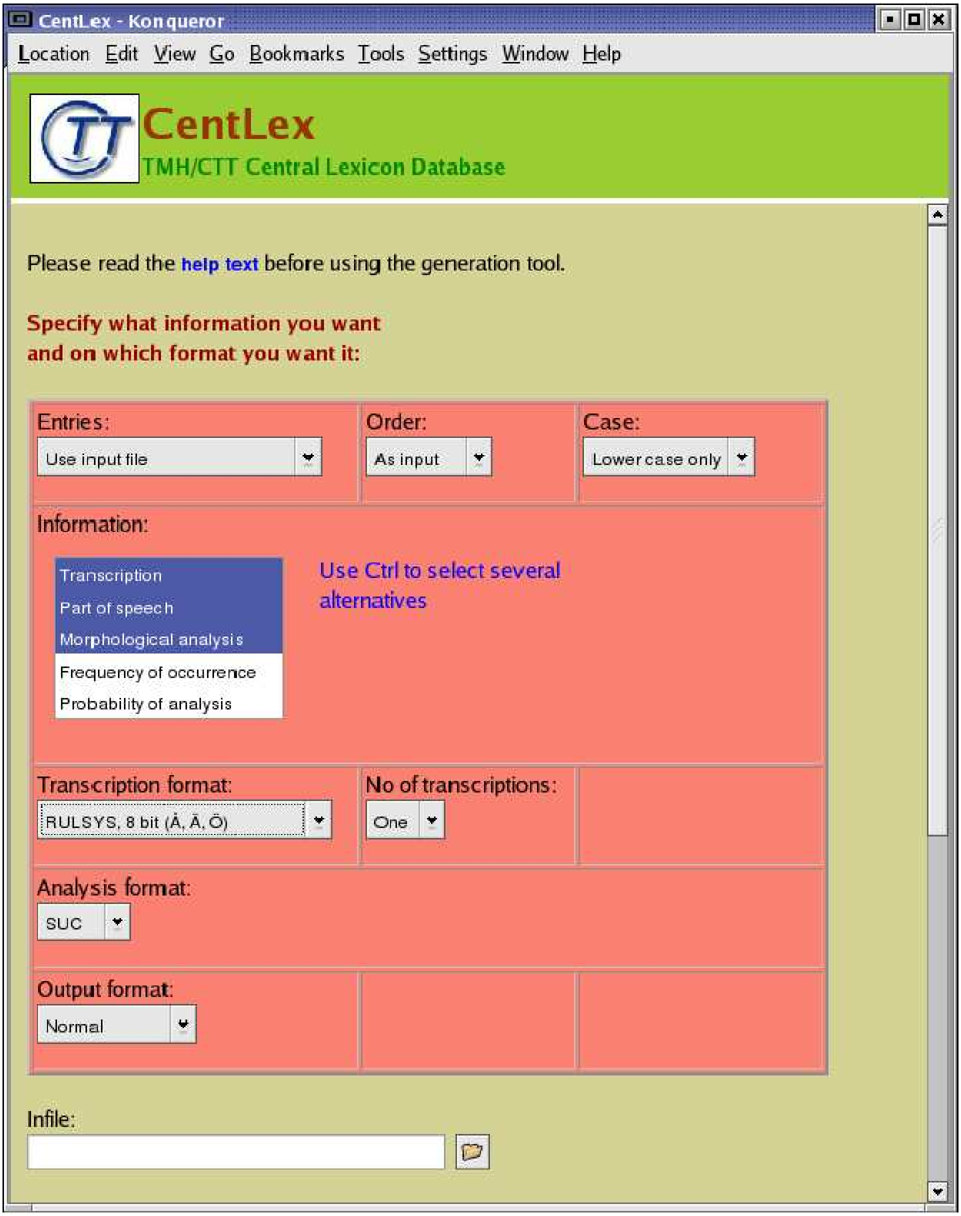
3.8 Availability and Continuous Development

An interface to the database on the internal web of the Department of Speech, Music and Hearing at KTH makes it possible to search the lexicon and to generate purpose-specific lexica with the set of information requested on several different output formats. Figure 3.2 shows the search interface and the result of searching the word *apa 'ape/monkey'*. Figure 3.3 shows the lexicon generation view, where lexicon information and format can be specified. The lexicon can be generated based on a specified input word list, or it can be based on some search criterion, e.g. all proper names. The database has been built to support project tags, so that it will be possible to check out a lexicon containing the entries or words associated with a specific project.


CentLex - Konqueror
Location Edit View Go Bookmarks Tools Settings Window Help
TT
CentLex
TMH/CTT Central Lexicon Database
New Search
Front page
Help

#	Word	Transcription	PoS	Morphology
1	apa	~A:PA	NN	UTR_SIN_IND_NOM
2	apa	~A:PA	VB	IMP_AKT
3	apa	~A:PA	VB	INF_AKT

Figure 3.2: *The* CentLex *Web interface—the results of searching the word* apa. *The highest ranking entry with this orthographic word is a noun analysis,* 'ape/monkey'. Apa *can also be used as a verb in the imperative or the infinitive in reflexive constructions such as* apa sig 'act like a monkey', *lit.* 'ape onself'. *Swedish has two word stress patterns, or* accents. *In the* CentLex *database, accent II primary stress is denoted by the ∼ symbol.*



CentLex

TMH/CTT Central Lexicon Database

Please read the help text before using the generation tool.

Specify what information you want
and on which format you want it:

Entries: Use input file

Order: As input

Case: Lower case only

Information:

- Transcription
- Part of speech
- Morphological analysis
- Frequency of occurrence
- Probability of analysis

Use Ctrl to select several alternatives

Transcription format: RULSYS, 8 bit (Å, Ä, Ö)

No of transcriptions: One

Analysis format: SUC

Output format: Normal

Infile:

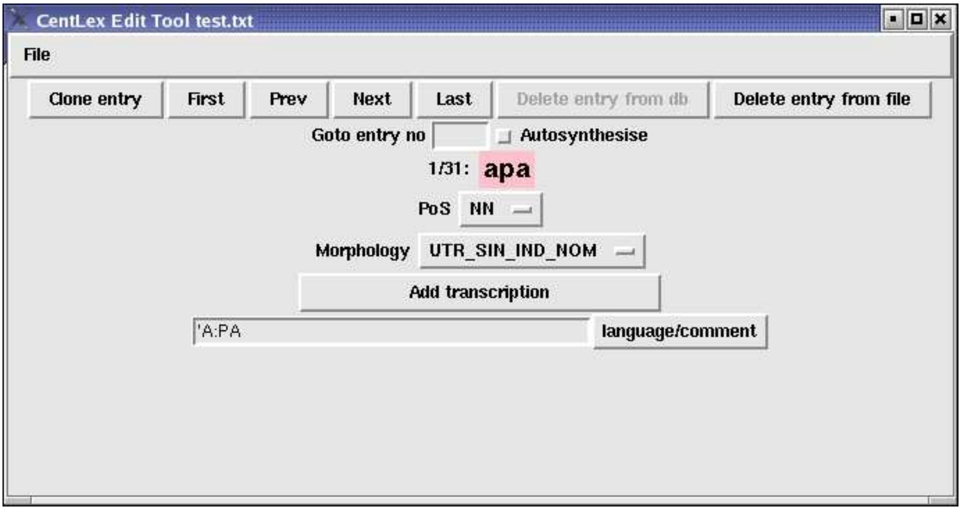
Figure 3.3: *The* CentLex *web interface—lexicon generation view. When the interface is used for creating a purpose-specific lexicon based on* CentLex, *the group of entries and the types of information to be included in the lexicon can be selected. There are also various formatting options.*

Selected users also have the possibility to edit the lexicon via the web interface, to stimulate continuous lexicon expansion and improvement of existing data. Changes are logged on a format enabling changes to be reversed. The web interface is not suited for large-scale changes of the database, so a stand-alone annotation/correction tool has been developed for lexicon development on a larger scale. This tool is described in Section 3.8.1.

The editing tools allow the lexicon to be incrementally built and the latest version of the lexicon is always available at a central location. As discussed above, some of the information first included in the database has been automatically generated and the information was initially integrated with automatic methods. The data thus has to be checked with respect to quality, which is done continuously. Subsequently added information is, however, mostly information which has been manually obtained or checked. Each lexicon entry is annotated with information about whether it has been manually checked/corrected, by whom and when, to separate information of different quality.

3.8.1 The CentLex Edit Tool

An annotation/correction tool has been developed for lexicon development on a larger scale. This tool stores information on a CentLex import format, so that it can be easily incorporated with the database. Figure 3.4 shows the CentLex Edit Tool interface.



CentLex Edit Tool test.txt

Goto entry no

□ Autosynthesise

1/31: apa

PoS NN

Morphology UTR_SIN_IND_NOM

'A:PA

Figure 3.4: *The* CentLex *Edit Tool interface.*

The CentLex Edit tool has been developed mainly by the author, building on a skeleton application written by Harald Berthelsen at *Södermalms Taltekno-*

logiservice (STTS). The tool has been continually improved in accordance with suggestions from the CTT CentLex project group (cf. Section 3.9) and users at *Acapela Group* and CTT.

The input to the tool can be a simple word list (*addition mode*) or a list of rows containing all information connected to an entry (*check mode*), or any subset of information minimally containing an orthographic word. If a list of words is to be included in CentLex, grammatical analyses, pronunciation representations, comments, etc. can be added manually. If any type of information has been automatically generated, the information can be manually checked and corrected and missing information added. A set of entries can be collected from CentLex and checked and corrected using the editing tool.

Using the KTH text-to-speech system (Carlson and Granström, 1976) integrated into the CTT toolbox[8], the CentLex Edit tool suggests pronunciation representations based on the orthographic word forms. The user can listen to synthesised versions of the pronunciation representations, choosing from a list of diphone voices or a formant synthesiser.

An arbitrary number of pronunciation representations can be added and an arbitrary number of language tags and comments can be attached to each representation. Pronunciation representations and language tags and comments can also be deleted. All pronunciation representations are parsed and the user is warned if the representation is not a valid CentLex pronunciation representation. It is not possible to advance to the next entry or to save the file until all pronunciation representations are on the valid format, as to avoid adding erroneous data to CentLex.

The user can choose from a list of all well-formed morphological tag strings for each valid Part of Speech. The list of valid morphological tag strings is dynamically updated to match the selected Part of Speech. Entries can be added or deleted from the current list of entries. If a set of entries has been checked out from the CentLex database for editing, the user can mark an entry for deletion from the database, if necessary. An auxiliary computer program reads the CentLex Edit tool output format (CentLex input format), converts it into a sequence of sql statements and updates the database. During this process, the program also performs a format integrity check.

3.9 Co-Operation

The work on CentLex has partly been conducted in co-operation with others. The initiative to building CentLex was taken by Rolf Carlson from the Department of Speech, Music and Hearing at KTH and the Centre for Speech Technology (CTT). The main part of the development has been conducted by the author, initially in co-operation with Jens Edlund at the department.

[8] Tcl/Snack tools for speech technology

Later, CentLex developed into a project within the CTT and a special project group was formed, with members from the Department of Speech, Music and Hearing at KTH and participating companies and organisations. The most active partners have been the Swedish Library of Talking Books and Braille (TPB), Acapela Group, Phoneticom, Södermalms Talteknologiservice (STTS) and Dolphin Audio Publishing/Labyrinten.

The project group has discussed the CentLex standards and e.g. made decisions on the phoneme set. Project participants have also been involved in augmenting the lexical information and in compiling mappings from the CentLex meta-format to different commonly used pronunciation representation formats.

Acapela Group has provided a set of the most frequent entries in a large newspaper corpus missing from CentLex. The entries were complete with manually checked grammatical analyses and pronunciation representations. Kjell Gustafson (Acapela Group/CTT) has been one of the most active project group participants and has been involved in e.g. developing the CentLex xenophone set and in the beta-testing of tools developed for lexicon work. Kjell Gustafson has also manually checked and corrected a set of words with automatically generated grammatical analyses and pronunciation representations prior to their integration with CentLex.

3.10 Applications

Thus far, the CentLex database has been used as a lexicon in an experimental speech synthesis system (used in various research-oriented applications at the department of Speech, Music and Hearing at KTH) and in a large vocabulary speech recognition system. CentLex has also been used for training grapheme-to-phoneme conversion rules for commercial speech synthesis and as a lexicon for commercial speech synthesis applications. It has further been used as a reference in the development of a system for production of talking books with synthetic speech for visually impaired and dyslectic university students. Finally, CentLex has been used for annotation in research projects aimed at context-sensitive prosody prediction and phone-level pronunciation prediction, the latter being the main focus of this thesis.

3.11 Summary

In this chapter, the development of CentLex, the central lexicon database for the Department of Speech, Music and Hearing at KTH and the Centre for Speech Technology (CTT), has been described. The lexicon was designed to meet the specific demands of the phone-level pronunciation modelling project which is the focus of this thesis, as well as general demands from speech technology research and application development. Tools for facilitating access and continuous, co-operative editing of the lexicon database have been developed.

The next chapter will present the result of an evaluation of the *coverage* of CentLex over a collection of texts of different types and of the *accuracy* of the pronunciation representations included in the lexicon.

Chapter 4

Pronunciation Lexicon Evaluation

This chapter gives a more detailed account of the contents of CentLex and reports the results of evaluations of the coverage of CentLex over a variety of text types and of the quality of the pronunciation representations included in CentLex.

4.1 Lexicon Contents

To obtain an unchanging version of CentLex to be used during the evaluation of the coverage and accuracy described in sections 4.2 to 4.5 below, the database was dumped to a text file. This 'frozen' version of CentLex contained 410,326 entries and 332,626 unique word forms. Out of these word forms, 56,130 occurred in more than one entry (i.e., had more than one grammatical analysis). A set of 12,902 entries had more than one pronunciation representation attached to it. Table 4.1 shows the number of entries having more than one pronunciation representation, grouped by number of representations.

Table 4.1: *The number of* CentLex *entries having more than one pronunciation representation, grouped by number of pronunciation representations.*

Number of representations	2	3	4	5	6	7	8	9	$\sum$
Number of entries	11,760	953	141	25	17	2	3	1	12,902

Not all entries had a pronunciation representation. More precisely, there were 23,099 entries for which there were no pronunciation representations. These entries were mostly irregular compounds, abbreviations and foreign names collected from a tagged text corpus. The entries lacking pronunciation representations were not analysable by twol, i.e., twol supplied neither an analysis nor a pronunciation representation for the word forms of the entries.

The KTH text-to-speech system (Carlson and Granström, 1976) could be used for generating pronunciation representations for the words, but since the words are mostly irregular forms not following normal Swedish phonotactic rules, it is hypothesised that automatically generated pronunciation representations would be of low standard. Thus, before such pronunciation representations are included in CentLex, they need to be manually checked and corrected. When tese words are manually checked, some of them will probably turn out not to be suitable for the lexicon and may thus be excluded.

Table 4.2 shows the distribution of entries over different Parts of Speech. As can be seen, the proper names constitute the largest Part of Speech group. This is on the account of the Onomastica lexicon, included in CentLex, containing only personal names and place names (both tagged as *proper names*, PM).

Table 4.2: *The distribution of* CentLex *entries over different Parts of Speech.*

Part of Speech	Part of Speech tag	Number of entries
Proper name	PM	196,094
Noun	NN	128,955
Verb	VB	28,573
Adjective	JJ	27,481
Participle	PC	23,537
Adverb	AB	2,797
Foreign word	UO	1,609
Interjection	IN	245
Pronoun	PN	235
Cardinal number	RG	215
Preposition	PP	166
Verb particle	PL	88
Conjunction	KN	67
Ordinal number	RO	60
Possessive pronoun	PS	54
Determiner	DT	53
WH-adverb	HA	39
Subjunction	SN	30
WH-pronoun	HP	16
Possessive WH-pronoun	HS	5
WH-determiner	HD	4
Infinitival marker	IE	3
$\sum$		410,326

Since ‘foreign word’ is not an actual Part of Speech, it has been decided to avoid this class in CentLex. However, the foreign word PoS tag (UO) occurs in the data used to build CentLex and there are currently UO entries in CentLex.

In a lexicon, it is better to specify how a specific ‘foreign word’ is used in Swedish (as a loan word) than to use a ‘foreign word’ PoS tag. However, since a tagger trained on suc is used in many applications, it may be useful to keep also the UO tags in the lexicon (i.e., to create multiple entries for loan words). If a specific word cannot be generally considered a loan word in Swedish and it is not a proper name, it should be excluded from the lexicon.

CentLex currently includes cardinal numbers and ordinal numbers (in text form, not in numeral form). Since numbers is not a closed set and can be easily handled by rules, it would perhaps be better to let rules supply the pronunciation representations for these, once they have been recognised as numbers (e.g. by a tagger). However, not all applications involve a tagger and it may be useful to have pronunciation representations for the most commonly occurring cardinal numbers (RG) and ordinal numbers (RO) stored in the lexicon. In any case, it is not harmful to have the forms in the lexicon, so there is no reason to remove the RG and RO entries that are already in CentLex. There are some archaic cardinal number forms in CentLex and these may also be useful to have stored in the lexicon.

There is only one infinitival marker in standard Swedish. Nevertheless, there are three infinitival markers in CentLex—there is the standard form *att* and a form which can be used when representing a colloquial speech form in text, *å*. There is also a form *at* in CentLex. This is probably a misspelled *att* which has still received the IE tag in some of the text resources used to build CentLex. It is highly likely that *at* is a common misspelling of *att* and it could be argued that this is a reason to keep the misspelled form in the lexicon (and giving it a special 'misspelled' tag). Currently, there is no standard for how to handle common misspelled forms of words in CentLex.

4.2 Coverage and Accuracy

If the intention is to use a lexicon for providing the pronunciation representations for the words encountered in a text, e.g. when the text is converted into speech using speech synthesis, it is interesting to know the approximate share of words in the text for which a lexicon can provide a pronunciation representation. It is also interesting to know the share of correct pronunciation representations in the set of provided representations. Since text-to-speech conversion is one of the intended uses of CentLex, the *coverage* and *accuracy* of CentLex has been investigated.

In this context, *coverage* is thus the share of words (or combinations of a word and a grammatical analysis) covered by the lexicon, i.e., the share of units where the lexicon can provide a pronunciation representation. The *accuracy* is the share of accurate pronunciation representations from the total of the representations supplied by the lexicon.

The coverage of CentLex is calculated across a set of text corpora to give a hint of what the coverage will be for different text types. A randomly selected subset of the corpus has been manually checked by the author and by another experienced linguist/phonetician (Kjell Gustafson, CTT/Acapela Group) to give an estimate of the accuracy. The experts corrected the erroneous entries encountered, which means that the evaluation could also be used to improve the lexicon. The evaluation statistics reported are, of course, based on the state of the lexicon prior to this improvement.

The pre and post improvement versions of the randomly chosen sets of lexicon entries also allow more detailed studies of the lexicon accuracy. For example, the types of discrepancies between the two versions and the presence of systematic errors (which can be corrected automatically) have been investigated.

4.3 Calculating Coverage

The coverage of CentLex over a set of text corpora has been assessed. Table 4.3 gives an overview of the corpora used in the coverage evaluation. The size of each corpus is described by three measures: *tokens*—the word count, i.e., the number of word tokens included in the corpus (punctuation excluded), *entry forms*—the number of CentLex entry forms, i.e., the number of different combinations of full form word and grammatical description (Part of Speech and morphology) included in the corpus and *type*—the number of word types (different full-form words) included in the corpus.

Table 4.3: *Text corpora used in the* CentLex *coverage evaluation.*

Corpus	Origin	Text type	Tokens	Entry forms	Types
DN	Newspapers	News text	81,928,359	1,632,273	1,407,071
TPB	Books	Student literature	9,343,445	325,444	289,481
GOV	Internet	Public information	3,943,147	46,961	41,565
RD	Internet	Public information	3,902,653	91,529	83,732
EU	Internet	Public information	51,367	8,987	8,425
JK	Internet	Public information	787,012	33,704	30,676
DOM	Internet	Public information	320,636	20,359	18,702
FMN	Internet	Public information	383,809	17,874	15,958
ALL	Miscellaneous	Miscellaneous	100,660,428	1,793,810	1,543,128

Dagens nyheter (DN) is a daily morning newspaper and *Expressen* is a daily tabloid. The DN corpus consists of printed text from *DN* 1992–1995 and printed text from *Expressen* 1990–1995.

The Swedish Library of Talking Books and Braille (TPB) is a State library in charge of e.g. supplying talking book student literature for dyslectic and visually impaired students. The TPB corpus is a collection of scrambled sentences from university student literature dealing with a variety of subjects. A randomly selected percentage of sentences from each book converted into a talking book by TPB has been added to the TPB corpus. The sentences are kept in their original form, but stored in random order in the corpus.

The GOV corpus is the publicly available Swedish text from the website of the Swedish government, *www.regeringen.se*. The RD corpus is the publicly available Swedish text from the website of the Swedish parliament, *www.riksdagen.se*. The EU corpus is the publicly available Swedish text from the website of the European parliament *www.europarl.eu.int* (changed to *www.europarl.europa.eu* after the site was downloaded).

Justitiekanslern (JK) 'Chancellor of Justice' is the Swedish Attorney General and the JK corpus is the publicly available Swedish text from the website of the Chancellor of Justice, *www.jk.se.* The DOM corpus is the publicly available Swedish text from the website of the Swedish court authority and Swedish courts, *www.dom.se.*

Fastighetsmäklarnämnden (FMN) 'the Real Estate Broker Agency' is a government authority for questions regarding registration and supervision of real estate brokers. The FMN corpus is the publicly available Swedish text from the *FMN* website, *www.fastighetsmaklarnamnden.se.* The texts originating from from websites were downloaded in June 2006.

4.3.1 Creating the Text Corpora

The sentences from student literature included in the TPB corpus and the websites of the European and Swedish parliaments, the Swedish government and different government authorities are all examples of texts available as synthesised speech.

The TPB corpus was provided by the Swedish Library of Talking Books and Braille (TPB) and the books from which the sentences in the TPB corpus originate have been converted into talking books with synthetic speech by TPB. All texts downloaded from the Internet were downloaded (and cleaned from hypertext markup, java script code etc.) by PHONETICOM, a company specialising in talking websites. PHONETICOM provides talking website solutions for the downloaded sites. As partners in the *Centre for Speech Technology (CTT)*, PHONETICOM and TPB have access to CENTLEX and use the lexicon in their work.

All texts were tokenised and tagged with the TnT part-of-speech and morphological tagger (Brants, 2000) trained by Megyesi (2001, 2002a) on the SUC corpus (Ejerhed et al., 1992). The DN corpus had already been tokenised and tagged in this manner for a different project (Rydin, 2002) and the tagged version of this corpus was re-used in the current evaluation.

The tagged texts were cleaned from non-Swedish text on a sentence basis (excluding quotes in foreign languages etc.) by excluding the entire sentence if more than 30% of the words were tagged as foreign words by the tagger[1] or if more than 30% of the words occurred in one of three respective lists of common English, French and German words. The English list contained 9,379 common English words, the French list contained 1,178 common French words and the German list contained 3,776 common German words.

The HTML documents downloaded from a website often included several versions of the same page, containing the same or approximately the same information. No attempts have been made to control the uniqueness of the pages downloaded. Web pages from a particular site may contain some part that is common to all or most pages, e.g. an address. Frequently occurring strings particular to a specific website

[1] As mentioned, *foreign word* is included as a special 'Part of Speech' tag in the SUC corpus and thus for the tagger trained on SUC.

have been excluded when manually detected. However, no systematic search for such strings has been conducted.

4.4 Coverage Results

As mentioned, the coverage is calculated as the share of *words* and as the share of *CentLex entry forms*, respectively, from a corpus for which CentLex can provide at least one pronunciation representation. Whether the pronunciation representation is correct or not is not assessed here. The quality assessment reported in Section 4.5 will address the accuracy issue.

The coverage is calculated both over words and over CentLex entry forms (the combination of full form word and grammatical analysis). Since the CentLex entry forms are derived from automatically generated tags, it should be noted that the tagger used has been reported to produce 6.45% errors when tagging with the full set of Part of Speech and morphological tags (Megyesi, 2001, 2002a), as in the current case. Further, the 6.45% error rate was on text more similar to the text the tagger was trained on than the text in the current text corpora (a tenfold cross validation experiment). Thus, the CentLex entry form type and token estimations should be seen as rough estimates of coverage rather than actual coverage over the corpora.

4.4.1 Coverage per Corpus

Table 4.4 presents the coverage of CentLex over the different text corpora and for the combined text from all corpora (ALL). When coverage is calculated over all text corpora, values are recalculated treating the collections of texts as a single corpus. This may affect the hapax legomenon status (cf. below) of certain word units and it may also affect the assignment of words to different frequency groups (cf. below). Since the DN corpus includes more than 80% of the tokens in the combined text corpus, coverage values for the combined corpus are close to those of the DN corpus.

It should be noted that the DN corpus constituted a part of the news text material included in the set of texts used to construct the Dragon lexicon, now part of CentLex. This may mean that the coverage of CentLex over the DN corpus is not directly transferable to the news text genre, but is most likely an over-estimation for the genre.

A *hapax legomenon* is a word or a CentLex entry form occurring only once in the corpus. Swedish is a compounding language, with many lexicalised compound words. However, compounding can be seen as part of the grammar and previously unseen compounds frequently occur in both news text and literature. Many of the hapax legomena found are temporary 'grammatical' compound constructions. Other sources for hapax legomena are misspellings, temporary abbreviations etc.

If we assume that the bulk of hapax legomena are due to mistakes and grammatical constructions, the probability of these units occurring ever again is very low, and the usefulness of having these units stored in a lexicon is questionable. The set of possible entries for a lexicon is not finite and there has to be some threshold for probability of occurrence for including a unit into a lexicon for the lexicon size to be manageable (from the viewpoint of being able to control the quality of the lexicon entries).

In the CentLex case, this means that most hapax legomenon units found in the evaluation text corpora will be of the sort that is not desired in the lexicon. To give an idea about the impact of hapax legomenon units on the coverage values, a coverage statistic calculated with hapax legomena excluded from the corpora is presented in Table 4.4 within brackets after the value based on all tokens. The coverage statistics for CentLex entry form types and word types increase significantly when the hapax legomena are excluded, while the coverage values for tokens are only marginally increased, since each hapax legomenon type by definition only occurs once.

Table 4.4: *Coverage (per cent) of* CentLex *over a set of text corpora, including coverage statistics based on all* CentLex *entry types and words, respectively, and based on the corpora with hapax legomena excluded (-hl).*

	CentLex entry form		Word	
Corpus	Types (-hl)	Tokens (-hl)	Types (-hl)	Tokens (-hl)
DN	11.56 (24.99)	94.05 (95.13)	11.91 (25.21)	95.06 (95.98)
TPB	34.72 (59.27)	94.98 (96.54)	36.54 (60.83)	95.79 (97.12)
GOV	56.82 (63.97)	91.18 (91.36)	61.60 (68.17)	93.80 (93.95)
RD	54.09 (66.98)	94.96 (95.53)	55.37 (68.25)	96.22 (96.74)
EU	78.22 (88.85)	93.35 (95.89)	79.00 (89.40)	94.09 (96.43)
JK	61.84 (74.22)	94.55 (95.47)	62.95 (74.57)	95.36 (96.16)
DOM	62.40 (73.87)	87.90 (89.05)	63.57 (75.11)	91.83 (92.94)
FMN	65.50 (76.75)	91.72 (92.57)	67.77 (78.14)	92.58 (93.28)
ALL	10.75 (23.03)	94.04 (95.00)	11.10 (23.30)	95.11 (95.92)

The share of CentLex entry form types and word types not in the lexicon grows as a function of corpus size. Since the bulk of word tokens in a text will always be a relatively closed set of common words that will not change as a corpus grows, the share of infrequent types will grow with increasing corpus size. If coverage is measured for a lexicon including only entries which are commonly accepted lexical items and no temporary compounds and text-specific terminology, the coverage of types will thus always decrease with the size of the corpus over which the coverage is calculated. The token coverage is thus the interesting property of the lexicon, since it is less affected by the size of the corpus over which the coverage is calculated, while the type coverage values should be regarded as references.

Across all corpora, 94.0% of the CentLex entry forms and 95.1% of the words in the texts can receive a pronunciation representation if CentLex is used as the only source for pronunciation representations. As previously discussed, this is only interesting in combination with an assessment of the pronunciation representation accuracy and such an assessment is presented in Section 4.5.

4.4.2 Coverage per Corpus and Frequency Group

Although the type coverage values are not very interesting when measured across entire text corpora, the type coverage is interesting when less frequent CentLex entry forms or words are disregarded and we are left with only mid to high frequency units—i.e., the types of units that are interesting to store in the lexicon. To be able to study the coverage of CentLex entry forms and words in different frequency groups, a procedure for dividing the corpus into a *high frequency* group, a *mid frequency* group and a *low frequency* group was developed.

Under this procedure, the CentLex entry forms and words in a corpus are assigned to frequency groups by ordering the entry forms and words, respectively, after descending frequency. The high frequency group is then formed by assigning entry forms/words to the group, starting with the most frequent entry form/word, while the share of corpus tokens covered by the entry forms/words in the group is less than 50%. When 50% of the corpus token coverage is exceeded, entry forms/words are instead assigned to the mid frequency group. Units are assigned to the mid frequency group while the corpus token coverage is less than 90%. When 90% is exceeded, the remaining entry forms/words are assigned to the low frequency group.

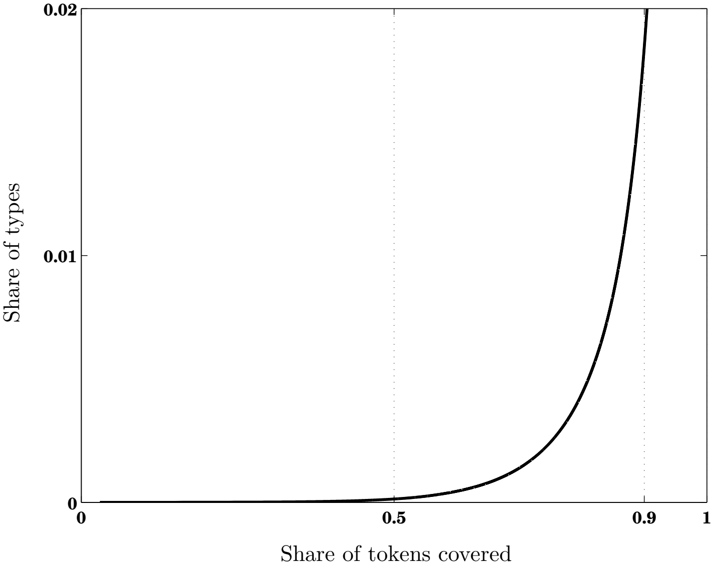


Figure 4.1: *Share of* CentLex *entry form types in the combined text corpus (sorted by decreasing frequency) plotted against share of corpus tokens covered by the types. The frequency group thresholds are marked by dotted lines.*

Figure 4.1 illustrates this procedure. The y axis shows the share of (frequency sorted) CentLex entry form types of the combined text corpus and the x axis shows the share of the corpus tokens covered by the types. The separate corpora have distributions similar to the distribution of the combined corpus.

Table 4.5: *The number of types in the three frequency groups over the different corpora.*

	High frequency types		Mid frequency types		Low frequency types	
Corpus	Entry form	Word	Entry form	Word	Entry form	Word
DN	250	205	33,987	28,545	1,598,036	1,378,321
TPB	159	131	19,842	16,650	305,443	272,700
GOV	197	170	6,470	5,407	40,294	35,988
RD	111	95	6,029	5,145	85,389	78,492
EU	150	130	3,762	3,428	5,075	4,867
JK	123	107	4,877	4,235	28,704	26,334
DOM	120	106	4,088	3,517	16,151	15,079
FMN	94	82	2,677	2,304	15,103	13,572
ALL	245	201	32,798	27,436	1,760,767	1,515,491

Since high frequency types per definition cover more tokens than low frequency words, there will be few entry form/word types in the high frequency group, while there will be many types in the low frequency group. For the combined corpus, there are 245 CentLex entry form types in the high frequency group, covering about 50% of the entry form tokens in the corpus. The number of types in each frequency group is shown in Table 4.5 and the exact shares of tokens in the frequency groups is shown in Table D.1 in Appendix D.

Table 4.6: *Type coverage over text corpora (per cent) in three frequency groups.*

	High frequency types		Mid frequency types		Low frequency types	
Corpus	Entry form	Word	Entry form	Word	Entry form	Word
DN	99.20	100.00	94.19	95.76	9.79	10.17
TPB	100.00	100.00	95.24	96.17	30.75	32.87
GOV	94.42	98.24	87.05	90.31	51.78	57.11
RD	97.30	100.00	94.41	96.02	51.19	52.65
EU	98.00	98.46	88.52	89.26	70.01	71.26
JK	97.56	99.07	91.96	93.34	56.57	57.92
DOM	88.33	94.34	85.71	88.43	56.30	57.56
FMN	93.62	93.90	89.20	90.54	61.13	63.75
ALL	99.18	100.00	94.11	95.78	9.18	9.55

Table 4.6 shows the type coverage over the three respective frequency groups and Table 4.7 shows the token coverage over the frequency groups. As discussed above, in the high and mid frequency groups, both type and token coverage values are interesting, since the types found in these frequency groups are mostly of the sort that are wanted in a lexicon. However, in the low frequency group, the type coverage values depend on the size of the evaluation corpus and do not reflect properties of the lexicon that are interesting in the current context.

As can be seen from tables 4.6 and 4.7, 100% of the word types and word tokens in the high frequency group are covered by CentLex for the combined corpus. For CentLex entry forms, the corresponding coverage values are 99.2% and 99.8%, respectively.

Table 4.7: *Token coverage over text corpora (per cent) in three frequency groups.*

	High frequency tokens		Mid frequency tokens		Low frequency tokens	
Corpus	Entry form	Word	Entry form	Word	Entry form	Word
DN	99.80	100.00	97.38	98.32	52.02	57.32
TPB	100.00	100.00	97.63	98.57	59.32	63.68
GOV	97.14	99.05	90.51	93.38	64.12	69.25
RD	99.32	100.00	94.80	96.58	73.82	75.94
EU	99.18	99.35	91.85	93.03	70.22	72.07
JK	99.30	99.74	95.32	96.32	67.80	69.62
DOM	91.52	96.13	89.64	93.14	62.85	65.10
FMN	95.83	96.06	91.82	92.96	70.83	73.68
ALL	99.80	100.00	97.17	98.27	52.66	57.98

The DN and TPB corpora are larger and contain text on various subjects, while the Internet corpora are smaller and highly specialised. These facts are reflected in the coverage values for the high and mid frequency groups of these corpora. There is a larger share of the high and mid frequency entry forms and words that are not covered by CentLex in the corpora based on texts downloaded from the Internet[2].

One cause of this is that there is much site-specific terminology, which is perhaps better handled by site specific lexica than by a general pronunciation lexicon such as CentLex, although some of the terms not included in CentLex are of general interest and thus could be included in the lexicon.

A cause of the low entry form coverage in relation to word coverage in the mid frequency groups of the Internet corpora is that the Internet corpora include texts with special hypertext properties, such as hypertext lists of links. The tagger has not been trained on this type of text and often misclassifies many times highly frequent units, such as hypertext links included on many pages, because of their anomalous context. If the tag is incorrect, the entry form will, of course, not be found in CentLex, although the correctly tagged entry exists in the lexicon.

4.5 Accuracy

As mentioned, the coverage of a lexicon is only interesting in association with an assessment of the quality of the information in the lexicon. This section reports an evaluation of the accuracy of the pronunciation representations in CentLex.

[2] The coverage for the combined corpus can still be 100%, since the high frequency group of the combined corpus is not the intersection of the high frequency groups of the individual corpora, but created using the same algorithm used for the separate corpora, but on the entire set of texts.

Of the 245 entries in the high frequency group from the combined text corpus (cf. Table 4.5), 243 (99.18%) occurred in CentLex. The pronunciation representations for these entries were manually checked and corrected (when necessary) by two trained phoneticians. Samples matching the high frequency entries in number (243 entries) from the mid frequency group and from the high frequency group, respectively, were checked and corrected in the same manner.

The samples were collected using random sampling without replacement, to ensure that the same entry could not be chosen more than once. For the high frequency group, all entries have thus been checked, while for the mid and low frequency groups, only small samples have been checked. The statistics for the mid and low frequency groups are thus accuracy estimates, while the statistics for the high frequency groups constitute the actual accuracy for the group. However, as will be discussed in Section 4.5.4, noise associated with the use of an automatic tagger affected the results.

An entry can have an arbitrary number of pronunciation representations attached to it and the representations are ordered according to their estimated frequency of use. In the analysis presented here, only the highest ranking pronunciation representation for each entry is considered. This is the pronunciation that would in most cases be used in a speech synthesis application, if there is no other information than orthographic word form and an automatically obtained Part of Speech tag to base the selection of pronunciation representations on.

However, there may not be a Part of Speech tag available for the words of a text in a particular application using speech synthesis. If no grammatical analysis is available, the best way to minimise the number of mispronunciations is to always select the highest ranking pronunciation representation of the highest ranking entry[3]. To investigate the accuracy in the absence of PoS tags, the highest ranking pronunciation representation of the highest ranking entry was also manually checked and corrected.

In the high frequency group, the entries under investigation were always the highest ranking entries given the word. For the 243 randomly selected entries from the mid frequency group, there were 27 entries that were *not* the highest ranking entry given the word and among the low frequency entries, there were 40 entries *not* the highest ranking.

4.5.1 Discrepancy Types

During manual checking and correction of the 243 high frequency words and the randomly selected words from the mid and low frequency groups, it was obvious that a simple correct/incorrect dichotomy was hard to establish. Instead, discrepancies from the manually supplied pronunciation representations were divided into three types: *formal*, *different* and *erroneous.*

[3] Since the entries have information about their probability given the orthographic word, entries sharing an orthographic word are mutually ranked.

The representations assigned to the *formal* discrepancy group were of the kind that would not affect the performance of an ASR system having the representation in its lexicon, and that would give no or only minor, sub-phonemic, effects in a speech synthesis setting. The most commonly occurring discrepancy of this type was an extra compound boundary.

Example 4.1 shows an entry from the low frequency group assigned to the formal discrepancy category because of an extra compound boundary. The pronunciation representation in its original form and in its corrected form, both in the CentLex internal format (modified STA, cf. Table A.1 in Appendix A), are presented in the example. In the pronunciation representations, a `hy` denotes a compound boundary.

The first compound constituent *bistånd ‘development assistance’* is etymologically a compound consisting of a prefix and a noun and from a diachronic perspective, it could be argued that *bistånd* is a compound. However, in modern usage, the word would not be considered a compound. The inserted compound boundary does not affect the pronunciation of the word.

Example 4.2 shows an entry from the low frequency group assigned to the formal discrepancy category because of an misplaced compound boundary. In this example, the [t] (`T`) should be aspirated when the compound boundary is correctly placed, but unaspirated for the pronunciation representation with the misplaced boundary and this may have an impact in a speech synthesis setting.

biståndsinsatser ‘development assistance contributions’
NN UTR PLU IND NOM
Formal discrepancy: `B"I:hySTÅNDShyINhyS‘ATSEOR`
Corrected: `B"I:STÅNDShyINhyS‘ATSEOR` ex. 4.1

minnestal ‘commemorative speech’ NN NEU SIN IND NOM
Formal discrepancy: `M"INEOhyST‘A:L`
Corrected: `M"INEOShyT‘A:L` ex. 4.2

The representations assigned to the *different* discrepancy group were also not considered erroneous. The most common type of discrepancy of the *different* category was that words with the common endings *-iskt* and *-igt* were represented by a reduced form, and not by the canonical form. This is due to a conscious decision made to adapt the pronunciation in lexica included in CentLex to running speech. However, for CentLex, it was decided to use the canonical form (defined as the most detailed pronunciation possible in practice, the ‘citation form’) as the first representation in the lexicon.

In a speech synthesis setting or an ASR setting, the reduced form would probably be appropriate more often than the canonical form, and a reduced form can be included in the list of pronunciation representations in cases where simple rules cannot supply these forms. Since reduced form pronunciation representations are (or will be) tagged with a special *reduced* tag, it is easy to select the reduced form instead of the highest ranking (canonical) form. Depending on the application, reduced forms can be chosen always when available, for function words only, or depending on specified context criteria.

In short, the *different* discrepancy is not an error from a practical perspective, although the pronunciation representation does not strictly follow the agreed upon standard for CentLex. Example 4.3 shows a discrepancy of this kind collected from the randomly selected entries from the mid frequency group.

samiskt 'sami' JJ POS NEU SIN IND NOM
Different discrepancy: `S'A:MIST`
Corrected: `S'A:MISKT` ex. 4.3

Another type of discrepancy that is included in the *different* category is when the pronunciation representation is an acceptable one, but another pronunciation is judged to be more common. In some cases, the more common pronunciation was actually present in the list of pronunciation representations for the entry, and in these cases the error was an error in the order of the representations (as assessed by the phoneticians checking the entries).

Pronunciation representations placed in the *erroneous* discrepancy group are representations that are incorrect and would give an erroneous pronunciation if used in a speech synthesis setting. The particular error or errors of a representation can be more or less grave and of different types. The error types encountered were *erroneous stress position*, *erroneous word accent* and *erroneous phoneme string.* Example 4.4 shows a phoneme string error from the low frequency group (the vowel length is incorrect). This is a word of foreign origin, but there were also words with Swedish origin in the *erroneous* group, as shown in Example 4.5, collected from the mid frequency group. In this case, there is a word accent error (`'` denotes accent I primary stress and and `"` accent II primary stress).

gentilt 'stylish' JJ POS NEU SIN IND NOM
Erroneous discrepancy: `SJANGT'ILT`
Corrected: `SJANGT'I:LT` ex. 4.4

duger 'is good enough' VB PRS AKT
Erroneous discrepancy: `D"U:GEOR`
Corrected: `D'U:GEOR` ex. 4.5

4.5.2 Discrepancy Statistics

The results from the investigation of pronunciation representation accuracy are summarised in Table 4.8. In the set of randomly selected low frequency entries, there were altogether 31 discrepancies. That is, there were 31 entries for which the highest ranking pronunciation representation differed from the manually corrected representation in some way. Out of these discrepancies, four were *formal* discrepancies, eleven were *different* discrepancies and 16 were *erroneous* discrepancies. Out of the *erroneous* discrepancies, eleven were names of foreign origin, three were names of Swedish origin (one first name, one family name and one nickname) and the remaining two discrepancies were a noun and an adjective, both well-established loan words of French origin (the adjective was *gentilt*, shown in Example 4.4).

Table 4.8: *Discrepancies found in the three groups of manually checked entries, presented as the number of entries with the highest ranking pronunciation representation differing from the manually corrected representation and as the share (per cent) of entries with a discrepancy.*

	High frequency		Mid frequency		Low frequency	
	Discrepancies	Share	Discrepancies	Share	Discrepancies	Share
Formal	2	0.82	1	0.41	4	1.65
Different	11	4.53	10	4.11	11	4.53
Erroneous	0	0.00	4	1.65	16	6.58
$\sum$	13	5.35	15	6.17	31	12.76

In the group of checked mid frequency entries, there were 15 discrepancies, of which one was a *formal* discrepancy, ten were *different* discrepancies and four were *erroneous* discrepancies. Out of the *erroneous* discrepancies, one was a family name of foreign origin, one was a first name of Swedish origin, one was a noun and one was a verb (*duger*, shown in Example 4.5).

Among the high frequency entries, there were 13 discrepancies, of which two were *formal* discrepancies and eleven were *different* discrepancies. There were no discrepancies classified as *erroneous*. All but one of the *different* discrepancies were common function words having a reduced pronunciation representation as the highest ranking representation. In most cases, the canonical version was also included in the set of pronunciation representation for the entry.

4.5.3 Accuracy without Access to the Grammatical Analysis

As mentioned, there may not be a grammatical tag available for the words of a text in a particular application using speech synthesis and in such cases, the highest ranking pronunciation representation of the highest ranking entry with a specific orthographic word is the representation that will probably be selected for the word.

In an evaluation of the accuracy without access to the grammatical analysis, i.e., when the orthographic word is the only criterion used to select a pronunciation representation, there are three questions that can be asked:

1. In how many cases is the pronunciation representation correct[4] for the randomly selected entry?

2. In how many cases is the pronunciation representation correct for the entry to which it is associated?

[4] For a representation to be *correct* in this context, it must be the most commonly occurring canonical representation for the entry.

3. In how many cases is the pronunciation representation the most commonly occurring one given the orthographic word?

The answers to questions 2 and 3 may differ if the ranking of analyses given the word is judged to be incorrect in one or more cases. Since the evaluation method involves randomly selecting 243 particular entries from the mid and low frequency group, it is mainly these entries that are of interest in the current context and the focus has thus been put on question 1.

To answer this first question, each of the 27 mid frequency entries and the 40 low frequency entries that were *not* the most probable ones given their orthographic word had their highest ranking pronunciation representation substituted for the highest ranking representation of the highest ranking entry given the word. It turned out that this introduced no changes; the pronunciation representations were the same for the randomly selected entry and for the highest ranking entry in all cases. Thus, the number and types of discrepancies were the same as those presented in Table 4.8 and that table can serve also as the answer to the first question presented above.

Since the manually corrected pronunciation representations were the same for both the randomly selected and the highest ranking entries in all cases, the answer to question 2 above can also be answered with the values presented in Table 4.8. The third question was not explicitly addressed, but it is considered highly unlikely that there are entries that should attain higher ranks for the orthographic words checked and that these entries should be pronounced differently from those checked. Thus, with all probability, we can consider also question 3 answered by Table 4.8.

4.5.4 High Frequency Entries not in CentLex

There were two entries in the high frequency group from the combined text corpus that did not occur in CentLex and it was of interest to investigate why the entries were missing from the lexicon. The results of this investigation showed that there was a discrepancy between the standard used for grammatical analyses in CentLex and the standard used for the version of suc, on which the tagger was trained.

The tag scheme described in Ejerhed et al. (1992) and used for CentLex was used for suc version 1.0. However, the tagger is trained on a later and only partly documented (and thus not easily re-usable) version of suc (Källgren, 1998). In this version, the set of well-formed tag strings has been updated. Thus, the version of the suc corpus on which the tagger used to tag the text corpora was trained includes tag sequences that are not listed as well-formed in Ejerhed et al. (1992). From looking at the tag strings occurring in the later version of suc, it seems that the differences in tag schemes are not large. The main thing that has happened between the versions is that the set of well-formed tag strings has been extended to accommodate some special cases not covered by the original set of well-formed tag strings. Unfortunately, words with special grammatical properties are almost exclusively high frequency words.

For example, one of the two high frequency entry forms missing from CentLex is not a well-formed SUC tag sequence, according to Ejerhed et al. (1992). This entry form is *flera 'several'* JJ POS UTR/NEU PLU IND NOM (cf. Table C.1 in Appendix C for an explanation of the tags), and since this is a special adjective only occurring in the indefinite form this is a reasonable analysis. However, in CentLex the analysis JJ POS UTR/NEU PLU IND/DEF NOM, which is a well-formed tag sequence, according to Ejerhed et al. (1992), is used instead. The difference in strategies for tag assignment in the SUC corpus (on which the tagger is trained) and CentLex is thus responsible for part of the missing entry types in CentLex.

The other of the two high frequency entry forms missing from CentLex is *egen 'own'* JJ POS UTR SIN IND/DEF NOM—also not a well-formed SUC tag in Ejerhed et al. (1992). However, in this case it is more unclear why the new tag has been introduced. The adjective *egen* in singular is always indefinite and there is a special definite form *egna*. The analysis JJ POS UTR SIN IND NOM, included in CentLex, thus seems to be the only reasonable analysis.

From looking at the set of grammatical tag strings occurring in SUC and comparing them to the well-formed SUC tags presented in Ejerhed et al. (1992), it seems unlikely that the discrepancy in tag sets between the tagger and CentLex affects more than a few entry form/word types. However, since the words involved are relatively frequent, there may be a significant number of tokens affected, mostly in the mid frequency group.

Further, as previously discussed, automatic tagging will not give 100 per cent correct results. The particular tagger used produced 6.45% errors (looking at the entire tag sequence) in a tenfold cross validation experiment (Megyesi, 2001, 2002a) and the error rate is probably higher over the text corpora used in the CentLex coverage and accuracy evaluations.

There are thus several sources of noise present in the data used for calculating the entry form coverage and the values presented must be interpreted as rough estimates, not only for the text types/genes investigated, but also for the particular corpora used. In contrast, the word coverage values, both regarding types and tokens, are not affected by these noise sources and are thus the actual coverage values for the corpora. However, when generalising to text types/genes, they are of course still estimates.

4.6 Strategies for Increasing Coverage and Accuracy

The main work with the CentLex database has been to combine different lexica on different formats and of different types into a single lexicon, to build the database and tools for accessing and editing the lexicon database. Some work has also been aimed at increasing the coverage of the lexicon and at improving the quality of the lexicon. The evaluations presented above showed that the coverage and accuracy are generally high, but that there is some room for improvement. The work on improving CentLex has only begun and is expected to proceed continuously henceforth.

The high frequency words investigated in the evaluations were manually corrected and the corrections have been used to update CentLex. It would take relatively little time and effort to go through all entries with function word tags not included in the high frequency group and make sure that the pronunciation representations are correctly ordered and that reduced forms are tagged with a special *reduced* tag. Further, the text corpora created for the evaluation can be used for extending CentLex with words from the mid frequency range. Missing entries from this range could be automatically transcribed and manually corrected relatively fast.

4.7 Summary

In this chapter, evaluations of the coverage of CentLex over a set of different text types and of the accuracy of the pronunciation representations in CentLex have been presented. The average coverage over the texts was 94.0% of the CentLex entry types (combinations of an orthographic word and a grammatical analysis) and 95.1% of the orthographic words.

The evaluation of pronunciation representation quality showed that among high frequency entry types, no pronunciation representations were obviously erroneous, although some differed from the agreed upon CentLex standard. Among mid and low frequency entry types the estimated shares of erroneous pronunciation representations were 1.7% and 6.6%, respectively.

CentLex can be seen as a model of the canonical pronunciation of words in central standard Swedish. The next chapter will present the speech data used for discourse context-dependent pronunciation modelling and some of the annotation methods employed for this pronunciation modelling effort. The canonical pronunciation representations from CentLex are used as the basis for much of the annotation.

Chapter 5

Annotation Method

A requirement of the data-driven approach taken to pronunciation modelling is, of course, data. In the current approach, the data consists of the annotation of spoken language, where the annotation is aimed at describing the discourse context of a phoneme from high-level linguistic variables such as speaking style, down to the articulatory feature level. It is important to have data that is accurate and also to have a sufficient amount of data. Mainly automatic methods are used for annotation, to make annotation fast in comparison to manual annotation and, thus, making it practically possible to obtain a sufficient amount of data. The price of using automatic methods is that the result may not always be as accurate as the result of manual annotation would have been. This chapter describes the speech data used for the work on pronunciation modelling presented in this thesis and the system and methods used for annotating the speech data.

5.1 Speech Data

The speech data used for the pronunciation modelling research described in this thesis consists of three speech databases: the VaKoS database, a Radio Interview database and a Radio News database.

Table 5.1: *Speech databases.*

Database	Origin	Type
VaKoS	Recording for phonological study	Elicited informal monologue
Radio Interview	Radio broadcast	Elicited formal dialogue
Radio News	Radio broadcast	Scripted formal monologue

The VaKoS database was originally constructed by Bannert and Czigler (1999) for a phonological study of variation in consonant clusters. The Radio Interview database and the Radio News database consist of recordings originating from *Sveriges radio* (Swedish public service radio) and have previously been used in the

Grog project, which was aimed at modelling the structuring of speech in terms of prosodic boundaries and groupings, cf. Carlson et al. (2002).

Table 5.2: *The distribution of speech from different speakers in the* VaKoS *database. Speakers 1–10 are informants for a phonological study.*

Speaker	Duration (s)	Number of Words	Number of Phonemes
1	647	1,642	6,082
2	629	1,593	6,145
3	624	1,548	6,081
4	623	2,003	6,994
5	622	1,517	5,964
6	618	1,408	5,617
7	617	1,442	5,604
8	600	1,217	5,000
9	599	1,619	6,112
10	586	1,187	4,478
$\sum$	6,165	15,176	58,077

The VaKoS database is a set of elicited monologues; ten speakers talk about some suggested topic or topics to a recording assistant (who is silent). About ten minutes from each speaker is included in the database. The VaKoS database includes some manual annotation at different levels. The parts of the annotation re-used for the purpose of pronunciation modelling are the orthographic transcripts, the word-level segmentation, prosodic boundary annotation, focal stress annotation, and annotation of word fragments (interrupted words), and filled pauses.

Table 5.3: *The distribution of speech from different speakers in the* Radio Interview *database. Speakers 11 and 12 are interviewees, speakers 13 and 14 are interviewers and speaker 15 is a radio announcer.*

Speaker	Duration (s)	Number of Words	Number of Phonemes
11	1,230	3,081	12,638
12	1,080	3,418	13,750
13	331	1,060	4,523
14	297	1,028	4,249
15	20	27	159
$\sum$	2,958	8,614	35,319

The Radio Interview database is a set of two 25-minute radio broadcast interviews, each including speech mainly from three speakers, the interviewee and two interviewers. The interviewees are experienced public speakers (politicians) and are allowed to answer questions in length, rarely being interrupted. The Radio News database includes two radio news broadcasts, including speech from altogether three studio news announcers and eight reporters. Only studio environment recordings are included in the Radio News database. The radio broadcast databases include orthographic transcripts and manual annotation of prosodic boundaries originating from the Grog project. For one of the interviews, focally stressed words were also annotated in the Grog project. This information was re-used in the annotation for pronunciation modelling purposes.

Table 5.4: *The distribution of speech from different speakers in the* Radio News *database. Speakers 16–18 are news announcers and speakers 19–26 are news reporters (recorded in a studio environment).*

Speaker	Duration (s)	Number of Words	Number of Phonemes
16	189	428	2,294
17	159	420	2,121
18	107	269	1,297
19	77	195	960
20	71	155	830
21	55	177	814
22	54	163	740
23	47	122	584
24	45	109	517
25	41	113	529
26	31	72	362
$\sum$	876	2,223	11,048

All speech data are digital studio recordings sampled at 16 kHz. Table 5.1 gives a brief overview of the speech data and Tables 5.2–5.4 give the details of the distribution of speech from different speakers in the respective databases.

5.2 A Multi-Layer Annotation System

The annotation used for pronunciation modelling is organised in six layers: 1) a discourse layer, 2) an utterance layer, 3) a phrase layer, 4) a word layer, 5) a syllable layer, and 6) a phoneme layer. The layers are segmented into units, which are linguistically meaningful and can be synchronised to the speech signal. The segmentation of each layer is strictly sequential, i.e., every part of the signal belongs to some unit at all layers and there is no overlap between units within a layer.

Durational boundaries are inherited from higher order layers to lower order layers, so that a discourse boundary is always also an utterance boundary, a phrase boundary, a word boundary, a syllable boundary and a phoneme boundary. The layers are thus hierarchically ordered so that a higher order unit serves as the parent of all lower order units within its segmental bounds. An arbitrary amount of information can be supplied for each unit in each layer. Figure 5.1 shows an excerpt of a sound file with some aligned example annotation. In Figure 5.1, phonetic transcripts are in the Swedish Technical Alphabet (STA) format. Table A.1 in Appendix A shows the STA symbols and their IPA equivalents.

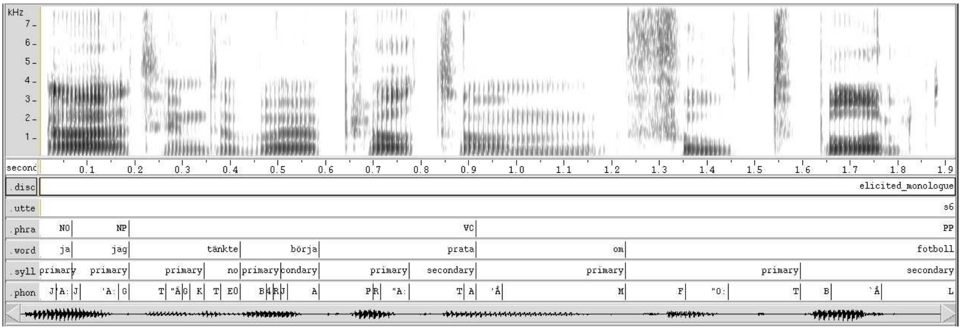

kHz
7
6
5
4
3
2
1
seconc
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9
.disc
elicited_monologue
.utte
s6
.phra
NO NP VC PP
.word
ja jag tänkte börja prata om fotboll
.syll
primary primary primary no primary condary primary secondary primary primary secondary
.phon
J A: J 'A: G T "Ä G K T E0 B Ö R J A P R "A: T A 'Å M F "O: T B `Å L

Figure 5.1: *Annotation layers with example annotation aligned to the speech signal.*

Since the layers are hierarchically ordered and the units in each layer are sequentially ordered, the annotation is organised as an hierarchical tree structure in six levels superimposed on a duration-based segmental structure. The time dimension can thus be excluded, so that the annotation is disconnected from the signal and forms a proper duration-independent tree structure. Figure 5.2 shows the annotation in Figure 5.1 as a tree structure.

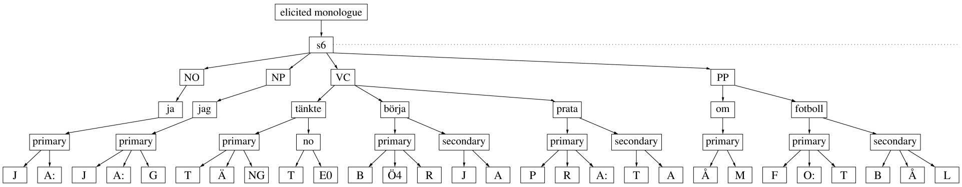


Figure 5.2: *The annotation in Figure 5.1 as a tree structure.*

The most important feature of this system of annotation is that information can be unambiguously inherited from units on higher layers by units on the layers below. A unit can thus pass on its information to all the units within its bounds in the lower order layers. Consequently, information connected to syllable, word, phrase, utterance and discourse layer units, respectively, as well as to the phoneme layer units, is accessible from the phoneme layer. This is important since the pronunciation models will use phoneme-sized units as input. Sequential context information, i.e., properties of the units adjacent to the current unit at the respective layers, is used at model induction together with information connected to the current units. Having the information stored in different layers enables easy access to the sequential context information.

5.3 Segmentation

The annotation process begins by a segmentation of each annotation layer into its respective type of unit. The next step is to retrieve, calculate or estimate the information to be associated with each unit. With some minor exceptions, automatic methods are used for segmentation, however with manual supervision to improve accuracy at some intermediate stages.

In the current context, each VaKoS monologue, each radio interview and each radio news broadcast is considered a separate discourse and an utterance is defined as a discourse turn uttered by a single speaker. This means that a monologue discourse is treated as a single utterance. For dialogues, the discourses were manually segmented into utterances. During utterance segmentation, pauses between utterances are included in the utterance to the right.

The speech data used is in one channel, and thus, the speech from different speakers cannot be separated if overlapping in time. The annotation system could be extended to accommodate overlapping speech. However, some of the information included in the annotation are measures calculated from automatically obtained f_0 extracts and phoneme durations (cf. sections 5.4 and 5.5 below) and automatic phonetic transcription based on the signal is employed (cf. Section 5.7). It is neither possible to calculate signal-dependent measures nor to estimate phone labels based on the signal for overlapping speech with any degree of certainty.

Thus, the overlapping parts of the speech signal are treated as special non-analysable units. During segmentation, overlapping speech *between utterances* is given the special utterance unit tag <*overlap*>, but no other information is associated with the unit. An <*overlap*> utterance unit is extended to the nearest word boundary, so that the entire word is included in the <*overlap*> utterance unit, even if the word is only partially overlapped.

Overlapping speech *within an utterance* (i.e., where the utterance has started before the overlap and continues through and after the overlap) is not annotated on the utterance layer (it is on the word layer, however). The speech segments annotated as overlapping on the utterance layer are given <*overlap*> tags also at the word layer and a <*junk*> tag on the phoneme layer, but otherwise no information is included for lower order layers.

In the Radio News database, there are some instances of speech overlapping with music. The parts of the speech overlapping with music are also annotated with <*overlap*> tags. The amount of overlapping speech in the speech data used is very low, and only a very small part of the data is affected.

Automatic segmentation begins at the word level. Automatic speech recognition can be used to facilitate orthographic transcription. However, for the currently used databases, the orthographic string, including annotation of filled pauses and non-speech sounds, has been manually supplied. The orthographic annotation was included in the databases used and thus inherited from the VaKoS and Grog projects. Only minor corrections were made during the work on pronunciation varaition reported in this thesis. Special consideration was taken to supplying an accurate word sequence, since the automatic alignment is highly dependent on the orthographic string.

Manually obtained word layer segmentation followed the VaKoS database and one of the radio interviews. The other radio interview and the Radio News database were segmented into word units using the nAlign automatic aligner (Sjölander, 2001; Sjölander, 2003; Sjölander and Heldner, 2004), with forced word boundaries at utterance boundaries. Manual correction of the word layer segmentation is performed, since all succeeding annotation depends on this segmentation. Manual supervision at this level is relatively fast, and increases in the word layer segmentation accuracy give large improvements in accuracy in successive annotation.

The phrase layer is segmented using the Spark parser (Aycock, 1998) with a context-free grammar for Swedish constructed by Megyesi (2002b,a) operating on a string of tags produced by the TnT part-of-speech and morphological tagger (Brants, 2000) trained by Megyesi (2001, 2002a) on the suc corpus (Ejerhed et al., 1992). Only phrase chunk information is used and the phrases are aligned to the signal using the word boundaries. The parser was created for parsing written text, but it is robust and produces parses also for tagged orthographic transcripts of spoken language.

During phrase layer segmentation, only maximal phrases are considered. A noun phrase can include modifiers of different types, e.g. nouns, adjective phrases and prepositional phrases. The entire maximal projection of the noun phrase is counted as a single phrase and the identity and boundaries of any constituent phrases are ignored. Similarly, conjoined adjective phrases (e.g. *'very interesting and nice'*) are counted as a single adjective phrase.

Some word units do not belong to any phrase chunk (mostly conjunctions). For phrase segmentation purposes, these words are given a *no phrase* tag and are treated as one-word phrases. Verb phrases are not included in the analysis. Verbs are instead parts of either a *verb cluster* or an *infinitive phrase*. A *verb cluster* is a single verb or a continuous sequence of verbs belonging to the same verb phrase (e.g. *'would have been'*) and an *infinitive phrase* is a verb in the infinitive proceeded by an infinitive particle. The infinitive phrase may contain adverb phrases and/or verb particles, e.g. *'to go out'*. The full set of phrase types produced by the parser can be seen in table 6.3 in the next chapter.

The phoneme layer is segmented word-by-word using the word boundaries and canonical phonemic pronunciation representations as input to the automatic aligner. The phonemic representations are collected from the CentLex pronunciation lexicon (cf. chapters 3 and 4; Jande, 2006), if the word occurs in the lexicon. Words not occurring in the lexicon receive phonemic representations generated by a grapheme-to-phoneme conversion algorithm included in the Rulsys text-to-speech system (Carlson and Granström, 1975, 1976; Carlson et al., 1982). The speech databases contain some instances of interrupted words (i.e., parts of words). In the cases where these are not correctly handled by the grapheme-to-phoneme rules, the phoneme representations are corrected manually, for consistency.

On the phoneme layer, the synchronisation of units to the signal is more abstract than on the higher order layers; not all phonemes in the canonical phonemic representations have overt correspondents in the speech signal, but they nevertheless will have a duration in the annotation. As will be seen in Section 5.4, the abstract nature of the phoneme boundaries is exploited in phoneme duration-based measures.

The most sonorant phone of the syllable constitutes the nucleus of the syllable. A minimal syllable consists of a nucleus phone only. However, a syllable may also contain a sequence of phones preceding the nucleus, the syllable *onset*, and/or a sequence succeeding the nucleus, the syllable *coda*. Syllable boundary allocation is based on the phonotactic constraints of the language. For central standard Swedish, these constraints have been described by Sigurd (1965). Onset and coda sequences in Swedish syllables follow the sonority hierarchy (cf. e.g. Jespersen, 1904) with some exceptions. However, there is no single standard for syllable boundary allocation in Swedish. Several strategies have been proposed (cf. e.g. Sigurd, 1965; Gårding, 1967).

The strategy chosen for the current annotation was to use lists of phonotactically allowed onset and coda consonant sequences based on Sigurd (1965) and Elert (1970, pp. 89–90) to exclude impossible syllable boundaries. When it is allowed to place the syllable boundary at more than one location in a consonant sequence between two vowels, the coda of the first syllable is maximised if the vowel is a short stressed vowel, and the onset of the second syllable is maximised otherwise. Further, syllable boundaries are forced at word boundaries and at compound constituent boundaries (compound boundaries are included in the phonemic representations collected from the pronunciation lexicon or generated by RULSYS). The syllable boundaries are synchronised to the signal using the phoneme boundaries.

Some units with special characteristics have been introduced at the word layer to ensure that parts of the signal that are not speech (or non-analysable speech) can be annotated. The special unit types are <*overlap*> (overlapping speech), <*pause*> (including pauses, inhalation and exhalation sounds), <*non-speech*> (including laughter, smacks, clicks, coughs and hawking sounds etc.), and <*filled pause*> (e.g. 'hesitation' sounds resembling /ə/, /eː/, /əm/, /ɜːm/ or /m/). The information supplied for normal word units is *not* included for these units. Within the boundaries of one of the special word layer units, a <*sil*> (for pauses) or a <*junk*> special phoneme unit is used and no additional annotation is supplied on the phoneme and syllable layers.

5.4 Mean Phoneme Duration Measures

As previously discussed, speech rate is an important factor for the phonetic realisation of words. Speech rate can be defined e.g. as the number of phonemes per time unit, which is the inversion of mean phoneme duration. In the current speech annotation, several measures of mean phoneme duration are calculated, including measures of mean Z-normalised phoneme duration. When normalised, duration values can be zero. Hence, converting the mean normalised phoneme duration to a speech rate measure may in these cases give rise to infinite speech rates as an artefact of the normalisation process. For this reason, the mean phoneme duration measures are used in the annotation rather than speech rate measures. *Mean phoneme duration* is measured globally, over the entire discourse, and locally over each utterance, phrase, word and syllable.

5.4.1 Phoneme Durations

The mean phoneme duration measures are based on the automatic segmentation of the phoneme layer, conducted through automatic alignment of canonical phonemic representations of words to the speech signal. *Mean phoneme duration* is thus an abstract measure and coincides with the more concrete measure *mean phone duration* when all phonemes in the phonemic representation are realised. The measure thus constitutes an estimate of what the mean phone duration would be if all phonemes in the canonical pronunciation representation were realised over a certain unit of fixed duration.

Mean *phone* duration cannot be used for prediction, since the phone string is the variable to be predicted by the pronunciation model. The exact number of phones is thus not known in advance when the model is used. The abstract nature of the mean phoneme duration measure is likely to make it a strong predictor of phone-level pronunciation; high speech rate is generally a good predictor of phonological assimilation and reduction processes and mean phoneme duration emphasises sections of the speech signal with high speech rates more than a measure corresponding to phones per time unit. It will also emphasise sections of the speech signal with high speech rates more than measures corresponding to syllables or words per time unit. As will be discussed in Chapter 10, Section 10.2, for the mean phoneme duration measure to be usable in the absence of a speech signal, a prosodic model estimating the durations of phonemes (and hence, of units on higher order layers) is necessary.

During the phone layer segmentation discussed above, the aligner had the advantage of being forced to align the phonemic representation of a word to the part of the signal between the manually supplied word boundaries, which makes the alignment very close to optimal. This is important for the mean phoneme duration measures over the syllable to be reliable.

5.4.2 Duration Normalisation

Different phonemes have different inherent durations (cf. e.g. Elert, 1964) and additionally, central standard Swedish has phonologically long and phonologically short vowels. The duration of a phoneme is also dependent on e.g. the phoneme context. Further, central standard Swedish has a complementary distribution of phoneme duration. Simply put, in a closed syllable (containing a coda consonant or consonant cluster), a *long vowel* will always be followed by a *short consonant* and a *short vowel* will always be followed by a *long consonant* or a consonant cluster, although length is not generally regarded as a phonological property of the consonant.

Further, vowels in stressed syllables are generally longer than vowels in unstressed syllables. These differences may be regarded as variation in speech rate on the local level: the stressed syllable is pronounced more slowly than the unstressed syllable. However, here it is assumed that this difference is not primarily a speech rate difference, and hence the difference is treated as a property of the vowel: the vowel in stressed position is inherently longer than the vowel in unstressed position.

Neither inherent length nor phonological length/complementary length have anything to do with speech rate. A one-syllable word with a phonologically long vowel may have a longer duration than a word with a phonologically short vowel. However, this does not reflect a difference in speech rate between the words. If mean phoneme duration is calculated over larger units, such as the phrase or the utterance, differences due to inherent length and phonological length will to a large degree even out. However, when speech rate is calculated locally over words and syllables, they mostly will not.

For this reason, measures based on *normalised* phoneme duration are included in the annotation alongside measures based on absolute phoneme duration. During normalisation, the duration of each phoneme token is related to the mean duration of the particular phoneme type using the normal transformation (cf. Equation 5.1, where Z is the normalised duration value, x is the phoneme duration, μ is the mean duration of the phoneme type over the database and σ is the standard deviation of the phoneme duration for the type). During normalisation, phonologically long phonemes (including consonants) are separated from phonologically short phonemes, and vowels serving as nuclei in stressed syllables are separated from their phonologically identical counterparts in unstressed syllables.

$$Z = \frac{x - \mu}{\sigma} \qquad \text{eq. 5.1}$$

5.4.3 Measures Calculated

A variant of the mean phoneme duration measure included in the annotation is the *mean vowel duration.* For this measure, all segments except vowels are ignored under the assumption that the perceived speech rate may be better modelled by vowel duration alone than by general segment duration.

The mean phoneme duration measures and the mean vowel duration measures are calculated both from duration on a linear scale and from duration on a logarithmic scale. Since small differences in speech rate probably have larger effects on phone-level pronunciation when the speech rates compared are high than when the speech rates are low, the relative size of small differences in duration is increased through transferring the phoneme durations to the logarithmic scale ($\log_e$).

To sum up, there are measures based on all phonemes and on vowels only; there are measures based on absolute duration and on normalised duration; and, finally, there are measures calculated on a linear time scale and on a logarithmic time scale. All combinations of variants are calculated, resulting in a total of eight *mean phoneme duration* measures.

5.5 Pitch Dynamics and Pitch Range Estimation

Pitch movement is correlated with emphasis; much pitch movement over a particular unit makes the unit stand out from its surroundings and signals that the unit is emphasised. Emphasis is also correlated with segmental pronunciation, in such a way that the pronunciation tends to be more similar to the canonical pronunciation for emphasised words than for non-emphasised words. This means that there is a correlation between pitch dynamics and phone-level pronunciation. The measures described below are included in the annotation to make use of this correlation.

The Esps pitch extraction algorithm incorporated in the Snack Sound Toolkit (Sjölander, 2004; Sjölander and Beskow, 2000) is used to extract the pitch contour from the speech data in 10 ms frames. The pitch extraction algorithm requires an a priori pitch range to be specified and it proved beneficial to use different ranges for different speakers. Dividing the speakers into two pitch register groups was sufficient for adequate pitch contours to be extracted (as determined by audio-visual manual assessments of random samples of the speech signal using the WaveSurfer speech tool (Sjölander and Beskow, 2000). A high register (90–600 Hz) group and a low register (60–300 Hz) group were thus defined and each speaker manually assigned to the most appropriate group. Using the extracted pitch contours, measures of *pitch range* and *pitch dynamics* (‘liveliness’), respectively, are calculated over each utterance, phrase and word unit.

Pitch range is defined as the difference between the largest pitch maximum and the smallest pitch minimum contained by a unit. The first and the last voiced sample of the unit, over which the pitch-based measures are calculated, are counted as extreme values. *Pitch dynamics* measures are based on the absolute distances of maximum and minimum points or plateaus from a base frequency. Two base frequencies are used: 1) the median pitch over the unit and 2) a base frequency estimating liveliness variation as perceived by human listeners.

Observations made by Traunmüller and Eriksson (1995a) suggest that the best correlation with human perception of liveliness variation is when the base frequency is located ∼1.5 standard deviations below the mean pitch of the speaker. Thus, the base frequency estimating human liveliness perception, f_b, is calculated separately for each speaker with equation 5.2, where $\bar{x}_{f_0}$ is the mean fundamental frequency of a speaker over the available recordings (i.e., the mean of the voiced pitch samples obtained by the pitch extraction algorithm) and σ_{f_0} is the standard deviation of the speaker’s f_0.

$$f_b = \bar{x}_{f_0} - 1.5\sigma_{f_0} \qquad \text{eq. 5.2}$$

The absolute distances of maximum and minimum points or plateaus from the respective base frequencies are summed up over a unit, and based on these sums, two different pitch dynamics measures are calculated for each base frequency. First, the sums are divided by the number of minimum or maximum points or plateaus contained by the unit, to obtain a measure of pitch dynamics differentiating between units with pitch extremes with *large* average deviations from the base frequency and units with pitch extremes with *small* average deviations from the base frequency.

Equations 5.3 and 5.4 show how the pitch dynamic measures divided with the number of extreme points ($f^d_{e,p_{50}}$ and f^d_{e,f_b}, respectively) are calculated. In these equations, E is the extreme point count over a unit, f^e_i is the frequency of the i:th extreme point, p_{50} is the median (the 50th percentile) and f_b is the Traunmüller-Eriksson base frequency.

$$f^d_{e,p_{50}} = \frac{\sum_{i=1}^{E} |f^e_i - p_{50}|}{E} \qquad \text{eq. 5.3}$$

$$f^d_{e,f_b} = \frac{\sum_{i=1}^{E} |f^e_i - f_b|}{E} \qquad \text{eq. 5.4}$$

Second, the sums are divided by the number of (non-zero) pitch samples contained within the unit, resulting in a measure differentiating between units with *fast* average pitch movement and units with *slow* average pitch movement. Equations 5.5 and 5.6 show how the pitch dynamic measures divided with the number of pitch samples contained by a unit, S, are calculated. The measures are denoted $f^d_{s,p_{50}}$ and f^d_{s,f_b}, respectively.

$$f^d_{s,p_{50}} = \frac{\sum_{i=1}^{E} |f^e_i - p_{50}|}{S} \qquad \text{eq. 5.5}$$

$$f^d_{s,f_b} = \frac{\sum_{i=1}^{E} |f^e_i - f_b|}{S} \qquad \text{eq. 5.6}$$

Equal differences in pitch measured in Hz are not perceptually equivalent across different pitch levels. Hence, three scales constructed to mirror the response of the human auditory system (psychoacoustic scales) are used for measuring pitch in addition to the linear Hz frequency scale. The three psychoacoustic scales used are the MEL scale (Stevens and Volkman, 1940), the equivalent rectangular bandwidth

(ERB) scale (More and Glasberg, 1983; Hermes and Gestel, 1991) and the semitone scale.

The MEL scale and the semitone scale are aimed at mimicking the way in which human listeners perceive differences in pitch height. Equal distances in Mel or semitones are thus perceived as equal by humans across pitch registers. The ERB scale is designed to mimic the frequency selectivity of the human auditory system and an equal ERB-rate will give an equal perceived prominence of pitch movements for speakers of different pitch registers (Nooteboom, 1997). The semitone scale has been shown to give the best results in terms of perceptual equivalence of pitch distance by e.g. Traunmüller and Eriksson (1995b) and Nolan (2003). The pitch sample values are converted to the three respective scales using equations 5.7 through 5.9, where f is the (fundamental) frequency in Hz.

$$Mel = 1127.01048\ log_e\ (1 + \frac{f}{700}) \qquad \text{eq. 5.7}$$

$$ERB\text{-}rate = 16.7\ log_{10}\ (1 + \frac{f}{165.4}) \qquad \text{eq. 5.8}$$

$$Semitone = 12\ log_2\ \frac{f}{100} \qquad \text{eq. 5.9}$$

Pitch range is thus estimated on four different scales, resulting in a total of four different pitch range measures. There are two different measures of *pitch dynamics* focusing on either average deviation from the base frequency or the average speed of pitch movements. Each of these measures is estimated from two different base frequencies and on four different scales, resulting in a total of sixteen different pitch dynamics measures.

5.6 Word Predictability and Related Measures

The predictability of a word has been shown to be important for the realisation of the word (cf. e.g. Fosler-Lussier and Morgan, 1999; Jurafsky et al., 2001). Many variables influence the predictability of a word in context. Measures related to word predictability included in the annotation described here are *collocation frequency*, *word repetitions*, *lexeme repetitions*, *the position of the word in a phrase*, *Part of Speech*, *the position of the word in a frequent collocation* and *global word frequency.* A special measure termed *word predictability* is also included in the annotation.

The word predictability statistic is the weighted combination of trigram probability, bigram probability and unigram probability, as shown by Equation 5.10. Here, P_w is the word predictability statistic, $p_e(w_n|w_{n-2}, w_{n-1})$ is the estimated probability of a word given the two preceding words (cf. Equation 5.11) and $p_e(w_n|w_{n-1})$

is the estimated probability of a word given the preceding word (cf. Equation 5.12). In Equation 5.11, $c(w_{n-2}, w_{n-1}, w_n)$ is the trigram frequency and $c(w_{n-2}, w_{n-1})$ is the frequency of the bigram preceeding the current word. In Equation 5.12, $c(w_{n-1}, w_n)$ is the bigram frequency and $c(w_{n-1})$ is the frequency of the unigram preceding current word.

$$P_w = \lambda_\alpha p_e(w_n|w_{n-2}, w_{n-1}) + \lambda_\beta p_e(w_n|w_{n-1}) + \lambda_\gamma p_e(w_n) \qquad \text{eq. 5.10}$$

$$p_e(w_n|w_{n-2}, w_{n-1}) = \frac{c(w_{n-2}, w_{n-1}, w_n)}{c(w_{n-2}, w_{n-1})} \qquad \text{eq. 5.11}$$

$$p_e(w_n|w_{n-1}) = \frac{c(w_{n-1}, w_n)}{c(w_{n-1})} \qquad \text{eq. 5.12}$$

The trigram weight (λ_α) is set to 0.6, the bigram weight (λ_β) is set to 0.3 and the unigram weight (λ_γ) is set to 0.1. The specific weights used are arbitrarily chosen, but with the main weight on the trigram statistic, which is assumed to be the best single estimator of word predictability from the three statistics. A combination of three statistics is used under the hypothesis that this gives a better estimator for word predictability than using the trigram statistic only (this will often be 0, which may give an unfairly low predictability estimation) or the bigram statistic or the unigram statistic only (these use less context and thus are less precise estimators).

Unigram, bigram and trigram probabilities were collected from a formatted version of the Göteborg Spoken Language Corpus (GSLC) (Allwood, 1999; Allwood et al., 2000, 2002). GSLC contains orthographic transcripts of spoken language from a variety of communicative situations. After formatting, exclusion of some types of non-word units and convertion of transcripts to standard orthography, the size of the corpus is approximately 1.3 million words. Probabilities are calculated utterance-by-utterance by introducing two utterance boundary symbols in between each two consecutive utterances before calculating trigram statistics and one utterance boundary symbol before calculating the bigram statistics. Simple full-form word probabilities were used for the unigram probability.

The estimated *global word probability* is sometimes used as a rough estimator of word predictability (e.g. in Fosler-Lussier and Morgan, 1999). Since an estimate of global word probability from GSLC is available, it is included in the annotation. The position of a word in its phrase or in a collocation affects the predictability of the word, and the positions of a word in the phrase and in a collocation, respectively, are included in the annotation as three-way classifications: *initial*, *medial* or *final*, where *initial* is the default value used for one-word phrases.

Collocations are, in the current context, defined as trigrams occurring at least four times in GSLC or bigrams occurring at least three times. Ir would be possible to improve the list of collocations by adding *lexicalised phrases* (cf. e.g. Lindberg, 1999). Most lexicalised phrases will not occur in a corpus very often. However, the co-occurrence of the words in a lexicalised phrase will generally be very high and words at the end of a lexicalised phrase will thus be highly predictable from the preceding words. However, no lexicalised phrase lexicon or detection is used for the current annotation. Since the lexicalised phrases occur quite sparsely, including lexicalised phrase detection would probably make very little difference in the current context.

Two measures of the *number of word repetitions* are included in the annotation, the number repetitions of the full-form word thus far in the discourse and the number of repetitions of the *lexeme* thus far in the discourse. PCkimmo (SIL International, 1995; Antworth, 1990, 1995), the SIL implementation of Koskenniemis's two-level morphology system (Koskenniemi, 1983; Karttunen, 1983) with lexica and rules for Swedish compiled by Ridings (2002)[1] is used for finding the lemma form of each word. The combination of the lemma form and the Part of Speech is used to define a lexeme. For some input, the PCkimmo/Ridings system cannot produce a lemma form. The back-up strategy in these cases is to use the full-form word to define the lexeme. Since only very few and infrequently occurring words do not receive a lemma form from the system, this strategy works well in practise.

5.7 Automatic Phonetic Transcription

Phonetic identity is the variable to be estimated by the pronunciation models and hence, the phonetic annotation is used as the key during model training. Manual phonetic annotation of speech, especially of conversational speech, is a time-consuming and thus expensive task. A system for automatic phonetic transcription has been built to facilitate the current annotation. The automatic transcription system is a hybrid phonetic decoder using statistical decoding and a set of a posteriori correction rules. The task of the system is to supply the context-dependent realisation of each phoneme in the canonical pronunciation representation collected from a lexicon. The realisation can be ∅ (*'no realisation'*). The phone label set is the same as the phoneme label set and includes 23 vowel symbols and 23 consonant symbols. There is also a place filler ∅ label in the phone label set that occupies a phoneme position with no realisation in the phonetic string.

5.7.1 Background

Automating the task of phonetic transcription and alignment as far as possible is important for any project involving phonetic transcription of spoken language data. A number of automatic phonetic transcription systems have been reported. Mostly, the systems have a two-fold task, *segmentation* and *labelling*, in contrast to the auto-transcription system used for the annotation described in this thesis,

[1] Based on the Parole lexicon used in the Swedish part of the le-parole project, cf. Toporowska Gronostaj (2005).

which is a labelling system only (the system actually produces both labels and label boundaries, although only the labels are used in the annotation, and thus only the label part of the output is evaluated).

An automatic phonetic transcription and segmentation system for German called MAUS (Munich AUtomatic Segmentation system) has been developed at the Department of Phonetics and Speech Communication (IPSK) at the University of Munich. This system uses a set of pronunciation variation rules derived from manually transcribed speech data (Wesenick, 1996) to create graphs representing all presumed pronunciation variants of an utterance from the canonical phonemic representation found in a pronunciation lexicon. The graph is used with monophone HMMs for Viterbi alignment of the best state sequence to the signal. A rule system is used to refine the segment boundary allocation (Kipp et al., 1996; Schiel, 1999). Data-driven pronunciation rule generation and statistical matching was shown to outperform an knowledge-based rule system (Kipp et al., 1997; Schiel et al., 1998; Wesenick and Kipp, 1996).

Different corpora and different transcribers yield very different results in terms of inter-labeller agreement. Kipp et al. (1996) report an average inter-labeller agreement of 93.8% for three labellers and an average system-labeller agreement of 87.9% for the same data. Using another speech corpus and, presumably, other labellers Kipp et al. (1997) report an average inter-labeller agreement of 80.4% for three labellers and an average system-labeller agreement of 78.5% for the same data. However, the low agreement values seem to be due to one labeller's decisions standing out. If the deviant labeller is excluded, the values are 82.6% and 80.3%, respectively. All values are based on the best match between phone sequences according to a dynamic programming algorithm.

The phone HMMs used by the MAUS system are trained on manually segmented and transcribed speech. Models for automatic speech recognition (ASR) are often trained on segmentations based on canonical phonemic representations. Since these are the forms in the recognition lexicon, this is optimal for ASR. However, in automatic transcription, phone HMMs which in themselves model as little as possible of the pronunciation variation are optimal, since the mismatches between canonical forms and actually uttered phone sequences in this case is model contamination (Kessens and Strik, 2001; Schiel, 2004).

There is also an iterative version of MAUS, which adapts pronunciation graph transition probabilities to the specific target material to be segmented and transcribed (Beringer and Schiel, 1999), and a version adapting the phone HMMs (Schiel, 2004). Thus, no new training data is required when using the system on a new speaking style. Further, specific rule sets for different regional variants of German have been induced from annotated data (Beringer and Neff, 2000b). Beringer and Neff (2000a); Beringer (2003b) and Beringer (2003a) also report experiments with segmenting and transcribing Japanese and English speech using the MAUSER system (based on the MAUS system).

Kessens et al. (1998); Wester et al. (1998a) and Wester et al. (2001b) report compiling a set of five pronunciation variation rules reflecting the most common phone-level pronunciation variation in Dutch: /n/ elision, /r/ elision, /t/ elision, /ə/ elision and /ə/ insertion. A set of utterances from a spontaneous dialogue database were selected and at each possible rule application, nine listeners gave a binary judgement of whether the rule had applied or not, comparing the canonical form of the word and the actual pronounced phone sequence. An automatic speech recognition system in forced recognition mode was used to make the same decision by choosing between all possible pronunciations of each word according to the rules. The average inter-listener agreement for this limited task was 82% and the average ASR-listener agreement was 78%. Kessens et al. (2000b,a) report data-driven rule generation. The data-derived rules generated less pronunciation variants, but resulted in somewhat poorer performance.

Binnenpoorte and Cucchiarini (2003) report an experiment with three different automatic transcription methods. The first method was to simply use canonical phonemic representations from a lexicon. The second method was applying a set of static assimilation rules for word boundaries on the canonical representations (Cucchiarini et al., 2001). In the third method, multiple representations were generated from the canonical representations with a small set of deletion and insertion rules. Forced alignment automatic speech recognition was then used to select the optimal phonetic representation of each word. The transcripts resulting from the three methods were dynamically aligned and compared to a gold standard compiled by two phoneticians in consensus. The use of pronunciation modelling improved the results compared to using the canonical representations. Using forced recognition resulted in lower substitution and insertion error rates than using static rules. However, since the recogniser required phone segments of at least 30 ms in duration, many segments were not detected in spontaneous speech and the elision error rate was high. This made the over all results of the static rules better than the forced recognition results. The best automatic transcription results were a 19.4% phone error rate (PER) for interview type speech and a 26.8% PER for spontaneous speech. Four human transcribers showed PERs between 10.1% and 11.0% for interviews and between 13.4% and 15.7% for spontaneous speech, compared to the same gold standard (Binnenpoorte et al., 2003; Cucchiarini and Binnenpoorte, 2002). Automatically extracting pronunciation variants from a transcribed training corpus and using these for forced recognition instead of rule-generated variants improved the results of the third method slightly (Binnenpoorte et al., 2004).

Demuynck and Laureys (2002); Demuynck et al. (2002) and Demuynck et al. (2004) report results from experiments on automatic phonetic transcription and segmentation of Dutch speech. In these experiments, phonetic realisation alternatives were collected from a multiple phonemic representation lexicon (including foreign word lexica for English, German and French). Phonemic representations of novel compounds and inflections (not present in the lexicon) were generated with rules from representations of their constituents. Phonemic pronunciation representations for words not in the lexicon, and whose constituents also could not be found in the lexicon, were generated with a grapheme-to-phoneme conversion decision tree. An assimilation rule system for cross-word phenomena was used for generating more alternative pronunciations. The possible pronunciations of each sentence (according to the system) were described as a pronunciation network (Demuynck et al., 1997). The path through the network (of phones described by context-dependent HMMs) best matching the acoustic signal was used to determine the output phoneme string. Compared to a manually corrected phonetic transcript (based on another automatic transcription algorithm), the transcription system showed an 8.25% PER for conversational speech. The acoustic models were trained on read speech with canonical phonemic representations. There was thus a mismatch between speaking styles between the training data and the conversational speech test data. Also, the acoustic models were somewhat contaminated, since canonical representations were used. Finally, the minimal duration constraint of the context-dependent phone models was 30 ms, while phones in conversational speech can be considerably shorter.

Torre Toledano et al. (1998) report a PER of 2.65% for a small corpus of Castilian Spanish spontaneous speech from one speaker, using canonical phonemic descriptions and rules for alternative pronunciations. Torre Toledano et al. focus on segmentation and no detailed description of the phoneme sequence detection procedure is given in the paper. The explanation given for the low PER is the two stage procedure used. Instead of using context-independent HMMs (which give good time resolution but poor phoneme sequence resolution), the phoneme sequence is estimated using context-dependent HMMs (which give poor time resolution but good phoneme sequence resolution). In a second stage, the phoneme boundary positions are subsequently refined using statistical cancellation of systematic segmentation errors and a set of fuzzy logic rules.

Vorstermans et al. (1996) used artificial neural nets (ANNs) for segmentation and classification of phonetic segments. The ANNs were originally trained on a Flemish continuous speech corpus and the strength of the system was that the ANNs could easily be adapted to new languages without large or manually segmented speech databases for the new languages. System performance on new languages (English, Danish and Italian) was comparable to or better than previously reported systems trained for a particular language.

Vereecken et al. (1997) report being able to significantly improve the result of a phonetic transcription system through dividing larger paragraphs into prosodic phrases using silences, breaths and clicks prior to the automatic annotation. This system was evaluated on three languages, English, Flemish and Italian.

Chang et al. (2000) report using artificial neural nets (ANNs) to classify each 10 ms frame of the speech signal in terms of articulatory phonetic features and subsequently mapping the features to phonetic segment labels. The system does not require an orthographic transcript. When tested on a database of excerpts of spontaneous American English speech, including mostly addresses and phone numbers, the system showed an 19.3% PER. Further experiments with this system on American English (Chang et al., 2001) and Dutch (Wester et al., 2001a) spontaneous telephone speech showed PERs of 38.5% and 32%, respectively. These results were obtained in spite of a much lower articulatory feature classification error rate for frames. Non-phoneme-based descriptions of words were suggested by the authors.

5.7.2 A Gold Standard Transcript

A small sample of the VaKoS speech database was manually transcribed to be used as a gold standard, using the WaveSurfer speech tool (Sjölander and Beskow, 2000). The first minute of speech from five randomly selected speakers was transcribed. Transcription stopped at the first word boundary more than 60 s from the start of the sound file. Altogether, there were 2,765 phoneme positions in the canonical representations associated with the speech portions transcribed.

During manual transcription, the phone boundaries produced during the automatic alignment of the canonical phonemic representation to the signal were shown to the transcriber. For manual transcription, the aim was to supply the phone string that gave the best possible match to the signal, using the available phone set and to and align the phonetic transcription to the phoneme boundaries, also in such a way that the closest match to the signal was obtained. The transcriber was thus forced to used the phoneme boundaries when aligning the phones to the signal. However, it was possible to use $\emptyset$ symbols in the phonetic transcripts, to signal that a certain phoneme had no overt realisation. The canonical phonemic representations were shown at manual transcription.

5.7.3 Creating Realisation Lists

As mentioned, in the annotation of the speech data used for the pronunciation modelling effort described in this thesis, the automatic phonetic transcription was performed by a hybrid system using statistical decoding and a set of a posteriori correction rules. A list of possible realisations for each phoneme was derived empirically. Based on studies of pronunciation variation in Swedish (Gårding, 1974; Bruce, 1986; Eliasson, 1986; Bannert and Czigler, 1999; Jande, 2003a,b) and general knowledge of the target language, a tentative realisation list with all generally possible realisations was compiled for each phoneme.

A tentative pronunciation net was then created using these tentative realisation lists. The Snack Sound Toolkit (Sjölander and Beskow, 2000) was used for building a finite state transition network from the pronunciation net and a set of HMM monophone models (Sjölander and Beskow, 2000; Sjölander, 2003). Snack tools were then used for Viterbi decoding (probability maximisation) given the observation sequence defined by the parameterised speech.

The output best matching phone sequence was compared to a three minute portion of the gold standard and the phone error rate (PER) was calculated. PER is the share of misclassified phones in per cent. PER is thus calculated with equation 5.13, where N is the total number of phone classifications performed by the statistical decoder and E is the number of misclassified phones (as compared to the gold standard). Since the $\emptyset$ symbol is treated as any other phoneme symbol, a phone error can be either a substitution or a deletion. When going from a canonical pronunciation-dictionary representation of a word to a representation corresponding to spontaneous speech, it is extremely rare to find inserted segments in central standard Swedish and the current transcription system thus does not handle insertions.

$$PER = 100\,\frac{E}{N} \tag*{eq. 5.13}$$

The realisation lists were updated in several steps to minimise the PER. That is, realisations whose inclusion in a list gave rise to more errors in the resulting automatic transcriptions than the errors its inclusion corrected (in relation to the canonical representation) were excluded. Some minor deviations from this general rule were made when it was judged that the results would not generalise from the small sample gold standard to the entire data set. The final list of realisations for each phoneme is shown in Table 5.5.

Table 5.5: *Sets of possible realisations of phonemes.*

Cons.	Realisations	Vowel*	Realisations	Vowel†	Realisations
p	p	ə	ə		
t	∅, ʈ, t	a	ə, a	a	a
k	∅, k	ɑː	ə, a, ɑː	ɑː	a, ɑː
b	b	e	ə, e	e	e
d	∅, ɹ, ɖ, d	eː	e, eː	eː	e, eː
g	∅, g	ɪ	ə, ɪ	ɪ	ɪ
f	∅, f	iː	ɪ, iː	iː	ɪ, iː
v	v	ʊ	ə, ʊ	ʊ	ʊ
s	s, ʂ	uː	ə, ʊ, uː	uː	ʊ, uː
ɧ	ɧ	ɵ	∅, ɵ	ɵ	ɵ
ç	ç	ʉ̟ː	ɵ, ʉ̟ː	ʉ̟ː	ɵ, ʉ̟ː
h	∅, h	ʏ	ʏ	ʏ	ʏ
m	m	yː	ə, ʏ, yː	yː	ʏ, yː
n	ŋ, ɳ, m, n	ɔ	∅, ə, ɔ	ɔ	ɔ
ŋ	ŋ	oː	ə, ɔ, oː	oː	ɔ, oː
l	∅, ɭ, l	ɛ	∅, ə, ɛ	ɛ	ɛ
j	∅, j	ɛː	ə, ɛ, ɛː	ɛː	ɛ, ɛː
ɹ	∅, ɹ	æ	ə, æ	æ	æ
ʈ	ʈ	æː	ə, æ, æː	æː	æ, æː
ɖ	ɖ	œ	∅, ə, œ	œ	œ
ɭ	∅, ɭ	øː	œ, øː	øː	œ, øː
ɳ	∅, ɳ	œ̞	∅, ə, œ̞	œ̞	œ̞
ʂ	∅, ʂ	œ̞ː	ə, œ̞, œ̞ː	œ̞ː	œ̞, œ̞ː

*In unstressed syllable, †In stressed syllable

The /ɹ/ phoneme is often denoted /r/ for central standard Swedish. However, although the realisation of the phoneme as a trill, [r], is possible and occurring, this is a relatively infrequent realisation. Other allophones, such as the approximant [ɹ], the tap [ɾ] or the retroflex fricative [ʐ] are more frequent. There is no phonemic difference between the realisations and the phonetic alphabet used only includes one [ɹ] allophone symbol. In this thesis, the allophone [ɹ], judged to be the most frequent realisation, is used to denote the phoneme class /ɹ/ in the IPA format.

5.7.4 Statistical Decoding

Finite state transition networks representing the possible realisations of a word are built using the empirically compiled context-insensitive list of possible realisations for each phoneme (cf. Table 5.5). Figure 5.3 is an example of a finite state transition network used for the current annotation. In the network, arc labels refer to phoneme realisations and state labels refer to phoneme positions relative to the word onset.

Statistical decoding is conducted in a word-by-word manner, forcing phoneme boundaries at the manually annotated word boundaries. The part of the speech signal corresponding to a specific word is parameterised to form a sequence of observations using the Snack sound toolkit (Sjölander and Beskow, 2000; Sjölander, 2003). Viterbi decoding is used to find the path through the network with the highest probability of having produced the observation sequence and the corresponding phone sequence (aligned to the signal) is the output of the statistical decoder. In a post processing step, the phone string is aligned to the phoneme string using phoneme position indices and ∅ *'null'* place filler phones.

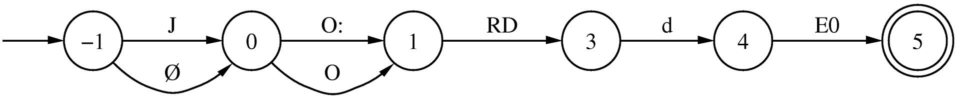


Figure 5.3: *Finite-state transition network representing the possible realisations of the word* gjorde 'did'*. The phone labels of the HMMs associated with the arcs between the states of the network are in the STA format (cf. Table A.1). A sequence of two HMMs is used for the phone* [ɖ]*, the first HMM* (RD) *representing the occlusion phase and the second HMM* (d) *representing the explosion phase. The acoustic models of all plosive consonants are composed of two separate HMMs.*

5.7.5 A Posteriori Correction Rules

The tentative phone string resulting from the statistical decoding process can be viewed as the result of a set of phonological transformation rules operating on the canonical phoneme string. A set of a posteriori rules inverting some of these phonological rules under certain conditions has been developed to correct some systematic errors made by the statistical decoder. The a posteriori correction rule set also includes some phonological rules, e.g. elision rules. This means that the sets of possible realisations of phonemes resulting from the hybrid system in some cases were larger than the sets listed in Table 5.5.

Both the phonological rules and the inverted phonological rules can utilise phoneme context (including word stress annotation) and tentative phone context. They can also use estimated phoneme duration and tentative phone duration as context. Some special rules for high frequency function words even use the orthography as context. A rule may be duration-independent or duration-dependent. A duration-independent rule is applied regardless of the estimated phoneme duration and phone duration and a duration-dependent rule is only applied when the estimated durational context is appropriate. By separating duration-independent and duration-dependent processes, the a posteriori correction rules are able to utilise the information from the statistical decoding maximally to improve the phonetic transcripts.

The rules were compiled using same three minutes of the manually transcribed gold standard that was used for realisation list development as a development corpus. For each phoneme in the canonical representation, the gold standard phone and the phone produced by the statistical decoder were compared and each type of deviation from the gold standard was investigated thoroughly. Rules were written to minimise PER, however with the restriction that the rules should be generally applicable.

5.7.6 Transcription System Evaluation

The automatic transcription system was evaluated against a small manually transcribed gold standard, including the first minute of speech from five randomly selected speakers from the VaKoS database. The gold standard transcripts from three speakers (2, 5 and 6) were used during correction rule development. The gold standard transcripts from the remaining two speakers (1 and 4) were used at evaluation only. The evaluation results show similar PERs and error distributions for the evaluation gold standard as for the development gold standard, both generally and when separating different speakers.

Table 5.6 shows the results from the evaluation. It can be seen that statistical decoding alone gives a higher phone error rate (PER) than estimating the phonetic transcript with the phoneme string. However, the errors made by the statistical decoder are *systematic* to a high degree and this fact is utilised at correction rule application. The final hybrid transcription system produces an overall PER of 15.5%, which is an error reduction by 40.4% compared to using the phoneme string for estimating the phone string.

Since manual transcription is restricted by a relatively small set of phone symbols, some decisions about phone identity are not obvious, most notably many cases of choosing between a full vowel and a [ə]. Defaulting to the system decision whenever a human transcriber is forced to make ad hoc decisions would increase the speed of manual transcript checking and correction considerably without lowering

Table 5.6: *Phone error rates (PER) when estimating the phone string with the phoneme string, the statistical decoder, and the hybrid automatic transcription system, respectively, and reduction in PER when switching from the phoneme string to the hybrid system output.*

Speaker index	$\text{PER}_{\text{Phoneme}}$	$\text{PER}_{\text{Statistical}}$	$\text{PER}_{\text{Hybrid}}$	Error reduction
2	26.51%	30.96%	**14.95%**	43.62%
5	21.68%	30.07%	**15.21%**	29.84%
6	29.47%	27.62%	**14.73%**	50.00%
1	25.14%	32.82%	**13.24%**	47.33%
4	27.37%	36.84%	**19.12%**	30.13%
$\bar{x}$	26.01%	31.69%	**15.50%**	40.42%

the quality of the resulting transcript. It is worth noting that if this strategy had been used for compiling the gold standard transcript, the PER would have been somewhat lower. The 15.5% PER is thus a slight under-estimation of the system performance. Manual correction of the automatically obtained transcripts was not possible for practical reasons. However, manual correction will most likely result in more accurate pronunciation models.

5.7.7 Phone Confusions

Tables 5.7 and 5.8 are confusion matrices showing the distributions of realisations produced by the hybrid transcription system for each gold standard phone (calculated over the complete set of gold standard transcripts).

As discussed, attempts have been made to reduce each type of confusion with a posteriori correction rules. However, minimising one type of error mostly increases errors in the opposite direction. There is thus a trading relation between e.g. [a]-for-[ə] confusions and [ə]-for-[a] confusions. The strategy has been to find the set of rules resulting in the best over-all correspondence between the gold standard transcripts and the hybrid transcription system output.

The errors remaining in the output of the final hybrid transcription system are errors that proved hard to avoid. For example, the phones [g] and [j] are often erroneously elided. These phones are hard for the statistical decoder to detect and their correct realisation in the automatic transcript depends heavily on the correction rules. To include the position in the word and in the syllable as context in the correction rules and to exploit information about word identity more may be beneficial for reducing these errors further.

For example, [g] should very seldom be elided in word initial position or generally in syllable onsets. A canonical /g/ is often not realised in suffixes such as *-ligt*, *-igt* and *-igen* and is also often elided in many function words, e.g. *jag 'I'*, *något 'something'* and *någon, 'someone'*.

Another type of error apparent from Table 5.7 is that retroflex consonants are often substituted for their dental counterparts.

Table 5.7: *Consonant confusion matrix for the hybrid automatic transcription system. Each row corresponds to a gold standard phone and each column shows the share of times the transcription system gave the phone output of the column for a particular gold standard phone. No realisation (∅) can be confused with both vowels and consonants. The approximant [j] has a .03 share of confusions with the vowel [ɪ] and a .03 share of confusions with the vowel [ə]. Otherwise, if row sums are not 1, it is because the numbers are rounded.*

phone	∅	p	t	k	b	d	g	f	v	s	ɧ	ç	h	m	n	ŋ	l	j	ɹ	ʈ	ɖ	ɭ	ɳ	ʂ
∅	.62		.06	.01		.03			.01				.02	.02	.02		.02	.01	.06					
p		1																						
t	.04		.95																	.01				
k				1																				
b					1																			
d	.10		.03			.81													.07					
g	.29						.71																	
f	.05							.95																
v	.10								.90															
s	.01									.99														.01
ɧ											1													
ç												1												
h	.07												.91					.02						
m	.05													.94	.01									
n	.01													.01	.96	.01							.02	
ŋ																1								
l	.01																.99							
j	.25																	.69						
ɹ	.04					.06													.90					
ʈ			.40			.20														.40				
ɖ						.31													.12		.56			
ɭ																	1							
ɳ															.67								.33	
ʂ										.11														.89

From table 5.8, it can be seen that the phones [ʏ], [æː] and [œ̞ː] are correctly predicted by the transcription system over the small gold standard to 100%. However, as can be seen from Table E.1 in Appendix E (showing the number of instances of each phone in the gold standard transcript), there are relatively few instances in the gold standard transcript of one or both vowels from the long-short pairs including these vowels. It is thus not possible to draw any general conclusions about them.

Table 5.8 also shows that confusions between long vowels and their short counterparts are common. Long vowels are often substituted for their short counterpart and, although to a lesser degree, short vowels are substituted for their long counterparts. An exception from this general rule is the short vowel [a] and its long counterpart [ɑː]; an [ɑː] is never substituted with an [a]. Both [a] and [ɑː] are fre-

Table 5.8: *Vowel confusion matrix for the hybrid automatic transcription system. Each row corresponds to a gold standard phone and each column shows the share of times the transcription system gave the phone output of the column for a particular gold standard phone. No realisation (∅) can be confused with both vowels and consonants. Otherwise, if row sums are not 1, it is because the numbers are rounded.*

phone	∅	ə	a	ɑː	e	eː	ɪ	iː	ʊ	uː	ɵ	ʉ̟ː	ʏ	yː	ɔ	oː	ɛ	ɛː	æ	æː	œ	øː	œ̞	œ̞ː
∅	.62	.06			.01		.01								.02									
ə	.01	.79	.09	.02	.02		.04								.03	.01	.01							
a		.16	.79	.04												.01								
ɑː		.02		.98																				
e		.17			.78	.03	.02																	
eː		.03			.11	.84													.03					
ɪ		.14					.80	.07																
iː							.09	.91																
ʊ		.08		.08					.75	.08														
uː									.15	.85														
ɵ		.05									.90	.05												
ʉ̟ː		.04									.04	.92												
ʏ													1											
yː													1											
ɔ	.04	.08													.82	.05								
oː															.17	.82								
ɛ		.15															.79	.03		.03				
ɛː		.07															.07	.64	.07	.14				
æ		.20															.10		.50	.20				
æː																				1				
œ																					.57		.43	
øː																					.33	.67		
œ̞																					.22		.78	
œ̞ː																								1

quently occurring phonemes, which indicates that their separation is not due to chance.

Vowel length is a phonological feature in Swedish. However, a length difference is mostly accompanied by a shift in articulatory position. This shift is most noticeable for the [a]-[ɑː] pair. The two vowels differ not only in tongue position, but also in lip rounding. This explains the fact that [ɑː] is clearly separable from [a] by the statistical decoder.

A common type of vowel confusion is that between a full vowel and [ə]. The phones [a], [e], [ɪ], [ɛ] and [æ] are the vowels most often substituted for a [ə] and [ə] is often substituted for [a], [ɪ] or [ɔ].

The differences between [ɛ] and [æ] and between [œ] and [œ̞], respectively, are not phonemic in Swedish, but rather allophonic. The same goes for the long counterparts of these vowels, [ɛː] and [æː] and [øː] and [œ̞ː], respectively. In the canonical pronunciation representations used in the current work, the more open allophones are used in pre-/ɹ/ position and the more closed allophones are used otherwise. However, in colloquial speech, a canonical /ɹ/ may be elided while the preceding vowel keeps its open quality to a higher of lesser degree, indirectly signalling the presence of an /ɹ/.

This means that there is often a degree of uncertainty for a human transcriber of how to label a vowel somewhere on a continuous scale between e.g. [œ] and [œ̞], preceding an /ɹ/ with no overt realisation. Both the human labeller and the transcription system is thus operating in a grey area when transcribing these speech sounds and this may account for the confusions between these allophones with a more open and a more close articulation, respectively. However, since the number of instances of these phones in the gold standard transcripts is low, these confusions may not be as frequent as they appear.

Further, also when an [ɹ] is clearly audible in the speech signal, the [ɹ] sound is often superimposed on an adjacent vowel or on adjacent vowels in Swedish and during the compilation of a representation of pronunciation using a sequential string of separate phoneme symbols, it can often be debated whether an [ɹ] is present or not. Both erroneous [ɹ]-for-∅ and ∅-for-[ɹ] substitutions are common in the transcription system output over the gold standard. Table E.1 in Appendix E shows that ∅ and [ɹ] are high frequency units in the gold standard transcript.

Other frequent substitutions are [ɹ] for [d] and [d] for [ɹ] and this may be due to the fact that a /d/ is sometimes pronounced as an [ɹ] in reduced speech, especially for high frequency function words such as *den 'it'*, *det 'it'* and *du, 'you'* and especially following a word ending in an /ɹ/ (which often happns, since verbs in present tense end with an /ɹ/). Once again, there is a gray area between a prototypical [d] and a prototypical [ɹ] and the scale between the two is continuous.

5.8 Summary

In this chapter, methods used for annotating speech data for discourse context-dependent pronunciation modelling have been presented. Speech databases, a multi-layer annotation system and methods employed for segmenting the layers into their respective utterance types have been described. Annotation is associated with units in six linguistically motivated layers, 1) a discourse layer, 2) an utterance layer, 3) a phrase layer, 4) a word layer, 5) a syllable layer, and 6) a phoneme layer.

The methods for calculating mean phoneme duration measures, pitch dynamics measures, pitch range measures and word predictability measures have been described. Mean phoneme duration measures are calculated over units in all layers but the phoneme layer. The measures are based on absolute and normalised duration, respectively, on linear and logarithmic duration, respectively, and on the durations of all phonemes and on vowels only.

The pitch dynamics measures are based on the distance between f_0 peaks and valleys and either the median f_0 over the unit for which the measure is calculated or a *base frequency* located 1.5σ below the mean frequency of the particular speaker. Pitch range is the difference between the highest f_0 peak and the lowest f_0 valley contained by the unit over which the measure is calculated. The pitch measures are calculated from pitch measured in Hertz and on several psychoacoustic scales.

Word predictability is defined as the weighted combination of the trigram, bigram and unigram probabilities calculated from the orthographic transcript of a spoken language corpus.

Further, the development and evaluation of a hybrid system for automatic phonetic transcription has been presented. The system uses statistical decoding and a set of a posteriori correction rules and supplies a context-dependent realisation of each phoneme in a canonical pronunciation representation. The automatically obtained phones were generated to be used as keys during pronunciation model training. Compared to a small, manually transcribed gold standard, the automatic transcription system produced a phone error rate of 15.5%.

The next chapter gives a detailed description of the information associated with the units in each layer of annotation, among others, the variables described in this chapter.

Chapter 6

Information Included in the Annotation

Values for a set of variables hypothesised to be important for predicting the realisation of a phoneme in its discourse context is attached to each unit at each layer of annotation. This chapter presents the information attached to the units at each respective layer.

6.1 The Discourse Layer

In the discourse layer, variables which are constant over the discourse are annotated. A set of 'inverted speech rate' measures based on the global *mean phoneme duration* is attached to each discourse layer unit. The details of how these measures are calculated were explained in Chapter 5, Section 5.4. The discourse layer information also includes four speaking style-related variables: *number of discourse participants*, *degree of formality*, *degree of spontaneity* and *type of interaction*. Table 6.1 gives a summary of the discourse layer annotation.

Table 6.1: *Discourse layer annotation.*

Variable	Values
Number of discourse participants	*monologue*, *two-part dialogue*, *multi-part dialogue*
Type of interaction	*human-directed*, *computer-directed*
Degree of formality	*formal*, *informal*
Degree of spontaneity	*spontaneous*, *elicited*, *scripted*, *acted*, *read*
Mean phoneme duration	Several continuous measures, $\mathbb{R}$

The variable *number of discourse participants* can take one of three different values: *monologue*, *two-part dialogue* or *multi-part dialogue*. It is hypothesised that this *one*, *two* or *many* distinction will give a sufficient separation of the dimensions of speaking style originating from the number of discourse participants in the current context. The categories into which speaking style-related variables can be divided are by nature rather coarse and the general approach has been to use a small set of categories. The variable *type of interaction* simply divides interactions into *human-directed* and *computer-directed*. The variable *degree of formality* is a coarse division of discourses into two groups, *formal* and *informal*.

Degree of spontaneity is a five-way variable, taking the values *spontaneous*, *elicited*, *scripted*, *acted* and *read*. Spontaneous speech is, in this context, defined as completely free and uncontrolled, while elicited speech is somehow evoked, e.g. by an interviewer asking questions or a subject being asked to talk about some specific topic. Elicited speech is, however, not based on some written or spoken script. Scripted speech may be a subject retelling a written or spoken text, however not being forced to exactly follow the script. Acted speech is speech closely following a written script, although with acted emotion. Finally, read speech is the result of reading a written text aloud in a 'neutral' fashion.

The variables and their sets of values are designed to be usable for a wider range of speech material than that actually used in the models developed. There are thus discourse layer variables with values not used for the speech data annotated. For the variable *number of discourse participants*, only the values *monologue* and *multi-part dialogue* are actually used. Further, for the variable *type of interaction* only the value *human-directed* is used, making the variable currently redundant. For the variable *degree of spontaneity*, only the values *elicited* and *scripted* are used. This redundancy of values and variables is due to the fact that the original plan was to annotate more databases than were actually annotated.

Speech databases available from the Grog project, but not actually used, were radio entertainment with spontaneous dialogue originating from *Sveriges radio* (Swedish public service radio) and talking books of including 'acted' readings of children's books and 'neutral' readings of fact literature originating from the *Swedish Library of Talking Books and Braille*. The original plan was to also include the annotation of recorded speech data from Wizard of Oz and actual dialogue system interaction from various dialogue system projects centred at the Department of Speech, Music and Hearing and the Centre for Speech Technology (CTT) at KTH. Data is available from four such projects: Waxholm, Gulan, August, adApt (cf. e.g. Gustafson, 2002; Bell, 2003). Finally, annotation of the Swedish map task corpus Smtc (Helgason, 2006) was planned.

6.2 The Utterance Layer

In the utterance layer, mostly speaker attributes are annotated. Table 6.2 gives a summary of the utterance layer annotation. *Speaker pitch register* is a binary variable that differentiates speakers with a high pitch register (90–600 Hz) from speakers with a low pitch register (60–300 Hz). This variable may interplay with measures based on pitch movement.

Dialogue act classification has been shown to be beneficial for reducing the word error rate in automatic speech recognition, (cf. e.g. Jurafsky et al., 1997). For the annotation presented in this thesis, no detailed dialogue act classification is made. However, a coarse four-way division into *utterance types*, corresponding to basic dialogue acts, is used to take the influence of discourse structure into account in dialogue data. Utterances are classified as belonging to one of the four types *statement*, *question/request response* (a question or an utterance which is not an explicit question, but still a request for a response), *answer/response* (an answer to a question or a response to a request) or *feedback*. For monologues, the default utterance type is *statement*. A set of *mean phoneme duration* measures over the utterance and sets of *pitch range* and *pitch dynamics* 'speech liveliness' measures (cf. Chapter 5, Section 5.5) are also included in the utterance layer annotation.

Table 6.2: *Utterance layer annotation.*

Variable	Values
Speaker pitch register	*high*, *low*
Utterance type	*statement*, *question/request response*, *answer/response*, *feedback*
Pitch dynamics	Several continuous measures, ℝ
Pitch range	Several continuous measures, ℝ
Mean phoneme duration	Several continuous measures, ℝ

Speaker age, dialect and social factors all influence spoken language performance. However, the speakers used for the current pronunciation modelling project are all part of a very coherent group from the perspectives of dialect, sociolect and age. The speakers are all university-educated adults below the age of retirement and they are all speaking the central standard variety of Swedish (or a similar variety). Pronunciation variation due to dialectal, social and age factors are thus not modelled in the current effort.

6.3 The Phrase Layer

The possible values of the *phrase type* variable are shown in Table 6.3. The interpretation of some of these variables is more or less obvious, while other may require some further explanation. A short explanation of some of the phrase types are given below. For more information on the phrase types, cf. Megyesi (2002b). As explained in Chapter 5, Section 5.3, the phrase type *verb cluster* is a continuous sequence of verbs belonging to the same verb phrase and and an *infinitive phrase* is a verb in the infinitive proceeded by an infinitive particle.

Some word units (mostly conjunctions) do not belong to any phrase. These units are annotated with a *no phrase* tag in the *phrase type* annotation. Only maximal phrases are considered. A noun phrase can include modifiers of different types, e.g. nouns, adjective phrases, prepositional phrases. In these cases, the entire maximal projection of the noun phrase is counted as a single noun phrase and the identity and boundaries of the constituents are ignored. Similarly, conjoined adjective phrases are counted as a single adjective phrase.

A set of *phrase length* measures are calculated for each phrase unit: the number of *words*, *syllables*, and *phonemes*, respectively, contained by the phase. Also, a three-way length label (*long*, *medium*, *short*) based on the syllable count is included in the phrase layer annotation.

Two measures associated with the *prosodic weight* of each phrase are calculated: the number of *stressed syllables* and the number of *focally stressed words* included in the phrase. Focal stress annotation has been manually supplied for the VaKoS database and one of the Radio Interviews. In cases where no focal stress annotation is available, an 'unknown value' tag has been used. Automatic focal stress detection built on overall intensity and spectral emphasis (cf. e.g. Campbell, 1995; Strangert and Heldner, 1995; Fant et al., 2001; Heldner et al., 1999; Heldner, 2003) have been attempted, but no such system has been used for the annotation described in this thesis.

Pitch dynamics and *pitch range* measures are calculated over each phrase as described in Chapter 5, Section 5.5. Details about the *mean phoneme duration* measures are presented in Chapter 5, Section 5.4. The complete list of variables included in the phrase layer annotation and their possible values are shown in Table 6.3.

Table 6.3: *Phrase layer annotation.*

Variable	Values
Phrase type	*adverb phrase*, *adjective phrase*, *noun phrase*, *prepositional phrase*, *verb cluster*, *infinitive phrase*, *numeral expression*, *no phrase*
Phrase length (words)	Continuous, $\mathbb{Z}$
Phrase length (syllables)	Continuous, $\mathbb{Z}$
Phrase length (phonemes)	Continuous, $\mathbb{Z}$
Phrase length Label	*long*, *medium*, *short*
Prosodic weight (stresses)	Continuous, $\mathbb{Z}$
Prosodic weight (foci)	Continuous, $\mathbb{Z}$
Pitch dynamics	Several continuous measures, $\mathbb{R}$
Pitch range	Several continuous measures, $\mathbb{R}$
Mean phoneme duration	Several continuous measures, $\mathbb{R}$

6.4 The Word Layer

Since the word is the principal conveyor of meaning in language and also the principal syntactic unit, there is a large variety of variables that can be included in the word layer. The complete list of variables included in the word layer annotation and their possible values are shown in Table 6.4.

Table 6.4: *Word layer annotation.*

Variable	Values
Part of Speech	*adverb*, *determiner*, *wh-adverb*, *wh-determiner*, *wh-pronoun*, *possessive wh-pronoun*, *infinitival* marker, *interjection*, *adjective*, *conjunction*, *noun*, *participle*, *verb particle*, *proper name*, *pronoun*, *preposition*, *possessive pronoun*, *cardinal number*, *ordinal number*, *subjunction*, *foreign word*, *verb*
Morphology (gender)	*common*, *neutre*, *masculine*, *unspecified*, *no value*
Morphology (number)	*singular*, *plural*, *unspecified*, *no value*
Morphology (definiteness)	*indefinite*, *definite*, *unspecified*, *no value*
Morphology (case)	*nominative*, *genitive*, *no value*
Morphology (pronoun form)	*subject*, *object*, *unspecified*, *no value*
Morphology (tense/aspect)	*present*, *preterite*, *infinitive*, *imperative*, *supinum*, *perfect*, *no value*
Morphology (mood)	*conjunctive*, *no value*
Morphology (voice)	*active*, *passive/s-form*, *no value*
Morphology (degree)	*positive*, *comparative*, *superlative*, *no value*
Word type	*content word*, *function word*
Function word	*content word*, set of function words
Word predictability	Continuous, $\mathbb{R}$
Global word probability	Continuous, $\mathbb{R}$
Position in phrase	*initial*, *medial*, *final*
Position in collocation	*initial*, *medial*, *final*
Word repetitions (full-form)	Continuous, $\mathbb{Z}$
Word repetitions (lexeme)	Continuous, $\mathbb{Z}$
L-adjacent filled pause	*yes*, *no*
R-adjacent filled pause	*yes*, *no*
L-adjacent interrupted word	*yes*, *no*
R-adjacent interrupted word	*yes*, *no*
L-adjacent prosodic boundary	*strong*, *weak*, *no*
R-adjacent prosodic boundary	*strong*, *weak*, *no*
L-adjacent pause	*yes*, *no*
R-adjacent pause	*yes*, *no*
L-adjacent pause duration	Two continuous measures, $\mathbb{Z}$
R-adjacent pause duration	Two continuous measures, $\mathbb{Z}$
Word length (syllables)	Continuous, $\mathbb{Z}$
Word length (phonemes)	Continuous, $\mathbb{Z}$
Word length label	*long*, *medium*, *short*
Focal stress	*focally stressed*, *not focally stressed*, *unknown*
Dist. to previous focus (words)	Continuous, $\mathbb{Z}$
Dist. to next focus (words)	Continuous, $\mathbb{Z}$
Pitch dynamics	Several continuous measures, $\mathbb{R}$
Pitch range	Several continuous measures, $\mathbb{R}$
Mean phoneme duration	Several continuous measures, $\mathbb{R}$

Part of Speech and morphological information from the tagger (cf. Chapter 5, Section 5.3) is included in the annotation in the SUC format (Ejerhed et al., 1992). *Morphology* is included as a set of tags corresponding to different morphological dimensions. For morphological dimensions not used in the description of the Part of Speech of a particular word unit, a *no value* tag is used.

Based on the Part of Speech tags, a division of words into *word types* (*content word* or *function word*) is made. All words tagged as nouns, proper names, adverbs, adjectives, participles, cardinal numbers and ordinal numbers are classified as content words. Words tagged as verbs are classified as function words if they 1) belong to the closed class of copula verbs (defined by their orthographic form) or 2) are used as auxiliary verbs (i.e., if they are not the final verb in the verb clusters or infinitive phrases in which they occur). Otherwise, verbs are classified as content words. A word tagged as any other Part of Speech is classified as a function word.

A similar variable denoted *function word* has the entire closed set of function words and a generic 'content word' representation as its possible values. There are pronunciation variation strategies specific to certain function words and the *function word* variable should be a strong predictor of this behaviour.

A set of word predictability-related variables are included in the word layer annotation, including a measure simply called *word predictability* (based on a weighted combination of trigram, bigram and unigram probabilities), a measure estimating *global word (unigram) probability*, the position of the word in the *phrase* and in a *collocation*, respectively, the number of repetitions of the *full-form word* and of the *lexeme* thus far in the discourse. A detailed account of these word predictability-related measures was given in Chapter 5, Section 5.6.

The presence of a *filled pause*[1] immediately succeeding or preceding the current word may also be of importance for the pronunciation of the current word (cf. e.g. Jurafsky et al., 1998a). Information about the presence of a filled pause in these two positions is thus included in the annotation. Interrupted (not articulatorily completed) words and other types of "disfluencies" have been shown to have an effect on adjacent words (e.g. Shriberg, 1999; Eklund, 1999). For this reason, information about the presence of *interrupted words* immediately succeeding or preceding the current word is included in the annotation.

Prosodic boundaries are important for grouping coherent subunits in the speech signal. For listeners, this grouping facilitates parsing the sound stream. Speakers have a number of parameters at their disposal for signalling prosodic boundaries, e.g. f_0, segment duration, intensity and the use of silent pauses. For example, Horne et al. (1995); Heldner and Megyesi (2003); and Heldner et al. (2004) have described how prosodic boundaries can be automatically estimated using f_0 resets, final lengthening, pauses etc. Gustafson-Čapková and Megyesi (2002) studied i.a. the correlation between different kinds of perceived boundaries and acoustic pauses using several types of context, e.g. discourse boundaries, syntactic bound-

[1] A filled pause is a "hesitation sound" used by a speaker to signal that the utterance will continue, although the speaker has momentarily stopped speaking, e.g. to think.

aries and Part of Speech. It would thus be possible to use an automatic prosodic boundary guesser as a support for annotation. However, for the speech data annotated, manual boundary annotation has been supplied and this is the annotation used. In the annotation, prosodic boundaries can be of two types, *strong* and *weak*. The variables *adjacent prosodic boundary (left)* and *adjacent prosodic boundary (right)* can thus take the values *strong*, *weak*, and *no*.

Tabain et al. (2001) showed a correlation between adjacent boundaries of different types and phonetic realisation. In this study, the effects of boundaries on phoneme durations and formants were studied. However, it is likely that boundaries can also be used to predict changes in phone identity.

Information about the presence of *pauses* adjacent to the current word and about the duration of adjacent pauses may also be important for predicting the realisation of the word. Two *adjacent pause duration* measures are included in the annotation: absolute duration and normalised duration. The latter measure relates the pause duration to the mean duration of all pauses in the database and hence, to the speaking style. The normal transformation or Z normalisation used is shown in Equation 6.1, where x is the pause duration, μ is the mean pause duration over the data set and σ is the standard deviation of the pause duration. If there is no pause, the absolute pause duration is set to 0. The normalised pause duration depends on the mean and standard deviation of the pause durations in the data set. Thus, the minimum normalised pause duration is not an absolute minimum, but specific for the database. Zero durations (i.e., no pause) are not included when calculating the mean and standard deviation used for normalisation.

$$Z = \frac{x - \mu}{\sigma} \qquad \text{eq. 6.1}$$

Three *word length* measures are included in the annotation: a *syllable count* (the number of syllables contained by the word), a *phoneme count* (the number of phonemes contained by the word) and a three-way *word length label* based on the syllable count with the possible values *long*, *medium*, and *short*. The counts are based on the canonical phonemic word representations.

Focal stress may be an important variable for predicting word realisation, since placing stress on a word is to make it more salient; to make it stand out from the surrounding sound stream. In the focal stress dimension, each word is classified as either *focally stressed* or *not focally stressed*. For the VaKoS database and one of the radio interviews, manual *focal stress* annotation was available. This information was included in the annotation. As mentioned in Section 6.3, it would be possible to use automatic focal stress detection built on e.g. overall intensity and spectral emphasis (Campbell, 1995; Fant et al., 2001; Heldner et al., 1999; Heldner, 2003) to facilitate annotation when manual annotation is not available. However, no attempt has been made to build or use an automatic focal stress detector for the annotation reported here. Hence, for the remaining speech data, the value of the *focal stress* variable is set to *unknown*.

Since the realisation of focal stress is largely a question of contrast, the realisation of a word may be dependent on e.g. the focal stress status of adjacent words, and the position of the word relative to preceding and succeeding focally stressed words. Both stressed and unstressed words may be realised differently depending on their stress context. Thus, the distances to the preceding and to the succeeding focally stressed word (in number of words) are included in the word layer annotation.

Some measures of *pitch dynamics* and *pitch range* over each word unit are included in the annotation, cf. Chapter 5, Section 5.5. The *mean phoneme duration* over the word unit is measured in several ways, cf. Chapter 5, Section 5.4.

6.5 The Syllable Layer

The variables included in the syllable layer annotation are presented in Table 6.5. Information about the stress and accent of the current syllable is derived from the phonemic representations. Swedish has two different types of word stress, *accent I* and *accent II*. In central standard Swedish, *accent I* has a single stressed syllable while *accent II* has a primary and a secondary stress. There is also a special compound accent similar to *accent II*, with primary stress on the first compound constituent and a secondary stress on the last compound constituent. The *stress* annotation is a simple division between stressed and unstressed syllables, while the *stress type* annotation takes the word accent into account, thus making the *stress type* classification a division into finer stress type classes.

Table 6.5: *Syllable layer annotation.*

Variable	Values
Stress	*stressed*, *unstressed*
Stress type	*no stress*, *(primary) stress in accent I word*, *primary stress in accent II word or compound*, *secondary stress in accent II word*, *secondary stress in compound*
Dist. to prev. stress (syll:s)	Continuous, $\mathbb{Z}$
Dist. to prev. prim. stress (syll:s)	Continuous, $\mathbb{Z}$
Dist. to next stress (syll:s)	Continuous, $\mathbb{Z}$
Dist. to next prim. stress (syll:s)	Continuous, $\mathbb{Z}$
Syllable length (phonemes)	Continuous, $\mathbb{Z}$
Syllable nucleus	Vowel symbols (cf. Table A.1 in Appendix A)
Position in the word	*initial*, *medial*, *final*
Mean phoneme duration	Several Continuous measures, $\mathbb{R}$

For example Bruce (1986) and Greenberg (2003) argue for a syllable-centric view on pronunciation variation. Greenberg (2003) especially points out the prosodic prominence of the syllable and the ordinal position of phonemes/phones (phonetic constituents) in the syllable as important factors for phonetic realisation. The variables emphasised by Greenberg (2003) are included in the current annotation. Variables relating to syllable prominence are annotated here, in the syllable layer (stress, stress type, and distances to preceding and succeeding stresses) and in the word layer (focal stress). Information about the position of the phoneme in the syllable is included in the phoneme layer annotation.

The distances to the nearest *preceding stressed syllable* and to the nearest preceding syllable with *primary stress* (measured in number of syllables) are included in the syllable layer annotation. The distances to succeeding stresses are also included. The word stress positions are fixed for Swedish words. However, the realisation of word stress is relative to e.g. the stress context. The idea behind including the distances to previous and succeeding stresses is that this will give a picture of word stress with higher resolution than information about the stress of the current syllable alone can give.

Syllable length is measured as the number of phonemes in the canonical phonemic representation of the syllable. *Syllable nucleus* is included in the syllable layer annotation, since there is a chance that the pronunciation variation pattern of a syllable is dependent on its nucleus. On the phonemic description level, only vowels can constitute syllable nuclei in central standard Swedish.

The initial and final syllables of a word are generally less prone to syllable reduction than medial syllables, which makes the *position of the syllable in the word* an important variable to include in the annotation. The position is annotated as a three-way variable, where each syllable is categorised as either *initial*, *medial* or *final*. The value used for one-syllable words is *initial*. The *mean phoneme duration* over the syllable is calculated as described in Chapter 5, Section 5.4.

6.6 The Phoneme Layer

The variables in the phoneme layer annotation and their values are shown in Table 6.6. The *phoneme identity* is represented by a phoneme symbol from the available phoneme set (cf. Table A.1 in Appendix A). In the segmented phoneme layer, there may also be <sil> and <junk> labels. These are used for non-speech sounds and are not treated as phonemes (the models induced from the annotation are not trained to predict the realisation of <sil> or <junk>, but the labels are provided as context for the model to predict realisations of proper phonemes, as described in Chapter 7).

A set of articulatory features describing the phoneme is associated with each phoneme unit. Five feature dimensions, shared by consonants and vowels, are used. The *sonorant* and *phonological length* dimensions have values shared by consonants and vowels, while all other feature dimensions have separate sets of values for consonants and vowels, respectively. The *sonorant* variable can take the value *yes* (for vowels and semi-vowels) or *no* (for obstruents). Swedish has two phonological phoneme lengths: *long* and *short*. Vowel length is included in the canonical phonemic representations and consonant length is derived from the vowel length. In a syllable with a single coda consonant, the consonant will be short if the vowel is long and long if the vowel is short (cf. e.g. Elert, 1964). In the current annotation, onsets and coda clusters are treated as consisting of short consonants.

The *manner of articulation/frontness* dimension has the possible values *stop, fricative, nasal, approximant* and *lateral approximant* for consonants and the possible values *front, central* and *back* for vowels. The *place of articulation/openness* variable can take the values *bilabial, labiodental, alveolar, dental, retroflex, palatal, velar* or *glottal* for consonants and *close, close-mid, mid, open-mid* or *open* for vowels. The *voicing/lip rounding* dimension has the possible values *voiced* and *voiceless* for consonants and the values *rounded* and *unrounded* for vowels.

Table 6.6: *Phoneme layer annotation.*

Variable	Values
Phoneme identity	Phoneme set (cf. Table A.1 in Appendix A)
Sonorant	*yes*, *no*
Phonological length	*long*, *short*
Manner/frontness	*stop*, *fricative*, *nasal*, *approximant*, *lateral approximant*, *front*, *central*, *back*
Place/openness	*bilabial*, *labiodental*, *alveolar*, *dental*, *retroflex*, *palatal*, *velar*, *glottal*, *close*, *close-mid*, *mid*, *open-mid*, *open*
Voicing/lip rounding	*voiced*, *voiceless*, *rounded*, *unrounded*
Position in syllable	*onset*, *nucleus*, *coda*
Consonant cluster length	Continuous, $\mathbb{Z}$
Position in cluster	Continuous, $\mathbb{Z}$
Phone identity	Phoneme set (cf. Table A.1 in Appendix A), $\emptyset$

The *position of the phoneme in the syllable* has been shown to be an important factor for predicting the realisation of the phoneme (cf. e.g. Duez, 1998). Thus, information about in which part of the syllable (*onset*, *nucleus* or *coda*) the phoneme is located is included in the annotation.

For a consonant phoneme, the *length of the cluster* in which it appears and its *position in the cluster* may be important for its realisation. Hence, information about these variables is included in the phoneme layer annotation. The consonant cluster length value used for vowels is 0. Only consonants belonging to the same syllable as the current phoneme are counted as parts of the current cluster. That is, cluster boundaries are forced at syllable boundaries. The first position in the cluster is 1 and vowels receive the position value 0.

The phone identities are collected from the phone string supplied by the automatic transcription system. The phone label set is the same as the phoneme label set, with an additional $\emptyset$ label for phonemes with no overt realisation. If the identity of the current phoneme is represented by a <sil> label or a <junk> label, the respective label will be used also as the phone identity label. In the phone string, these labels only serve as place fillers indicating that the phoneme position is not occupied by an actual phone.

The accuracy of the phone identity labels could be improved by manually checking and correcting the phone string produced by the automatic transcription system. At manual correction, it would be possible to connect several phone units to a single phoneme unit to describe insertion phenomena (epenthesis). In such cases, the set of phones associated with a single phoneme position would be treated as a single multi-phone unit at model training. There will not be many occurrences of each multi-phone unit to train on, but since epenthesis occurs very infrequently in central standard Swedish, this does not pose a problem in practise.

6.7 Summary

This chapter has described the information associated with the units at each layer of annotation. Tables showing the variables included in the annotation of each layer and the possible values the variables were presented along with short descriptions of the variables and the motivation for including them in the annotation.

The next chapter will describe how this information is used for creating phoneme-sized training and validation instances, respectively, for decision tree induction and execution. The creation of decision tree pronunciation models from the training instances will also be described.

Chapter 7

Pronunciation Model Creation

Using the annotation from the speech databases, pronunciation models can be created with different types of machine learning methods. If a model is to be used for descriptive purposes, it must be transparent, i.e., it must contain information such that the model can be represented in a format interpretable by a human familiar with linguistic theory. A machine learning paradigm that creates transparent models and is suitable for the type of data at hand is the *decision tree induction* paradigm. A decision tree inducer commonly needs no ad hoc knowledge and can induce models directly from training data. It is thus easy to use once you have the data. For these reasons, the decision tree paradigm has been selected for creating the models reported in this thesis. It has not been tested whether the decision tree paradigm necessarily produces the best models. Other machine learning paradigms may be able to create more accurate models or models which meet certain application-specific demands.

7.1 Decision Tree Induction

A decision tree induction algorithm builds a tree level by level using training instances. Each training instance is a set of attribute values and a classification key. Induction starts from a root node containing all training instances. A number of tentative first tree levels are built, one for each attribute, by dividing the data set into a number of branches corresponding to the number of possible values of the specific attribute. The instances in the node of a specific branch thus have the same value on the attribute used for branching.

From the tentative first tree levels, the one that is optimal according to a given criterion (generally based on entropy minimisation) is selected. The process is then repeated for each node on the optimal tree level and a new level of nodes is thus created. When all examples in a node have the same classification or there is an insufficient number of instances in the node to continue the branching procedure or when some other stopping criterion is met, the tree is finished. Decision tree induction algorithms are thus greedy algorithms which always choose locally optimal sets of branches.

7.1.1 Training and Evaluation Data

For creating the decision tree pronunciation models presented in this thesis, training instances are compiled from the structured annotation. The training instances are phoneme-sized and can be seen as a set of *context-sensitive phonemes.* Each training instance includes a set of 516 attribute values and the phone realisation, which is used as the classification key. The features of the current unit at each layer of annotation are included as attributes in the training examples. Where applicable, information from the neighbouring units at each annotation layer is also included in the attribute sets. For example, the values of the Part of Speech and morphology variables of the words at positions $n \pm 4$ are included, n being the position of the current word in the word layer annotation. The values of the variables of the phonemes at positions $m \pm 4$, m being the position of the current phoneme in the phoneme layer annotation, are also included. For most other variables, a context range of ± 2 is used. Training instances are created for each unit in the phoneme layer annotation, except for the special units <sil> and <junk>. These units are, however, used in the phoneme context attributes.

The task of a finished decision tree model is to take instances in the same format as the training instances and make a decision about the appropriate phone realisation (which may be $\emptyset$) of each instance. The model will thus describe phone-level pronunciation only. The relation between a phoneme and its phone realisation can be seen as a phonological process. From a phonological point of view, the models describe processes affecting the presence or absence of phones and processes affecting the broad-phonetic phone identities. However, processes that do not change the broad-phonetic identities of phones, e.g. nasalisation and devoicing of certain phonemes in Swedish, are not handled by the models.

7.2 Pruning

Training data generally contain some degree of noise and a decision tree may be biased toward the particular noise in the data used for inducing the tree (overtrained). However, once a tree is constructed, it can be pruned to make it more generally applicable. The idea behind pruning is that the most common patterns are kept in the model, while less common patterns, with high probability of being due to noise in the training data, are disregarded. During pruning, a subtree of a particular node is replaced by a leaf (terminal node) with the most common class of the leaves governed by the subtree, when some criterion is met. A commonly used pruning criterion is that pruning should be performed if no deterioration of accuracy (on the training data) results from pruning. In performing this *basic pruning*, following “Occam’s razor,” the simplest model is selected if there are several models giving the same result.

Decision trees may also be subjected to some more advanced form of pruning, e.g. *confidence interval pruning*. Confidence interval pruning is performed by setting a confidence level, which is used to calculate a pessimistic estimation of the probability of erroneous classification in a sub-tree and in a tentative replacement leaf, respectively. If the tentative replacement leaf gives an equal or lower estimated error probability, then pruning is performed by replacing the sub-tree with the replacement leaf.

7.3 Decision Tree Inducer

A number of freely available decision tree inducer implementations have been evaluated (Jande, 2004, cf.) for the purpose of inducing a pronunciation variation from annotations of speech data. Of the evaluated implementations, the DTree program suite (Borgelt, 2004a) could produce the best performing trees when the available optimisation options for the different implementations were utilised. DTree also had the fastest inducer of the implementations tested and there was a useful tree visualisation tool, DTView (Borgelt, 2004b). The DTree program suite was thus selected for the pronunciation variation model induction. Thus, the DTree program suite (Borgelt, 1998, 2004a) was used for inducing the pronunciation models presented in this thesis.

7.4 Attribute Selection Measure and Optimisation Options

The DTree inducer can use both attributes with categorical values and attributes with continuous values. A categorical attribute has a finite number of unordered values. For categorical attributes, the tree branches into n branches, where n is the number of values for the attribute occurring in the training data set. Optionally, the inducer can be set to group categorical values to find the optimal number of branches. The inducer differentiates between integer and real number continuous attributes. For continuous values, the inducer finds a single optimal cut-off point and performs binary branching at this point. The inducer can handle unknown values for both categorical and continuous attributes.

A set of 30 different attribute selection measures were available for the DTree inducer. Also, some optimisation options could be made, e.g. to allow the inducer to group discrete values to obtain the optimal number of nodes. A test exploring the effects of different measures for selecting the attributes to be used for branching and of available optimisation options was conducted. During this test, trees were created using all combinations of attribute selection measures and relevant optimisation options (for optimisation options taking continuous values as input, a range of sample settings was used). A measure referred to as *symmetric information gain ratio* (Lopez de Mantaras, 1991; Borgelt and Kruse, 1998) was shown to yield the trees with the highest prediction accuracy over all combinations of optimisation options tested. Many attribute selection measures such as *information gain* and, to a lesser degree *information gain ratio* (Borgelt, 1998; Borgelt and Kruse, 1998) have a bias toward selecting categorical attributes with many possible values before attributes with fewer values, which may give sub-optimal models (Mitchell, 1997). The symmetric information gain ratio measure compensates for the number of values and calculates information gain values independent of the number of possible values of categorical attributes.

The best classification performance was obtained when selecting attributes with this measure, and allowing the inducer to group discrete values to obtain the optimal number of nodes at each level, and using the default values for all other optimisation options. This was thus the setting used for inducing the models evaluated in the next section.

7.5 Summary

This chapter has described the creation of decision tree models from annotated speech data. The steps described were the creation of training and evaluation instances, induction and pruning. The decision tree inducer implementation and the measure and optimisation options used during induction were also presented.

The training instances are phoneme-sized and can be seen as a set of context-sensitive phonemes. Each training instance includes a set of 516 attribute values and the phone realisation, which is used as the classification key. The DTree program suite was used with the *symmetric information gain ratio* measure to select attributes and with the possibility to group discrete values to obtain the optimal number of nodes at each tree level.

The next chapter describes a tenfold cross validation procedure for evaluating models created as described in this chapter and presents the results from the evaluation.

Chapter 8

Pronunciation Model Evaluation

In this chapter, an evaluation of models of the type described in Chapter 7 is presented. A tenfold cross-validation procedure is employed for model evaluation. Under this procedure, the data is divided into ten equally sized partitions using random sampling. Ten different decision trees are induced, each with one of the partitions held out during training. The partition not used during training is then used for validation.

A separate tenfold cross-validation process was performed for data from each of the three databases (VaKoS, Radio Interview and Radio News) and for the collapsed data set. Table 8.1 shows the number of training instances and the number of validation instances used for inducing and validating each created tree.

Table 8.1: *The number of training instances and evaluation instances, respectively, for each tree, for four different training data sets.*

Database	Number of training instances	Number of validation instances
VaKoS	52,263	5,807
Radio Interview	31,779	3,531
Radio News	9,936	1,104
All	93,996	10,444

The prosodic attributes (variables based on pitch and duration measures calculated from the signal) cannot be fully exploited in e.g. a speech synthesis context. Thus, it was interesting to investigate the influence of the prosodic information on model performance. For this purpose, a tenfold series of cross-validation experiments, during which the decision tree inducer did not have access to the prosodic information, was performed. As a baseline, an evaluation of trees induced from phoneme layer information only was also performed for each data set.

Thus, twelve different tenfold cross-validation experiments were performed. The models trained with access to different numbers of attributes were trained on the same instances, to make the resulting models comparable. That is, the random division of data into ten partitions was performed once for each data set, and the same partitions were used in each of the three experiments with access to different numbers of attributes. The attribute set including all information is denoted *attribute set A*, the set with prosodic attributes excluded is denoted *attribute set B* and the set with only phoneme layer attributes is denoted *attribute set C*.

Each tree created for the cross-validation experiment was pruned using a set of confidence intervals ranging from 0.01 to 0.99. For trees induced using *symmetric information gain ratio*, the confidence interval used did not affect the resulting tree and all pruned versions of a tree were thus the same. Hence, each data set gave rise to ten pairs of trees, each pair containing a *pruned* tree and the original, *unpruned* tree. Although referred to as *unpruned*, the original trees had been subjected to *basic pruning*, as explained in Section 7.2.

From each pair of trees, the optimal tree was selected to be used in the evaluation, where *optimal* was defined as producing the lowest phone error rate (PER) *on the validation data set*. If the pruned and the unpruned version of a tree produced equal PERs on the evaluation data, the pruned version was selected as the optimal tree.

8.1 Baselines

The results of the pronunciation models in terms of prediction accuracy can be compared to the results from estimating the phone string with the phoneme string. The phoneme string is the simplest baseline used. However, since there may be assimilation processes always occurring at word boundaries when words are put together, the phonemic representations for isolated words collected from a lexicon may not be a fair baseline. To explore this possibility, some phonological sandhi rules (word boundary rules) were constructed to adapt the phonemic representations for isolated words to their phonemic context.

In the sandhi rule system, three place assimilation rules were included: a *recursive rightward retroflex assimilation rule* (the [+retroflex] feature of a consonant to the left of a word boundary will recursively spread rightwards to [+dental] consonants to the right of the word boundary), a *leftward bilabial assimilation rule* (the [+bilabial] feature of a consonant to the right of the word boundary will spread to an /n/ to the left of the word boundary, changing it to an /m/) and a *leftward velar assimilation rule* (the [+velar] feature of a consonant to the right of the word boundary will spread to an /n/ to the left of the word boundary, changing it to an /ŋ/).

Also included in the sandhi rule system were *a leftward voice assimilation rule* (the [+/-voice] feature of a plosive consonant to the right of the word boundary will spread to a plosive with the same place of articulation to the left of the word boundary) and *a double consonant elision rule* (a consonant to the right of the word boundary will be elided if the same consonant occurs to the left of the word boundary).

The rules are applied in a strict order, but each rule can be set to either *on* or *off*, so that the effects of all combinations of rules resulting from either applying or not applying each rule can be explored. The combination of rules giving the adapted phoneme strings with the highest prediction accuracy (over the entire data set) is used as the second baseline.

When exploring the combinations of rules, specific rules were used rather than rules on the general format presented above. For example, the voice assimilation rule can affect three pairs of plosives, /p/-/b/, /t/-/d/ and /k/-/g/ and both the [+voice] feature and the [-voice] feature can spread leftwards. Thus, the voice assimilation rule is split into six rules, which can then be applied (or not applied) separately.

As mentioned, a third baseline used is the result of pronunciation models trained with access only to attributes originating from the phoneme layer annotation. This baseline can be used to show the effect of including variables above the phoneme layer when modelling pronunciation in discourse context. To sum up, the tree baselines used are:

- The phoneme string
- The phoneme string adapted with sandhi rules
- The output of models trained on phoneme layer attributes only

8.2 Phone Error Rates

Table 8.2 summarises the results from the cross-validation experiments. On average, we get a phone error rate (PER) of 8.2% when training on 90% of the collapsed data set and allowing the decision tree inducer to use all available information (type A tree). Using the phoneme string to estimate phone realisations gives a PER of 20.4%, which means that phone errors can be reduced by 60.0% by using an average pronunciation variation model instead of a phoneme string collected directly from a lexicon.

Table 8.2: *Mean and standard deviation of phone error rate (PER) for sets of decision trees. Means are presented in per cent and standard deviation in per cent units. Each mean and standard deviation is based on the ten optimal trees resulting from one of the twelve tenfold cross-validation experiments. Attribute set C contains only attributes from the phoneme layer, set B contains all non-prosodic attributes and set A contains all available attributes.*

Database	Set A		Set B		Set C	
	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ
VaKoS	9.08	0.31	14.99	0.33	15.53	0.49
Radio Interview	8.91	0.34	12.43	0.70	13.48	0.54
Radio News	9.24	0.72	10.70	0.95	11.53	0.93
All	8.17	0.25	13.18	0.36	14.15	0.38

Applying phonological sandhi rules to adapt the phonemic representations for isolated words to their context did not give rise to any large changes in the PER produced by the phoneme string. All combinations of applying or not applying each rule in the rule set described in Section 8.1 was tested. The combination of rules giving the largest decrease of PER compared to using the original phoneme string lowered the PER only 0.6 per cent units from 20.4% to 19.8% (the change is, however, statistically significant, p<0.01, using the McNemar test). The phonological sandhi rule set giving rise to a reduction of PER is shown in Figure 8.1.

$\mathrm{C}_\alpha \rightarrow \emptyset \;/\; \underline{\quad}\; \#_w \mathrm{C}_\alpha$
$\mathrm{n} \rightarrow \text{ŋ} \;/\; \underline{\quad}\; \#_w \mathrm{k}$
$\mathrm{n} \rightarrow \mathrm{m} \;/\; \underline{\quad}\; \#_w [\text{+bilabial}]$
$\mathrm{g} \rightarrow \mathrm{k} \;/\; \mathrm{k}\#_w \; \underline{\quad}$
$\mathrm{k} \rightarrow \mathrm{g} \;/\; \mathrm{g}\#_w \; \underline{\quad}$
$\mathrm{C}_\alpha \rightarrow \emptyset \;/\; \underline{\quad}\; \#_w \mathrm{C}_\alpha$

Figure 8.1: *The set of phonological sandhi rules giving rise to a reduction of phone error rate. In these rules, $\#_w$ denotes a word boundary and C_α denotes a specific consonant. The double consonant elision rule is applied first and then re-applied when all other rules have been applied.*

As can be seen from Table 8.2, we get a reduction of PER from 14.2% to 8.2% when switching from a model trained on phoneme level information only (type C tree) to a type A tree. This is an improvement by 42.2%, as can be seen in Table 8.3.

8.3 Data Size and Speaking Style

It is likely that the data presented in Table 8.2 reflects the fact that both the amount and the type of training data affects the performance of the models induced. Neither models trained on the VaKoS database nor models trained on the Radio News database have the lowest PER of the models trained on separate databases, although the VaKoS database has the largest number of training instances and the Radio News database has the most formal, strict type of speech. Instead, the models trained on the Radio Interview database show the lowest PER (for type A trees). The Radio Interview database has the advantages of having relatively formal speech in comparison with the VaKoS database, relatively few speakers, and many more training instances than the Radio News database.

Further, we can see from Table 8.3 that the improvement arising from making more attributes available for the decision tree inducer is greater for the VaKoS data than for the Radio Interview data and for the Radio News data. Models trained on the VaKoS database are thus more dependent on prosodic information and

Table 8.3: *Error reduction (per cent) as a consequence of using trees trained on all attributes compared to using trees trained on subsets of attributes. Type C trees are trained with access only to phoneme level attributes, type B trees are trained with access only to non-prosodic attributes and type A trees are trained with access to all attributes.*

Database	Error reduction switching from type B to type A trees	Error reduction switching from type C to type A trees
VaKoS	39.42	41.50
Radio Interview	28.33	33.93
Radio News	13.63	19.87
All	37.97	42.23

generally on information from layers above the phoneme, while the models trained on the Radio News database are less dependent on this type of information.

8.4 Phone Confusions

Tables 8.4 and 8.5 are confusion matrices summarising the confusions (errors) made by the type A trees trained on data from all databases. Tables F.1 and F.2 in Appendix F show the confusions broken down by source phoneme.

The consonant confusion statistics in Table 8.4 reveal that retroflex consonants are often confused with their dental counterparts and that [g] is often erroneously elided. The evaluation of the automatically obtained key transcripts against a manually supplied gold standard (cf. Section 5.7.7) revealed that these were also the most prominent consonant confusions made by the automatic transcription system. There was the high degree of confusion of retroflex consonants with dentals, including a 100% confusion of [ɭ] for [l] over the small gold standard transcript (however, based on 3 occurrences only). There was also the a high degree of confusion between [g] and ∅.

From Table 8.5, it can be seen that when long vowels are confused, it is manly with their short counterparts and, to a lesser degree, with [ə]. Short vowels are often confused with [ə], but also with their long counterparts. The exception is the short vowel [a] and its long counterpart [ɑː], which are never confused with each other. This is probably due to the low level of noise in the key transcripts, especially for [ɑː], as discussed in Section 5.7.7.

Some confusion involving the allophonic variants [ɛ]-[æ] and [ø]-[œ̝] can also be seen in Table 8.5. These trends could also be seen in the evaluation of the key transcripts against the gold standard.

It is thus safe to assume that this noise in the training keys is largely responsible for the large shares of erroneous classifications made by the decision tree model for certain phones. This noise also largely accounts for the large shares of confusions in the model output between [ə] and full vowels, between long and short vowels, between the [ɛ] and [æ] allophones, and between the [ø] and [œ̝] allophones.

Table 8.4: *Consonant confusion matrix for the ten optimal type A trees trained on all data. Each column shows the share of classifications corresponding to the class shown at the bottom of the column for the key phone of the row. No realisation (∅) can be confused with both vowels and consonants. Otherwise, if row sums are not 1, it is because the numbers are rounded.*

phone	∅	p	t	k	b	d	ɡ	f	v	s	ɦ	ç	h	m	n	ŋ	l	j	ɹ	ʈ	ɖ	ɭ	ɳ	ʂ
∅	.75					.03	.02	.01					.01	.01	.03		.01	.01	.06					
p	.01	.99																						
t	.01		.99																					
k	.01			.99																				
b	.01				.99																			
d	.04					.92													.03					
ɡ	.20						.80																	
f	.02							.98																
v									1															
s										.99														
ɦ											1													
ç												1												
h	.05												.95											
m	.01													.98										
n	.02														.97									
ŋ	.01														.02	.97								
l	.01																.99							
j	.05																	.95						
ɹ	.08					.08													.84					
ʈ			.06																	.94				
ɖ						.17													.01		.81			
ɭ	.64																					.36		
ɳ	.10														.27								.63	
ʂ	.01									.30														.70

In the key transcript, many ∅ realisations should actually have been [ɡ] phones, many dental consonants should have been retroflexes, many [ə] realisations should have been full vowels, and many short vowels should have been long and vice versa. As an effect of this, the share of retroflex consonant phonemes realised as dentals, the share of /ɡ/ phonemes realised as ∅, the share of full vowel phonemes realised as [ə] and the share of short vowels realised as long and vice versa are unproportionally large. Since the phoneme identity is an important predictor for the realisation of the phoneme, the error types found in the key transcript against the gold standard will propagate to the decision tree model.

A high degree of uncertainty about the identity of a key phone in relation to the gold standard will give rise to uncertainty about the phone in the decision tree predictions. It is the degree of confusability that propagates rather than the exact

Table 8.5: *Vowel confusion matrix for the ten optimal type A trees trained on all data. Each column shows the share of classifications corresponding to the class shown at the bottom of the column for the key phone of the row. No realisation (∅) can be confused with both vowels and consonants. Otherwise, if row sums are not 1, it is because the numbers are rounded.*

phone																								
∅	.75	.01									.01				.01								.01	
ə		.89	.03		.01		.04		.01						.01									
a		.06	.94																					
ɑː		.01		.99																				
e		.18			.77	.06																		
eː		.03			.03	.94																		
ɪ		.04					.93	.03																
iː							.03	.97																
ʊ		.17							.74	.09														
uː									.03	.97														
ɵ	.01	.01									.92	.05												
ʉ̟ː		.02									.01	.97												
ʏ													.98	.02										
yː		.03											.04	.94										
ɔ	.01	.02													.93	.04								
oː															.08	.92								
ɛ		.01															.90	.02	.07	.01				
ɛː		.01															.05	.93						
æ		.07															.09		.71	.13				
æː		.01															.01		.10	.88				
œ	.03																				.70	.09	.13	.04
øː																					.06	.94		
œ̞	.06	.02																			.01		.86	.06
œ̞ː	.01																				.01		.04	.94
	∅	ə	a	ɑː	e	eː	ɪ	iː	ʊ	uː	ɵ	ʉ̟ː	ʏ	yː	ɔ	oː	ɛ	ɛː	æ	æː	œ	øː	œ̞	œ̞ː

error types, but since variation is restricted, the errors in Tables 8.4 and 8.5 will largely mirror the errors in the confusion matrices for the automatic transcription system compared to the gold standard (Tables 5.7 and 5.8).

It should be noted, however, that since the same errors are present both in the training data and in the validation data, the PER calculated for the decision tree model output is most probably lower than it would have been if the models had been trained on the automatically obtained transcripts and evaluated by manually obtained transcripts (unfortunately, no such transcripts besides the small gold standard are available), although the PER is probably higher than it would have been if the models had been both trained on and validated against manually obtained transcripts.

An [ɭ] is erroneously elided by the decision tree models in the majority of cases.

However, as can be seen from Table G.1 in Appendix G, the [ɭ] phone is very infrequently occurring in the validation data (and, thus, in the training data). In fact, the phone only occurs 11 times in the key transcripts from all databases. The training instances for this phone are thus very sparse and it is not surprising that the relative error for this particular phone is high. However, since the number of instances is so low, these errors do not significantly contribute to the PER.

Table 8.6 shows the most frequent phone classification errors made by the trees as the share of the total number of errors. It can be seen that errors mostly go both ways. For example, the two most common errors made by the models are erroneous [ɹ] insertions and erroneous [ɹ] elisions. It can also be seen that the choice between a [ə] and a full vowel is a large source of errors.

Table 8.6: *The most frequently occurring phone classification errors.*

	Occurrences of error	Share of total errors	Phone$_{\text{key}}$	Phone$_{\text{model}}$
1	566	6.63%	∅	ɹ
2	504	5.90%	ɹ	∅
3	466	5.46%	ɹ	d
4	433	5.07%	ə	ɪ
5	389	4.56%	e	ə
6	375	4.39%	ə	a
7	320	3.75%	a	ə
8	294	3.44%	∅	d
9	260	3.05%	∅	n
10	226	2.65%	∅	g
11	178	2.09%	g	∅
12	173	2.03%	d	∅
13	144	1.69%	ɔ	oː
14	140	1.64%	n	∅
15	131	1.53%	∅	ɔ
16	129	1.51%	ə	e
17	125	1.46%	d	ɹ
			e	eː
			ɪ	ə
			∅	f
18	121	1.42%	ʂ	s
			∅	j
19	114	1.34%	ɪ	iː
20	112	1.31%	oː	ɔ
21	95	1.11%	∅	θ
22	91	1.07%	ɛː	æ
23	86	1.01%	∅	h
24	84	0.98%	ə	ɔ

Further, Table 8.6 shows that the confusions between [g] and ∅ discussed above do not only constitute a large proportional error for the [g] phone, but a large proportion of error type instances. From the total number of errors in the decision tree output, 2.7% were ∅ realisations erroneously classified as [g] and 2.1% were [g] realisations erroneously classified as ∅. Since there are many more instances of ∅ than of [g] in the key transcript, it is not obvious from Table 8.4 that there are more erroneous [g] insertions than erroneous [g] elisions, but this can be seen in Table 8.6.

The very frequent substitutions of [ɹ] for [d] (466 occurrences) and the relatively frequent substitutions of [d] for [ɹ] (125 occurrences) reflects the fact that /d/ is often pronounced [ɹ] in colloquial speech in central standard Swedish and that the variation in pronunciation is relatively free. The variation is perhaps more idiomatic than governed by general constraints for the language variety.

The errors in choosing between a [ə] and a full vowel are probably not only actual errors, but also artefacts of free variation. That is, a [ə] and a full vowel may be equally correct in many cases. If the model is used in a speech synthesis setting, such deviation from the key transcript due to free variation would neither affect the intelligibility nor the perceived naturalness of the resulting speech. In cases where the classification is actually erroneous, the error would probably not affect intelligibility in any critical way. A more serious type of error is erroneous vowel elision. Erroneous consonant elisions may also in many contexts affect the naturalness and/or intelligibility. Out of the total number of errors produced by the ten optimal models trained on all data, as many as 18.6% were erroneous elisions. However, only 1.6% were erroneous vowel elisions.

8.5 Attribute Ranking

Table 8.7 shows the 36 top ranking attributes over the ten optimal type A trees trained on the collapsed data set. The layer from which the attribute originates is used as a prefix in the attribute names. Attributes can refer to the current unit or to units at ±4 positions from the current unit in the specific annotation layer. Duration measures can be based on the duration of all *phonemes* or on the duration of *vowels* only, they can be based on *normalised* or *absolute* phoneme duration, and they can be based on duration on a *log* scale.

The ranking in the first column of Table 8.7 is based on the position of the attribute in the ten type A trees. In this case, the attribute governing the largest number of subtrees (leaves excluded) will get the highest rank (1). The ranking method of the second column weights the subtree count with the number of classifications involving the attribute (over the training data). For this measure, an attribute involved in many classifications can climb in rank even if it does not appear in the absolute top of the tree (near the root). The *phoneme identity* attribute appears in the top node of all trees. This means that it governs all subtrees and is involved in all classifications made by the trees. Hence, *phoneme identity* ends up at the top irrespective of ranking method.

Table 8.7: *The 36 top ranking attributes for trees trained on all information from all databases (type A trees), using two different ranking methods.*

	Subtree rank	Subtree · classification rank
1	phoneme_identity	phoneme_identity
2	phoneme_identity+1	phoneme_identity+1
3	word_duration_phonemes_absolute	word_duration_phonemes_absolute
4	word_function_word-1	word_function_word
5	word_function_word+1	word_function_word+1
6	phoneme_identity+4	word_function_word-1
7	phoneme_identity-2	phoneme_identity-1
8	word_function_word	word_duration_vowels_absolute
9	phoneme_identity-1	phoneme_identity+2
10	phoneme_identity+2	phoneme_identity-3
11	phoneme_identity-4	phoneme_identity+4
12	phoneme_identity+3	phoneme_identity+3
13	phoneme_identity-3	phoneme_identity-2
14	word_duration_vowels_absolute	phoneme_identity-4
15	syllable_stress_type	syllable_stress_type
16	syllable_nucleus	phrase_duration_phonemes_absolute
17	word_duration_vowels_normalised	word_duration_vowels_normalised
18	word_duration_vowels_log_absolute	syllable_nucleus
19	syllable_position_in_word	phoneme_feature_py+1
20	phoneme_feature_py+1	syllable_position_in_word
21	phrase_duration_phonemes_log_absolute	word_duration_vowels_log_absolute
22	phrase_duration_phonemes_absolute	word_duration_phonemes_log_normalised
23	phrase_duration_phonemes_log_normalised	phrase_duration_phonemes_log_absolute
24	syllable_duration_vowels_absolute	phrase_duration_phonemes_log_normalised
25	word_duration_phonemes_log_normalised	syllable_duration_vowels_absolute
26	phrase_duration_vowels_absolute	syllable_stress
27	discourse_duration_vowels_absolute	discourse_duration_vowels_absolute
28	word_part_of_speech-3	syllable_duration_phonemes_absolute
29	word_part_of_speech+2	phrase_duration_vowels_absolute
30	syllable_stress	word_duration_phonemes_log_absolute
31	syllable_duration_phonemes_absolute	word_part_of_speech-3
32	word_part_of_speech-2	word_part_of_speech-2
33	phrase_duration_vowels_log_absolute	phrase_duration_vowels_log_absolute
34	word_part_of_speech+3	word_morphology_tense_aspect
35	word_duration_phonemes_log_absolute	word_part_of_speech
36	word_part_of_speech-4	phrase_duration_phonemes_normalised

8.6 Attributes Used by the Models

Table 8.8 shows the probability of a variable from each of the six annotation layers showing up at a specific level of the type A decision trees trained on the collapsed data set. From this table, it can be seen that variables from all layers of annotation are used by the trees trained on all available information from all databases[1].

In fact, from 516 available attributes, as many as 449 were used at least once in the ten trees. However, the phoneme and word layer attributes are the attributes most commonly used in the higher levels of the trees. The top ranking utterance layer attribute shows up at rank 55 using the first ranking method and at rank 43 using the second ranking method. For the first method, the attribute is a phoneme-based duration measure and for the second method, the attribute is a vowel-based

[1] Six of the optimal trees were pruned, but four were unpruned (subjected to *basic pruning* only).

duration measure. The top discourse layer attribute is also a vowel-based duration measure and shows up at rank 27 for both ranking methods.

Table 8.8: *The probability (per cent) of a variable from an annotation level appearing at a specific level (1 being the top node) of the optimal type A trees trained on all data.*

Level	Phoneme layer	Syllable layer	Word layer	Phrase layer	Utterance layer	Discourse layer	∑
1	100.00	0.00	0.00	0.00	0.00	0.00	100.00
2	36.07	10.38	50.82	1.09	1.64	0.00	100.00
3	35.68	4.32	43.51	16.22	0.00	0.27	100.00
4	37.59	5.37	44.63	10.37	1.11	0.93	100.00
5	34.02	4.88	42.99	12.76	3.15	2.20	100.00
6	41.04	3.11	38.67	13.19	2.07	1.93	100.00
7	40.41	3.80	32.49	17.43	3.49	2.38	100.00
8	40.90	5.63	32.83	14.45	4.69	1.50	100.00
9	40.38	3.77	34.10	15.06	5.02	1.67	100.00
10	40.33	0.48	36.99	16.47	3.10	2.63	100.00
11	38.86	5.42	33.73	14.76	5.12	2.11	100.00
12	40.00	3.02	35.85	13.96	6.04	1.13	100.00
13	38.49	4.37	32.94	13.10	9.13	1.98	100.00
14	30.05	5.91	45.81	11.82	5.91	0.49	100.00
15	35.33	3.80	41.30	13.04	6.52	0.00	100.00
16	33.33	6.41	41.03	10.26	7.05	1.92	100.00
17	34.81	2.96	43.70	13.33	2.96	2.22	100.00
18	36.13	8.40	40.34	10.08	3.36	1.68	100.00
19	28.28	5.05	46.46	14.14	4.04	2.02	100.00
20	33.33	6.41	39.74	11.54	5.13	3.85	100.00
21	41.67	1.67	30.00	18.33	6.67	1.67	100.00
22	36.17	4.26	34.04	19.15	4.26	2.13	100.00
23	26.09	10.87	30.43	15.22	15.22	2.17	100.00
24	40.54	2.70	35.14	16.22	5.41	0.00	100.00
25	47.06	2.94	29.41	11.76	8.82	0.00	100.00
26	30.77	3.85	50.00	3.85	7.69	3.85	100.00
27	42.11	0.00	31.58	21.05	5.26	0.00	100.00
28	33.33	0.00	40.00	13.33	6.67	6.67	100.00
29	38.46	7.69	53.85	0.00	0.00	0.00	100.00
30	33.33	0.00	44.44	11.11	11.11	0.00	100.00
31	44.44	0.00	22.22	11.11	11.11	11.11	100.00
32	0.00	0.00	50.00	0.00	0.00	50.00	100.00

The *word frequency* and *word predictability* attributes both get relatively low ranks (*word frequency* is ranked 67 and 94 by the respective ranking methods and *word predictability* is ranked 91 and 133). The relatively weak predictive strength of these variables may be due to the fact that they are obscured by the *function word* variables, who get high ranks and, to a certain degree, contain information overlapping with the *word frequency* and *word predictability* variables. Also, the *word frequency* and *word predictability* measures are estimated from a corpus of transcribed speech, relatively small in comparison to standard text corpora. These measures would probably be improved if supplemented with data from text corpora.

A large variety of the duration and pitch based measures, respectively, are represented among the variables used by the optimal trees. The first measure based on pitch shows up at place 44 using the first ranking strategy and on place 47 using the second ranking strategy. The highest ranking pitch-based attributes are two different pitch dynamic measures calculated over the phrase. Most of the duration measures seem to be nearly equivalent in terms of predictive power, with vowel-based measures working somewhat better over durationally larger units than over smaller units.

Since higher order layer units are large in terms of duration, it is not possible to make exact predictions from these units only and attributes from these layers mostly end up in the lower levels of the decision trees, as a result of the 'greedy' induction algorithm used.

8.7 Model Complexity

The ranking of attributes is closer to optimal when using symmetric information gain ratio than when using other selection measures given the type of training data used and thus trees are generally smaller after basic pruning when the *symmetric information gain ratio* measure has been used to induce the tree. Symmetric information gain ratio thus gives both better predictions and less complex models than using e.g. information gain ratio for selecting attributes. For this reason, the effect of pruning on model performance was small for the decision trees evaluated. In most cases, pruning affected model performance (on the test data) positively.

Six pruned trees performed better than their unpruned counterparts. On average over the ten type A trees trained on all data, pruning decreased the PER only by 0.5%, but decreased the average number of attributes used by the models by 82.0% (from 302.8 to 54.5). The model complexity thus dropped significantly as a result of pruning: the average number of nodes decreased by 89.6% (from 4151.9 to 433.1) and the average number of tree levels decreased by 62.2% (from 32.3 to 12.2)[2]. Using the McNemar test, the difference in PER between pruned and unpruned models was shown *not* to be significant.

A pruned model is much simpler than an unpruned model and thus requires less input attributes to be obtained. Although the McNemar test showed that there is no gain in predictability associated with pruning, it also showed that there is no loss of predictability associated with pruning. Hence, a pruned model would be the choice in an application. However, it should be noted that although a pruned model uses less attributes than an unpruned model, there are still attributes from all annotation layers used in the pruned models.

[2] The node counts include leaf nodes and the level counts include levels containing only leaves.

8.8 Weighted Phone Error Rates

Since all classification errors do not have equal perceptual impact, the PER measure may not give an adequate measure of model performance. For example, replacing a plosive consonant with a vowel will under most conditions give a greater perceptual impact than replacing a full vowel with a [ə]. In representations of pronunciations of connected speech involving a small set of phone symbols, there are many gray areas and during manual phonetic transcription, many cases are encountered where ad hoc decisions about phone identity have to be made. If synthesised with appropriate prosody, a phone string with a full vowel and a string with a [ə], respectively, may not give rise to any perceived difference with regards to naturalness.

For a certain phoneme, a decision tree model of the type under discussion can only produce a realisation that it has encountered during training[3]. Thus, in the current case, the comparisons of phones from the auto-transcription system (used as the key in the tenfold cross-validation experiment) and phones produced by the decision tree model will always be comparisons of relatively similar phones. All possible phone distances are on average about equal (although context-dependent), so that the PER measure's uniform weight of the distances from a phone to all its possible misclassifications is actually justifiable.

However, although the distances of a key phone to all its possible misclassifications may be about equal, the distance between a certain phone and its possible misclassifications may not be equal to the distance between a another phone and *its* possible misclassifications. For this reason, a *weighted PER* measure, PER_w, is introduced. For calculating this measure, each comparison between a key phone and a model-produced phone classification is weighted with a static *phone distance weight*, λ ($0 \leq \lambda \leq 1$), resulting in a $\text{PER}_w \leq \text{PER}$.

If the key phone and the model phone are represented by the same symbol, λ will be 0. If a vowel is compared to a plosive consonant, λ will be 1. Thus, a 100 per cent PER_w will be the effect of replacing all vowels with plosive consonants and vice versa. A PER_w close to 100% will always be completely corrupted, while a PER close to 100% may theoretically still be at least partly understandable (and it may also be completely corrupted). The PER and PER_w measures thus move on different scales and cannot be compared.

The actual phone distance weights used for the PER_w measure are based on an unpublished version of a description of an algorithm for measuring the "phonetic distance" between written words (Brodda, 1966). The published and the unpublished versions differ mainly in that, in the unpublished version, there is an actual matrix with estimations of the "phonetic distances" between letters (and letters and $\emptyset$).

This matrix has been extended and adapted to form a phone distance matrix for the phone set used in the annotation described in this thesis. The distances

[3] Since discrete variables can be clustered at model induction, it is possible to induce a model that for a certain phoneme can produce a realisation never encountered for that particular phoneme in the training data.

in the original matrix were based mainly on differences in articulatory features and on relative distances between tongue positions during articulations of different vowels. Another strategy for constructing phoneme distance matrices is to base the distances on phoneme confusions in listening tests. For example Fant et al. (1966) constructed consonant distance matrices in this manner. Table H.1 in Appendix H shows the phone distances used in the current evaluation. The distances are specified on a scale from 0 to 8 and the phone distance weight λ is the phone distance divided by 8.

Table 8.9: *Mean and standard deviation of weighted phone error rate (PER$_W$) for sets of decision trees. Means are presented in per cent and standard deviation in per cent units. Each mean and standard deviation is based on the ten optimal trees resulting from one of the twelve tenfold cross-validation experiments. Attribute set C contains only attributes from the phoneme layer, set B contains all non-prosodic attributes and set A contains all available attributes.*

Database	Set A		Set B		Set C	
	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ
VaKoS	2.63	0.10	4.69	0.12	4.89	0.15
Radio Interview	2.54	0.14	3.73	0.22	4.03	0.17
Radio News	2.59	0.21	3.02	0.32	3.23	0.29
All	2.32	0.09	4.08	0.12	4.44	0.13

Of course, the perceived difference between two phones depends highly on the context, e.g. the position in the syllable, the phone duration, the length of the utterance and the word, whether there are phonotactic or semantic constraints on replacing one phone with the other, etc. The perceived phonetic distance between utterances is thus not a simple function of individual phone distances. Simultaneously, a pronunciation differing in phone-level realisation from that of a recorded utterance, all other things being equal, will not always be perceived as an erroneous or unnatural pronunciation—a certain degree of variation is allowed. Still, it is likely that the more fine-grained PER$_w$ measure will give a more accurate description of model performance and a better estimation of the perceived quality of the phone string at a hypothetical pronunciation than PER gives.

Table 8.9 shows the mean PER$_w$ and PER$_w$ standard deviations for the twelve tenfold cross-validation experiments reported above. For models trained on data from all databases and on all available attributes, the PER$_w$ is 2.3%.

As can be seen from Table 8.10, when using the PER$_w$ as the evaluation measure, the estimated dependence of the models on information above the phoneme level and on prosodic variables increases. The reduction in PER associated with making prosodic variables available for the tree inducer was 37.97% and the reduction associated with making attributes originating above the phoneme annotation layer was 42.23% for models trained on all databases. The corresponding reductions in PER$_w$ are 43.13% and 47.70%, respectively.

Given that the PER$_w$ measure is a better estimator of model accuracy, this makes it even more clear that the multi-layer information approach to pronunciation modelling does indeed improve predictions of phone-level pronunciation in discourse context, compared to using phoneme layer information only.

Table 8.10: *Error reduction based on weighted phone error rate (per cent) as a consequence of using trees trained on all attributes compared to using trees trained on subsets of attributes. Type C trees are trained with access only to phoneme level attributes, type B trees are trained with access only to non-prosodic attributes and type A trees are trained with access to all attributes.*

Database	Error reduction switching from type B to type A trees	Error reduction switching from type C to type A trees
VaKoS	43.95	46.27
Radio Interview	31.94	37.05
Radio News	14.22	19.84
All	43.13	47.70

8.9 Effects of Noise

The erroneous classifications possible for a phoneme are limited to the set of realisations for the phoneme found in the training data (except in some cases where the decision tree inducer has collapsed phoneme classes). Both training and evaluation data contain up to 15.5% errors on the phone level, as previously discussed (cf. Section 5.7). Since the phone string is generated by an automatic transcription system with a priori restrictions on the possible realisations of each phoneme, the range of variation is probably less than it would have been if the transcripts had been produced by a human. It is not immediately obvious whether this has translated into lower or into higher phone error rates in the cross-validation setting, than would have been the case if the phones in the training and validation data had been supplied by a human transcriber.

However, the correspondences between the confusion matrices comparing the gold standard transcript and the automatically generated key transcript on the one hand, and confusion matrices comparing the the key transcript and the model output on the other hand, indicate that the phones, for which there is a large discrepancy between the gold standard and the key transcript, are the phones for which there are more discrepancies between the key transcript and the decision tree model output.

It is thus relatively safe to assume that the restricted number of possible realisations of a phoneme used in the automatic transcription system and thus in the keys used during decision tree model training has affected the PER of the resulting models negatively, rather than positively. This suggests that less noise in the training data (training on manually supplied key transcripts) would produce more models with lower, not higher, PERs in a cross-validation setting.

The fact that less noisy key transcripts would probably produce models with lower PERs in a tenfold cross-validation setting than the current transcripts does not mean that less noisy key transcripts would give this effect compared to all types of more noisy key transcripts. For example, if all key phones in the training and validation data were set to the same symbol, e.g. /p/, trees induced from and validated against this data could never produce anything else than 100% correct decisions. This would, however, be in relation to the erroneous keys and have no practical value.

To sum up, the noise in the training data has probably affected the PERs in the tenfold cross-validation experiments negatively and access to less noisy training data would give an improvement not only in model performance in relation to gold standard transcripts but most probably also in a cross-validation setting.

8.10 Reliability Issues

To test the reliability of the evaluation method, the tenfold cross-validation experiment was repeated two more times with a new randomisation each time. As presented above, the first tenfold cross-validation run showed a mean PER of 8.17% and a standard deviation of 0.25 over the ten generated trees trained on all attributes from all databases. Six of the optimal trees were pruned trees and four were unpruned trees (trees with basic pruning). The first repetition gave a mean PER of 8.15% and a standard deviation of 0.26. Five of the optimal trees were pruned and five were not. The second repetition gave a mean PER of 8.16% and a standard deviation of 0.21 and seven of the optimal trees were pruned.

The results are thus relatively stable over different data randomisations, and it is safe to claim that a 8.2% PER is an actual mean for the given data in a tenfold cross-validation setting with randomly partitioned data.

Other issues related to reliability are the facts that trees are trained and evaluated on very similar data and that data from the same speakers occur in both training end evaluation data. It is thus likely that the phone error rates reported are lower than they would have been if an entire database had been held out during training and then used as validation data.

To investigate the performance of models for predicting the pronunciation of speakers not present in the training data, the tenfold cross-validation procedure was repeated once more. However, this time, instead of randomly choosing ten per cent of the instances to hold out from training, all instances originating from a single speaker from the VaKoS database were held out at each run. This meant that the number of instances held out at each run was not exactly the same, although the discrepancies were small. The entire Radio Interview and Radio News databases were always included in the training data set. Thus, the trees were always trained on more than ten per cent of the available instances and evaluated on less than ten per cent of the instances.

Table 8.11: *Mean and standard deviation of phone error rate (PER) for sets of decision trees trained on data from all databases. The two methods used for holding out instances at training is using random sampling and excluding all instances based on the speech from a specific speaker (speaker 1–10 from the* VaKoS *database). The values for the random method were included in Table 8.2 and are repeated here for comparison. Means are presented in per cent and standard deviation in per cent units. Attribute set C contains only attributes from the phoneme layer, set B contains all non-prosodic attributes and set A contains all available attributes.*

Hold-out method	Set A		Set B		Set C	
	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ
Random	8.17	0.25	13.18	0.36	14.15	0.38
Speaker	8.51	0.76	14.66	1.83	15.87	1.67

Table 8.11 shows the results of the cross-validation experiment with all instances originating from a single speaker held out at each run. The results from the proper tenfold cross-validation experiments using random sampling for holding out instances is also included in the table, for comparison. It can be seen that the average PERs over the trees created using the *speaker* hold-out method are slightly higher the PERs in the original tenfold cross-validation experiment.

The higher standard deviations are probably in part an artefact of the smaller evaluation set, but it is likely that they also reflect the fact that the different speakers are idiomatic to different degrees and that their specific pronunciations thus are differently hard to predict on the basis of the training data from other speakers of the same language variety.

In using the current pronunciation modelling method, the goal is not to get 100% correct predictions for single speakers. The goal is to get pronunciation representations that are correct from an *average* perspective rather than speaker-specific perspective. In this context, the PER thus is the sum of actual errors and adaptations from the speaker-specific pronunciation to an average pronunciation.

Very large differences in results between the hold-out methods would have indicated that variation in pronunciation is largely speaker-specific and only to a small degree general to the language variety. The fact that the differences are small shows that variation in pronunciation due to discourse context is mainly a language trait, but that there is also idiomatic variation.

8.11 Gold Standard Evaluation

Although it is hard to speculate about how the model performance would be affected by more accurate training data, the transcripts generated by the current models can be evaluated against actual target transcripts. When evaluated against the small gold standard consisting of five minutes of manually transcribed speech from the VaKoS database, the optimal type A trees trained on all data produced a PER of 17.7% (PERw = 5.3%), which means that the deterioration in performance when using the model instead of the automatic transcription system is only 12.3% (11.2% deterioration of PERw) and that the improvement using the model instead of the phoneme string is 32.1% (34.6% improvement of PERw).

8.12 Summary

This chapter has presented a tenfold cross-validation experiment, in which it was shown that including information from multiple layers can improve the performance of pronunciation models, most notably for spontaneous speech, where the predictive power of phonological and grammatical information is relatively low. Attributes from all layers of annotation were used in the models with the highest prediction accuracy. The optimal models produced an average phone error rate of 8.2%, which is an improvement of 60.0% compared to using the phoneme string for estimating phone-level realisation.

A comparison between models trained only on phoneme layer attributes and models trained on attributes from all layers showed that the prediction accuracy of pronunciation models could be improved by 42.2% by including information above the phoneme level. The multi-layer information approach to pronunciation modelling thus improves predictions of phone-level pronunciation in discourse context compared to using phoneme layer information only, although phoneme layer attributes are still the most important predictors.

The next chapter presents a pronunciation model trained on all available data in detail and discusses rules for the realisation of separate phonemes. Further, the realisation distribution and prediction accuracy of separate phones are presented.

Chapter 9

Phoneme-to-Phone Conversion

As stated in the introduction to this thesis, modelling pronunciation variation in discourse context is interesting for the description of a language variety. This chapter will describe some aspects of pronunciation in central standard Swedish derived from a model induced from the annotated speech data. The distributions of realisations for each phoneme, the shares of correct classifications and the phoneme-to-phone conversion rules employed by the model are discussed.

9.1 The Decision Tree Pronunciation Model

The evaluation based on tenfold cross-validation, presented in the preceding chapter, suggested that a model trained on all available data and subsequently pruned is optimal both from a performence perspective and from a model complexity perspective. Hence, this is the type of model used for describing central standard Swedish from a pronunciation variation perspective.

The results presented in Chapter 8 were based on means of ten trees trained on 90% of a particular data set and validated against the remaining 10% of data. Here, a single tree trained on all available data is studied. This final, pruned tree uses 57 of the 516 available attributes and has 427 nodes (including leaves) in 12 levels.

Tables I.1 to I.23 in Appendix I and Tables J.1 to J.23 in Appendix J present the distributions of realisations produced by the model on the training data for each phoneme. The tables also show the key phones for each phoneme and the share of correct classifications made by the model. As mentioned, the values in these tables are based on classifications made on the *training data*. Since the tree was trained on all available data, it could not be validated against data not used at training. However, the values will still give an idea about the approximate over-all distribution of phone realisations for each phoneme.

Over the training data, the tree output showed a 7.39% phone error rate, PER, and a 2.09% weighted phone error rate, PER_w (the corresponding means for the trees in the cross-validation experiment were 8.17% and 2.32%). A better model can be expected simply due to the fact that more training data is used. The relatively small difference in PER between the model trained on all data and the mean PER of the corresponding trees trained on 90% of the data is thus probably not only due to the fact that the training and validation data overlap, but also to the fact that there was more training data for the former tree.

The unpruned tree (i.e., the tree subjected to *basic pruning* only), however, is expected to produce a significantly lower PER when applied to the data from which it was induced. This was also the case; the unpruned tree showed a PER of only 0.61% using 329 attributes distributed over 4,605 nodes in 33 levels.

As discussed earlier, the PER reflects differences between the key transcript and the model output and these differences may be due to actual errors, but can also be a artefacts of a general model. The generalisation obtained through model pruning excludes most idiomatic variation specific to the speakers on whose speech annotation the model was trained, making the model a more applicable for the language variety. The generalisation procedure thus contributes to the PER, without giving rise to actual errors—the general nature is a desired property of the model.

In appendices I and J, each branch from the top node *phoneme identity* of the final tree is presented separately as a subtree, in affect showing the rule for converting each phoneme to its respective phone realisation depending on the discourse context. The tree is split for readability reasons, and the top node of each subtree corresponds to *the same* phoneme identity node in the complete tree.

Appendix I shows the subtrees for consonant conversions and discusses the realisation rules for each consonant in detail. The shares and frequencies of realisations are also presented. Appendix J shows the subtrees for vowel conversions. The realisation of vowels is less dependent on phoneme context and more dependent on duration-based attributes than the realisation of consonants. The realisation rules for vowels are in most cases more complex than the rules for consonants and are not discussed in the same detail as the consonant realisation rules.

Since the individual phoneme-to-phone conversion rules are interesting phenomena from a linguistic-phonological point-of-view, the trees illustrating these rules are all included in appendices I and J. In this chapter, regularities over different phoneme-to-phone conversions will also be discussed.

9.1.1 Unseen Contexts

Not all possible contexts are handled by the model, since all possible contexts were not available in the training data. If an unknown context attribute shows up in data to be classified by a pronunciation model, the model uses the majority class of the lowest node for which there is a known context attribute. In the figures in appendices I and J, only the majority class of each leaf is shown, but the model has access to the complete distribution of key phones in the training data for each leaf. From this information, the majority class of a higher node can be calculated.

For example, in the leftmost branch of the tree model illustrated in Figure 9.1 (also included in Appendix I), a /p/ should be realised as [p] if its right adjacent phoneme is /p/, /ɹ/ or /a/ and as ∅ if the adjacent phoneme is /d/, /ɑː/, /oː/, /æ/ or <sil> (silence). So, what happens if the right adjacent phoneme for a /p/, for which the model is to produce a realisation, is in fact an /s/?

Then, the model knows that in the training data, there were 5 instances with /p/, /ɹ/ or /a/ as their right neighbour and they should all be realised as [p] according to the key transcript. The model also knows that there were 89 instances in the training data preceding /d/, /ɑː/, /oː/, /æ/ or <sil> and that 87 of these should be realised as ∅ and 2 as [p], according to the key transcript. The total number of instances for the mother node of these two leafs is thus 94, of which 87 should be realised as ∅. This makes ∅ the majority class of the mother node and the realisation assigned to the /p/ in this case.

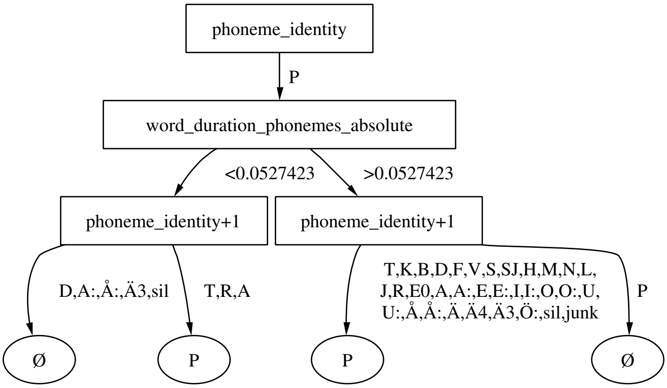

phoneme_identity
P
word_duration_phonemes_absolute
<0.0527423
>0.0527423
phoneme_identity+1
phoneme_identity+1
D,A:,Å:,Ä3,sil
T,R,A
T,K,B,D,F,V,S,SJ,H,M,N,L,
J,R,E0,A,A:,E,E:,I,I:,O,O:,U,
U:,Å,Å:,Ä,Ä4,Ä3,Ö:,sil,junk
P
Ø
P
P
Ø

Figure 9.1: *The realisations of the phoneme /p/ (phoneme representations in the figure are in STA format).*

9.2 General Discussion on Phoneme Realisation

In the pruned decision tree pronunciation model, only 57 of the 516 available attributes are used. However, as was discussed in Chapter 8, many of the attributes are nearly equivalent in terms of predictive power and depending on the part of the training data randomly excluded during training, different attributes could be selected to be included in a model, with very little effect on model performance. There is thus a degree of chance in the selection of the particular attributes, as aresult of the fact that there is much redundant information in the set of available attributes.

There are, however, some variables that are particularly informative. The identity of the current phoneme and the phoneme context ± 4 positions from the current phoneme are particularly strong predictors of phone identity, as could be expected. Also, the *function word* attribute and the function word context were strong predictors. Several duration-based measures also proved to be among the top predictors. It is clear that the mean absolute phoneme duration and the mean absolute vowel duration over the word were the best duration-based predictors.

9.2.1 Consonants

As can be seen from figures I.1 to I.23 in Appendix I, mean absolute phoneme duration is particularly important for determining if a particular consonant should be realised or elided. The break-off point is similar for most consonants, situated at about 35 ms. Since there is noise in the key transcript, it cannot be excluded that this is partly an artefact of properties of the statistical decoder and the a posteriori correction rules (which, to some extent, use durational properties as context). However, it probably also reflects an actual break-off point in the speech rate dimension, where it is no longer possible to realise all phonemes of a word.

The pruned model shows a simplified distribution of realisations for most phonemes. There are often only two alternative realisations in the model where there are three or more alternatives in the key transcript. The distribution of realisations is mostly biased towards the majority class, which is a natural consequence of pruning.

The approximants /ɹ/ and /j/ are relatively hard to handle for the model. This is partly because of the difficulty of the automatic transcription system to produce accurate keys for these consonant phonemes. However, it also reflects an innate ambiguous nature of these phonemes—they often merge with adjacent vowels, and it is very hard to handle this fact in a sequential-segmental description of the speech stream. An /ɹ/ may also merge with a successive dental consonant to form a retroflex consonant. The model produced 92.3% correct realisations for /j/ and only 89.0% correct realisations for /ɹ/.

Other consonants for which the realisations are hard to predict are /d/ and /g/. Neither the [ɹ] nor the ∅ realisation of /d/ is allowed by the model. This results in the fact that /d/ gets the lowest share of correct decisions of any consonant, 77.6%. The phoneme /g/ has many erroneous ∅ realisations in the key transcript and the [g] and ∅ realisations are often confused by the pronunciation model. The share of correct decisions made for /g/ was 79.7%, the second lowest for any consonant.

The relatively infrequent retroflex consonants /ɭ/, /ʂ/ and /ɳ/ also have low shares of correct realisations in the model output, 90.9%, 88.0% and 80.8%, respectively. The low energy fricatives /f/ and /h/ are relatively often confused with ∅ and have shares of correct classifications of 93.3% and 91.3%, respectively.

However, in general, the model is better at predicting the realisations of consonants than the realisations of vowels. The share of correct classifications made by the model for the plosives /p/, /t/, /k/, /b/ and /ɖ/ is around 99%. Also the nasal /ŋ/ has a 99% correct classification rate. Two other semi-vowels, /m/ and /l/ were correctly classified in 97.0% of the cases and in 98.2% of the cases, respectively.

The fricatives /v/, and /s/ and the plosive /ʈ/ also have high correct classification rates, 98.4%, 97.1% and 97.8%, respectively. The realisations of the high energy fricatives /ɧ/ and /ç/ are both correctly predicted by the model in 100% of the cases.

The remaining semi-vowel, /n/, however, is a phoneme that is especially prone to be affected by its context, which increases the variability in the data and lowers the share of correct classifications. The model produced 92.9% correct phone realisations for /n/.

9.2.2 Vowels

The realisation of vowels depends heavily on prosodic information. Mean vowel duration measures are used in high level nodes of the pronunciation model for most vowels. All syllables and some words (and one-word phrases) contain only one vowel. Short words are often high frequency words, and thus the realisations of vowels are often predicted from the duration of the vowel directly. In a speech synthesis context, much is gained if the phone-level pronunciation can be predicted from mean durations over larger units[1] or not using prosodic variables at all. The mean duration attributes and especially the mean vowel duration measures may thus not be fully usable if the model is to be used in a speech synthesis context.

However, the values presented in Table 8.2 in Chapter 8 showed that it was possible to predict the pronunciation in discourse context with tree models not trained on prosodic attributes. The accuracy was lower than when prosodic attributes were available during model training, but still significantly higher than when using the phoneme string to estimate the phone string.

In Swedish, vowels come in pairs, in such a way that there is a phonologically short and a phonologically long variant of each vowel, although the variants may also differ in tongue position or lip rounding. The shares of correct decisions made by the model for different vowels show that the realisations of long vowels are generally easier to predict than the realisations of short vowels. This in spite of the fact that there are generally more possible realisations for the long vowels and that there are more training examples for the short vowels.

The long vowels are more stable in their realisation and were easier to correctly classify by the automatic transcription system. Thus, there is a greater share of correct keys for these vowels (cf. tables 5.7 and 5.8 in Section 5.7.7). However, the fact that the realisation of phonologically long vowels are more predictable may also suggest that the long vowels are more rule-governed in their realisation, while there is a higher degree of free variation for the short vowels.

Some vowels occur more frequently than others and the short variant of a pair is mostly more frequent than the long variant. The exception to this general rule is when the vowel pair is infrequent, but the long vowel occurs in one or several of

[1] In Chapter 10, a method for using the output of a prosodic model as input to the phone-level pronunciation model is discussed.

the most frequent function words. The most clear example of this is the /æ/-/æː/ pair, where the long variant is much more frequent than the short variant, since the long variant occurs in the copula verb *är 'is'*. Table 9.1 presents the number of instances and the share of correct decisions made by the model for each vowel, also shown in tables J.1 to J.23 in Appendix J.

Table 9.1: *The number of instances and the share of correct decisions made by the model for each vowel phoneme.*

Short vowel	Instances	Share correct	Long vowel	Instances	Share correct
ə	3,956	99.44%			
a	7,644	91.55%	ɑː	3,036	98.35%
e	3,353	86.94%	eː	2,833	89.80%
ɪ	3,600	85.00%	iː	1,835	92.81%
ʊ	512	81.05%	uː	842	94.30%
ɵ	1,023	86.61%	ʉ̟ː	1,056	93.18%
ʏ	450	99.33%	yː	126	96.83%
ɔ	4,076	93.20%	oː	2,306	84.78%
ɛ	1,197	98.50%	ɛː	408	94.85%
æ	251	62.15%	æː	1,429	87.82%
œ	152	97.37%	øː	262	94.27%
œ̞	356	57.87%	œ̞ː	543	90.61%

There are some cases where allophones for what is actually the same phonemic class are included in the STA phonetic alphabet and thus in the set of phonemes used in the canonical pronunciation representations. These allophones have thus been treated as phonemes in the work described in this thesis.

The more open /æ/ and /æː/ phonemes are actually pre-/ɹ/ allophones of /ɛ/ and /ɛː/, respectively. The /œ̞/ and /œ̞ː/ phonemes are pre-/ɹ/ allophones of /œ/ and /øː/. In (non-canonical) continuous speech, the more open allophones can be used also when the /ɹ/ has been elided, as a remnant of the /ɹ/ phoneme. The open allophone can also assimilate to its more close allophone counterpart and the vowel may be realised anywhere on a continuous scale between prototypical instances of the allophones.

The phonemes in the front open-mid to mid region of the vowel space thus differ from the other vowels in several respects and are harder for the model to handle. It is obvious from the data presented in this chapter, and in appendices I and J, that certain phonemes are more prone to variable realisation than others, and that certain phonemes show a more free variation than others. Further, certain phonemes show more continuous variation than others, which makes the phone classification ambiguous.

In spite of the prediction problems associated with certain phonemes, the pronunciation models give relatively accurate predictions. Further, with more accurate training keys and more training data, it is expected that the prediction accuracy can be significantly increased. Since the model format is transparent, the model can also be manually changed and the effects of the changes tested. Thus, the data-driven and the knowledge-based approaches to pronunciation modelling can easily be combined using the decision tree paradigm. Where a sufficient amount of data is available, the models can extend linguistic knowledge and in cases where data is sparse, linguistic knowledge can be used to improve the models.

9.3 Summary

In this chapter and appendices I and J, a decision tree pronunciation model has been presented in detail. Rules for the realisation of separate phonemes have been discussed, as well as the realisation distributions and the prediction accuracies for the phonemes. A general discussion dealing with common patterns of classes of phonemes was also included in this chapter.

It could be seen that certain phonemes are more prone to variable realisation than others, i.e., certain phonemes show a more free variation than others and certain phonemes show more continuous variation than others, which makes the phone classification of these phonemes ambiguous. For these phonemes, it is harder for a pronunciation model to give correct realisation predictions. The realisations of certain other phonemes are hard to predict since the phonemes are infrequent and suffer from data sparsity problems to a higher degree than more frequent phonemes. However, in general, the model produces highly accurate predictions.

The next chapter will briefly discuss how pronunciation variation models of the kind described in this chapter and the preceding chapters 7 and 8 can be used in a speech technology application: *speech synthesis*. The next chapter also discusses the fact that the annotation methods and annotation described in chapters 5 and 6 can be used irrespective of particular pronunciation modelling paradigms.

Chapter 10

Pronunciation Modelling in Speech Synthesis

Using pronunciation models to predict the phone-level realisation of words in actual speech databases, as was the case in the cross-validation experiment described in the previous chapter, means that there is a single ‘correct’ realisation of each phoneme. In a speech synthesis setting, several alternative pronunciations can be equally natural-sounding. Also, naturalness is dependent on the entire phone string, not on each segment independently. Listening experiments with speech synthesised using pronunciation variation modelling are thus necessary to evaluate the performance of pronunciation variation models in a synthesis setting. This section will discuss how pronunciation modelling can be used in speech synthesis system to make the speech sound more dynamic and natural.

10.1 The Need for Natural-Sounding Speech Synthesis

As discussed, the background for the work presented in this thesis is that, in natural speech, the pronunciation of a word is not always the same. Instead, the pronunciation depends to a large degree on the context in which the word is uttered. This is also the case in speech generated by state-of-the-art speech synthesis systems. However, speech generated by present day synthesis systems is much less dynamic than natural speech and it is not easy to adapt the speaking style of the synthetic speech to different speaking situations.

Unnatural-sounding speech synthesis may bore or irritate a frequent user in any context. There are, however, some areas where more natural-sounding and adaptable speech synthesis is especially needed, e.g. for people using speech synthesis as a vocal aid and in language (pronunciation) training systems. Further, dialogue systems e.g. for booking tickets and checking timetables are getting increasingly ‘intelligent’, with barge-in, bigger vocabularies and better dialogue handling. Interaction with such systems is thus becoming increasingly like interaction with humans. A system will be perceived of as more coherent if the ‘intelligence’ of the dialogue system is reflected in the synthetic speech produced by the system.

Visually impaired persons using speech synthesis for reading may choose speed and clarity before naturalness. Dyslectic users may also prefer clarity before naturalness. However, synthesis at speech rates above the natural range may still be improved by pronunciation variation modelling, and using a phonetic representation corresponding to a fast-speaking human may be easier to process than a canonical (maximally detailed) phonemic representation (cf. e.g. Ogden et al., 2000).

10.2 Annotation and Speech Synthesis

In addition to increasing the speed of annotation, there was also another reason for using automatic methods when the speech databases used for training pronunciation models were annotated. This auxiliary reason for using automatic methods was to ensure that the corresponding information can be supplied automatically in e.g. a speech synthesis context. The system for annotation was also adapted to be usable irrespective of the presence of a speech signal. It is thus possible to separate the annotation from the signal, so that information derived from text can be represented in the same format as the information in the annotated speech data. Manual annotation and supervision was used in such a way that it was not required in a speech synthesis context. For example, in a speech synthesis context, the orthographic string and utterance boundaries are known and do not have to be manually supplied.

It is not possible to calculate any exact correspondences to the parts of the annotation based on f_0 contours and phoneme durations in a speech synthesis setting. However, such variables can be used to harmonise the prosodic values produced by the prosodic model of a synthesiser with the phone string. For the resulting synthesis to sound natural, it is important that the prosody and the phone string harmonise.

10.3 Pronunciation Modelling and Synthesis System Types

The pronunciation variation information is contained in the speech data annotation and when the information is to be used to improve a speech technology application, rather than for creating a descriptive or explanatory model, the transparity of the model used to represent the information is not critical. The annotation and the annotation methods can thus be used for pronunciation variation modelling in e.g. speech synthesis in other ways than using decision tree classification models.

The optimal way of integrating phone-level pronunciation modelling into a synthesis system depends on the type of system. A speech synthesis system typically consists of two parts, a *front-end* generating parameters which are subsequently used by a *back-end* to generate sound, either by using an articulatory or an acoustic model or through unit (selection and) concatenation.

In a parametric synthesiser or a diphone synthesiser, pronunciation modelling can be used in the front-end for parameter generation. A parametric synthesis front-end generates acoustic or articulatory parameters to the synthesiser back-end. A diphone synthesiser front-end generates phonetic strings used by the back-end to locate the appropriate concatenation units and in the typical case also durations and f_0 contours used to adapt the units. In standard parametric or diphone synthesiser front-ends, there are usually some phonological co-articulation rules operating on the synthesiser-internal phonological representation of the utterance being synthesised. However, such rules typically only take phoneme context into account. They may also make use of the function word/content word dichotomy.

A variable size unit selection system has co-articulation built into its concatenation units to a higher or lower degree. The variation in the data is thus implicit and a canonical phoneme string may be used to represent the speech string both in the database and at unit selection. Pronunciation modelling is here most efficiently used at unit selection. The speech data available for concatenation must then be annotated with context variables important for the phone-level pronunciation. At unit selection, the units are selected that best match the context criteria determined by the synthesis front-end (i.e., target costs) as well as concatenation constraints. Thus, the actual phonetic string is implicit in the speech unit selected. Combinations of context variables giving rise to similar effects on pronunciation will have to be clustered to make the abstractions necessary to use the speech data optimally and thus minimise the amount of data needed. To make good variable size unit selection synthesis of different speaking styles in this manner, speech representing different styles must be present in the database. The problem of finding optimal target cost measures with multiple context variables will need to be addressed.

Hidden Markov Model (HMM) synthesis is a type of fixed-size unit selection synthesis. The units are acoustic models corresponding to n-phones. Several context-specific units of the same n-phone can be created using a context-annotated database at model training. As in the variable size unit selection case, the units may be segmentally different, although these differences are implicit. An advantage of HMM synthesis compared to variable size unit selection synthesis is that models can be clustered and states can be tied using standard methods. Thus, generalisation from the training data is relatively simple. As for the unit selection case, a canonical phonemic representation is enough to represent units. Another advantage of HMM synthesis is that all speech data does not have to come from the same speaker, nor be recorded under the same conditions, since the models can be homogenized after training.

10.4 Using Pronunciation Models with an Existing Speech Synthesis System

Some initial attempts at making an existing diphone speech synthesis system sound more natural have been made, using a pronunciation variation model of the type described in chapters 7 to 9 of this thesis. In the synthesis system used, Rulsys (Carlson and Granström, 1975, 1976; Carlson et al., 1982) is used as the front-end and mbrola with the *Ingmar* voice, created for the commercial infovox 330 synthesis system, is used as the back-end.

One strategy used for generating discourse context-adpapted synthetic speech was to generate f_0 contours and phoneme durations through applying the prosodic model of the synthesiser to a canonical phoneme string and basing the prosodic attribute values for the pronunciation model on the prosodic output. Based on these values and attribute values derived from the orthography, the model produced a new phone-sequence. This sequence was then used to produce the actual synthetic speech. This strategy harmonises the phone sequence with the synthesiser prosody, but the naturalness is largely dependent on the prosodic model. For more information on prosody modelling for speech synthesis, cf. e.g. Bruce and Granström (1993); Horne and Filipsson (1996); Bruce et al. (1996, 2000); Eskénazi (1992); Zellner (1994); Frid (2003), and Fant and Kruckenberg (2002).

Another strategy tested was to simulate an optimal prosodic model by re-synthesising the prosody (f_0 contour and durations) of extracts from recorded speech and using the decision tree pronunciation model to produce phone-strings based on the actual prosody of the recorded speaker (with some adaptations to fit the synthesis system[1]). The re-synthesised prosody was used also for producing the synthetic speech.

This strategy did not produce optimal synthesis, mainly since the diphone database available was not designed for producing highly reduced speech. Some phone strings produced by the decision tree models were not possible to synthesise because diphones corresponding to phone sequences not present in canonical speech are not included in the diphone database. However, in actual speech and in model-produced phone strings, phones not occurring in sequence in canonical pronunciation representations may be paired. This problem is mostly caused by syllable elisions and may be "solved" trough [ə] epenthesis.

However, the largest problem was that only very little allophonic variation could be described using the small set of diphones available. For example, there was only the choice between a full vowel and a [ə] and to make the synthesised speech with re-synthesised prosody and model-generated phone strings truly sound natural, there is a need for allophonic variants in-between full vowels and [ə]. There were also problems with the speech rate; the actual speech rate could often not be re-synthesised, since diphones had to be of a certain duration for noise not to be introduced at concatenation.

Among others, Kohler (2000) and Prahallad et al. (2006) discuss the fact that the linear segmental approach to speech is not always successful at describing connected speech processes. The changes when going from a maximally detailed pronunciation of a word to a highly reduced form in connected speech mostly do not appear in

[1] The f_0 was adapted to the base frequency of the synthetic voice and the pitch contour was smoothed, since too high δf_0 values introduced noise during synthesis.

quantal steps, but there is more of a successive lenition where speech sounds become continuously (however, not necessarily linearly) less spectrally and durationally salient.

Further, the linear-sequential assumption used in standard phonological and segmental-phonetic descriptions of speech becomes less and less useful the faster and less formal the speech under study becomes. In going from citation form pronunciations to spontaneous continuous speech, certain features of phonemes can remain and transfer to adjacent phonemes, while units with durational properties (what is generally conceived of as separate phonetic segments) corresponding to these phonemes are no longer present (cf. e.g. Jakobson et al., 1963).

To catch this non-segmental property of pronunciation variation, there is a need for continuous models or more detailed discrete models including allophones describing various steps between the maximally detailed phone and elision, using some description below the phone-level. To make truly natural synthetic speech, it is necessary to include speech of many types, especially spontaneous speech, in concatenation databases or to develop methods for controlling the variation in continuous spontaneous speech in other ways.

10.5 Summary

This chapter has discussed how pronunciation modelling can be used in different types of speech synthesis systems and the problem of the small set of phonemes included in the phoneme set used for the annotation described in this thesis. To make speech synthesis sound truly natural, there is a need for having a phoneme set that can describe a wider range of variation. It is also necessary to include speech of many types, especially spontaneous speech, in concatenation databases or to develop methods for controlling the variation in continuous spontaneous speech in other ways.

The next chapter will briefly summarise each chapter of the thesis and give some general conclusions.

Chapter 11

Summary and Conclusions

The focus of this thesis has been the modelling of systematic, discourse context-induced variation in phone-level pronunciation inherent in the central standard Swedish language variety. The methods used can, however, easily be adapted for modelling other language varieties and languages. The aim has been to find patterns common to the language variety while idiomatic variation specific to individual language users is avoided.

A data-driven approach was used for this task and the work involved annotating spoken language with linguistic and related information on levels ranging from the discourse down to articulatory features, and machine learning was used to create pronunciation models from the annotation. An important part of the work was the development of an annotation scheme, so that data could be organized in a way that was theoretically and practically appropriate for the current purposes. Another important part of the work was the development of methods for data annotation.

The work described in this thesis was partly driven by an interest in human language processing and the factors involved in how humans choose to alter their speech over different situations. The work was also partly driven by an interest in using knowledge about human language performance to improve speech technology systems, such as synthetic speech.

To accommodate both of these aspects of pronunciation in discourse context, the decision tree induction paradigm was used for creating pronunciation models. This paradigm is not impeded by the fact that the data from which a model is to be induced is of disparate kinds, as it was in the annotation used for the work on pronunciation modelling described in this thesis. The decision tree paradigm also produces transparent models, which can easily be transformed into rules.

11.1 Pronunciation Lexicon Development

The point of departure for the data-driven pronunciation modelling method described in this thesis was a set of context independent pronunciation representations that correspond to phonemic descriptions of the type that can be found in a pronunciation lexicon. For the method to be successful, it was important that the phonemic pronunciation descriptions were of high and consistent quality. A part of the pronunciation modelling research reported in this thesis was thus aimed at developing a canonical pronunciation lexicon for Swedish.

This machine-readable pronunciation lexicon was called CentLex and based on lexical data resulting from a number of projects at the Department of Speech, Music and Hearing (TMH) and the Centre for Speech Technology (CTT) at KTH over the years. However, CentLex has been expanded beyond the original data.

Since a high quality pronunciation lexicon is of the essence for many areas of speech technology research and for most speech technology applications, CentLex was built to be a central lexicon database for the Department of Speech, Music and Hearing at KTH and the Centre for Speech Technology (CTT). The lexicon was designed to meet the specific demands of the phone-level pronunciation modelling project which was the focus of this thesis, as well as general demands from speech technology research and application development. Tools for facilitating access to the lexicon and for continuous, co-operative editing of the lexicon database were developed.

CentLex is a full-form lexicon, with each entry minimally containing an orthographic word form and a grammatical analysis (Part of Speech and morphology). An entry can also have an arbitrary number of phonemic representations, ordered by their probability of use. Each phonemic representation may be enriched with information about the intended context of the representation (e.g. *reduced form* or *foreign language*). Such information is added e.g. for proper names, since orthographically identical names may be pronounced differently depending on the native language environment of the person bearing the name. An entry also contains information about the probability of the particular grammatical analysis, given the orthographic word (estimated from a large automatically tagged text corpus).

11.2 Pronunciation Lexicon Evaluation

CentLex was evaluated for coverage and pronunciation representation quality. At the time of evaluation, CentLex contained 410,326 entries and 332,626 unique word forms. The coverage of CentLex was calculated over some different text types, tokenised and automatically tagged. The average coverage over the texts was 94.0% of the CentLex entry types (combinations of an orthographic word and a grammatical analysis) and 95.1% of the orthographic words.

The evaluation of pronunciation representation quality showed that among the high frequency entry types (defined as the set of the most frequent entry types covering 50% of the tokens in the texts used for evaluation), no pronunciation representations were obviously erroneous, although some differed from the agreed upon standard. In most cases, the discrepancy was that a reduced form of a function word was listed as the highest ranking alternative, although a canonical pronunciation should be in this position, according to the CentLex standard.

Among the mid frequency entry types (defined as the entry types from a frequency-sorted list covering 50 to 90 per cent of the tokens in the texts) and low frequency entry types (defined as the remainder of entry types, covering 90 to 100 per cent of the tokens), the estimated shares of erroneous pronunciation representations were 1.7% and 6.6%, respectively.

The CentLex database has been used as a lexicon in an experimental speech synthesis system and in a large vocabulary speech recognition system. CentLex has also been used for training grapheme-to-phoneme conversion rules for commercial speech synthesis and as a lexicon for commercial speech synthesis applications. It has further been used as a reference in the development of a system for production of talking books with synthetic speech for visually impaired and dyslectic university students. Finally, CentLex has been used for annotation in a research project aimed at context-sensitive prosody prediction and, of course, in for phone-level pronunciation prediction as described in this thesis.

11.3 Annotation Method

The data-driven approach to pronunciation modelling required annotation of spoken language with linguistic and related information, from which a machine learning algorithm could induce pronunciation models. Speech data of different types was annotated in six linguistically motivated layers, 1) a discourse layer, 2) an utterance layer, 3) a phrase layer, 4) a word layer, 5) a syllable layer, and 6) a phoneme layer. The layers were segmented into their specific unit types and linguistic information was asssociated with each unit of each layer.

Each monologue, interview and radio news broadcast was considered a separate discourse. The utterance layer was manually segmented and the word layer was segmented using an automatic alignment system, forcing word boundaries at the manually obtained utterance boundaries. The word boundaries were manually checked and corrected and the alignment system was used to segment the phoneme layer, forcing phoneme boundaries at the manually checked word boundaries. A part-of-speech tagger and a parser were used to chunk the word string into phrases and the phrase layer was segmented by aligning the phrases to the signal using the word boundaries. The phoneme string was clustered into syllables using rules and the syllable layer was segmented through aligning the syllables to the signal using the phoneme boundaries.

Mean phoneme duration measures were calculated based on absolute and normalised duration, respectively, on linear and logarithmic duration, respectively, and based on the duration of all phonemes and on the duration of vowels only. Pitch dynamics measures were calculated from the distance between f_0 peaks and valleys and either the median f_0 or a *base frequency* located 1.5σ below the mean frequency of the particular speaker. Pitch range, a measure defined as the difference between the highest f_0 peak and the lowest f_0 valley contained by a particular unit, was also calculated. The pitch-based measures were calculated with pitch measured in Hertz and on several psychoacoustic scales.

Further, *word predictability* and related measures were calculated. Word predictability was defined as the weighted combination of the trigram, bigram and unigram probabilities calculated from an the orthographic transcripts of a spoken language corpus.

Phonetic annotation was necessary, since this was used as the key during pronunciation model training. Manual phonetic annotation of speech, especially of conversational speech, is a time-consuming task and it was not possible to supply manual phonetic transcripts for all of the speech data used. Instead, a hybrid automatic transcription system using statistical decoding and a set of a posteriori correction rules was developed for supplying a context-dependent realisation of each phoneme in the canonical pronunciation representation. The automatically obtained phones were used as keys during pronunciation model training. Compared to a small, manually transcribed gold standard, the automatic transcription system produced a phone error rate of 15.5%.

11.4 Information Included in the Annotation

In the discourse layer, variables which are constant over the discourse were annotated. A set of *mean phoneme duration* measures and four speaking style-related variables: *number of discourse participants*, *degree of formality*, *degree of spontaneity* and *type of interaction* were attached to each discourse layer unit. Mean phoneme duration measures were calculated over the units in each annotation layer, except the phoneme layer.

In the utterance layer, the variables *speaker pitch register* and a coarse four-way division into *utterance types*, corresponding to basic dialogue acts, were included in the annotation. Sets of *pitch range* and *pitch dynamics* 'speech liveliness' measures were also included in the utterance layer annotation. Such measures were calculated over the utterance, the phrase and the word.

The units of the phrase layer were annotated with a *phrase type* tag and a set of *phrase length* measures. In the word layer annotation, *Part of Speech* and morphological information was included along with *word type* information (content word or function word), the particular function word or a generic 'content word' representation, a set of word predictability-related measures, the position of the word in the *phrase* and in a *collocation*, respectively, and the number of repetitions of the *full-form word* and of the *lexeme* thus far in the discourse.

The information included in the syllable layer annotation was the *stress* and *accent* of the current syllable, the distances to the nearest *stressed syllables* and to the nearest *primary stressed* syllables, respectively, the *syllable length*, the *syllable nucleus* and the *position of the syllable in the word*.

The variables included in the phoneme layer annotation were *phoneme identity*, a set of articulatory features describing the canonical phoneme, *the position of the phoneme in the syllable* (onset, nucleus or coda), *consonant cluster length* and *position in the cluster*. The identity of the automatically obtained phone was also included in the phoneme layer annotation.

11.5 Pronunciation Model Creation

The annotation was used by a decision tree induction machine learning algorithm to create models describing phoneme realisation in discourse context. The decision tree paradigm was used since the resulting models are transparent and since the induction algorithm is not impeded by the fact that the data from which a model is to be induced are of disparate kinds, as is the case for the annotation described in this thesis.

Training instances were compiled from the structured annotation. Using the phoneme as the primary unit, a set of training instances, essentially being context-sensitive phonemes, were created. Each instance contained information about the current phoneme, and about the current unit in all higher annotation layers. The instance also contained information about the sequential context of the current unit in each layer. In all, each training instance included a set of 516 attribute values and the key phone realisation.

The particular decision tree implementation used for pronunciation model induction was the DTree program suite (Borgelt, 2004a). The best classification performance was obtained when selecting attributes with the *symmetric information gain ratio* measure (Lopez de Mantaras, 1991; Borgelt and Kruse, 1998), and allowing the inducer to group discrete values to obtain the optimal number of nodes at each level.

Training data generally contains some degree of noise and a decision tree may be biased toward the particular noise in the data used for inducing the tree (overtrained). However, once a tree is constructed, it can be pruned to make it more generally applicable. The idea behind pruning is that the most common patterns are kept in the model, while less common patterns, with high probability of being due to noise in the training data, are disregarded.

11.6 Pronunciation Model Evaluation

In a set of tenfold cross-validation experiments, decision tree models were created from different types of speech and with access to different subsets of attributes. It was shown that including information from multiple layers improves the performance of the decision tree models, most notably for spontaneous speech, where the predictive power of phonological and grammatical information is relatively low. A comparison between models trained only on phoneme layer attributes and models trained on attributes from all layers showed that the prediction accuracy of pronunciation models could be improved by 42.2% by including information from above the phoneme level.

Attributes from all layers of annotation were used in the models with the highest prediction accuracy. The optimal models, trained on all available data and with access to all attributes, produced an average phone error rate of 8.2%, which is an improvement of 60.0% compared to using the phoneme string for estimating the phone-level realisation.

Repetitions of the experiment showed that the results are relatively stable over different data randomisations and thus that the method is reliable. Experiments excluding a particular speaker from training and evaluating the model on that speaker indicated that although there is a minor degree of idiomatic variation, variation in pronunciation due to discourse context is mainly a language trait.

The attributes involved in predicting the phone realisations in discourse context were ranked according to their position in the decision trees and according to their position weighted with the number of decisions they were involved in making (over the training data). As expected, the identity of the current phoneme was the highest ranking predictor for both methods of ranking. The identities of the phonemes at positions ± 4 in relation to the current phoneme, the *function word* attribute and the mean absolute phoneme duration over the word were other high ranking attributes.

11.7 Phoneme-to-Phone Conversion

Certain phonemes are more prone to variable realisation than others. That is, certain phonemes show a more free variation than others and certain phonemes show more continuous variation than others, which makes the phone classification of these phonemes ambiguous. For these phonemes, it is harder for a pronunciation model to give correct realisation predictions. The realisations of other phonemes are hard to predict since the phonemes are infrequent and suffer from data sparsity problems. However, in general, the models give highly accurate predictions.

The approximants /ɹ/ and /j/ were relatively hard to handle for the model. This is partly because of the difficulty of the automatic transcription system to produce accurate keys for these consonant phonemes. However, it also reflects an innate ambiguous nature of the phonemes—they often merge with adjacent vowels, and it is very hard to handle this fact in a sequential-segmental description of the speech stream. An /ɹ/ may also merge with a successive dental consonant to form a retroflex consonant. The model produced 92.3% correct realisations for /j/ and only 89.0% correct realisations for /ɹ/.

Other consonants that were hard to predict the realisations for were /d/ and /g/. Neither the [ɹ] nor the ∅ realisation of /d/ were allowed by the model, although both realisations were present in the key transcripts. This resulted in the fact that /d/ received the lowest share of correct decisions of any consonant, 77.6%. The phoneme /g/ had many erroneous ∅ realisations in the key transcript and the [g] and ∅ realisations were often confused by the pronunciation model. The share of correct decisions made for /g/ was 79.7%, the second lowest for any consonant.

The relatively infrequent retroflex consonants /ɭ/, /ʂ/ and /ɳ/ also had low shares of correct realisations in the model output, 90.9%, 88.0% and 80.8%, respectively. The low energy fricatives /f/ and /h/ were relatively often confused with ∅ and had shares of correct classifications of 93.3% and 91.3%, respectively. The nasal /n/ is a phoneme that is especially prone to be affected by its context, which increases the variability in the data and lowers the share of correct classifications. The model produced 92.9% correct phone realisations for /n/. For other consonants, the model produced between 97 and 100% correct classifications.

The realisation of vowels depends heavily on prosodic information, mostly mean vowel duration and mean phoneme duration over words and phrases. The shares of correct decisions made by the model for different vowels show that the realisations of long vowels are generally easier to predict than the realisations of short vowels, in spite of the fact that there are generally more possible realisations for the long vowels and that there are more training examples for the short vowels. This indicates that the long vowels are more stable and rule-governed in their realisation, while there is a higher degree of free variation for the short vowels.

In spite of prediction problems associated with certain phonemes, the pronunciation models give relatively accurate predictions and with more accurate training keys and more training data, it is expected that the prediction accuracy can be significantly increased. Since the decision tree model format is transparent, the models can also be manually changed and the effects of the changes tested. Thus, the data-driven and the knowledge-based approaches to pronunciation modelling can easily be combined using the decision tree paradigm. Where a sufficient amount of data is available, the models can extend linguistic knowledge and in cases where data is sparse, linguistic knowledge can be used to improve the models.

11.8 Pronunciation Modelling in Speech Synthesis

For speech synthesis used as a vocal aid, in pronunciation training systems, and in 'intelligent' human-computer dialogue systems, there is a need for more dynamic, human-sounding speech synthesis. Visually impaired persons using speech synthesis for reading may choose speed and clarity before naturalness. Dyslectic users may also prefer clarity before naturalness. However, synthesis at speech rates above the natural range may still be improved by pronunciation variation modelling and using a phonetic representation corresponding to a fast-speaking human may be easier to process than a canonical (maximally detailed) phonemic representation.

The pronunciation variation information is contained in the speech data annotation and when the information is to be used to improve a speech technology application rather than for creating a descriptive or explanatory model, the transparity of the model used to represent the information is not critical. The annotation and the annotation methods can thus be used for pronunciation variation modelling in e.g. speech synthesis in other ways than using decision tree classification models. The optimal way of integrating phone-level pronunciation modelling into a synthesis system depends on the type of system.

In a parametric synthesiser, pronunciation modelling can be used for parameter generation. In a diphone synthesiser, pronunciation modelling can be used for generating phonetic strings. A variable size unit selection system has co-articulation built into its concatenation units to a higher or lower degree. The variation in the data is thus implicit and a canonical phoneme string may be used to represent the speech string both in the database and at unit selection. Pronunciation modelling is here most efficiently used at unit selection. The speech data available for concatenation must then be annotated with context variables important for the phone-level pronunciation.

For an HMM synthesiser, several context-specific units of the same n-phone can be created using a context-annotated database at model training. As in the variable size unit selection case, the units may be segmentally different, although these differences are implicit. An advantage with HMM synthesis is that all speech data does not have to come from the same speaker nor be recorded under the same conditions, since the models can be homogenized after training.

Some initial attempts at using a pronunciation variation model of the type described in this thesis for making an existing diphone speech synthesis system sound more natural were made. It proved hard to incorporate a pronunciation variation model into the diphone synthesiser, since the diphones units were created for a pronunciation close to the canonical pronunciation. To make synthetic speech sound natural, there is a need for e.g. more allophones in-between full vowels and schwa.

Further, the linear-sequential assumption used in standard phonological and segmental-phonetic descriptions of speech becomes less and less useful the faster and less formal the speech under study becomes. In going from citation form pronunciations to spontaneous continuous speech, certain features of phonemes can remain and transfer to adjacent phonemes, while units with with durational properties (what is generally conceived of as separate phonetic segments) corresponding to these phonemes are no longer present.

To catch this non-segmental property of pronunciation variation, there is a need for continuous models or more detailed discrete models including allophones describing various steps between the maximally detailed phone and elision, using some description below the phone-level. To make truly natural synthetic speech, it is necessary to include speech of many types, especially spontaneous speech, in concatenation databases or to develop methods for controlling the variation in continuous spontaneous speech in other ways.

11.9 General Conclusions

The work described in this thesis was partly driven by an interest in human language processing and the factors involved in how humans choose to alter their speech over different situations. The work was also partly driven by an interest in using knowledge about human language performance to improve speech technology applications, such as speech synthesis systems.

The method used to model the phone-level pronunciation of words in discourse context seems to hold and gives highly accurate predictions, although more accurate training keys and more training data is expected to significantly increase the prediction accuracy. Since decision tree models are transparent, linguistic knowledge can also be used directly to improve the models.

Models of the type induced are expected to be usable for creating dynamic and highly natural-sounding speech synthesis. However, the phone set must be extended to contain more allophonic variation. If some form of concatenation synthesis system is used, the speech data used in the particular type of synthesis system must contain spontaneous speech.

Bibliography

Adda-Decker, M. and L. Lamel (2000). Modeling reduced pronunciations in German. In *PHONUS 5*, pp. 145–159. Saarbrücken: Institute of Phonetics, University of the Saarland.

Allwood, J. (1999). The Swedish spoken language corpus at Göteborg university. In *Proceedings from Fonetik*, volume 81 of *Gothenburg Papers in Theoretical Linguistics*, Göteborg, Sweden.

Allwood, J., M. Björnberg, L. Grönqvist, E. Ahlsén, and C. Ottesjö (2000). The spoken language corpus at the linguistics department, Göteborg university. *Forum: Qualitative Social Research*, 1(3).

Allwood, J., L. Grönqvist, E. Ahlsén, and M. Gunnarsson (2002). Göteborgskorpusen för talspråk (The Göteborg spoken language corpus). In *Nydanske Sprogstudier 30*, pp. 39–58. København: Akademisk Forlag.

Antworth, E. L. (1990). *PC-KIMMO: A two-level processor for morphological analysis.* Number 16 in Occasional Publications in Academic Computing. Dallas, Texas: Summer Institute of Linguistics.

Antworth, E. L. (1995). User's guide to PC-KIMMO version 2. ftp://ftp.sil.org/software/dos/pc-kimmo/guide.zip.

Aycock, J. (1998). Compiling little languages in Python. In *Proceedings of the International Python Conference*, Houston, Texas.

Bannert, R. and P. E. Czigler (1999). *Variations in consonant clusters in standard Swedish.* Phonum 7, Reports in Phonetics. Umeå: Umeå University.

Bates, R. and M. Ostendorf (2001). Modeling pronunciation variation in conversational speech using syntax and discourse. In *Proceedings of the ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, Red Bank, New Jersey.

Bates, R. and M. Ostendorf (2002). Modeling pronunciation variation in conversational speech using prosody. In *Proceedings of the ISCA Tutorial and Research*

Workshop Pronunciation Modeling and Lexicon Adaptation for Spoken Language (PMLA), Estes Park, Colorado.

Bates, R. A. (2003). *Speaker Dynamics as a Source of Pronunciation Variability for Continuous Speech Recognition Models.* PhD thesis, University of Washington.

Bell, A., D. Jurafsky, E. Fosler-Lussier, C. Girand, M. L. Gregory, and D. Gildea (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America*, 113(2):1001–1024.

Bell, L. (2003). *Linguistic Adaptations in Spoken Human-Computer Dialogues: Empirical Studies of User Behaviour.* PhD thesis, Stockholm: KTH, Department of Speech, Music and Hearing.

Bennett, C. L. and A. W. Black (2003). Using acoustic models to choose pronunciation variations for synthetic voices. In *Proceedings of Eurospeech*, pp. 2937–2940.

Bennett, C. L. and A. W. Black (2005). Prediction of pronunciation variations for speech synthesis: A data-driven approach. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 297–300, Philadelphia, Pennsylvania.

Beringer, N. (2003a). Boundary deviation of phonemes in automatic segmentation systems – A cross language study. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, pp. 1317–1320, Barcelona, Catalonia.

Beringer, N. (2003b). Rule-based categorial analysis of unprompted speech – A cross-language study. In *Proceedings of the Phonetics and Phonology in Iberia (PaPI) Conference*, Lisbon, Portugal.

Beringer, N. and M. Neff (2000a). Generation of pronunciation rule sets for automatic segmentation of American English and Japanese. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Beijing, China.

Beringer, N. and M. Neff (2000b). Regional pronunciation variants for automatic segmentation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Athens, Greece.

Beringer, N. and F. Schiel (1999). Independent automatic segmentation of speech by pronunciation modeling. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, pp. 1653–1656, San Francisco, California.

Binnenpoorte, D. and C. Cucchiarini (2003). Phonetic transcription of large speech corpora: How to boost efficiency without affecting quality. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, pp. 2981–2984, Barcelona, Catalonia.

Binnenpoorte, D., C. Cucchiarini, H. Strik, and L. Boves (2004). Improving automatic phonetic transcription of spontaneous speech through variant-based pronunciation variation modelling. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 681–684, Lisbon, Portugal.

Binnenpoorte, D., S. Goddijn, and C. Cucchiarini (2003). How to improve human and machine transcriptions of spontaneous speech. In *Proceedings of the ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, pp. 147–150, Tokyo, Japan.

Borgelt, C. (1998). A decision tree plug-in for DataEngine. In *Proceedings of the European Congress on Intelligent Techniques and Soft Computing (EUFIT)*, volume 2, pp. 1299–1303, Aachen (Aix-la-Chapelle), Germany.

Borgelt, C. (2004a). Dtree. http://fuzzy.cs.uni-magdeburg.de/∼borgelt/dtree.html.

Borgelt, C. (2004b). Dtview: Decision and regression tree visualization. http://fuzzy.cs.uni-magdeburg.de/∼borgelt/doc/dtview/dtview.html.

Borgelt, C. and R. Kruse (1998). Attributauswahlmaße für die induktion von entscheidungsbäumen: Ein Überblick (Attribute selection measures for decision tree induction: An overview). In Nakhaeizadeh, G. (editor), *Data Mining: Theoretische Aspekte und Anwendungen*, pp. 77–98. Heidelberg: Physica-Verlag.

Brants, T. (2000). TnT – A statistical Part-of-Speech tagger. In *Proceedings of the Applied Natural Language Processing Conference (ANLP)*, pp. 224–231, Seattle, Washington.

Brodda, B. (1966). *En algoritm för att bestämma »avståndet« mellan ord (An algorithm for determining the "distance" between words)*. Interim Report 4. Stockholm: Reseach Group for Quantitative Linguistics (KVAL).

Bruce, G. (1985). Fonologiska regler för elliptiskt tal (Phonological rules for elliptic speech). In Allén, S. (editor), *Svenskans beskrivning 15*, pp. 149–158. Göteborg: Göteborg University.

Bruce, G. (1986). Elliptical phonology. In Dahl, Ö. (editor), *Papers from the Scandinavian Conference on Linguistics*, pp. 86–95. Stockholm: Stockholm University.

Bruce, G., M. Filipsson, J. Frid, B. Granström, K. Gustafson, M. Horne, D. House, B. Lastow, and P. Touati (1996). Developing the modelling of Swedish prosody in spontaneous dialogue. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 1, pp. 370–373, Philadelphia, Pensylvania.

Bruce, G., M. Filipsson, J. Frid, B. Granström, K. Gustafson, M. Horne, and D. House (2000). Modelling of Swedish text and discourse intonation in a speech synthesis framework. In Botinis, A. (editor), *Intonation: Analysis, Modelling and Technology*, pp. 291–320. Dordrecht, Boston, London: Kluwer.

Bruce, G. and B. Granström (1993). Prosodic modelling in Swedish speech synthesis. *Speech Communication*, 13:63–73.

Byrne, W., M. Finke, S. Khudanpur, J. McDonough, H. Nock, M. Riley, M. Saraclar, C. Wooters, and G. Zavaliagkos (1998). Pronunciation modelling using a hand-labelled corpus for conversational speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 313–316, Seattle, Washington.

Campbell, N. (1995). Prosodic influence on segmental quality. In *Proceedings of Eurospeech*, pp. 1011–1014, Madrid, Spain.

Carlson, R. and B. Granström (1975). A phonetically oriented programming language for rule description of speech. In *Proceedings of the Speech Communication Seminar*, volume 2. Stockholm: Almqvist & Wiksell.

Carlson, R. and B. Granström (1976). A text-to-speech system based entirely on rules. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 686–688, Philadelphia, Pensylvania.

Carlson, R., B. Granström, M. Heldner, D. House, B. Megyesi, E. Strangert, and M. Swerts (2002). Boundaries and groupings – The structuring of speech in different communicative situations: A description of the GROG project. In *Proceedings of Fonetik*, pp. 65–68, Stockholm, Sweden.

Carlson, R., B. Granström, and S. Hunnicutt (1982). A multi-language text-to-speech module. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 3, pp. 1604–1607, Paris, France.

Carlson, R., B. Granström, and A. Lindström (1990). Methods to generate the pronunciation of proper names in Swedish. In *Proceedinge of Fonetik*, pp. 74–77, Umeå, Sweden.

Chang, S., S. Greenberg, and M. Wester (2001). An elitist approach to articulatory-acoustic feature classification. In *Proceedings of Eurospeech*, pp. 1725–1728, Aalborg, Denmark.

Chang, S., L. Shastri, and S. Greenberg (2000). Automatic phonetic transcription of spontaneous speech (American English). In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Bejing, China.

Chung, G., C. Wang, S. Seneff, E. Filisko, and M. Tang (2004). Combining linguistic knowledge and acoustic information in automatic pronunciation lexicon generation. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 1457–1460, Jeju Island, Korea.

Cucchiarini, C. and D. Binnenpoorte (2002). Validation and improvement of automatic phonetic transcriptions. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 313–316, Denver, Colorado.

Cucchiarini, C., D. Binnenpoorte, and S. Goddijn (2001). Phonetic transcriptions in the spoken Dutch corpus: How to combine efficiency and good transcription quality. In *Proceedings of Eurospeech*, pp. 1679–1682, Aalborg, Denmark.

Cucchiarini, C. and H. Strik (2003). Automatic phonetic transcription: An overview. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, pp. 347–350, Barcelona, Catalonia.

Demuynck, K., J. Duchateau, and D. Van Compernolle (1997). A static lexicon network representation for cross-word context dependent phones. In *Proceedings of Eurospeech*, pp. 143–146, Rhodes, Greece.

Demuynck, K. and T. Laureys (2002). A comparison of different approaches to automatic speech segmentation. In *Proceedings of the International Conference on Text, Speech and Dialogue (TSD)*, pp. 277–284, Brno, Czech Republic.

Demuynck, K., T. Laureys, and S. Gillis (2002). Automatic generation of phonetic transcriptions for large speech corpora. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 333–336, Denver, Colorado.

Demuynck, K., T. Laureys, P. Wambacq, and D. Van Compernolle (2004). Automatic phonemic labeling and segmentation of spoken Dutch. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 61–64, Lisbon, Portugal.

Duez, D. (1998). Consonant sequences in spontaneous French speech. In *Proceedings of the ESCA Sound Patterns of Spontaneous Speech (SPoSS) workshop*, pp. 63–68, La Baume-les-Aix, France.

Duez, D. (2001). Reduction and assimilatory effects in conversational French speech. In Keller, E., G. Bailly, A. Monaghan, J. Terken, and M. Huckvale (editors), *Improvements in speech synthesis*, pp. 228–236. Chichester: John Wiley & Sons.

Ejerhed, E., G. Källgren, O. Wennstedt, and M. Åström (1992). *The Linguistic Annotation System of the Stockholm-Umeå Corpus Project.* Umeå: Department of Linguistics, Umeå University.

Eklund, R. (1999). A comparative study of disfluencies in four Swedish travel dialogue corpora. In *Proceedings of the ICPhS Sattelite Workshop on Disfluency in Spontaneous Speech (DiSS)*, pp. 3–6, Berkeley, California.

Eklund, R. and A. Lindström (2001). Xenophones: An investigation of phone set expansion in Swedish and implications for speech recognition and speech synthesis. *Speech Communication*, 35(1–2):81–102.

Elert, C.-C. (1964). *Phonologic Studies of Quantity in Swedish: Based on material from Stockholm speakers*, volume 27 of *Monografier utgivna av Stockholms kommunalförvaltning.* Uppsala: Almquist & Wiksell. Ph.D. Thesis, Uppsala University.

Elert, C.-C. (1970). *Ljud och ord i svenskan (Sounds and Words in the Swedish Language).* Stockholm: Almqvist & Wiksell.

Eliasson, S. (1986). Sandhi in peninsular Scandinavian. In Andersen, H. (editor), *Sandhi phenomena in the languages of Europe*, pp. 271–300. Berlin: Mouton de Gruyter.

Engstrand, O. (1988). Articulatory correlates of stress and speaking rate in Swedish VCV utterances. *Journal of the Acoustical Society of America (JASA)*, 83(5):1863–1875.

Engstrand, O. (1992). Systematicity of phonetic variation in natural discourse. *Speech Communication*, 11:337–346.

Engstrand, O. (1999). *Handbook of the International Phonetic Association*, chapter Illustrations of the IPA: Swedish, pp. 140–142. Cambridge: Cambridge University Press.

Engstrand, O. and D. Krull (1988). On the systematicity of phonetic variation in spontaneous speech. *Phonetic Experimental Research, Institute of Linguistics, University of Stockholm (PERILUS)*, 8:34–47.

Eskénazi, M. (1992). Changing speech styles: Strategies in read speech and casual and careful spontaneous speech. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 755–758.

Fant, G. and A. Kruckenberg (2002). A new approach to intonation analysis and synthesis of Swedish. In *Proceedings of Speech Prosody*, pp. 283–286, Aix-en-Provence, France.

Fant, G., A. Kruckenberg, J. Liljencrants, and A. Botinis (2001). Prominence correlates. a study of Swedish. In *Proceedings of Eurospeech*, pp. 657–660, Aalborg, Denmark.

Fant, G., B. Lindblom, and A. de Serpa-Leitão (1966). Consonant confusions in english and swedish. a pilot study. *STL-QPSR*, 7(4):31–34.

Finke, M. and A. Waibel (1997). Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. In *Proceedings of Eurospeech*, pp. 2379–2382, Rhodes, Greece.

Fosler-Lussier, E. and N. Morgan (1998). Effects of speaking rate and word predictability on conversational pronunciations. In *Proceedings of the ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pp. 35–40, Kerkrade, the Netherlands.

Fosler-Lussier, E. and N. Morgan (1999). Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Communication*, 29(2–4):137–158.

Frid, J. (2003). *Lexical and Acoustic Modelling of Swedish Prosody.* Number 45 in Travaux de l'institut de linguistique de Lund. Department of Linguistics and Phonetics, Lund University.

Fukada, T., T. Yoshimura, and Y. Sagisaka (1998). Automatic generation of multiple pronunciations based on neural networks and language statistics. In *Proceedings of the ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pp. 41–46, Kerkrade, the Netherlands.

Gårding, E. (1967). *Internal juncture in Swedish*, volume 6 of *Travaux de l'Institute de Phonétique de Lund.* Lund: C.W.K Gleerup.

Gårding, E. (1974). Sandhiregler för svenska konsonanter (Sandhi rules for Swedish consonants). In Platzack, C. (editor), *Svenskans beskrivning 8*, pp. 97–106. Lund: Lund University, Department of Nordic Languages.

Garlén, C. (editor) (2003). *Svenska språknämndens uttalsordbok ('Pronunciation dictionary of the Swedish Language Council').* Stockholm: Norstedts Akademiska Förlag, first edition.

Greenberg, S. (2003). Pronunciation variation is key to understanding spoken language. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, pp. 219–222, Barcelona, Catalonia.

Greenberg, S., H. Carvey, and L. Hitchcock (2002). The relation between stress accent and pronunciation variation in spontaneous American English discourse. In *Proceedings of the Speech Prosody conference*, Aix-en-Provence, France.

Greenberg, S. and E. Fosler-Lussier (2000). The uninvited guest: Information's role in guiding the production of spontaneous speech. In *Proceedings of the Crest Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling*, Kloster Seeon, Bavaria.

Gregory, M. L., W. Raymond, A. Bell, E. Fosler-Lussier, and D. Jurafsky (1999). The effects of collocational strength and contextual predictability in lexical production. *Proceedings of the Chicago Linguistics Society (CLS)*, 35:151–166.

Grice, H. P. (1975). Logic and conversation. In Cole, P. and J. L. Morgan (editors), *Syntax and Semantics, Volume 3: Speech Acts*, pp. 41–58. New York: Academic Press.

Grice, H. P. (1989). Logic and conversation. In Grice, P. (editor), *Studies in the Way of Words*, pp. 22–40. Cambridge: Harvard University Press.

Gustafson, J. (1995a). Transcribing names with foreign origin in the ONOMASTICA project. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, volume 2, pp. 318–321, Stockholm, Sweden.

Gustafson, J. (1995b). Using two-level morphology to transcribe Swedish names. In *Proceedings of Eurospeech*, volume 3, pp. 2231–2234, Madrid, Spain.

Gustafson, J. (1996). A swedish name pronunciation system. Licenciate thesis, Stockholm: KTH, Department of Speech, Music and Hearing.

Gustafson, J. (2002). *Developing Multimodal Spoken Dialogue Systems: Empirical Studies of Spoken Human-Computer Interaction.* PhD thesis, Stockholm: KTH, Department of Speech, Music and Hearing.

Gustafson-Čapková, S. and B. Megyesi (2002). Silence and discourse context in read speech and dialogues in swedish. In *Proceedings of the Speech Prosody conference*, pp. 363–366.

Handke, J. (1995). *The Structure of the Lexicon: Human versus Machine.* Number 5 in the Natural Language Processing series. Berlin, New York: Mounton de Gruyter.

Hawkins, S., J. House, M. Huckvale, J. Local, and R. Ogden (1998). ProSynth: An integrated prosodic approach to device-independent, natural-sounding speech synthesis. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 5, pp. 1707–1710, Sydney, Australia.

Hazen, T. J., I. L. Hetherington, H. Shu, and K. Livescu (2002). Pronunciation modeling using a finite-state transducer representation. In *Proceedings of the ISCA Tutorial and Research Workshop Pronunciation Modeling and Lexicon Adaptation for Spoken Language (PMLA)*, pp. 99–104, Estes Park, Colorad.

Heldner, M. (2003). On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. *Journal of Phonetics*, 31(1):39–62.

Heldner, M., J. Edlund, and T. Björkenstam (2004). Automatically extracted f_0 features as acoustic correlates of prosodic boundaries. In *Proceedings of Fonetik*, pp. 52–55, Stockholm, Sweden.

Heldner, M. and B. Megyesi (2003). The acoustic and morpho-syntactic context of prosodic boundaries in dialogs. In *Proceedings of Fonetik*, pp. 117–120, Lövånger, Sweden.

Heldner, M., E. Strangert, and T. Deschamps (1999). A focus detector using overall intensity and high frequency emphasis. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, volume 2, pp. 1491–1493, San Francisco, California.

Helgason, P. (2006). SMTC: A Swedish map task corpus. In *Proceedings of Fonetik*, pp. 57–60, Lund, Sweden.

Hermes, D. J. and J. C. Gestel (1991). The frequency scale of speech intonation. *Journal of the Acoustical Society of America (JASA)*, 90(1):97–102.

Horne, M. and M. Filipsson (1996). Implementation and evaluation of a model for synthesis of Swedish intonation. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 1848–1851, Philadelphia, Pensylvania.

Horne, M., M. Filipsson, M. Ljungqvist, and A. Lindström (1993). Referent tracking in restricted texts using a lemmatized lexicon: Implications for generation of prosody. In *Proceedings Eurospeech*, pp. 2011–2014, Berlin, Germany.

Horne, M., M. Filipsson, M. Ljungqvist, and A. Lindström (1994). Computational modelling of contextual coreference: implications for swedish text-to-speech. In *Proceedings of the conference on Focus and natural language processing*, pp. 103–112, Heidelberg, Germany.

Horne, M., E. Strangert, and M. Heldner (1995). Prosodic boundary strength in Swedish: Final lengthening and silent interval duration. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, volume 1, pp. 70–173, Stockholm, Sweden.

Jakobson, R. C., G. M. Fant, and M. Halle (1963). *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates.* Cambridge: MIT Press, fourth edition.

Jande, P.-A. (2003a). Evaluating rules for phonological reduction in Swedish. In *Proceedings of Fonetik*, pp. 149–152, Lövånger, Sweden.

Jande, P.-A. (2003b). Phonological reduction in Swedish. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, pp. 2557–2560, Barcelona, Catalonia.

Jande, P.-A. (2004). Pronunciation variation modelling using decision tree induction from multiple linguistic parameters. In *Proceedings of Fonetik*, pp. 12–15, Stockholm, Sweden.

Jande, P.-A. (2006). Integrating linguistic information from multiple sources in lexicon development and spoken language annotation. In *Proceedings of the LREC workshop on merging and layering linguistic information*, Genoa, Italy.

Jespersen, O. (1904). *Phonetische Grundfragen.* Leipzig: Teubner.

Jurafsky, D., R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, P. Taylor, and C. Van Ess-Dykema (1997). Automatic detection of discourse structure for speech recognition and underatanding. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 88–95, Santa Barbara, California.

Jurafsky, D., A. Bell, E. Fosler-Lussier, C. Girand, and W. Raymond (1998a). Reduction of English function words in Switchboard. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 3111–3114, Sydney, Australia.

Jurafsky, D., A. Bell, M. Gregory, and W. Raymond (2001). The effect of language model probability on pronunciation reduction. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pp. 2118–2121, Salt Lake City, Utah.

Jurafsky, D., E. Shriberg, B. Fox, and T. Curl (1998b). Lexical, prosodic, and syntactic cues for dialog acts. In *Proceedings of the ACL/COLING Workshop on Discourse Relations and Discourse Markers*, pp. 114–120.

Karttunen, L. (1983). Kimmo: A general morphological processor. In *Texas Linguistics Forum 22*, pp. 165–186.

Kessens, J. M. and H. Strik (2001). Lower WERs do not guarantee better transcriptions. In *Proceedings of Eurospeech*, pp. 1721–1724, Aalborg, Denmark.

Kessens, J. M., H. Strik, and C. Cucchiarini (2000a). A bottom-up method for obtaining information about pronunciation variation. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 274–277, Beijing, China.

Kessens, J. M., H. Strik, and C. Cucchiarini (2002). Modeling pronunciation variation for ASR: Comparing criteria for rule selection. In *Proceedings of the ISCA Tutorial and Research Workshop Pronunciation Modeling and Lexicon Adaptation for Spoken Language (PMLA)*, pp. 18–23, Denver, Colorado.

Kessens, J. M., M. Wester, C. Cucchiarini, and H. Strik (1998). The selection of pronunciation variants: Comparing the performance of man and machine. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 6, pp. 2715–2718, Sydney, Australia.

Kessens, J. M., M. Wester, and H. Strik (1999). Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation. *Speech Communication*, 29(2–4):193–207.

Kessens, J. M., M. Wester, and H. Strik (2000b). Automatic detection and verification of Dutch phonological rules. In *PHONUS 5*, pp. 117–128. Saarbrücken: Institute of Phonetics, University of the Saarland.

Kipp, A., M.-B. Wesenick, and F. Schiel (1996). Automatic detection and segmentation of pronunciation variants in German speech corpora. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, Pensylvania.

Kipp, A., M.-B. Wesenick, and F. Schiel (1997). Pronunciation modelling applied to automatic segmentation of spontaneous speech. In *Proceedings of Eurospeech*, pp. 1023–1026, Rhodes, Greece.

Kohler, K. J. (2000). Investigating unscripted speech: Implications for phonetics and phonology. *Phonetica*, 57:85–94.

Koskenniemi, K. (1983). *Two-level morphology: A general computational model of word-form recognition and production.* PhD thesis, Helsinki: University of Helsinki, Department of General linguistics.

Koval, S., N. Smirnova, and M. Khitrov (2002). Modelling pronunciation variability for ASR tasks. In *Proceedings of the ISCA Tutorial and Research Workshop Pronunciation Modeling and Lexicon Adaptation for Spoken Language (PMLA)*, Estes Park, Colorado.

Källgren, G. (1998). Documentation of the stockholm-umeå corpus. http://www.ling.su.se/staff/sofia/suc/manual.pdf.

Landauer, T. K. and S. T. Dumais (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104:211–240.

Landauer, T. K., P. W. Foltz, and D. Laham (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Lindberg, J. (1999). Automatic detection of lexicalised phrases in Swedish. In *Proceedings of Nordiska datalingvistdagarna (NoDaLiDa)*, pp. 103–114.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America (JASA)*, 35(11):1773–1781.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In Hardcastle, W. J. and A. Marchal (editors), *Speech Production and Speech Modeling*, pp. 403–439. Dordrecht, Boston, London: Kluwer.

Lindström, A. (2003). Non-native elements in spoken swedish. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, pp. 2353–2356, Barcelona, Catalonia.

Lopez de Mantaras, R. (1991). A distance-based attribute selection measure for decision tree induction. *Machine Learning*, 6(1):81–92.

Lundgren, A. (2005). HMM-baserad talsyntes: An HMM-based text-to-speech system applied to Swedish. Master's thesis, Stockholm: KTH, Department of Speech, Music and Hearing.

Magnuson, T., B. Granström, R. Carlson, and F. Karlsson (1990). Phonetic transcription of a Swedish morphological analyzer. In *Proceedings of Fonetik*, pp. 58–61, Lövånger, Sweden.

Megyesi, B. (2001). Comparing data-driven learning algorithms for PoS tagging of swedish. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 151–158, Pittsburgh, Pensylvania.

Megyesi, B. (2002a). *Data-Driven Syntactic Analysis – Methods and Applications for Swedish.* PhD thesis, Stockholm: KTH, Department of Speech, Music and Hearing.

Megyesi, B. (2002b). Shallow parsing with PoS taggers and linguistic features. *Journal of Machine Learning Research*, 2:639–668.

Miller, C. A. (1998a). Individuation of postlexical phonology for speech synthesis. In *Proceedings of the ESCA/COCOSDA Workshop on Speech Synthesis*, pp. 133–136, Jenolan Caves, Australia.

Miller, C. A. (1998b). *Pronunciation Modeling in Speech Synthesis.* PhD thesis, Pennsylvania: University of Pennsylvania, Department of Linguistics.

Mitchell, T. (1997). *Machine Learning.* WCB/McGraw-Hill.

Moon, S.-J. and B. Lindblom (1994). Interaction between duration, context, and speaking style in English stressed vowels. *Journal of the Acoustical Society of America*, 96(1):40–55.

More, B. C. J. and B. R. Glasberg (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America (JASA)*, 74(3):750–753.

Nakajima, H., I. Hirano, Y. Sagisaka, and K. Shirai (2001). Pronunciation variant analysis using speaking style parallel corpus. In *Proceedings of Eurospeech*, pp. 65–68, Aalborg, Denmark.

Nolan, F. (2003). Intonational equivalence: an experimental evaluation of pitch scales. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, pp. 771–774, Barcelona, Catalonia.

Nooteboom, S. (1997). The prosody of speech: Melody and rythm. In Hardcastle, W. J. and J. Laver (editors), *The Handbook of Phonetic Sciences*, chapter 21, pp. 641–673. Oxford: Blackwell Publishers, first edition.

Ogden, R., S. Hawkins, J. House, M. Huckvale, J. Local, P. Carter, J. Dankovicová, and S. Heid (2000). ProSynth: An integrated prosodic approach to device-independent, natural-sounding speech synthesis. *Computer Speech and Language*, 14:177–210.

Ordelman, R. J. F., A. J. van Hessen, and D. A. van Leeuwen (1999a). Dealing with phrase level co-articulation (PLC) in speech recognition: A first approach. In *Proceedings of the ESCA Tutorial and Research Workshop on Accessing Information in Spoken Audio*, pp. 64–68, Cambridge, United Kingdom.

Ordelman, R. J. F., A. J. van Hessen, and D. A. van Leeuwen (1999b). Improving recognition performance using co-articulation rules on the phrase level: A first approach. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 1641–1644, Denver, Colorado.

Ostendorf, M., B. Byrne, M. Bacchiani, M. Finke, A. Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A. Waibel, B. Wheatley, and T. Zeppenfeld (1996). Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 1039–1042, Philadelphia, Pensylvania.

Ostendorf, M., B. Byrne, M. Bacchiani, M. Finke, A. Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A. Waibel, B. Wheatley, and T. Zeppenfeld (1997). Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode. In *Research Annote 24, 1996 LVCSR Summer Research Workshop Technical Report*. Baltimore: Johns Hopkins University, Center for Language and Speech Processing.

Pastor-i-Gadea, M. and F. Casacuberta (2001). Automatic learning of finite state automata for pronunciation modeling. In *Proceedings of Eurospeech*, pp. 2297–2300, Aalborg, Denmark.

Prahallad, K., A. W. Black, and R. Mosur (2006). Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 853–856, Toulouse, France.

Reimers, B., S. Modin, and S. Åberg (1995). Beskrivning och utvärdering av projekt DragonDictate© för Windows (description and evaluation of project DragonDictate© for Windows). Internal report, TeleNova.

Ridings, D. (2002). Swedish resources for language engineering. http://folk.uio.no/danielr/swedish.html.

Riley, M., W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters, and G. Zavaliagkos (1999). Stochastic pronunciation modeling from hand-labelled phonetic corpora. *Speech Communication*, 29(2–4):209–224.

Rydin, S. (2002). Building a hyponymy lexicon with hierarchical structure. In *Unsupervised Lexical Acquisition: Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, volume 9, pp. 26–33, Philadelphia, Pensylvania.

Schiel, F. (1999). Automatic phonetic transcription of non-prompted speech. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, pp. 607–610, San Francisco, California.

Schiel, F. (2004). MAUS goes iterative. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 1015–1018, Lisbon, Portugal.

Schiel, F., A. Kipp, and H.-G. Tillmann (1998). Statistical modeling of pronunciation: It's not the model, it's the data. In *Proceedings of the ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pp. 31–36, Kerkrade, the Netherlands.

Seneff, S. and C. Wang (2002). Modelling phonological rules through linguistic hierarchies. In *Proceedings of the ISCA Tutorial and Research Workshop Pronunciation Modeling and Lexicon Adaptation for Spoken Language (PMLA)*, Estes Park, Colorad.

Seneff, S. and C. Wang (2005). Statistical modeling of phonological rules through linguistic hierarchies. *Speech Communication*, 46(2):204–216.

Shriberg, E. (1999). Phonetic consequences of speech disfluency. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, volume 1, pp. 619–622, San Francisco, California.

Sigurd, B. (1965). *Phonotactic structures in Swedish.* Lund: Uniskol. Ph.D. Thesis, Lund University.

SIL International (1995). PCKIMMO. http://www.sil.org/pckimmo/.

Sjölander, K. (2001). Automatic alignment of phonetic segments. In *Proceedings of Fonetik*, pp. 140–143, Lund, Sweden.

Sjölander, K. (2003). An HMM-based system for automatic segmentation and alignment of speech. In *Proceedings of Fonetik*, pp. 93–96, Lövånger, Sweden.

Sjölander, K. (2004). The Snack sound toolkit. http://www.speech.kth.se/snack/.

Sjölander, K. and J. Beskow (2000). WaveSurfer – a public domain speech tool. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume IV, pp. 464–467, Bejing, China.

Sjölander, K. and M. Heldner (2004). Word level precision of the NALIGN automatic segmantation system. In *Proceedings of Fonetik*, pp. 116–119, Stockholm, Sweden.

Stevens, S. S. and J. Volkman (1940). The relation of pitch to frequency: A revised scale. *American Journal of Psychology*, 53:329–353.

Strangert, E. and M. Heldner (1995). The labelling of prominence in Swedish by phonetically experienced transcribers. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, volume 4, pp. 204–207.

Strik, H. and C. Cucchiarini (1998). Modeling pronunciation variation for ASR: Overview and comparison of methods. In *Proceedings of the ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pp. 137–144, Kerkrade, the Netherlands.

Strik, H. and C. Cucchiarini (1999). Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication*, 29(2-4):225–246.

Su, T.-T. and P. Basset (1998). Language dependent and independent spontaneous speech phenomena. In *Proceedings of the Sound Patterns of Spontaneous Speech (SPoSS) workshop*, pp. 55–58.

Sundaram, S. and S. Narayanan (2002). Spoken language synthesis: Experiments in synthesis of spontaneous monologues. In *Proceedings of the IEEE Speech Synthesis Workshop*, Santa Monica, California.

Tabain, M., G. Rolland, and C. Savariaux (2001). Coarticulatory effects at prosodic boundaries: Some acoustic results. In *Proceedings of Eurospeech*, pp. 963–966.

Tajchman, G., E. Fosler, and D. Jurafsky (1995a). Building multiple pronunciation models for novel words using exploratory computational phonology. In *Proceedings of Eurospeech*, Madrid, Spain.

Tajchman, G., D. Jurafsky, and E. Fosler (1995b). Learning phonological rule probabilities from speech corpora with exploratory computational phonology. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pp. 9–15, Cambridge, Massachusetts.

Tokuda, K., H. Zen, and A. W. Black (2002). An HMM-based speech synthesis system applied to english. In *Proceedings of the IEEE Speech Synthesis Workshop*, Santa Monica, California.

Toporowska Gronostaj, M. (2005). The Swedish PAROLE lexicon. http://spraakbanken.gu.se/parole/lexikon/swedish.parole.lexikon.html.

Torre Toledano, D., M. A. Rodríguez Crespo, and J. G. Escalada Sardina (1998). Trying to mimic human segmentation of speech using HMM and fuzzy logic post-connection rules. In *Proceedings of the ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, Australia.

Trancoso, I. (1995). The ONOMASTICA inter-language pronunciation lexicon. In *Proceedings of Eurospeech*, volume 1, pp. 829–832, Madrid, Spain.

Traunmüller, H. and A. Eriksson (1995a). The frequency range of the voice fundamental in the speech of male and female adults. http://www.ling.su.se/staff/hartmut/f0_m&f.pdf.

Traunmüller, H. and A. Eriksson (1995b). The perceptual evaluation of F0-excursions in speech as evidenced in liveliness estimations. *Journal of the Acoustical Society of America (JASA)*, 97(3):1905–1915.

Välimaa-Blum, R. (1998). What is deleted in spontaneous Finnish: Segmental interaction with word stress, vowel harmony and moras. In *Proceedings of the Sound Patterns of Spontaneous Speech (SPoSS) workshop*, pp. 47–50, La Baume-les-Aix, France.

Van Bael, C. P. J., H. van den Heuvel, and H. Strik (2004). Investigating speech style specific pronunciation variation in large spoken language corpora. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 586–589, Jeju Island, Korea.

Van Eynde, F. and D. Gibbon (editors) (2000). *Lexicon development for speech and language processing*. Number 12 in the Text, Speech and Language Technology series. Dordrecht, Boston, London: Kluwer.

Vereecken, H., A. Vorstermans, J.-P. Martens, and B. Van Coile (1997). Improving the phonetic annotation by means of prosodic phrasing. In *Proceedings of Eurospeech*, pp. 179–182, Rhodes, Greece.

Vorstermans, A., J.-P. Martens, and B. Van Coile (1996). Automatic segmentation and labelling of multi-lingual speech data. *Speech Communication*, 19(4):271–293.

Werner, S., M. Wolff, M. Eichner, and R. Hoffman (2004a). Modeling pronunciation variation for spontaneous speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pp. 673–676, Montreal, Canada.

Werner, S., M. Wolff, M. Eichner, and R. Hoffman (2004b). Toward spontaneous speech synthesis – utilizing language model information in TTS. *IEEE Transactions on Speech and Audio Processing*, 12(4):436–445.

Wesenick, M.-B. (1996). Automatic generation of German pronunciation variants. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, Pensylvania.

Wesenick, M.-B. and A. Kipp (1996). Estimating the quality of phonetic transcriptions and segmentations of speech signals. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, Pensylvania.

Wester, M., S. Greenberg, and S. Chang (2001a). A Dutch treatment of an elitist approach to articulatory-acoustic feature classification. In *Proceedings of Eurospeech*, pp. 1729–1732, Aalborg, Denmark.

Wester, M., J. M. Kessens, C. Cucchiarini, and H. Strik (1998a). Selection of pronunciation variants in spontaneous speech: Comparing the performance of man and machine. In *Proceedings of the ESCA Sound Patterns of Spontaneous Speech (SPoSS) workshop*, pp. 157–160, La Baume-les-Aix, France.

Wester, M., J. M. Kessens, C. Cucchiarini, and H. Strik (2001b). Obtaining phonetic transcriptions: A comparison between expert listeners and a continuous speech recognizer. *Language and Speech*, 44(3):377–403.

Wester, M., J. M. Kessens, and H. Strik (1998b). Improving the performance of a Dutch CSR by modeling pronunciation variation. In *Proceedings of the ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pp. 145–150, Kerkrade, the Netherlands.

Wester, M., J. M. Kessens, and H. Strik (2000). Using Dutch phonological rules to model pronunciation variation in ASR. In *PHONUS 5*, pp. 105–116. Saarbrücken: University of the Saarland, Institute of Phonetics.

Yoshimura, T., K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proceedings of European Conference on Speech Communication and Technology*, pp. 2347–2350, Budapest, Hungary.

Zellner, B. (1994). Pauses and the temporal structure of speech. In Keller, E. (editor), *Fundamentals of speech synthesis and speech recognition*, pp. 41–62. Chichester: John Wiley.

Zheng, J., H. Franco, and A. Stolcke (2000). Rate-dependent acoustic modeling for large vocabulary conversational speech recognition. In *Proceedings of the NIST Speech Transcription Workshop*, College Park, Maryland.

Öhman, S. E. G. (1967). Numerical model of coarticulation. *Journal of the Acoustical Society of America (JASA)*, 41(2):310–320.

Appendicies

Appendix A

The Swedish Technical Alphabet (STA)

Table A.1: *The phoneme symbols of the Swedish Technical Alphabet (STA) and their equivalents in the International Phonetic Alphabet (IPA). In the annotation used for pronunciation modelling in this thesis, the retroflex consonants have symbols different from those in the original STA. These symbols are shown in brackets after the original STA symbol.*

Vowels

STA symbol	IPA symbol
`E0`	ə
`A`	a
`A:`	ɑː
`E`	e
`E:`	eː
`I`	ɪ
`I:`	iː
`O`	ʊ
`O:`	uː
`U`	ɵ
`U:`	ʉ̟ː
`Y`	ʏ
`Y:`	yː
`Å`	ɔ
`Å:`	oː
`Ä`	ɛ
`Ä:`	ɛː
`Ä4`	æ
`Ä3`	æː
`Ö`	œ
`Ö:`	øː
`Ö4`	œ̞
`Ö3`	œ̞ː

Consonants

STA symbol	IPA symbol
`P`	p
`T`	t
`K`	k
`B`	b
`D`	d
`G`	ɡ
`F`	f
`V`	v
`S`	s
`SJ`	ɧ
`TJ`	ç
`H`	h
`M`	m
`N`	n
`NG`	ŋ
`L`	l
`J`	j
`R`	ɹ
`2T (RT)`	ʈ
`2D (RD)`	ɖ
`2L (RL)`	ɭ
`2N (RN)`	ɳ
`2S (RS)`	ʂ

Table A.2: *The non-phoneme symbols of the Swedish Technical Alphabet (STA).*

Symbol	Description
`hy`	Compound boundary marker
`'`	Accent I stress marker
`"`	Accent II or compound accent primary stress marker
`` ` ``	Compound accent secondary stress marker
`*`	Marker signalling that strings corresponding to multi-character segment labels should be interpreted as two consecutive single-character lables

Appendix B

Xenophones in CentLex

Table B.1: *The set of xenophone symbols ('foreign phonemes') included in the* CentLex *pronunciation representation meta-format. In the left hand sub-table, vowel xenophones are shown, and in the right hand sub-table, the consonant xenophones are shown. Each xenophone is illustrated with an example (source language, orthographic word and pronunciation representation in* CentLex *format).*

Vow.	**Example**	
`a:`	German *Bahn*	`B'a:N`
`Av`	English *but*	`B'AvT`
`Åa`	English *not*	`N'ÅaT`
`o:`	English *hall*	`H'o:Le`
`er`	English *bird*	`B'erD`
`A9`	French *blanc*	`BL'A9`
`O9`	French *non*	`N'O9`
`E9`	French *fin*	`F'E9`
`Ö9`	French *brun*	`BRf'Ö9`
`Ei`	English *mail*	`M'EiLe`
`Ai`	English *fine*	`F'AiN`
`Oi`	English *boy*	`B'Oi`
`I@`	English *beer*	`B'I@`
`E@`	English *fair*	`F'E@`
`U@`	English *poor*	`P'U@`
`Au`	Swedish *aula*	`'AuLA`
`au`	English *brown*	`BRe'auN`
`Eu`	Swedish *Europa*	`EuR'O:PA`
`@u`	English *mode*	`M'@uD`
`Ou`	Finnish *Oulu*	`'OuLO`

Cons.	**Example**	
`th`	English *thing*	`th'ING`
`dh`	English *this*	`dh'IS`
`Z`	German *Sohn*	`Z'Å:N`
`Sh`	English *ship*	`Sh'IP`
`zh`	French *Jean*	`zh'A9`
`C`	German *Licht*	`L'ICT`
`X`	German *Bach*	`B'AX`
`Lj`	Italian *Oglio*	`'ÅLjÅ`
`Nj`	French *ligne*	`L'INj`
`Jy`	French *lui*	`LJy'I:`
`Le`	English *ball*	`B'o:Le`
`Re`	English *red*	`Re'EOD`
`Rf`	French *rouge*	`Rf'O:zh`
`W`	English *wet*	`W'ET`
`Q`	German *Beamte*	`BEOQ'AMTEO`
`Ts`	German *Zug*	`Ts'O:K`
`ch`	English *chin*	`ch'IN`
`Dz`	Italian *Zacchi*	`Dz'AKI`
`dZ`	English *James*	`dZ'EiMZ`
`Pf`	German *Pferd*	`Pf'E:RfT`
`Ls`	English *beetle*	`B'I:TLs`
`Ms`	English *prism*	`PRe'IZMs`
`Ns`	English *burden*	`B'erDNs`
`Rs`	Croatian *Krk*	`K'RsK`

Appendix C

The SUC Tag Set

Table C.1: *The set of Part of Speech (PoS) and morphological tags used in the Stockholm-Umeå Corpus (*SUC*) annotation scheme. There are also four "unspecified" morphological tags, UTR/NEU, IND/DEF, SIN/PLU and SUB/OBJ.*

PoS tag	Description
AB	Adverb
DT	Determiner
HA	WH-adverb
HD	WH-determiner
HP	WH-pronoun
HS	Possessive WH-pronoun
IE	Infinitival marker
IN	Interjection
JJ	Adjective
KN	Conjunction
MAD	Major delimiter
MID	Minor delimiter
NN	Noun
PAD	Parenthetical delimiter
PC	Participle
PL	Verb particle
PM	Proper name
PN	Pronoun
PP	Preposition
PS	Possessive pronoun
RG	Cardinal number
RO	Ordinal number
SN	Subjunction
UO	Foreign word
VB	Verb

Morph. tag	Tag type	Description
UTR	Gender	Common
NEU	Gender	Neutre
MAS	Gender	Masculine
SIN	Number	Singular
PLU	Number	Plural
IND	Definiteness	Indefinite
DEF	Definiteness	Definite
NOM	Case	Nominative
GEN	Case	Genitive
SMS	Case	Compound
SUB	Pronoun form	Subject
OBJ	Pronoun form	Object
PRS	Tense/Aspect	Present
PRT	Tense/Aspect	Preterite
INF	Tense/Aspect	Infinitive
IMP	Tense/Aspect	Imperative
SUP	Tense/Aspect	Supinum
PRF	Tense/Aspect	Perfect
KON	Mood	Conjunctive
AKT	Voice	Active
SFO	Voice	Passive/s-form
POS	Degree	Positive
KOM	Degree	Comparative
SUV	Degree	Superlative
AN	Abbreviation	Abbreviation

Appendix D

The Share of Tokens in Three Frequency Groups

Table D.1: *The share of tokens (per cent) in the three frequency groups over different text corpora.*

	High frequency tokens		Mid frequency tokens		Low frequency tokens	
Corpus	Entry form	Word	Entry form	Word	Entry form	Word
DN	49.98	49.97	40.02	40.03	10.00	10.00
TPB	49.99	49.96	40.01	40.04	10.00	10.00
GOV	49.97	50.00	40.02	40.00	10.00	10.00
RD	49.94	49.90	40.06	40.10	10.00	10.00
EU	49.96	49.92	40.04	40.08	10.00	10.00
JK	49.91	49.99	40.08	40.01	10.00	10.00
DOM	49.95	49.89	40.05	40.11	10.00	10.00
FMN	49.95	50.00	40.05	40.00	10.00	10.00
ALL	49.97	49.98	40.03	40.02	10.00	10.00

Appendix E

Phone Instances in the Gold Standard Transcript

Table E.1: *The number of phone instances in the gold standard transcript.*

Elision	Instances
∅	341

Consonant	Instances
p	39
t	178
k	110
b	31
d	72
ɡ	17
f	42
v	49
s	165
ɧ	8
ç	3
h	45
m	118
n	156
ŋ	23
l	84
j	65
ɹ	159
ʈ	5
ɖ	16
ɭ	3
ɳ	6
ʂ	9

Vowel	Instances
ə	317
a	112
ɑː	48
e	64
eː	37
ɪ	74
iː	23
ʊ	12
uː	20
ɵ	21
ʉ̟ː	24
ʏ	19
yː	1
ɔ	113
oː	40
ɛ	39
ɛː	14
æ	10
æː	14
œ	7
øː	3
œ̞	9
œ̞ː	3

Appendix F

Phone Confusion Matrices

Table F.1: *Matrix showing the confusions between ∅ and different consonants made by the tree models trained on all data, broken down over source phonemes. Each column shows the share of classifications corresponding to the class shown at the bottom of the column for the key ∅ of the row derived from the particular subscripripted consonant phoneme.*

phone$_{phoneme}$

	∅	p	t	k	b	d	g	f	v	s	ɧ	ç	h	m	n	ŋ	l	j	ɹ	ʈ	ɖ	ɭ	ɳ	ʂ
$\emptyset_{\text{p}}$	.87	.13																						
$\emptyset_{\text{t}}$	.98		.02																					
$\emptyset_{\text{k}}$	.98			.02																				
$\emptyset_{\text{b}}$	.70				.30																			
$\emptyset_{\text{d}}$	.48					.51													.01					
$\emptyset_{\text{g}}$	.77						.23																	
$\emptyset_{\text{f}}$	.43							.57																
$\emptyset_{\text{v}}$	.89								.11															
$\emptyset_{\text{s}}$	.81									.19														.01
$\emptyset_{\text{h}}$	.70												.30											
$\emptyset_{\text{m}}$	.89													.11										
$\emptyset_{\text{n}}$	.54													.02	.42	.01								
$\emptyset_{\text{ŋ}}$																1								
$\emptyset_{\text{l}}$	.60																.40							
$\emptyset_{\text{j}}$	.84																	.16						
$\emptyset_{\text{ɹ}}$	.76																		.24					
$\emptyset_{\text{ʈ}}$	.14																			.86				
$\emptyset_{\text{ɖ}}$																					1			
$\emptyset_{\text{ɭ}}$	.55																					.45		
$\emptyset_{\text{ɳ}}$	.38																						.62	
$\emptyset_{\text{ʂ}}$	.05																							.95

Table F.2: *Consonant confusion matrix for tree models trained on all data, with confusions broken down over source phonemes. Each column shows the share of classifications corresponding to the class shown at the bottom of the column for the key phone of the row derived from the particular subscripripted consonant phoneme.*

$phone_{phoneme}$

p_p	.01	.99																						
t_t	.01		.99																					
k_k	.01			.99																				
b_b	.01				.99																			
d_d	.04					.92													.03					
g_g	.20						.80																	
f_f	.02							.98																
v_v									1															
s_s										.99														
$\text{ɧ}_\text{ɧ}$											1													
$\text{ç}_\text{ç}$												1												
h_h	.05												.95											
m_m	.01													.99										
m_n	.04													.88	.07									
n_n	.02														.97									
ŋ_n	.02														.14	.85								
$\text{ŋ}_\text{ŋ}$	.01															.99								
l_l	.01																.99							
j_j	.05																	.95						
ɹ_d	.02					.88													.11					
$\text{ɹ}_\text{ɹ}$	.09																		.91					
ʈ_t			.88																	.12				
$\text{ʈ}_\text{ʈ}$																				1				
ɖ_d	.02					.88													.06		.04			
$\text{ɖ}_\text{ɖ}$																					1			
$\text{ɭ}_\text{ɭ}$	.64																					.36		
ɳ_n	.01														.91								.08	
$\text{ɳ}_\text{ɳ}$	.14																						.86	
ʂ_s										.88														.12
$\text{ʂ}_\text{ʂ}$	.01																							.99
	∅	p	t	k	b	d	g	f	v	s	ɧ	ç	h	m	n	ŋ	l	j	ɹ	ʈ	ɖ	ɭ	ɳ	ʂ

Table F.3: *Matrix showing the confusions between ∅ and different vowels made by the tree models trained on all data, broken down over source phonemes. Each column shows the share of classifications corresponding to the class shown at the bottom of the column for the key ∅ of the row derived from the particular subscripripted vowel phoneme.*

$phone_{phoneme}$

$\emptyset_{\text{ə}}$	.65	.35																						
$\emptyset_{\text{a}}$	.39	.44	.16																					
$\emptyset_{\text{ɑː}}$				1																				
$\emptyset_{\text{e}}$					1																			
$\emptyset_{\text{eː}}$						1																		
$\emptyset_{\text{ɪ}}$		.22					.78																	
$\emptyset_{\text{iː}}$	.91	.09																						
$\emptyset_{\text{ʊ}}$	.33	.50							.17															
$\emptyset_{\text{uː}}$	.67								.33															
$\emptyset_{\text{ɵ}}$	.19										.81													
$\emptyset_{\text{ʉ̞ː}}$		.50										.50												
$\emptyset_{\text{ʏ}}$													1											
$\emptyset_{\text{ɔ}}$	.32	.02													.66									
$\emptyset_{\text{oː}}$																1								
$\emptyset_{\text{ɛ}}$	.29	.43															.29							
$\emptyset_{\text{æː}}$		1																						
$\emptyset_{\text{œ}}$	.73																				.27			
$\emptyset_{\text{øː}}$																						1		
$\emptyset_{\text{œ̞}}$	.39	.05																			.01		.55	
$\emptyset_{\text{œ̞ː}}$	.12																						.25	.62
	∅	ə	a	ɑː	e	eː	ɪ	iː	ʊ	uː	ɵ	ʉ̞ː	ʏ	yː	ɔ	oː	ɛ	ɛː	æ	æː	œ	øː	œ̞	œ̞ː

Table F.4: *Matrix showing the confusions between* [ə] *and different vowels made by the tree models trained on all data, broken down over source phonemes. Each column shows the share of classifications corresponding to the class shown at the bottom of the column for the key* [ə] *of the row derived from the particular subscripripted vowel phoneme.*

phone$_{phoneme}$	∅	ə	a	ɑː	e	eː	ɪ	iː	ʊ	uː	ɵ	ʉ̟ː	ʏ	yː	ɔ	oː	ɛ	ɛː	æ	æː	œ	øː	œ̞	œ̞ː
ə$_{\text{ə}}$		1																						
ə$_{\text{a}}$	.01	.81	.19																					
ə$_{\text{ɑː}}$		.95		.05																				
ə$_{\text{e}}$		.91			.09																			
ə$_{\text{eː}}$		.97			.01	.02																		
ə$_{\text{ɪ}}$	.01	.20					.80																	
ə$_{\text{iː}}$		.94					.03	.03																
ə$_{\text{ʊ}}$		.58							.42															
ə$_{\text{uː}}$		.11							.67	.22														
ə$_{\text{ɵ}}$		.78									.22													
ə$_{\text{ʉ̟ː}}$		.81									.03	.16												
ə$_{\text{ʏ}}$													1											
ə$_{\text{yː}}$														1										
ə$_{\text{ɔ}}$	.01	.92													.08									
ə$_{\text{oː}}$		.91													.06	.03								
ə$_{\text{ɛ}}$		.70															.30							
ə$_{\text{ɛː}}$		.94															.06							
ə$_{\text{æ}}$		.49																	.39	.12				
ə$_{\text{æː}}$		.94															.01		.04	.01				
ə$_{\text{œ}}$	1																							
ə$_{\text{øː}}$																						1		
ə$_{\text{œ̞}}$	.17	.06																					.78	
ə$_{\text{œ̞ː}}$		.33																					.33	.33

Table F.5: *Vowel confusion matrix for tree models trained on all data, with confusions broken down over source phonemes. Each column shows the share of classifications corresponding to the class shown at the bottom of the column for the key phone of the row derived from the particular subscripripted vowel phoneme.*

$phone_{phoneme}$

	∅	ə	a	ɑː	e	eː	ɪ	iː	ʊ	uː	ɵ	ʉ̟ː	ʏ	yː	ɔ	oː	ɛ	ɛː	æ	æː	œ	øː	œ̞	œ̞ː
a_{a}		.06	.94																					
$\text{a}_{\text{ɑː}}$				1																				
$\text{ɑː}_{\text{ɑː}}$		.01		.99																				
e_{e}		.16			.84																			
$\text{e}_{\text{eː}}$		.33			.14	.53																		
$\text{eː}_{\text{eː}}$		.03			.03	.94																		
$\text{ɪ}_{\text{ɪ}}$		.04					.96																	
$\text{ɪ}_{\text{iː}}$		.01					.72	.26																
$\text{iː}_{\text{iː}}$							.03	.97																
$\text{ʊ}_{\text{ʊ}}$		.21							.79															
$\text{ʊ}_{\text{uː}}$		.02							.56	.42														
$\text{uː}_{\text{uː}}$									.03	.97														
$\text{ɵ}_{\text{ɵ}}$	.01										.98													
$\text{ɵ}_{\text{ʉ̟ː}}$		.18									.06	.75												
$\text{ʉ̟ː}_{\text{ʉ̟ː}}$		.02									.01	.97												
$\text{ʏ}_{\text{ʏ}}$													1											
$\text{ʏ}_{\text{yː}}$													.25	.75										
$\text{yː}_{\text{yː}}$		.03											.04	.94										
$\text{ɔ}_{\text{ɔ}}$	.01	.02													.97									
$\text{ɔ}_{\text{oː}}$		.01													.79	.20								
$\text{oː}_{\text{oː}}$															.08	.92								
$\text{ɛ}_{\text{ɛ}}$		.01															.99							
$\text{ɛ}_{\text{ɛː}}$		.01															.71	.28						
$\text{ɛ}_{\text{æ}}$		.03															.17		.75	.05				
$\text{ɛ}_{\text{æː}}$		.05															.27		.53	.15				
$\text{ɛː}_{\text{ɛː}}$		.01															.05	.93						
$\text{æ}_{\text{æ}}$		.11															.10		.73	.07				
$\text{æ}_{\text{æː}}$		.06															.08		.70	.16				
$\text{æː}_{\text{æː}}$		.01															.01		.10	.88				
$\text{œ}_{\text{œ}}$	.02																				.98			
$\text{œ}_{\text{øː}}$																					.31	.69		
$\text{œ}_{\text{œ̞}}$	.10																				.07		.83	
$\text{œ}_{\text{œ̞ː}}$	.07																				.07		.29	.57
$\text{øː}_{\text{øː}}$																					.06	.94		
$\text{œ̞}_{\text{œ̞}}$	.12	.04																			.02		.82	
$\text{œ̞}_{\text{œ̞ː}}$																							.89	.10
$\text{œ̞ː}_{\text{œ̞ː}}$	.01																				.01		.04	.94

Appendix G

Phone Instances in the Evaluation Data

Table G.1: *The number of instances of each consonant (and ∅) in the key transcript for each database, sum over all evaluation data sets.*

Phone	VaKoS	Radio Interview	Radio News	All
∅	5,796	3,193	753	9,742
p	740	330	157	1,229
t	4,123	2,858	704	7,685
k	1,943	1,044	384	3,372
b	640	329	194	1,164
d	2,175	1,479	345	3,999
ɡ	473	307	130	910
f	932	621	201	1,755
v	1,353	692	262	2,308
s	3,083	2,012	672	5,768
ɧ	189	241	58	488
ç	85	30	19	134
h	995	519	116	1,630
m	2,110	1,324	395	3,830
n	3,370	2,128	747	6,248
ŋ	502	244	124	870
l	2,281	1,112	440	3,833
j	840	475	139	1,454
ɹ	3,240	2,157	700	6,097
ʈ	171	108	52	331
ɖ	120	89	50	259
ɭ	4	3	4	11
ɳ	116	72	63	251
ʂ	231	122	54	407

Table G.2: *The number of instances of each vowel in the key transcript for each database, sum over all evaluation data sets.*

Phone	VaKoS	Radio Interview	Radio News	All
ə	7,308	3,646	1,070	12,027
a	2,796	2,037	756	5,591
ɑː	1,257	731	272	2,261
e	1,136	839	242	2,219
eː	730	645	171	1,546
ɪ	1,781	1,271	442	3,494
iː	635	360	105	1,100
ʊ	258	140	43	441
uː	320	305	110	735
ɵ	512	302	87	902
ʉ̟ː	466	347	83	896
ʏ	288	143	28	459
yː	40	40	32	112
ɔ	2,181	1,156	302	3,639
oː	840	419	146	1,405
ɛ	741	530	122	1,393
ɛː	143	103	28	274
æ	247	114	21	382
æː	276	288	60	624
œ	122	62	28	212
øː	116	75	38	229
œ̞	255	138	56	449
œ̞ː	110	130	35	275

Appendix H

Phone Distance Matrix

Table H.1: *Phone distance matrix, including a distance between each phone and ∅. Phone distances are specified on a scale from 0 to 8.*

	∅	p	t	k	b	d	ɡ	f	v	s	ɧ	ç	h	m	n	ŋ	l	j	ɹ	ʈ	ɖ	ɭ	ɳ	ʂ	ə	a	ɑː	e	eː	ɪ	iː	ʊ	uː	ɵ	ʉ̟ː	ʏ	yː	ɔ	oː	ɛ	ɛː	æ	æː	œ	øː	œ̞	œ̞ː
∅	0	3	3	3	2	2	2	3	2	2	3	4	1	2	2	4	2	1	2	4	4	2	3	3	4	5	6	5	6	5	6	5	6	5	6	5	6	5	6	5	6	5	6	5	6	5	6
p	3	0	2	2	1	3	4	3	2	4	5	3	4	2	4	5	4	4	4	2	3	4	4	4	5	7	8	7	8	7	8	7	8	7	8	7	8	7	8	6	7	6	7	6	7	6	7
t	3	2	0	2	3	1	3	4	4	2	4	2	4	4	4	4	3	4	4	1	2	4	4	3	6	7	8	7	8	7	8	7	8	7	8	7	8	7	8	7	8	7	8	7	8	7	8
k	3	2	2	0	4	3	1	4	4	4	2	4	3	4	4	3	4	4	4	3	4	5	4	4	6	7	8	7	8	7	8	6	7	7	8	6	7	7	8	7	8	7	8	7	8	7	8
b	2	1	3	4	0	2	3	2	1	4	5	4	3	2	3	4	3	3	3	3	2	3	3	4	5	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7
d	2	3	1	3	2	0	2	4	3	4	5	4	3	3	3	4	3	3	3	2	1	3	3	4	5	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7
ɡ	2	4	3	1	3	2	0	4	3	4	3	5	3	3	3	2	3	3	3	4	3	4	3	5	5	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7
f	3	3	4	4	2	4	4	0	1	2	4	2	4	4	4	5	4	4	4	3	4	5	5	4	5	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7
v	2	2	4	4	1	3	3	1	0	3	4	3	3	3	3	4	3	3	3	4	3	4	4	5	4	5	6	6	7	6	7	6	7	3	4	5	6	5	6	5	6	5	6	5	6	5	6
s	2	4	2	4	4	4	4	2	3	0	3	1	2	4	4	5	3	3	4	3	4	5	5	1	5	5	6	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7
ɧ	3	5	4	2	5	5	3	4	4	3	0	2	2	5	5	3	4	3	4	4	4	5	5	2	6	7	8	7	8	7	8	6	7	7	8	7	8	6	7	7	8	7	8	7	8	6	7
ç	4	3	2	4	4	4	5	2	3	1	2	0	4	5	5	5	3	4	3	3	4	5	5	2	6	7	8	7	8	6	7	7	8	7	8	6	7	7	8	7	8	7	8	7	8	6	7
h	1	4	4	3	3	3	3	4	3	2	2	4	0	3	3	5	3	2	3	3	4	5	5	3	3	5	6	4	5	5	6	5	6	5	6	5	6	4	5	4	5	4	5	4	5	4	5
m	2	2	4	4	2	3	3	4	3	4	5	5	3	0	1	2	3	3	3	4	3	3	2	4	5	5	6	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7
n	2	4	4	4	3	3	3	4	3	4	5	5	3	1	0	1	2	3	2	4	3	3	1	4	4	5	6	6	7	6	7	6	7	6	7	6	7	5	6	5	6	5	6	5	6	5	6
ŋ	4	5	4	3	4	4	2	5	4	5	3	5	5	2	1	0	3	3	4	5	4	3	2	5	3	5	6	4	5	5	6	4	5	5	6	5	6	5	6	5	6	5	6	4	5	4	5
l	2	4	3	4	3	3	3	4	3	3	4	3	3	3	2	3	0	3	2	3	2	3	1	4	5	5	6	6	7	6	7	6	7	6	7	5	6	6	7	6	7	6	7	6	7	6	7
j	1	4	4	4	3	3	3	4	3	3	3	4	2	3	3	3	3	0	3	4	3	3	3	4	4	5	6	4	5	3	4	5	6	5	6	3	4	5	6	5	6	5	6	5	6	5	6
ɹ	2	4	4	4	3	3	3	4	3	4	4	3	3	3	2	4	2	3	0	3	3	2	2	3	4	5	6	6	7	6	7	6	7	6	7	6	7	5	6	5	6	4	5	5	6	4	5
ʈ	4	2	1	3	3	2	4	3	4	3	4	3	3	4	4	5	3	4	3	0	1	4	3	2	5	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7	6	7
ɖ	4	3	2	4	2	1	3	4	3	4	4	4	4	3	3	4	2	3	3	1	0	3	2	3	4	5	6	5	6	5	6	5	6	5	6	5	6	5	6	5	6	5	6	5	6	5	6
ɭ	2	4	4	5	3	3	4	5	4	5	5	5	5	3	3	3	3	3	2	4	3	0	2	4	5	5	6	6	7	6	7	6	7	6	7	5	6	6	7	6	7	6	7	6	7	6	7
ɳ	3	4	4	4	3	3	3	5	4	5	5	5	5	2	1	2	1	3	2	3	2	2	0	3	4	5	6	6	7	6	7	6	7	6	7	6	7	5	6	5	6	5	6	5	6	5	6
ʂ	3	4	3	4	4	4	5	4	5	1	2	2	3	4	4	5	4	4	3	2	3	4	3	0	6	6	5	7	8	7	8	7	8	7	8	7	8	7	8	7	8	7	8	7	8	7	8
ə	4	5	6	6	5	5	5	5	4	5	6	6	3	5	4	3	5	4	4	5	4	5	4	6	0	4	5	2	3	2	3	1	2	1	2	2	3	2	3	2	3	3	4	1	3	2	4
a	5	7	7	7	6	6	6	6	5	5	7	7	5	5	5	5	5	5	5	6	5	5	5	6	4	0	1	3	4	4	5	3	4	4	5	4	5	2	3	2	3	3	4	4	5	5	6
ɑː	6	8	8	8	7	7	7	7	6	6	8	8	6	6	6	6	6	6	6	7	6	6	6	5	5	1	0	4	3	5	4	4	3	5	4	5	4	3	2	3	2	4	3	5	4	6	5
e	5	7	7	7	6	6	6	6	6	6	7	7	4	6	6	4	6	4	6	6	5	6	6	7	2	3	4	0	1	1	2	4	5	3	4	3	4	4	5	0	1	1	2	2	3	3	4
eː	6	8	8	8	7	7	7	7	7	7	8	8	5	7	7	5	7	5	7	7	6	7	7	8	3	4	3	1	0	2	1	5	4	4	3	4	3	5	4	2	1	2	1	3	2	4	3
ɪ	5	7	7	7	6	6	6	6	6	6	7	6	5	6	6	5	6	3	6	6	5	6	6	7	2	4	5	1	2	0	1	4	5	3	4	1	2	4	5	2	3	3	4	2	3	3	4
iː	6	8	8	8	7	7	7	7	7	7	8	7	6	7	7	6	7	4	7	7	6	7	7	8	3	5	4	2	1	1	0	5	4	4	3	2	1	5	4	3	2	4	3	3	2	4	3
ʊ	5	7	7	6	6	6	6	6	6	6	6	7	5	6	6	4	6	5	6	6	5	6	6	7	1	3	4	4	5	4	5	0	1	1	2	2	3	1	2	4	5	3	4	1	2	2	2
uː	6	8	8	7	7	7	7	7	7	7	7	8	6	7	7	5	7	6	7	7	6	7	7	8	2	4	3	5	4	5	4	1	0	2	1	3	2	2	1	5	4	4	3	2	1	3	2
ɵ	5	7	7	7	6	6	6	6	3	6	7	7	5	6	6	5	6	5	6	6	5	6	6	7	1	4	5	3	4	3	4	1	2	0	1	1	2	5	6	4	5	5	6	2	3	3	4
ʉ̟ː	6	8	8	8	7	7	7	7	4	7	8	8	6	7	7	6	7	6	7	7	6	7	7	8	2	5	4	4	3	4	3	2	1	1	0	2	1	6	6	5	4	6	5	3	2	4	3
ʏ	5	7	7	6	6	6	6	6	5	6	7	6	5	6	6	5	5	3	6	6	5	5	6	7	2	4	5	3	4	1	2	2	3	1	2	0	1	4	5	3	4	4	5	1	2	2	3
yː	6	8	8	7	7	7	7	7	6	7	8	7	6	7	7	6	6	4	7	7	6	6	7	8	3	5	4	4	3	2	1	3	2	2	1	1	0	5	4	4	3	5	4	2	1	3	2
ɔ	5	7	7	7	6	6	6	6	5	6	6	7	4	6	5	5	6	5	5	6	5	6	5	7	2	2	3	4	5	4	5	1	2	5	6	4	5	0	1	4	5	5	6	4	5	5	6
oː	6	8	8	8	7	7	7	7	6	7	7	8	5	7	6	6	7	6	6	7	6	7	6	8	3	3	2	5	4	5	4	2	1	6	6	5	4	1	0	5	4	6	5	5	6	6	5
ɛ	5	6	7	7	6	6	6	6	5	6	7	7	4	6	5	5	6	5	5	6	5	6	5	7	2	2	3	0	2	2	3	4	5	4	5	3	4	4	5	0	1	1	2	2	3	4	6
ɛː	6	7	8	8	7	7	7	7	6	7	8	8	5	7	6	6	7	6	6	7	6	7	6	8	3	3	2	1	1	3	2	5	4	5	4	4	3	5	4	1	0	2	1	3	3	4	5
æ	5	6	7	7	6	6	6	6	5	6	7	7	4	6	5	5	6	5	4	6	5	6	5	7	3	3	4	1	2	3	4	3	4	5	6	4	5	5	6	1	2	0	1	4	5	3	4
æː	6	7	8	8	7	7	7	7	6	7	8	8	5	7	6	6	7	6	5	7	6	7	6	8	4	4	3	2	1	4	3	4	3	6	5	5	4	6	5	2	1	1	0	4	3	4	2
œ	5	6	7	7	6	6	6	6	5	6	7	7	4	6	5	4	6	5	5	6	5	6	5	7	1	4	5	2	3	2	3	1	2	2	3	1	2	4	5	2	3	4	4	0	1	1	2
øː	6	7	8	8	7	7	7	7	6	7	8	8	5	7	6	5	7	6	6	7	6	7	6	8	3	5	4	3	2	3	2	2	1	3	2	2	1	5	6	3	3	5	3	1	0	2	1
œ̞	5	6	7	7	6	6	6	6	5	6	6	6	4	6	5	4	6	5	4	6	5	6	5	7	2	5	6	3	4	3	4	2	3	3	4	2	3	5	6	4	4	3	4	1	2	0	1
œ̞ː	6	7	8	8	7	7	7	7	6	7	7	7	5	7	6	5	7	6	5	7	6	7	6	8	4	6	5	4	3	4	3	2	2	4	3	3	2	6	5	6	5	4	2	2	1	1	0

Appendix I

Consonant Realisations

The Realisation of /p/

The phoneme /p/ can be realised either as [p] or as ∅ (i.e., have no overt realisation; be elided). As can be seen from Table I.1, /p/ is realised as [p] in a little more than 93% of the cases and elided in just under 7% of the cases both according to the automatically obtained key transcript and according to the decision tree model.

Table I.1: *The realisations of the phoneme /p/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	91	6.89%	91	6.89%	89	97.80%
p	1,229	93.11%	1,229	93.11%	1,227	99.84%
$\sum$	1,320	100.00%	1,320	100.00%	1,316	99.70%

Figure I.1 shows the part of the decision tree model handling the realisations of the /p/ phoneme. From this figure, it can be seen that /p/ is elided when the mean phoneme duration over the word is less than 52.7 ms (corresponding to a ‘canonical’ speaking rate faster than 19.0 phonemes per second), unless its right hand neighbouring phoneme is /p/, /ɹ/ or /a/. A canonical /p/ is also elided at slower speech rates when its right adjacent phoneme is also a /p/. Otherwise, /p/ is realised as [p].

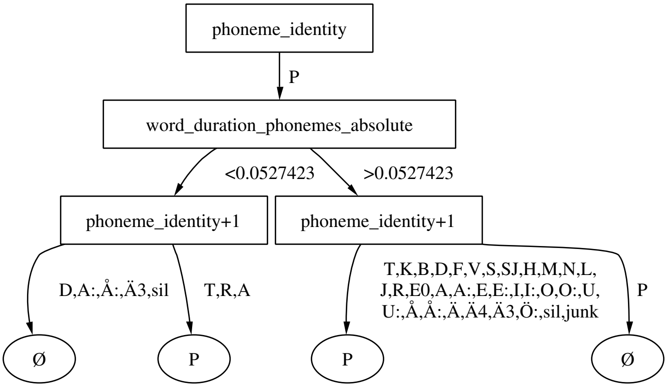

phoneme_identity
P
word_duration_phonemes_absolute
<0.0527423
>0.0527423
phoneme_identity+1
phoneme_identity+1
D,A:,Å:,Ä3,sil
T,R,A
T,K,B,D,F,V,S,SJ,H,M,N,L,
J,R,E0,A,A:,E,E:,I,I:,O,O:,U,
U:,Å,Å:,Ä,Ä4,Ä3,Ö:,sil,junk
P
Ø
P
P
Ø

Figure I.1: *The realisations of the phoneme /*p*/ (phoneme representations in the figure are in STA format).*

The Realisation of /t/

A /t/ can be realised as [t] or it can lack a realisation. Although the final model does not allow this realisation, in the key transcript, there are also a few /t/ phonemes realised as [ʈ]. Table I.2 shows the distribution of /t/ realisations.

Table I.2: *The realisations of the phoneme* /t/ *in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	1,090	12.39%	1,166	13.25%	1,079	98.99%
t	7,686	87.34%	7,634	86.75%	7,599	98.87%
ʈ	24	0.27%	0	0.00%	0	0.00%
$\sum$	8,800	100.00%	8,800	100.00%	8,678	98.61%

Figure I.2 shows that a canonical /t/ is always elided by the model when its right adjacent phoneme is /p/, /t/, /b/ or /s/. It is also elided if the phoneme located three positions before the /t/ is a /ʏ/ and the phoneme four positions before the /t/ is an /m/. Otherwise, /t/ is realised as [t].

The left branch of the tree, specifying a context of /m/ and /ʏ/ at four and three phoneme positions, respectively, preceding the /t/, effectively handles the elision of /t/ in the common word *mycket ‘much’*, mostly used as an adverb. The pronunciation of this word could have been handled e.g. by using the *function word* attribute, but for the current training data, it was optimal to use the *phoneme identity* attribute.

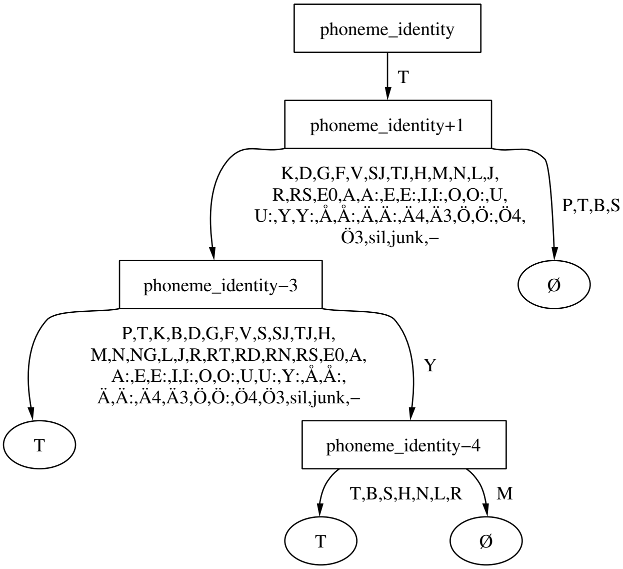

phoneme_identity
T
phoneme_identity+1
K,D,G,F,V,SJ,TJ,H,M,N,L,J,
R,RS,E0,A,A:,E,E:,I,I:,O,O:,U,
U:,Y,Y:,Å,Å:,Ä,Ä:,Ä4,Ä3,Ö,Ö:,Ö4,
Ö3,sil,junk,–
P,T,B,S
phoneme_identity–3
Ø
P,T,K,B,D,G,F,V,S,SJ,TJ,H,
M,N,NG,L,J,R,RT,RD,RN,RS,E0,A,
A:,E,E:,I,I:,O,O:,U,U:,Y:,Å,Å:,
Ä,Ä:,Ä4,Ä3,Ö,Ö:,Ö4,Ö3,sil,junk,–
Y
T
phoneme_identity–4
T,B,S,H,N,L,R
M
T
Ø

Figure I.2: *The realisations of the phoneme /*`t`*/ (phoneme representations in the figure are in STA format).*

The Realisation of /k/

A /k/ in the canonical pronunciation representation can be realised as [k] or be elided. Both in the key transcript and in the decision tree output, it is elided in about 22% of the cases, as can be seen in Table I.3.

Table I.3: *The realisations of the phoneme /*k*/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	935	21.71%	943	21.89%	915	97.86%
k	3,372	78.29%	3,364	78.11%	3,344	99.17%
$\sum$	4,307	100.00%	4,307	100.00%	4,259	98.89%

Figure I.3 shows that the model always produces the realisation [k] for the phoneme /k/ for all words except *och 'and'*. All other words with a /k/ in their canonical pronunciation representations are covered by the set of function words and the generic content word representation (*content*) of the *function word* attribute. The set of function words includes auxiliary verbs, interjections, adverbs and verb particles.

There are also a few words in the set that are not actual function words, but have been misclassified by the tagger[1]. The noun loan from English *country 'country'* (referring to the music genre), the noun compound *diktaturländer 'dictatorial countries'*, and the verbs *frakta 'freight'* and *sjunka 'sink'* are obvious examples. However, in this particular case, the misclassifications do not matter, since the distinction is really between *och 'and'* and all other words.

In the word *och*, /k/ is elided if the mean phoneme duration over the word is less than 135.9 ms or, otherwise, if the mean normalised vowel duration over the phrase is more than 1.63 (the mean of this normalised measure is 0).

[1] Auxiliary verbs are defined partly by their context and for some words in the set, it may be the right adjacent word that has been misclassified rather than the current word. The error may also be due to a parsing/chunking error.

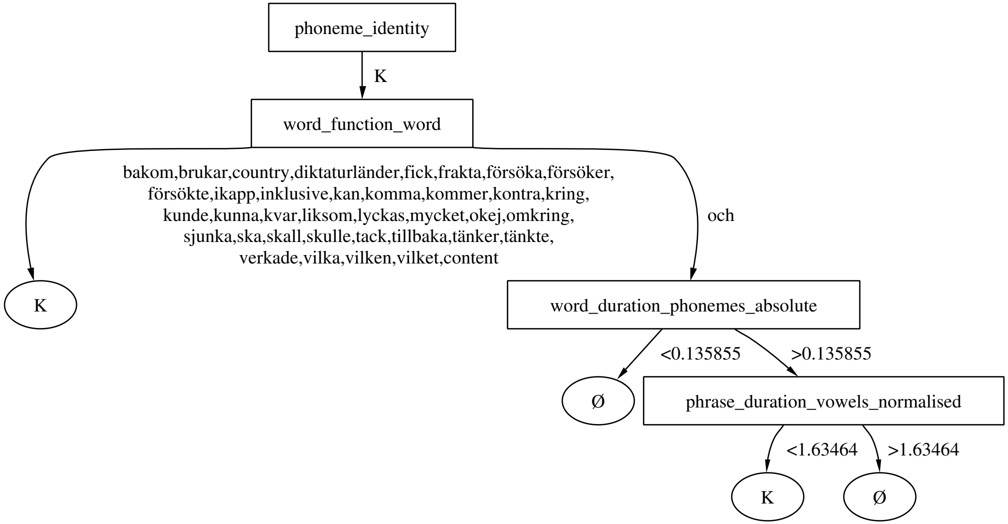

phoneme_identity
K
word_function_word
bakom,brukar,country,diktaturländer,fick,frakta,försöka,försöker,
försökte,ikapp,inklusive,kan,komma,kommer,kontra,kring,
kunde,kunna,kvar,liksom,lyckas,mycket,okej,omkring,
sjunka,ska,skall,skulle,tack,tillbaka,tänker,tänkte,
verkade,vilka,vilken,vilket,content
och
K
word_duration_phonemes_absolute
<0.135855
>0.135855
Ø
phrase_duration_vowels_normalised
<1.63464
>1.63464
K
Ø

Figure I.3: *The realisations of the phoneme /*k*/ (phoneme representations in the figure are in STA format).*

The Realisation of /b/

A /b/ is seldom realised as anything except [b], although it is elided in about 3% of the cases in both the key transcript and the model output. Table I.4 shows the exact numbers and shares of realisations for the phoneme /b/.

Table I.4: *The realisations of the phoneme /b/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	37	3.08%	35	2.91%	30	81.08%
b	1,164	96.92%	1,166	97.09%	1,159	99.57%
$\sum$	1,201	100.00%	1,201	100.00%	1,189	99.00%

Figure I.4 shows that /b/ is elided when the mean phoneme duration over the word is less than 34.5 ms. If the mean phoneme duration over the word is more than 34.5 ms, /b/ is elided if the mean phoneme duration over the *phrase* is less than 31.6 ms and realised as [b] otherwise.

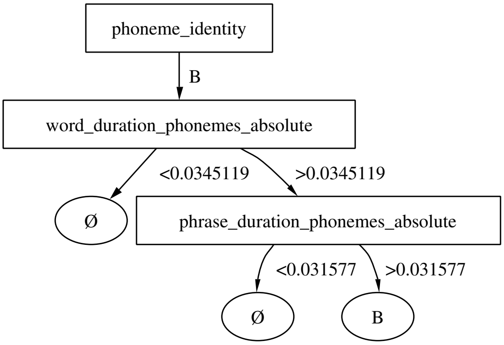

phoneme_identity
B
word_duration_phonemes_absolute
<0.0345119
>0.0345119
Ø
phrase_duration_phonemes_absolute
<0.031577
>0.031577
Ø
B

Figure I.4: *The realisations of the phoneme /b/ (phoneme representations in the figure are in STA format).*

The Realisation of /d/

In the key transcript, /d/ can be realised as [d], [ɹ] or [ɖ], and it can be elided. The decision tree model only allows the [d] realisation of /d/, as can be seen from Table I.5 and Figure I.5.

The reason for this is probably that most of the realisations of /d/ as [ɹ], [ɖ] and ∅ produced by the unpruned model were incorrect realisations. As can be seen from tables F.1 and F.2 in Appendix F, for the optimal trees in the tenfold cross-validation experiment still allowing the [ɹ], [ɖ] and ∅ realisations of /d/, the classification of /d/ as ∅ was correct in 48% of the cases, the classification of /d/ as [ɹ] was correct in 11% of the cases and the classification of /d/ as [ɖ] was correct in only 4% of the cases. The classification of /d/ as [d] was correct in 92% of the cases. This, in turn, is probably due to insufficient data or irregular realisation (at least given the available context attributes).

Table I.5: *The realisations of the phoneme* /d/ *in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	573	11.12%	0	0.00%	0	0.00%
d	3,999	77.59%	5,154	100.00%	3,999	100.00%
ɹ	532	10.32%	0	0.00%	0	0.00%
ɖ	50	0.97%	0	0.00%	0	0.00%
∑	5,154	100.00%	5,154	100.00%	3,999	77.59%

At pruning, the possibility for correct realisations of /d/ as [ɹ], [ɖ] or ∅ is excluded along with the possibility for erroneous classifications as these phones (or no realisation), since the model cannot discriminate between correct and erroneous classifications.

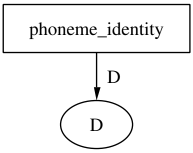


Figure I.5: *The realisations of the phoneme* /d/ *(phoneme representations in the figure are in STA format).*

The Realisation of /g/

The phoneme /g/ is elided in about 50% of the cases in the key transcript and in the model output, as shown in Table I.6. As could be seen in Table 5.7 in Chapter 5, Section 5.7, there are 29% erroneous [g] elisions (all corresponding to /g/ phonemes) in the automatically obtained key transcript over the part covered by the gold standard transcript. There is thus a large degree of uncertainty regarding the key realisations for this particular phoneme.

Table I.6: *The realisations of the phoneme /g/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	997	52.28%	945	49.55%	777	77.93%
g	910	47.72%	962	50.45%	742	81.54%
$\sum$	1,907	100.00%	1,907	100.00%	1,519	79.65%

Figure I.6 shows the contexts in which /g/ is realised as [g] and ∅, respectively. Given that the /g/ occurs in the small set of function words associated with the leftmost branch of the tree in Figure I.6, it should always be elided. If the /g/ does not occur in one of the function words of the leftmost branch of the tree, it is elided if the mean log phoneme duration over the phrase is less than -3.08 and realised as [g] otherwise.

It is unlikely that generally deleting [g] in the words *vågar*[2] *'dare'* and *angående, 'concerning'* would lead to a better model performance in relation to a manually transcribed standard. However, since errors are relatively common for /g/ in the key transcripts, it increases performance in relation to the key. Although relatively frequently occurring, *vågar* and *angående* are not in the absolute top of the list of high frequency words. Hence, their inclusion in the list of words for which /g/ should always be elided will give rise to less errors than if the words had been very high frequency words erroneously included in the list.

Because of the transparency of the decision tree models, the induced models can be manually edited and the effects of changes can be evaluated through executing the original and the edited models on the same data set.

[2] *Included in the function word set since it can be used as an auxiliary verb. However, the orthographic form can also be a main verb or a noun.*

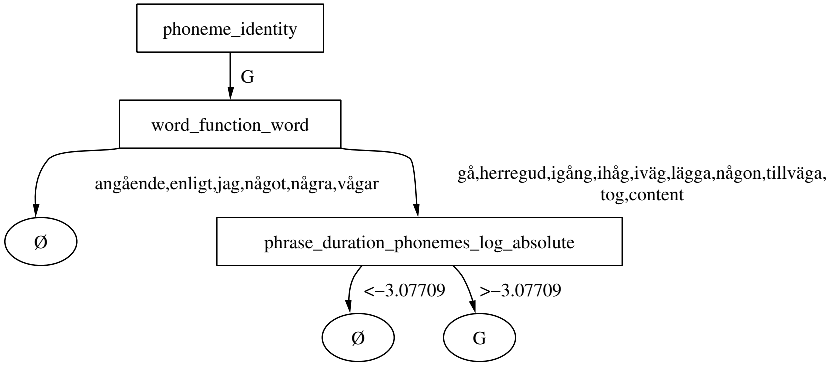

phoneme_identity
G
word_function_word
angående,enligt,jag,något,några,vågar
gå,herregud,igång,ihåg,iväg,lägga,någon,tillväga,
tog,content
Ø
phrase_duration_phonemes_log_absolute
<−3.07709
>−3.07709
Ø
G

Figure I.6: *The realisations of the phoneme /g/ (phoneme representations in the figure are in STA format).*

The Realisation of /f/

The phoneme /f/ is realised as [f] in the majority of cases in both the key transcript and in the model output. However, as can be seen from Table I.7, it is realised as [f] to a greater extent in the model output. The alternative is that /f/ is elided.

Table I.7: *The realisations of the phoneme /*f*/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	218	11.05%	88	4.46%	87	39.91%
f	1,755	88.95%	1,885	95.54%	1,754	99.94%
$\sum$	1,973	100.00%	1,973	100.00%	1,841	93.31%

If the mean phoneme duration over the word is less than 36.7 ms and the word following the word in which /f/ occurs is not a proper name (PM), the /f/ is elided by the pronunciation model and otherwise, it is realised as [f], as illustrated in Figure I.7. This is an interesting pattern, for which there is no immediate analytical explanation. However, it likely not a general rule, but an artefact of data sparsity at model training.

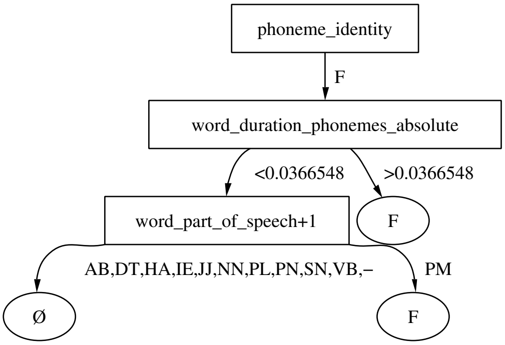


Figure I.7: *The realisations of the phoneme /*f*/ (phoneme representations in the figure are in STA format).*

The Realisation of /v/

About 14% of the occurrences of /v/ are realised as [v] and 86% as ∅ both according to the key transcript and according to the model output, as shown in Table I.8.

Table I.8: *The realisations of the phoneme /v/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	392	14.52%	363	13.44%	356	90.82%
v	2,308	85.48%	2,337	86.56%	2,301	99.70%
$\sum$	2,700	100.00%	2,700	100.00%	2,657	98.41%

The tree model for the realisation of /v/ is shown in Figure I.8. A /v/ is always elided if the mean phoneme duration over the word is less than 35.1 ms. It is also elided if the mean phoneme duration over the word is between 35.1 and 38.2 ms and the /v/ occurs in one of the function words *av 'of'*, *vad 'what'*, *vara 'be'*, *vi 'we'* or *vid 'at'*.

However, when the mean phoneme duration over the word is in the 35.1 to 38.2 ms interval and the /v/ occurs in any other word (explicitly, a content word or the function word *blev 'became'*), it is realised as [v]. During pruning, two leaves with the same realisation have been created (cf. Figure I.8). This is, of course, equivalent to having one leaf with the realisation [v] and a common arc for the function word *blev 'became'* and the generic content word representation.

If the mean phoneme duration over the word is more than 38.2 ms, /v/ is realised as [v], unless when followed by another /v/, in which case it is elided.

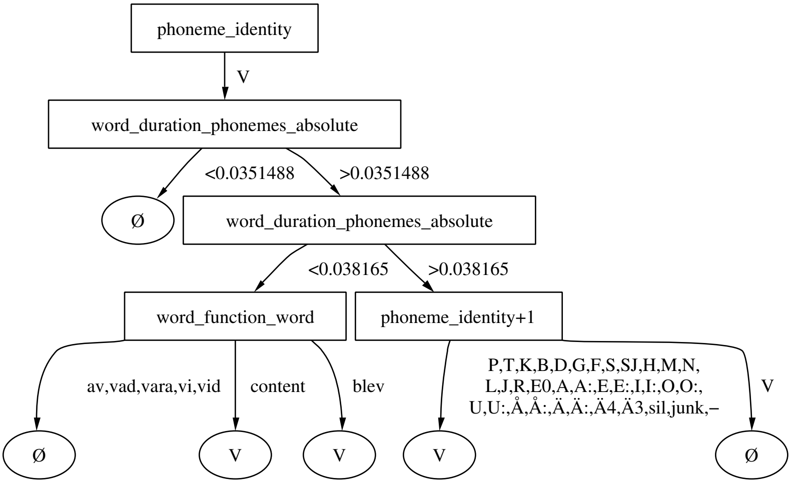

phoneme_identity
V
word_duration_phonemes_absolute
<0.0351488
>0.0351488
Ø
word_duration_phonemes_absolute
<0.038165
>0.038165
word_function_word
phoneme_identity+1
av,vad,vara,vi,vid
content
blev
P,T,K,B,D,G,F,S,SJ,H,M,N,
L,J,R,E0,A,A:,E,E:,I,I:,O,O:,
U,U:,Å,Å:,Ä,Ä:,Ä4,Ä3,sil,junk,–
V
Ø
V
V
V
Ø

Figure I.8: *The realisations of the phoneme /v/ (phoneme representations in the figure are in STA format).*

The Realisation of /s/

In the key transcript, /s/ can be realised as [s], [ʂ] or ∅. In the model output, only the [s] and ∅ realisations remain and [s] constitutes a larger share of the realisations. As can be seen in Table I.9, more than 97% of the instances of /s/ are realised as [s] according to the model predictions.

Table I.9: *The realisations of the phoneme /s/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	197	3.23%	166	2.72%	162	82.23%
s	5,768	94.53%	5,936	97.28%	5,765	99.95%
ʂ	137	2.25%	0	0.00%	0	0.00%
∑	6,102	100.00%	6,102	100.00%	5,927	97.13%

According to the pronunciation tree, shown in Figure I.9, /s/ is always elided if succeeded by another /s/. If the mean phoneme duration over the word is less than 36.4 ms, the /s/ is elided if preceding /t/, /l/, /e/, /eː/, /ɔ/ or /oː/ and otherwise realised as [s]. This phoneme context corresponds to several function words with a word initial /s/. An /s/ is also realised as [s] if the mean phoneme duration over the word is more than 36.4 ms (and the /s/ is not followed another /s/).

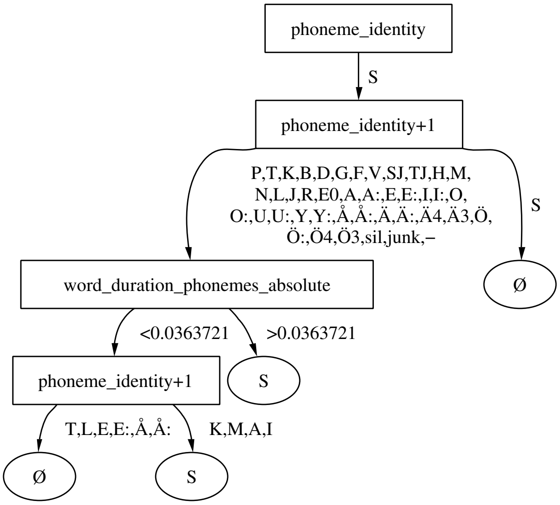


Figure I.9: *The realisations of the phoneme /s/ (phoneme representations in the figure are in STA format).*

The Realisation of /ɧ/

A /ɧ/ is always realised as [ɧ] according both to the key transcript and to the pronunciation model, as shown in Table I.10 and Figure I.10.

Table I.10: *The realisations of the phoneme /ɧ/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
ɧ	488	100.00%	488	100.00%	488	100.00%

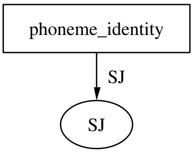


Figure I.10: *The realisations of the phoneme /ɧ/ (phoneme representations in the figure are in STA format).*

The Realisation of /ç/

The rule for the realisation of the phoneme /ç/ shows the same simplicity as the rule for /ɧ/: a /ç/ is always realised as [ç]. Table I.11 and Figure I.11 illustrate this fact.

Table I.11: *The realisations of the phoneme /ç/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
ç	134	100.00%	134	100.00%	134	100.00%

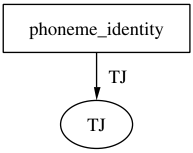


Figure I.11: *The realisations of the phoneme /ç/ (phoneme representations in the figure are in STA format).*

The Realisation of /h/

An /h/ is mostly realised as [h], but may also be elided. Table I.12 shows that /h/ is elided in about 15% of the cases according to the key transcript and in about 13% of the cases according to the model predictions.

Table I.12: *The realisations of the phoneme /h/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	284	14.84%	245	12.80%	181	63.73%
h	1,630	85.16%	1,669	87.20%	1,566	96.07%
$\sum$	1,914	100.00%	1,914	100.00%	1,747	91.27%

Figure I.12 shows the realisation rules for /h/. It can be seen that an /h/ is never elided in a syllable in word initial position (i). However, if the syllable is in final (f) or medial (m) position, the mean log phoneme duration over the phrase must be less than -2.54 and the phoneme two positions to the left cannot be /p/, /ɧ/, /n/, /j/, /ɑː/, /eː/, /ʊ/ or /ɛ/ for the /h/ to be elided. This phoneme identity context probably targets a specific group of words and it likely that the /h/ realisation would have been handled differently if more data had been available for training the model. If the mean log phoneme duration over the phrase is more than -2.54, /h/ is always realised as [h].

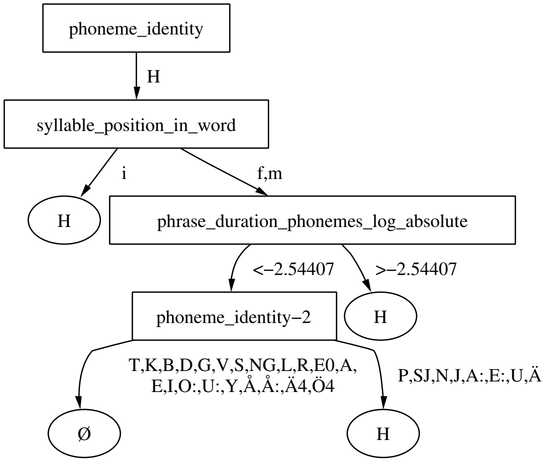


Figure I.12: *The realisations of the phoneme /h/ (phoneme representations in the figure are in STA format).*

The Realisation of /m/

For /m/, the key transcript and the model transcript contain almost equal shares of realisations: in about 90.5% of the cases, /m/ is realised as [m] and in the reminder of cases, it is elided. Table I.13 shows this distribution of realisations.

Table I.13: *The realisations of the phoneme /m/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	379	9.40%	385	9.55%	322	84.96%
m	3,653	90.60%	3,647	90.45%	3,590	98.28%
$\sum$	4,032	100.00%	4,032	100.00%	3,912	97.02%

Figure I.13 shows that /m/ is elided when the mean phoneme duration over the word is less than 35.1 ms or, otherwise, when the /m/ is followed by another /m/. If the mean phoneme duration over the word is more than 35.1 ms and the /m/ is not followed by /m/, it is realised as [m].

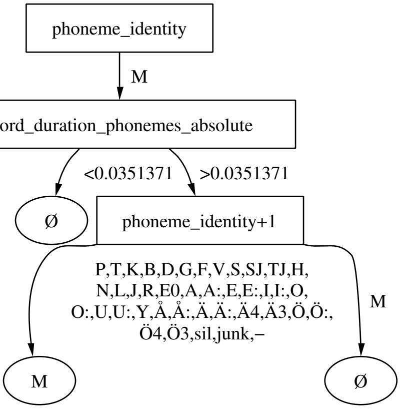

phoneme_identity
M
word_duration_phonemes_absolute
<0.0351371
>0.0351371
Ø
phoneme_identity+1
P,T,K,B,D,G,F,V,S,SJ,TJ,H,
N,L,J,R,E0,A,A:,E,E:,I,I:,O,
O:,U,U:,Y,Å,Å:,Ä,Ä:,Ä4,Ä3,Ö,Ö:,
Ö4,Ö3,sil,junk,–
M
M
Ø

Figure I.13: *The realisations of the phoneme /m/ (phoneme representations in the figure are in STA format).*

The Realisation of /n/

Table I.14 shows that the set of possible realisations for /n/ is larger than the sets of possible realisations for previously discussed consonants. Although realised as [n] in about 86% of the cases according to the key transcript and in about 91% of the cases according to the model output, /n/ can also be realised as [m], [ŋ] or [ɳ] or be elided.

Table I.14: *The realisations of the phoneme /n/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	612	8.46%	364	5.03%	291	47.55%
m	178	2.46%	206	2.85%	171	96.07%
n	6,248	86.42%	6,554	90.65%	6,155	98.51%
ŋ	117	1.62%	105	1.45%	101	86.32%
ɳ	75	1.04%	1	0.01%	1	1.33%
$\sum$	7,230	100.00%	7,230	100.00%	6,719	92.93%

Figure I.14 illustrates the realisation rules for /n/. An /n/ is realised as [m] preceding a /p/ or a /b/ and as [ŋ] if preceding a /k/. These are common and well-known place assimilation rules for Swedish connected speech. An /n/ is, further, elided preceding another /n/.

When the succeeding phoneme is any other than the above mentioned, /n/ is always realised as [n] when the mean phoneme duration over the word is more than 36.2 ms. At faster speaking rates, /n/ is realised as [ɳ] in the function word *ned 'down'*. From Table I.14, it is apparent that this rule only applies once for the entire data set, and the probability of this rule being an artefact of sparse data is thus large.

When the succeeding phoneme is not /p/, /b/, /k/ or /n/ and the mean phoneme duration over the word is less than 36.2 ms, /n/ is elided in the function words *ens 'ones'*, *ni 'you'*, *någon 'someone'*, *något 'something'*, *några 'some'*, *nåt 'something'* (not standard spelling, but a spelling indicating an abbreviated/reduced pronunciation), *sånt 'such'*, *under 'under'* and *än 'yet'*.

If the /n/ occurs in one of the function words *bland 'among'*, *den 'the'*, *en 'a'*, *man 'one'*, *men 'but'* or *nån 'someone'* (not standard spelling, but a spelling indicating an abbreviated/reduced pronunciation) or in a content word, it is realised as [n] if the normalised mean phoneme duration over the word is more than -0.55 and elided otherwise.

If the /n/ occurs in one of the function words *från 'from'*, *han 'he'*, *hans 'his'*, *hon 'she'*, *nej 'no'* or *när 'when'*, it is realised as [n] if the normalised mean phoneme duration over the phrase is more than -1.17 and elided otherwise.

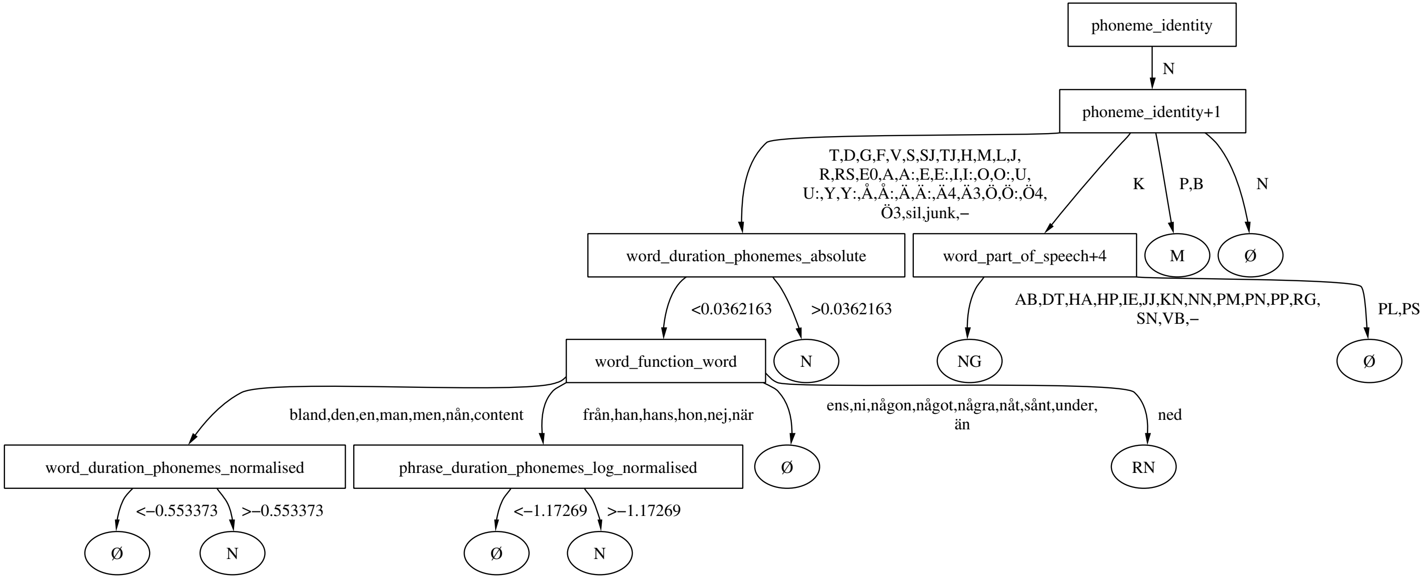

phoneme_identity
N
phoneme_identity+1
T,D,G,F,V,S,SJ,TJ,H,M,L,J,
R,RS,E0,A,A:,E,E:,I,I:,O,O:,U,
U:,Y,Y:,Å,Å:,Ä,Ä:,Ä4,Ä3,Ö,Ö:,Ö4,
Ö3,sil,junk,–
K
P,B
N
word_duration_phonemes_absolute
word_part_of_speech+4
M
Ø
<0.0362163
>0.0362163
AB,DT,HA,HP,IE,JJ,KN,NN,PM,PN,PP,RG,
SN,VB,–
PL,PS
word_function_word
N
NG
Ø
bland,den,en,man,men,nån,content
från,han,hans,hon,nej,när
ens,ni,någon,något,några,nåt,sånt,under,
än
ned
word_duration_phonemes_normalised
phrase_duration_phonemes_log_normalised
Ø
RN
<–0.553373
>–0.553373
<–1.17269
>–1.17269
Ø
N
Ø
N

Figure I.14: *The realisations of the phoneme /n/ (phoneme representations in the figure are in STA format).*

The Realisation of /ŋ/

An /ŋ/ is realised as [ŋ] in more than 99% of the cases and elided in the reminder of the cases in the key transcript, as can be seen in Table I.15. In the decision tree model, the option of deleting /ŋ/ is not included, as seen in Table I.15 and Figure I.15.

Table I.15: *The realisations of the phoneme /ŋ/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	7	0.92%	0	0.00%	0	0.00%
ŋ	753	99.08%	760	100.00%	753	100.00%
∑	760	100.00%	760	100.00%	753	99.08%

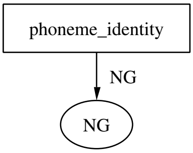


Figure I.15: *The realisations of the phoneme /ŋ/ (phoneme representations in the figure are in STA format).*

The Realisation of /l/

An /l/ is mostly realised as [l] according to the key transcript and also to the model output, as shown in Table I.16. The /l/ is elided slightly more often in the key transcript than in the model output (95.9% vs. 97.4%).

Table I.16: *The realisations of the phoneme /l/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	164	4.10%	104	2.60%	98	59.76%
l	3,833	95.90%	3,893	97.40%	3,827	99.84%
∑	3,997	100.00%	3,997	100.00%	3,925	98.20%

As for most consonants, an /l/ is elided in the pronunciation model if followed by an identical consonant. Double consonants may occur at word boundaries when canonical pronunciation representations are concatenated. However, at pronunciation in context, these double consonants are mostly realised as a single consonant.

From Figure I.16, it is also clear that /l/ is elided preceding several phoneme sequences starting with an /s/, but that it is realised as [l] in most cases also when preceding an /s/. When the phoneme following the /l/ is not another /l/ or an /s/, /l/ is elided when occurring in the function words *vilka 'which'*, *vilken 'what'* and *vilket 'what'* and realised as [l] otherwise.

It is likely that the phoneme sequence following the /l/, involving *phoneme identity+2* and *phoneme identity+3* is not the primary predictor for the realisation of /l/ as such. The phoneme sequences probably to some extent target specific words and the rules with the specified phoneme contexts may not be generally applicable. However, the phoneme sequence following the /l/ plays a part for the realisation and, given the training data and the 'greedy' training algorithm used, the *phoneme identity* attributes are the best of the predictors involved in the decision in the unpruned tree. Thus, the *phoneme identity* attributes are the predictors left in the pruned model. If more data had been available at training, other attributes may have been used.

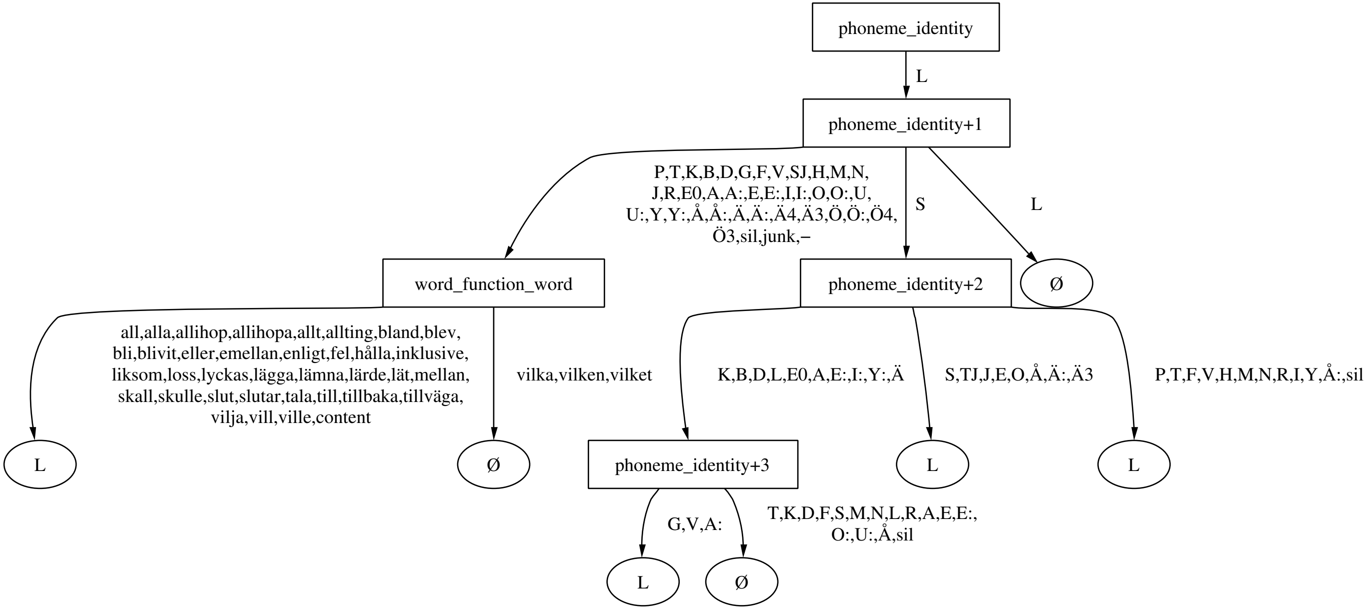

phoneme_identity
L
phoneme_identity+1
P,T,K,B,D,G,F,V,SJ,H,M,N,
J,R,E0,A,A:,E,E:,I,I:,O,O:,U,
U:,Y,Y:,Å,Å:,Ä,Ä:,Ä4,Ä3,Ö,Ö:,Ö4,
Ö3,sil,junk,–
S
L
word_function_word
phoneme_identity+2
Ø
all,alla,allihop,allihopa,allt,allting,bland,blev,
bli,blivit,eller,emellan,enligt,fel,hålla,inklusive,
liksom,loss,lyckas,lägga,lämna,lärde,lät,mellan,
skall,skulle,slut,slutar,tala,till,tillbaka,tillväga,
vilja,vill,ville,content
vilka,vilken,vilket
K,B,D,L,E0,A,E:,I:,Y:,Ä
S,TJ,J,E,O,Å,Ä:,Ä3
P,T,F,V,H,M,N,R,I,Y,Å:,sil
L
Ø
phoneme_identity+3
L
L
G,V,A:
T,K,D,F,S,M,N,L,R,A,E,E:,
O:,U:,Å,sil
L
Ø

Figure I.16: *The realisations of the phoneme /l/ (phoneme representations in the figure are in STA format).*

The Realisation of /j/

Table I.17 shows that /j/ may be realised as [j] or elided. Table 5.7 in Chapter 5, Section 5.7 indicated that the key transcript contains a large share of [j] erroneously classified as ∅. The [j] phones in the gold standard transcript almost exclusively originate from /j/ and it is thus likely that the decision tree model elides /j/ unproportionally often when compared to a phonetic transcript of speech with higher quality than the automatic transcription system output.

Table I.17: *The realisations of the phoneme /j/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	751	34.06%	649	29.43%	615	81.89%
j	1,454	65.94%	1,556	70.57%	1,420	97.66%
$\sum$	2,205	100.00%	2,205	100.00%	2,035	92.29%

In the model, a /j/ is always realised as [j] unless it is succeeded by /g/, /f/, /ɧ/, /h/, /j/, /a/, /ɑː/, /iː/ or /oː/. If succeeded by another phoneme, it is elided if the mean phoneme duration over the phrase is less than 35.2 ms. If the mean duration is longer, /j/ is elided if the preceding phoneme is /g/, /ɹ/, /e/, /ɛ/, <sil> or <junk> and the phoneme two positions to the right is not /g/, /n/, /ɹ/, /ʊ/ or /æ/. The part of the pronunciation model handling the realisations of /j/ is shown in Figure I.17.

The attributes used by the model for determining the realisation of /j/ are mainly phoneme context attributes. The hypothesis is that the current model would give poor realisation predictions for /j/ if evaluated against a manually transcribed standard and that more data and especially more accurate training keys would facilitate finding better rules for the realisation of /j/.

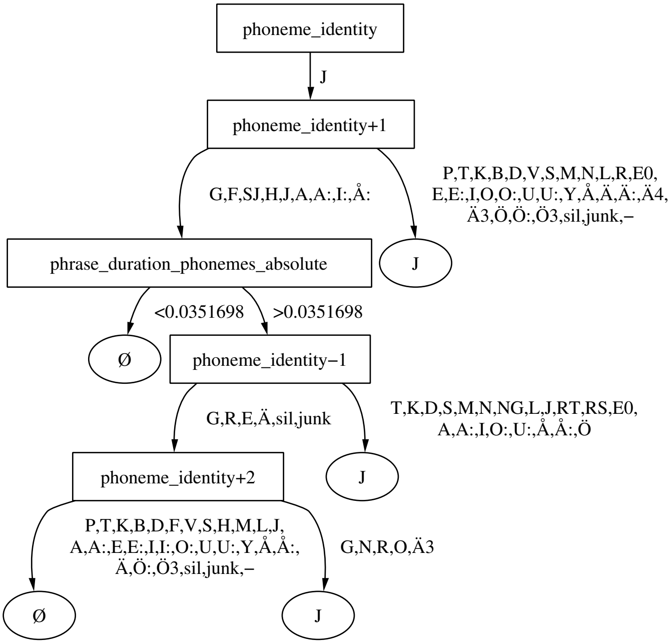


Figure I.17: *The realisations of the phoneme /*j*/ (phoneme representations in the figure are in STA format).*

The Realisation of /ɹ/

Table I.18 shows that /ɹ/ is either realised as [ɹ] or elided. The /ɹ/ is elided more often in the key transcript than in the model output.

Table I.18: *The realisations of the phoneme /ɹ/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	2,317	29.40%	1,813	23.00%	1,630	70.35%
ɹ	5,565	70.60%	6,069	77.00%	5,382	96.71%
$\sum$	7,882	100.00%	7,882	100.00%	7,012	88.96%

Approximants were hard to handle by the automatic transcription system and the keys corresponding to approximant phonemes thus contain a larger share of errors than the keys for most other phonemes. The realisation of approximants may also be less rule-governed than the realisation of other consonants. Further, both the phone identity and the presence of these phones in actual spoken language recordings can often be debated.

In the tree governing the realisation of /ɹ/, shown in Figure I.18, the highest ranking attribute, *phoneme feature py+1*[3] separates the right adjacent phoneme context into vowels and /h/ (and no phoneme, i.e., <sil>, <junk> or discourse boundary) on one side and all consonants but /h/ on the other side.

If the mean phoneme duration over the word is less than 30.6 ms, /ɹ/ is elided if the phoneme two position to the left is /f/, /h/ or /j/ and the right adjacent phoneme is /iː/, /ʉ̟ː/, /æː/, <sil> or <junk>. The /ɹ/ is also elided in the case where the mean phoneme duration over the word is less than 30.6, the phoneme two position to the left is *not* /f/, /h/ or /j/ and the current syllable contains less than 3.5 phonemes. Otherwise, when the *phoneme feature py+1* attribute targets a vowel or /h/, an /ɹ/ is realised as [ɹ].

An /ɹ/ is, further, elided preceding /d/, /ɧ/, /ç/, /m/, /n/, /l/, /ɹ/, or /ɖ/. Preceding other consonants (except /h/), the /ɹ/ is elided if the mean phoneme duration over the word is less than 36.6 ms and the phoneme two positions to the left is *not* /d/, /f/, /h/, /l/, /j/ or /ɹ/. If the phoneme two positions to the left is one of these consonants, /ɹ/ is still elided if the mean phoneme duration over the following word is more than 41.6 ms and if it is less than 41.6 ms and the number of phones in the phrase is less than 2.5[4]. Otherwise, when not preceding /d/, /ɧ/, /ç/, /m/, /n/, /h/, /l/, /ɹ/ or /ɖ/, an /ɹ/ is realised as [ɹ].

Finally, if /ɹ/ does not precede /d/, /ɧ/, /ç/, /m/, /n/, /h/, /l/, /ɹ/ or /ɖ/ and the mean phoneme duration over the word is more than 36.6 ms, the realisation of

[3] *py* is the combination of the *place of articulation* feature for consonants and the *tongue position in the y (open-close) dimension* feature for vowels.

[4] Under the phrase chunking method employed, subjunctions and conjunctions were treated as one-word phrases and this rule seems to target such units specifically.

/ɹ/ depends on the word to the right of the current word (cf. Figure I.18 to see which particular function words render a specific realisation). The *function word+1* node has four branches. However, again the pruning procedure has created two branches with the same leaf ([ɹ]) which are equivalent to a single [ɹ] branch. The remaining branches gives an ∅ realisation and leads to a *syllable nucleus* node, respectively. If the syllable nucleus is /ɑː/ or /e/, and the type of the phrase preceding the current phrase is *not* VC (verb cluster), the /ɹ/ is realised as ∅ and otherwise as [ɹ].

Both from the *syllable nucleus* and the *phrase type-1* node, there are branches with identical leaves, which are equivalent to a single leaf. One of the branches from the *syllable nucleus* node is for "no syllable nucleus" (–). All normal syllables contain a nucleus, which in central standard Swedish is always a vowel. However, some interrupted words contain interrupted syllables. For example, an interrupted word may consist of only a word initial consonant. In the syllable layer, these relatively rare interrupted syllable units are treated as syllables with no nuclei.

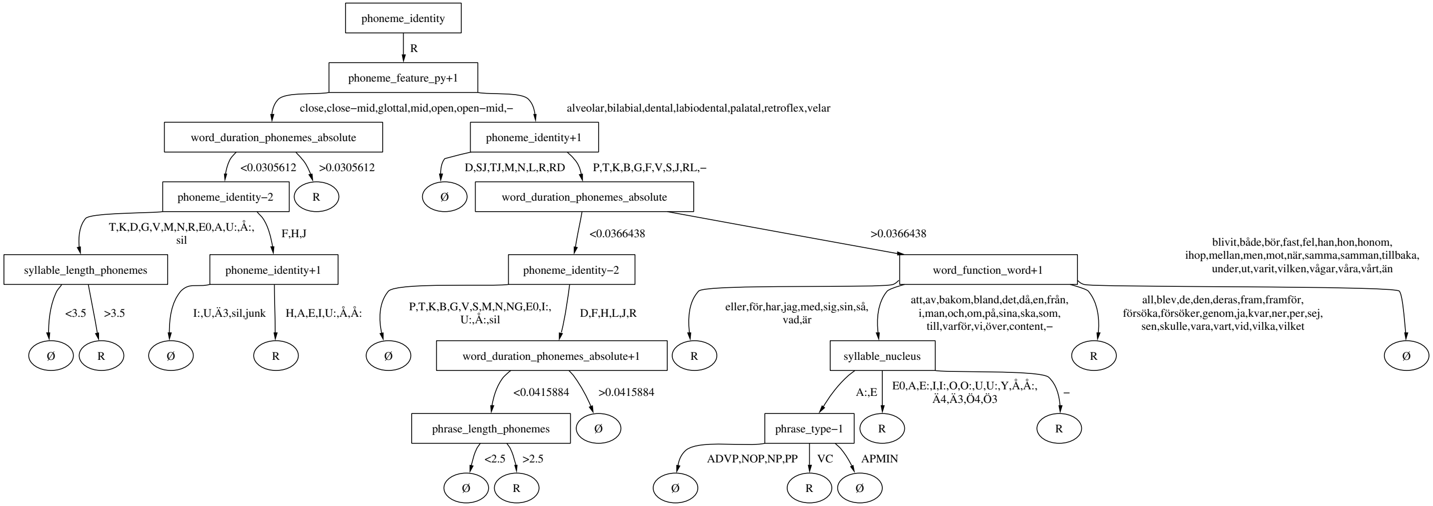

phoneme_identity
R
phoneme_feature_py+1
close,close−mid,glottal,mid,open,open−mid,−
alveolar,bilabial,dental,labiodental,palatal,retroflex,velar
word_duration_phonemes_absolute
<0.0305612
>0.0305612
R
phoneme_identity+1
D,SJ,TJ,M,N,L,R,RD
P,T,K,B,G,F,V,S,J,RL,−
Ø
word_duration_phonemes_absolute
phoneme_identity−2
T,K,D,G,V,M,N,R,E0,A,U:,Å:,
sil
F,H,J
syllable_length_phonemes
<3.5
>3.5
Ø
R
phoneme_identity+1
I:,U,Ä3,sil,junk
H,A,E,I,U:,Å,Å:
Ø
R
<0.0366438
>0.0366438
phoneme_identity−2
P,T,K,B,G,V,S,M,N,NG,E0,I:,
U:,Å:,sil
D,F,H,L,J,R
Ø
word_duration_phonemes_absolute+1
<0.0415884
>0.0415884
phrase_length_phonemes
<2.5
>2.5
Ø
R
Ø
word_function_word+1
eller,för,har,jag,med,sig,sin,så,
vad,är
R
att,av,bakom,bland,det,då,en,från,
i,man,och,om,på,sina,ska,som,
till,varför,vi,över,content,−
syllable_nucleus
A:,E
E0,A,E:,I,I:,O,O:,U,U:,Y,Å,Å:,
Ä4,Ä3,Ö4,Ö3
−
phrase_type−1
ADVP,NOP,NP,PP
VC
APMIN
Ø
R
Ø
R
R
all,blev,de,den,deras,fram,framför,
försöka,försöker,genom,ja,kvar,ner,per,sej,
sen,skulle,vara,vart,vid,vilka,vilket
R
blivit,både,bör,fast,fel,han,hon,honom,
ihop,mellan,men,mot,när,samma,samman,tillbaka,
under,ut,varit,vilken,vågar,våra,vårt,än
Ø

Figure I.18: *The realisations of the phoneme /ɹ/ (phoneme representations in the figure are in STA format).*

The Realisation of /ʈ/

In the key transcript, a /ʈ/ is realised as [ʈ] in almost 98% of the cases and realised as ∅ in the few remaining cases, as shown in Table I.19. The decision tree model does not include the option of eliding /ʈ/, as seen in Table I.19 and Figure I.19.

Table I.19: *The realisations of the phoneme /ʈ/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	7	2.23%	0	0.00%	0	0.00%
ʈ	307	97.77%	314	100.00%	307	100.00%
∑	314	100.00%	314	100.00%	307	97.77%

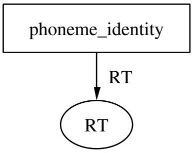


Figure I.19: *The realisations of the phoneme /ʈ/ (phoneme representations in the figure are in STA format).*

The Realisation of /ɖ/

In the key transcript, a /ɖ/ is realised as [ɖ] in all cases but one, where it is elided, as can be seen in Table I.20. The pruned decision tree model always realises /ɖ/ as [ɖ], as seen in Table I.20 and Figure I.20.

Table I.20: *The realisations of the phoneme /ɖ/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	1	0.48%	0	0.00%	0	0.00%
ɖ	209	99.52%	210	100.00%	209	100.00%
∑	210	100.00%	210	100.00%	209	99.52%

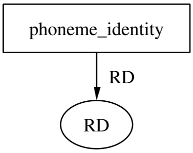


Figure I.20: *The realisations of the phoneme /ɖ/ (phoneme representations in the figure are in STA format).*

The Realisation of /ɭ/

An /ɭ/ is realised as [ɭ] and as ∅, respectively, equally often in the key transcript. In the model output, /ɭ/ is elided in about 59% of the cases, as can be seen in Table I.21. Since there were very few training examples, the selection of attributes during model training was probably highly dependent on chance. The model may not be very good for predicting the realisations of /ɭ/ generally, since it is probably biased towards the particular data used for training the model.

Table I.21: *The realisations of the phoneme /ɭ/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	11	50.00%	13	59.09%	11	100.00%
ɭ	11	50.00%	9	40.91%	9	81.82%
∑	22	100.00%	22	100.00%	20	90.91%

As illustrated by the tree in Figure I.21, the pitch dynamic measure defined as the sum of extreme point distances from the median f_0 over the utterance in Hz divided by the number of extremes (minimum or maximum points or plateaus) contained by the utterance (i.e., the average distance between the median and the extreme point frequencies) is used as the first attribute under the *phoneme identity* node for /ɭ/.

If the pitch dynamic value is higher than 33.25, /ɭ/ is always realised as [ɭ]. If the pitch dynamic value is lower than 33.25, /ɭ/ is realised as ∅ if the mean normalised logarithmic phoneme duration over the right adjacent syllable is less than 0.02 and as [ɭ] otherwise.

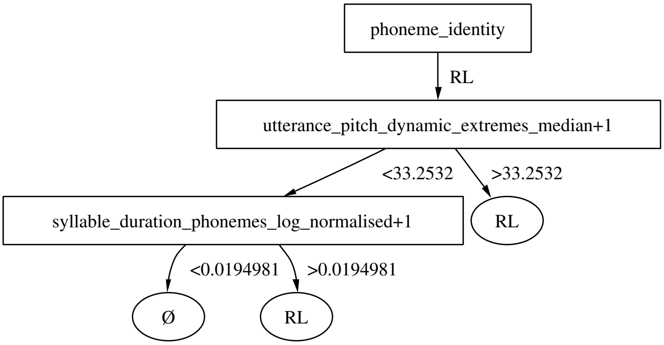

phoneme_identity
RL
utterance_pitch_dynamic_extremes_median+1
<33.2532
>33.2532
syllable_duration_phonemes_log_normalised+1
RL
<0.0194981
>0.0194981
Ø
RL

Figure I.21: *The realisations of the phoneme /ɭ/ (phoneme representations in the figure are in STA format).*

The Realisation of /ɳ/

The /ɳ/ phoneme is realised as [ɳ] in somewhat more than 70% of the cases and elided in the remainder of the cases, both according to the key transcript and to the decision tree model, as shown in Table I.22.

Table I.22: *The realisations of the phoneme /ɳ/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	63	26.36%	69	28.87%	43	68.25%
ɳ	176	73.64%	170	71.13%	150	85.23%
∑	239	100.00%	239	100.00%	193	80.75%

In the model, if the mean normalised log phoneme duration over the word is less than -0.21, /ɳ/ is elided and otherwise, it is realised as [ɳ], as illustrated in Figure I.22.

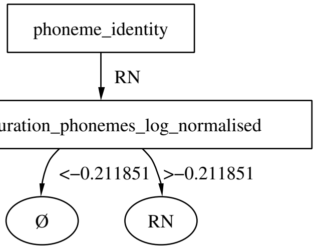


Figure I.22: *The realisations of the phoneme /ɳ/ (phoneme representations in the figure are in STA format).*

The Realisation of /ʂ/

In the key transcript, a /ʂ/ is realised as [ʂ] in about 88% of the cases and otherwise elided, as can be seen in Table I.23. The pruned decision tree model always realises /ʂ/ as [ʂ], as shown by Table I.23 and Figure I.23.

Table I.23: *The realisations of the phoneme /ʂ/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	37	12.05%	0	0.00%	0	0.00%
ʂ	270	87.95%	307	100.00%	270	100.00%
∑	307	100.00%	307	100.00%	270	87.95%

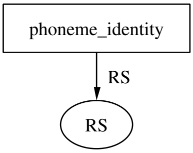


Figure I.23: *The realisations of the phoneme /ʂ/ (phoneme representations in the figure are in STA format).*

Appendix J

Vowel Realisations

The Realisation of /a/

An /a/ can be realised as [a] or [ə] or, although very seldom, be elided. Table J.1 shows the details of the distribution of realisations of the /a/ phoneme.

Table J.1: *The realisations of the phoneme /a/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	61	0.80%	49	0.64%	30	49.18%
ə	1,997	26.13%	1,879	24.58%	1,623	81.27%
a	5,586	73.08%	5,716	74.78%	5,345	95.69%
$\sum$	7,644	100.00%	7,644	100.00%	6,998	91.55%

As can be seen from Figure J.1, the realisation of /a/ is dependent on several different duration-based attributes: 1) the mean *normalised vowel* duration over the *word*, 2) the mean *log vowel* duration over the *word*, 3) the mean *phoneme* duration over the *word*, 4) the mean *normalised log phoneme* duration over the *phrase*, 5) the mean *vowel duration* over the *discourse*, and 6) the mean *log phoneme duration* over *the following word*. The length of the preceding phrase (measured in number of phonemes) is also utilised by the model.

The syllable *stress type* variable is used high up in the tree structure. Here, it is used to group syllables with primary stress in accent II words and compounds (*prim2*) with syllables with secondary stress in compounds (*secondComp*) on the one hand and syllables with stress in accent I words (*prim1*) with syllables with secondary stress in accent II words (*second2*) and unstressed syllables (*no*) on the other hand.

Typically, function words are accent I words and their word stress is not realised in practice, since function words mostly have an unstressed pronunciation. Nevertheless, in the canonical pronunciation representation, the word stress is marked.

Thus, since function words are so common, the *prim1* syllables will be equivalent to *no* in the majority of cases in practice. The secondary stress in accent II words is often much less pronounced than the primary stress. Thus, although theoretically disparate, the *prim1*, *second2* and *no* syllables in practice make up a likely cluster.

The realisation of /a/ is also to a high degree dependent on the *function word* attribute and the *function word+1* attribute and an /a/ has a greater probability of being reduced to a [ə] in a syllable in word final position than if the /a/ is the nucleus of a word initial or word medial syllable.

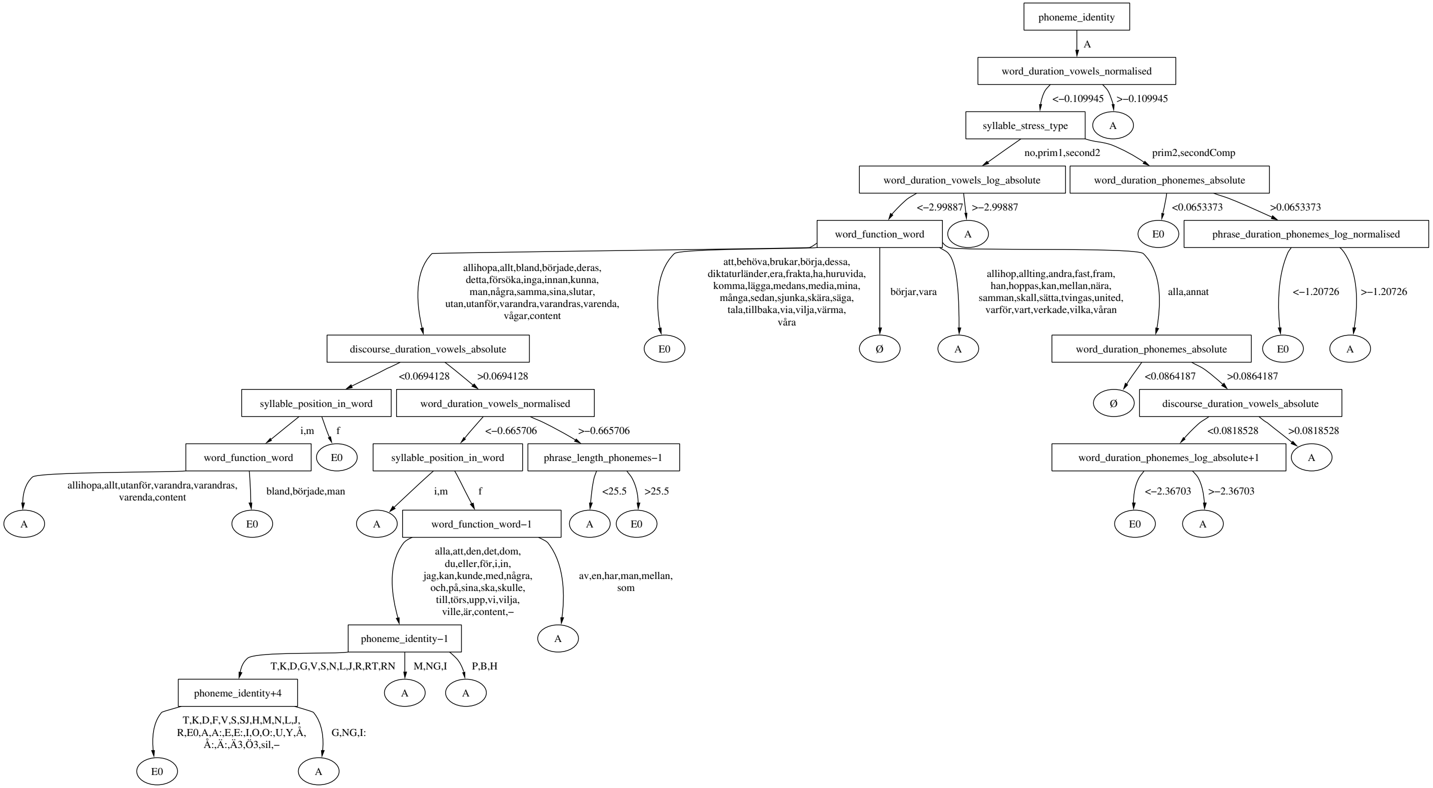

phoneme_identity
A
word_duration_vowels_normalised
<−0.109945
>−0.109945
syllable_stress_type
A
no,prim1,second2
prim2,secondComp
word_duration_vowels_log_absolute
word_duration_phonemes_absolute
<−2.99887
>−2.99887
<0.0653373
>0.0653373
word_function_word
A
E0
phrase_duration_phonemes_log_normalised
allihopa,allt,bland,började,deras, detta,försöka,inga,innan,kunna, man,några,samma,sina,slutar, utan,utanför,varandra,varandras,varenda, vågar,content
att,behöva,brukar,börja,dessa, diktaturländer,era,frakta,ha,huruvida, komma,lägga,medans,media,mina, många,sedan,sjunka,skära,säga, tala,tillbaka,via,vilja,värma, våra
börjar,vara
allihop,allting,andra,fast,fram, han,hoppas,kan,mellan,nära, samman,skall,sätta,tvingas,united, varför,vart,verkade,vilka,våran
alla,annat
<−1.20726
>−1.20726
discourse_duration_vowels_absolute
E0
Ø
A
word_duration_phonemes_absolute
E0
A
<0.0694128
>0.0694128
<0.0864187
>0.0864187
syllable_position_in_word
word_duration_vowels_normalised
Ø
discourse_duration_vowels_absolute
i,m
f
<−0.665706
>−0.665706
<0.0818528
>0.0818528
word_function_word
E0
syllable_position_in_word
phrase_length_phonemes−1
word_duration_phonemes_log_absolute+1
A
allihopa,allt,utanför,varandra,varandras, varenda,content
bland,började,man
i,m
f
<25.5
>25.5
<−2.36703
>−2.36703
A
E0
A
word_function_word−1
A
E0
E0
A
alla,att,den,det,dom, du,eller,för,i,in, jag,kan,kunde,med,några, och,på,sina,ska,skulle, till,törs,upp,vi,vilja, ville,är,content,−
av,en,har,man,mellan, som
phoneme_identity−1
A
T,K,D,G,V,S,N,L,J,R,RT,RN
M,NG,I
P,B,H
phoneme_identity+4
A
A
T,K,D,F,V,S,SJ,H,M,N,L,J, R,E0,A,A:,E,E:,I,O,O:,U,Y,Å, Å:,Ä:,Ä3,Ö3,sil,−
G,NG,I:
E0
A

Figure J.1: *The realisations of the phoneme /a/ (phoneme representations in the figure are in STA format).*

The Realisation of /ɑː/

The realisation of /ɑː/ is restricted to [ɑː] and [ə] in the pronunciation model, although the key transcript also includes five instances of [a] and a ∅ realisation, as shown in Table J.2.

Table J.2: *The realisations of the phoneme /ɑː/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	1	0.03%	0	0.00%	0	0.00%
ə	768	25.30%	726	23.91%	725	94.40%
a	5	0.16%	0	0.00%	0	0.00%
ɑː	2,262	74.51%	2,310	76.09%	2,261	99.96%
$\sum$	3,036	100.00%	3,036	100.00%	2,986	98.35%

Figure J.2 shows that the realisation of /ɑː/ is dependent on several duration-based attributes, on the identity of the preceding phoneme, on the *function word* identity and on the *Part of Speech* of the word in which the /ɑː/ appears.

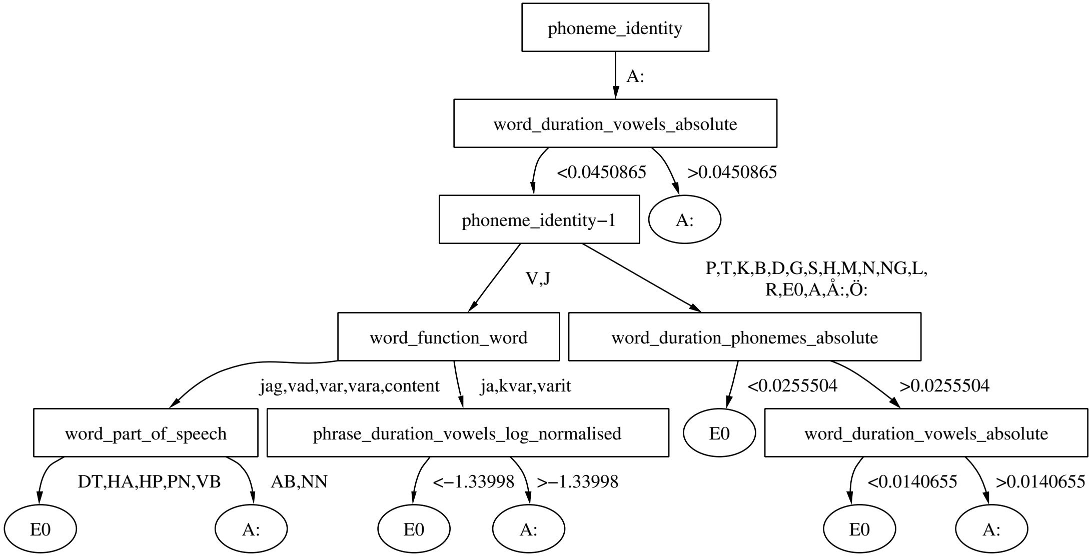


Figure J.2: *The realisations of the phoneme /ɑː/ (phoneme representations in the figure are in STA format).*

The Realisation of /ə/, /e/ and /eː/

The phonemes /ə/, /e/ and /eː/ share an arc from the top *phoneme identity* node, but already at the next tree level, the tree is split into a separate branch for each of the phonemes, as can be seen in Figure J.3. This odd structure occurred since a greater symmetric information gain at the first split was obtained when the tree inducer grouped these similar phonemes than when separate branches were created for them. However, at the second split, the greatest gain in information was obtained by splitting the phoneme group.

Table J.3: *The realisations of the phoneme /ə/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	51	1.29%	43	1.09%	36	70.59%
ə	3,905	98.71%	3,913	98.91%	3,898	99.82%
$\sum$	3,956	100.00%	3,956	100.00%	3,934	99.44%

The realisation of a /ə/ is dependent on the mean phoneme duration over the word and over the phrase, as can be seen in Figure J.3. A /ə/ in the canonical pronunciation representation is almost exclusively realised as a [ə], but is elided in about 1% of the cases, as shown in Table J.3.

Table J.4: *The realisations of the phoneme /e/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	2	0.06%	0	0.00%	0	0.00%
ə	1,367	40.77%	1,681	50.13%	1,306	95.54%
e	1,984	59.17%	1,672	49.87%	1,609	81.10%
$\sum$	3,353	100.00%	3,353	100.00%	2,915	86.94%

The realisation of /e/ depends on the *function word* attribute[1], the *syllable stress type* and the mean phoneme duration over the word, as can be seen in Figure J.3. In the key transcript, /e/ is realised as [e] in the majority of cases, elided in 2 cases and realised as [ə] in the remainder of cases. In the model output, /e/ is realised as [e] and as [ə] about equally often, as shown in Table J.4.

The realisation of /eː/ is dependent on the mean vowel duration over the word, the *function word* attribute and the mean phoneme duration over the word, as can be seen in Figure J.3. An /eː/ is mostly realised as [eː] or [ə], but can also be realised as [e]. In the model output, the [e] realisation occurs less often than in the key transcript. The ∅ realisation, which occurs once in the key transcript, is not allowed by the decision tree model, as can be seen in Table J.5 and Figure J.3.

[1] N.b.: the proper name *tv-fem 'TV five'* is erroneously included in the list of function words.

Table J.5: *The realisations of the phoneme /eː/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	1	0.04%	0	0.00%	0	0.00%
ə	1,051	37.10%	1,120	39.53%	1,025	97.53%
e	235	8.30%	93	3.28%	42	17.87%
eː	1,546	54.57%	1,620	57.18%	1,477	95.54%
$\sum$	2,833	100.00%	2,833	100.00%	2,544	89.80%

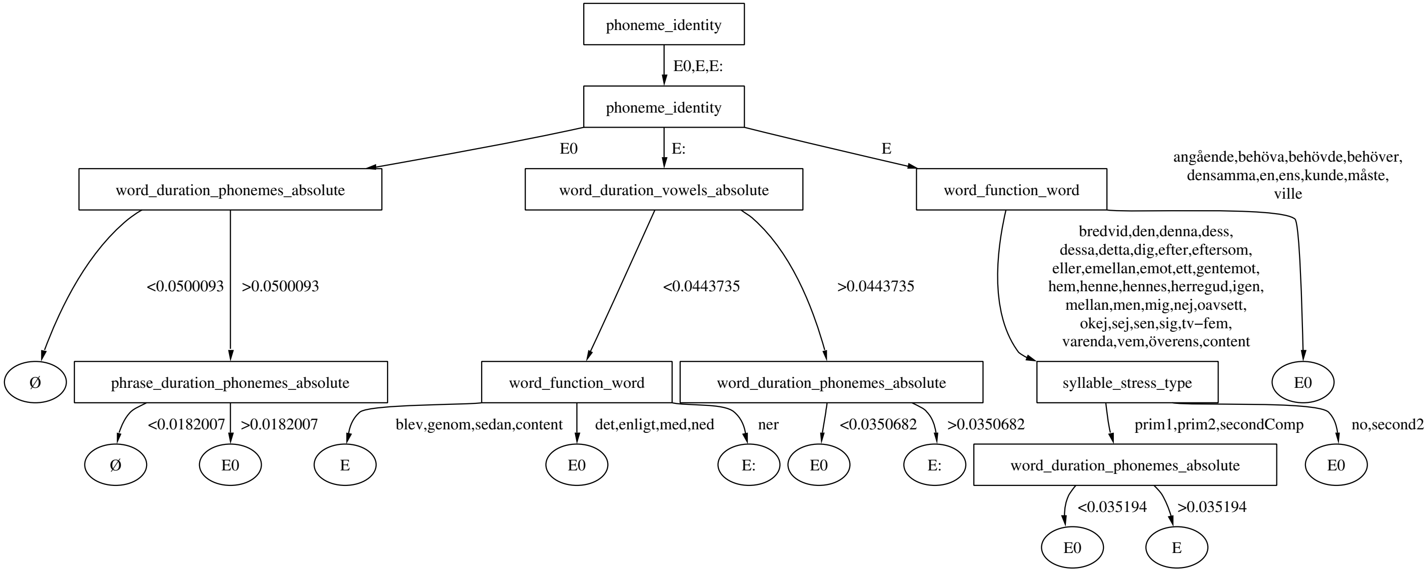

phoneme_identity
E0,E,E:
phoneme_identity
E0
E:
E
word_duration_phonemes_absolute
word_duration_vowels_absolute
word_function_word
angående,behöva,behövde,behöver,
densamma,en,ens,kunde,måste,
ville
<0.0500093
>0.0500093
<0.0443735
>0.0443735
bredvid,den,denna,dess,
dessa,detta,dig,efter,eftersom,
eller,emellan,emot,ett,gentemot,
hem,henne,hennes,herregud,igen,
mellan,men,mig,nej,oavsett,
okej,sej,sen,sig,tv−fem,
varenda,vem,överens,content
Ø
phrase_duration_phonemes_absolute
word_function_word
word_duration_phonemes_absolute
syllable_stress_type
E0
<0.0182007
>0.0182007
blev,genom,sedan,content
det,enligt,med,ned
ner
<0.0350682
>0.0350682
prim1,prim2,secondComp
no,second2
Ø
E0
E
E0
E:
E0
E:
word_duration_phonemes_absolute
E0
<0.035194
>0.035194
E0
E

Figure J.3: *The realisations of the phonemes /ə/, /e/ and /eː/ (phoneme representations in the figure are in STA format).*

The Realisation of /ɪ/

An /ɪ/ is always realised as [ɪ] in the pronunciation model, as apparent from Figure J.4 and Table J.6. In the key transcript, /ɪ/ can also be realised as [ə] or be elided. It is likely that the simplification of the /ɪ/ realisation rules resulting from pruning the decision tree model does not increase the prediction performance for /ɪ/ realisations on new data.

Table J.6: *The realisations of the phoneme /ɪ/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	9	0.25%	0	0.00%	0	0.00%
ə	531	14.75%	0	0.00%	0	0.00%
ɪ	3,060	85.00%	3,600	100.00%	3,060	100.00%
$\sum$	3,600	100.00%	3,600	100.00%	3,060	85.00%

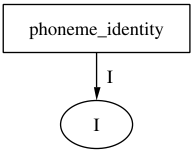


Figure J.4: *The realisations of the phoneme /ɪ/ (phoneme representations in the figure are in STA format).*

The Realisation of /iː/

An /iː/ can be realised as [iː], [ɪ], [ə] or ∅ and has similar realisation distributions in the key transcript and the model output. However, as can be seen from Table J.7, the decision tree model produces a greater share of [iː] realisations and a lesser share of [ɪ] realisations than the automatic transcription system.

Table J.7: *The realisations of the phoneme /iː/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	11	0.60%	13	0.71%	11	100.00%
ə	290	15.80%	286	15.59%	283	97.59%
ɪ	434	23.65%	324	17.66%	317	73.04%
iː	1,100	59.95%	1,212	66.05%	1,092	99.27%
$\sum$	1,835	100.00%	1,835	100.00%	1,703	92.81%

Figure J.5 shows that the realisation of /iː/ depends on duration-based measures over the word and over the phrase, on the *function word* attribute and on the *Part of Speech* of the word in which the /iː/ occurs. For verbs (VB), the number of times the lexeme (the word base form or any inflected form of the word) has been repeated thus far in the discourse, is used as a predictor. If the duration-based criteria match and the verb has been repeated more than 7.5 times, and thus is likely to convey *given* information, the /iː/ is realised as [ɪ] rather than as [iː], as it is when the verb has been repeated less than 7.5 times.

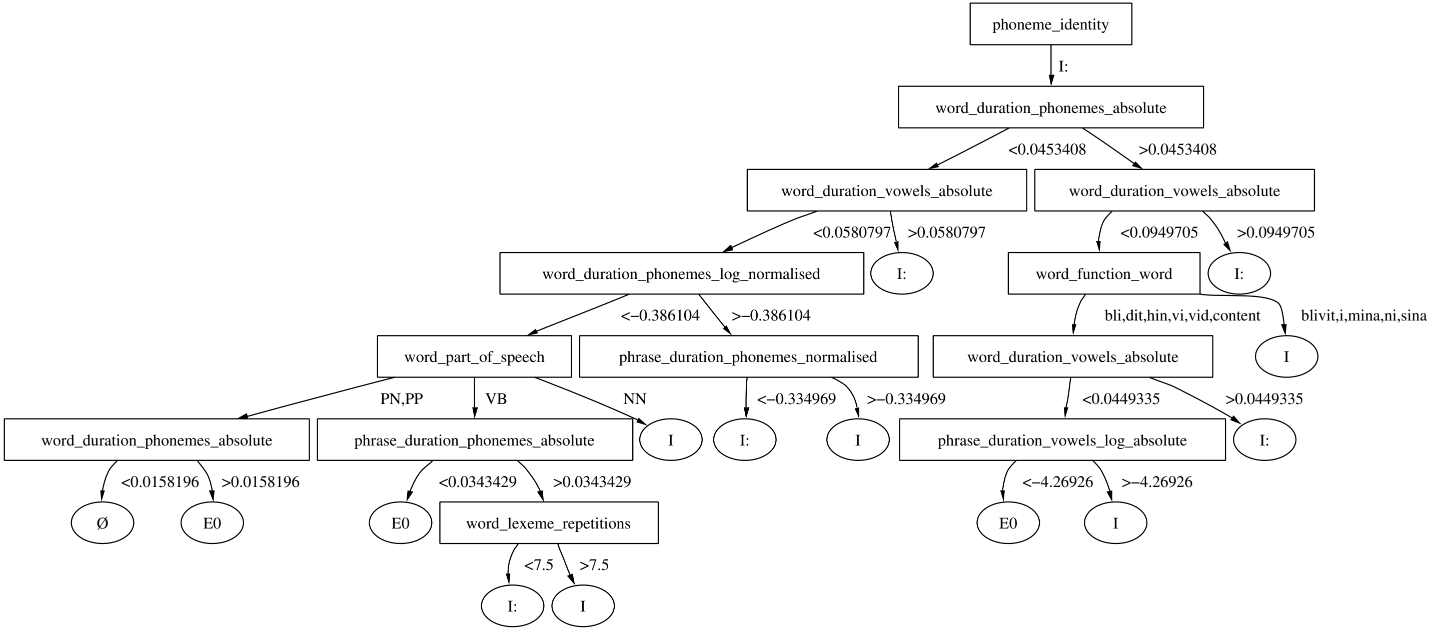


Figure J.5: *The realisations of the phoneme /***i:***/ (phoneme representations in the figure are in STA format).*

The Realisation of /ʊ/

The phoneme /ʊ/ has very similar realisation distributions in the key transcript and in the model output, respectively. The phoneme can be realised as [ʊ], [ə] or ∅, as can be seen in Table J.8.

Table J.8: *The realisations of the phoneme /ʊ/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	6	1.17%	6	1.17%	6	100.00%
ə	160	31.25%	171	33.40%	117	73.12%
ʊ	346	67.58%	335	65.43%	292	84.39%
$\sum$	512	100.00%	512	100.00%	415	81.05%

The realisation of /ʊ/ is not dependent on duration-based attributes in the pruned decision tree model, shown in Figure J.6. Instead, the realisation is dependent on phoneme context and word predictability (a weighted sum of trigram probability, bigram probability and unigram probability).

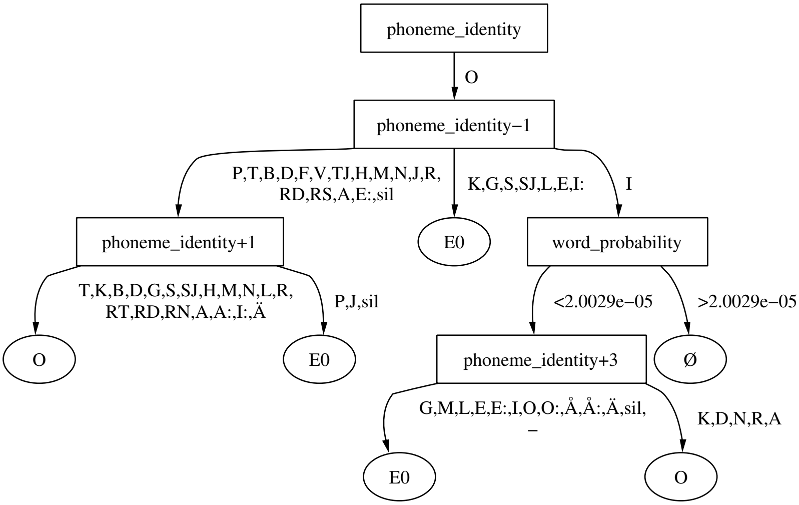


Figure J.6: *The realisations of the phoneme /ʊ/ (phoneme representations in the figure are in STA format).*

The Realisation of /uː/

An /uː/ is mainly realised as [uː], but may also be realised as [ʊ] and, in some cases, as [ə] or ∅. All of these four realisations occur in both the key transcript and in the model output, however with slightly different distributions. Table J.9 gives the details of the realisation distributions for /uː/.

Table J.9: *The realisations of the phoneme /uː/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	3	0.36%	3	0.36%	3	100.00%
ə	9	1.07%	2	0.24%	2	22.22%
ʊ	95	11.28%	81	9.62%	65	68.42%
uː	735	87.29%	756	89.79%	724	98.50%
$\sum$	842	100.00%	842	100.00%	794	94.30%

As can be seen from Figure J.7, the realisation of /uː/ in the pronunciation model depends on the mean vowel duration over the word, on the identity of the phoneme two positions to the left of the /uː/ and on the identity of the word preceding the word in which the /uː/ occurs.

It is hypothesised that the realisations of both /ʊ/ and /uː/ suffer from data sparsity and that these particular realisation rules are to some degree specific to the particular training data.

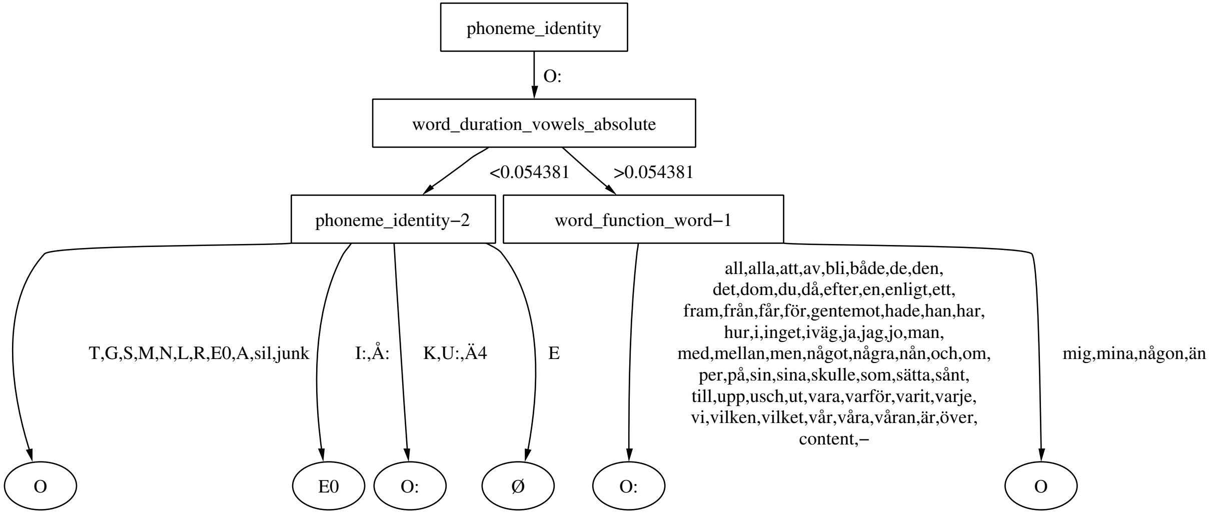

phoneme_identity
O:
word_duration_vowels_absolute
<0.054381
>0.054381
phoneme_identity−2
word_function_word−1
T,G,S,M,N,L,R,E0,A,sil,junk
I:,Å:
K,U:,Ä4
E
all,alla,att,av,bli,både,de,den, det,dom,du,då,efter,en,enligt,ett, fram,från,får,för,gentemot,hade,han,har, hur,i,inget,iväg,ja,jag,jo,man, med,mellan,men,något,några,nån,och,om, per,på,sin,sina,skulle,som,sätta,sånt, till,upp,usch,ut,vara,varför,varit,varje, vi,vilken,vilket,vår,våra,våran,är,över, content,−
mig,mina,någon,än
O
E0
O:
Ø
O:
O

Figure J.7: *The realisations of the phoneme /uː/ (phoneme representations in the figure are in STA format).*

The Realisation of /ɵ/

An /ɵ/ can be realised as [ɵ] or [ə] or be elided according to the key transcript. The model does not allow for the /ɵ/ be elided and very seldom produces a [ə] realisation for /ɵ/, as can be seen in Table J.10. Only the mean vowel duration over the word is used by the model to determine the realisation of /ɵ/, as shown by Figure J.8.

Table J.10: *The realisations of the phoneme /ɵ/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	118	11.53%	0	0.00%	0	0.00%
ə	68	6.65%	49	4.79%	49	72.06%
ɵ	837	81.82%	974	95.21%	837	100.00%
∑	1,023	100.00%	1,023	100.00%	886	86.61%

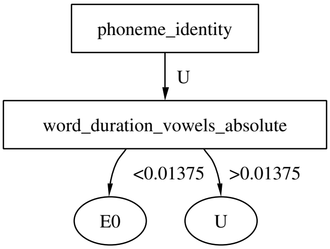


Figure J.8: *The realisations of the phoneme /ɵ/ (phoneme representations in the figure are in STA format).*

The Realisation of /ʉ̟ː/

As shown in Table J.11, /ʉ̟ː/ is realised as [ʉ̟ː] in the majority of cases and the majority class is more pronounced in the model output than in the key transcript, as it is for the realisations of most phonemes. Other realisations that occur in both the key transcript and the model output are [ɵ], [ə] and ∅.

Table J.11: *The realisations of the phoneme /ʉ̟ː/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	2	0.19%	1	0.09%	1	50.00%
ə	93	8.81%	89	8.43%	83	89.25%
ɵ	65	6.16%	5	0.47%	5	7.69%
ʉ̟ː	896	84.85%	961	91.00%	895	99.89%
$\sum$	1,056	100.00%	1,056	100.00%	984	93.18%

The realisation rules for /ʉ̟ː/ are illustrated by the tree in Figure J.9. Here, it can be seen that the realisation of /ʉ̟ː/ depends on duration-based measures over the word, on the Part of Speech of the word in which the /ʉ̟ː/ occurs and on the phoneme context, however not the adjacent phonemes.

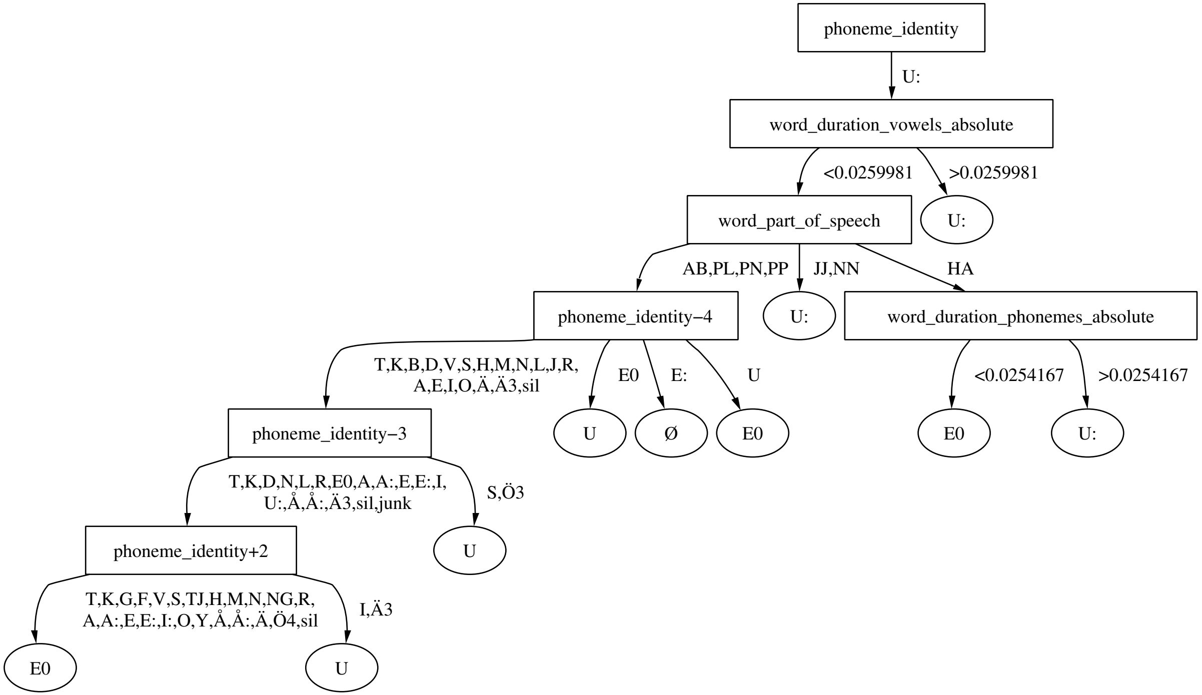


Figure J.9: *The realisations of the phoneme /ʉ̟ː/ (phoneme representations in the figure are in STA format).*

The Realisation of /ʏ/

An /ʏ/ can be realised as [ʏ], [ə], or ∅ in the key transcript. As can be seen in Table J.12, the [ə] and ∅ realisations are very rare and in the model output, only the [ʏ] realisation occurs. Figure J.10 shows that this is the only possible output from the model.

Table J.12: *The realisations of the phoneme /ʏ/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	2	0.44%	0	0.00%	0	0.00%
ə	1	0.22%	0	0.00%	0	0.00%
ʏ	447	99.33%	450	100.00%	447	100.00%
$\sum$	450	100.00%	450	100.00%	447	99.33%

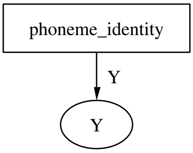


Figure J.10: *The realisations of the phoneme /ʏ/ (phoneme representations in the figure are in STA format).*

The Realisation of /yː/

The phoneme /yː/ is realised as [yː], [ʏ] or [ə] and have similar realisation distributions in the key transcript and in the model output, respectively. Table J.13 shows that the [ə] realisation is infrequent.

Table J.13: *The realisations of the phoneme /yː/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
ə	2	1.59%	4	3.17%	2	100.00%
ʏ	12	9.52%	11	8.73%	10	83.33%
yː	112	88.89%	111	88.10%	110	98.21%
$\sum$	126	100.00%	126	100.00%	122	96.83%

In the pronunciation model, the realisation of /yː/ depends on the identity of the phoneme four positions to the right of the /yː/ and on duration-based measures over the phrase and over the syllable. It is likely that the phoneme context attribute targets specific words and that the realisation rule set for /yː/ may not be fully generalisable to new data as a result of data sparsity.

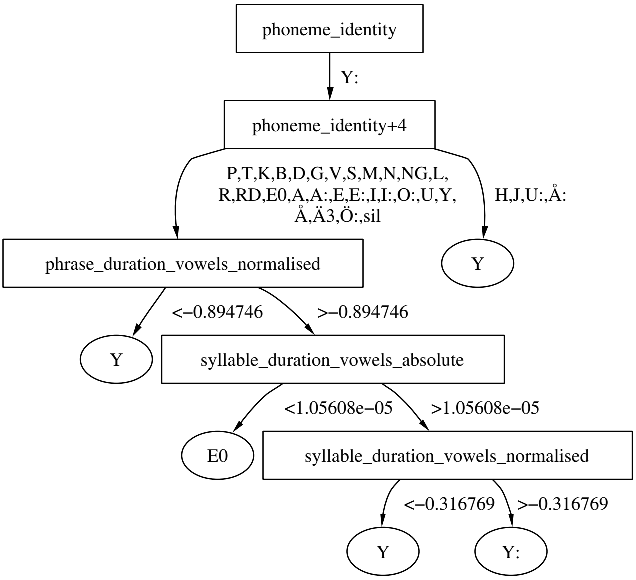

phoneme_identity
Y:
phoneme_identity+4
P,T,K,B,D,G,V,S,M,N,NG,L, R,RD,E0,A,A:,E,E:,I,I:,O:,U,Y, Å,Ä3,Ö:,sil
H,J,U:,Å:
phrase_duration_vowels_normalised
Y
<–0.894746
>–0.894746
Y
syllable_duration_vowels_absolute
<1.05608e–05
>1.05608e–05
E0
syllable_duration_vowels_normalised
<–0.316769
>–0.316769
Y
Y:

Figure J.11: *The realisations of the phoneme /yː/ (phoneme representations in the figure are in STA format).*

The Realisation of /ɔ/

Table J.14 shows that /ɔ/ is mainly realised as [ɔ] or [ə] and sometimes elided. The share of elisions is lower in the model output than in the key transcript.

Table J.14: *The realisations of the phoneme /ɔ/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	198	4.86%	64	1.57%	52	26.26%
ə	946	23.21%	888	21.79%	855	90.38%
ɔ	2,932	71.93%	3,124	76.64%	2,892	98.64%
$\sum$	4,076	100.00%	4,076	100.00%	3,799	93.20%

The phoneme /ɔ/ occurs in many high frequency function words, and is always elided if occurring in one of the words *eftersom ‘since’*, *någon ‘someone’* or *något ‘something’*, as can be seen in Figure J.12. If the /ɔ/ occurs in one of the function words *liksom ‘like’*, *och ‘and’* or *som ‘that’*, duration-based measures and the identity of the preceding phoneme are used to chose between an [ɔ] and a [ə] realisation. If occurring in any other word than the above mentioned, /ɔ/ is always realised as [ɔ].

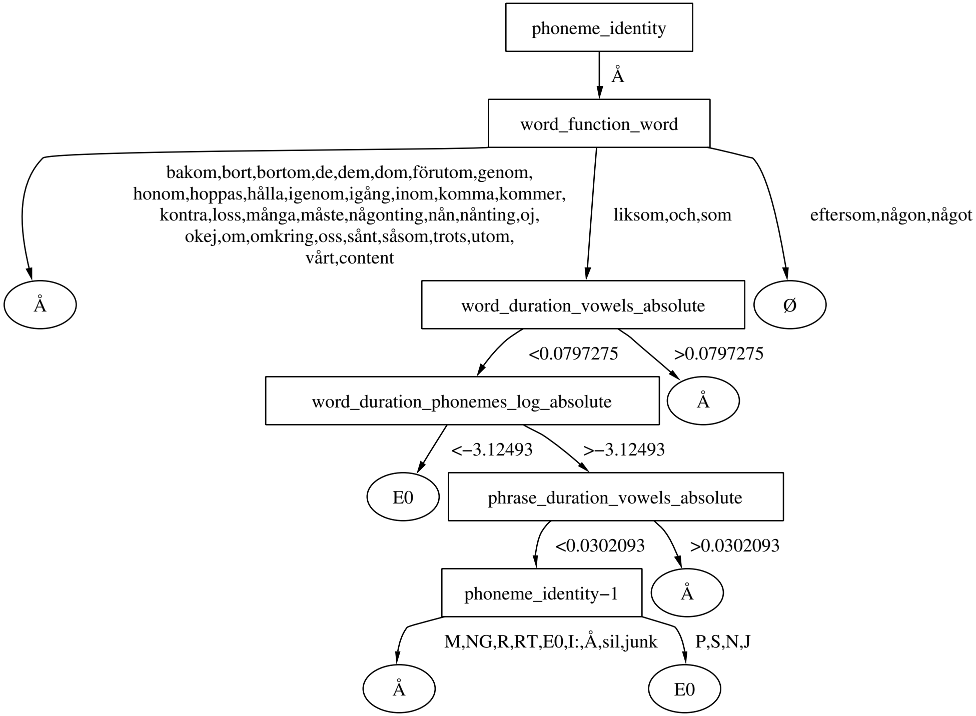

phoneme_identity
Å
word_function_word
bakom,bort,bortom,de,dem,dom,förutom,genom,
honom,hoppas,hålla,igenom,igång,inom,komma,kommer,
kontra,loss,många,måste,någonting,nån,nånting,oj,
okej,om,omkring,oss,sånt,såsom,trots,utom,
vårt,content
liksom,och,som
eftersom,någon,något
Å
word_duration_vowels_absolute
Ø
<0.0797275
>0.0797275
word_duration_phonemes_log_absolute
Å
<−3.12493
>−3.12493
E0
phrase_duration_vowels_absolute
<0.0302093
>0.0302093
phoneme_identity−1
Å
M,NG,R,RT,E0,I:,Å,sil,junk
P,S,N,J
Å
E0

Figure J.12: *The realisations of the phoneme /ɔ/ (phoneme representations in the figure are in STA format).*

The Realisation of /oː/

An /oː/ can be realised as [oː], [ɔ] or [ə] by the pronunciation model, as can be seen in Table J.15. In the key transcript, there is also one case of /oː/ elision. An /oː/ is realised as [oː] in about 60% of the cases according to both the key transcript and the model output. The [ɔ] realisation is more common in the model output than in the key transcript at the expense of the [ə] realisation.

Table J.15: *The realisations of the phoneme /oː/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	1	0.04%	0	0.00%	0	0.00%
ə	193	8.37%	117	5.07%	114	59.07%
ɔ	707	30.66%	795	34.48%	578	81.75%
oː	1,405	60.93%	1,394	60.45%	1,263	89.89%
$\sum$	2,306	100.00%	2,306	100.00%	1,955	84.78%

Figure J.13 shows that /oː/ is realised as [oː] if the mean vowel duration over the word is more than 54.8 ms and that duration-based measures are also used to select between the [ɔ] and [ə] realisations when the mean vowel duration over the word is less than 54.8 ms.

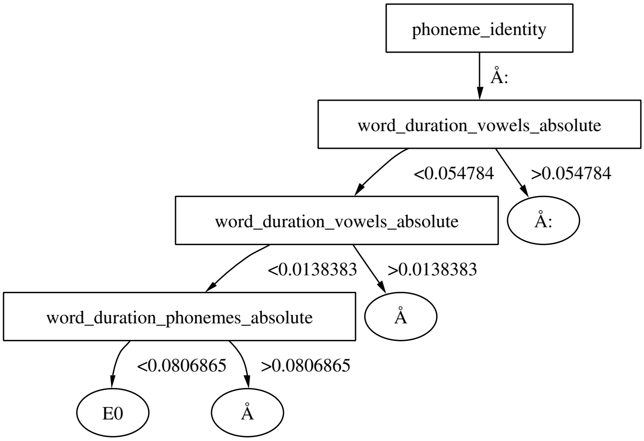

phoneme_identity
Å:
word_duration_vowels_absolute
<0.054784
>0.054784
word_duration_vowels_absolute
Å:
<0.0138383
>0.0138383
word_duration_phonemes_absolute
Å
<0.0806865
>0.0806865
E0
Å

Figure J.13: *The realisations of the phoneme /oː/ (phoneme representations in the figure are in STA format).*

The Realisation of /ɛ/

The phoneme /ɛ/ is mostly realised as [ɛ], but is sometimes reduced to a [ə]. As can be seen in Table J.16 and Figure J.14, the model never elides a /ɛ/, although there are some elided /ɛ/ phonemes according to the key transcript. The fact that there is a high concentration of /ɛ/ realisations to [ɛ] in the key transcript makes the realisation easier to predict and the tree model makes the correct decision about phone identity in 98.5% of the cases.

Table J.16: *The realisations of the phoneme /ɛ/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	7	0.58%	0	0.00%	0	0.00%
ə	23	1.92%	34	2.84%	20	86.96%
ɛ	1,167	97.49%	1,163	97.16%	1,159	99.31%
$\sum$	1,197	100.00%	1,197	100.00%	1,179	98.50%

Figure J.14 shows that an /ɛ/ is always realised as [ə] in unstressed syllables and can also be realised as [ə] in stressed syllables, if the mean phoneme duration over the phrase is less than 27.2 ms. Otherwise, /ɛ/ is realised as [ɛ].

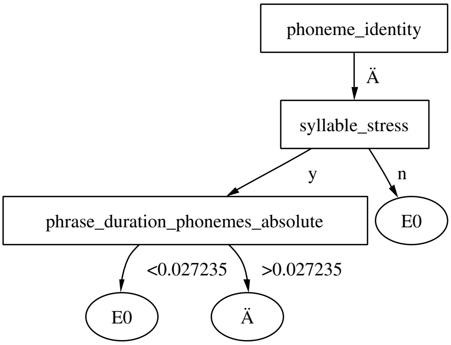


Figure J.14: *The realisations of the phoneme /ɛ/ (phoneme representations in the figure are in STA format).*

The Realisation of /ɛː/

The /ɛː/ phoneme is realised as [ɛː], [ɛ] or [ə] with very similar distributions in the key transcript and the model output, as shown in Table J.17.

Table J.17: *The realisations of the phoneme /ɛː/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
ə	52	12.75%	53	12.99%	52	100.00%
ɛ	82	20.10%	73	17.89%	67	81.71%
ɛː	274	67.16%	282	69.12%	268	97.81%
$\sum$	408	100.00%	408	100.00%	387	94.85%

The part of the decision tree handling the /ɛː/ phoneme is relatively complex and includes several duration-based attributes, stress, phoneme context, Part of Speech context and function word context. Figure J.15 shows the details of the rules employed for determining the realisation of /ɛː/.

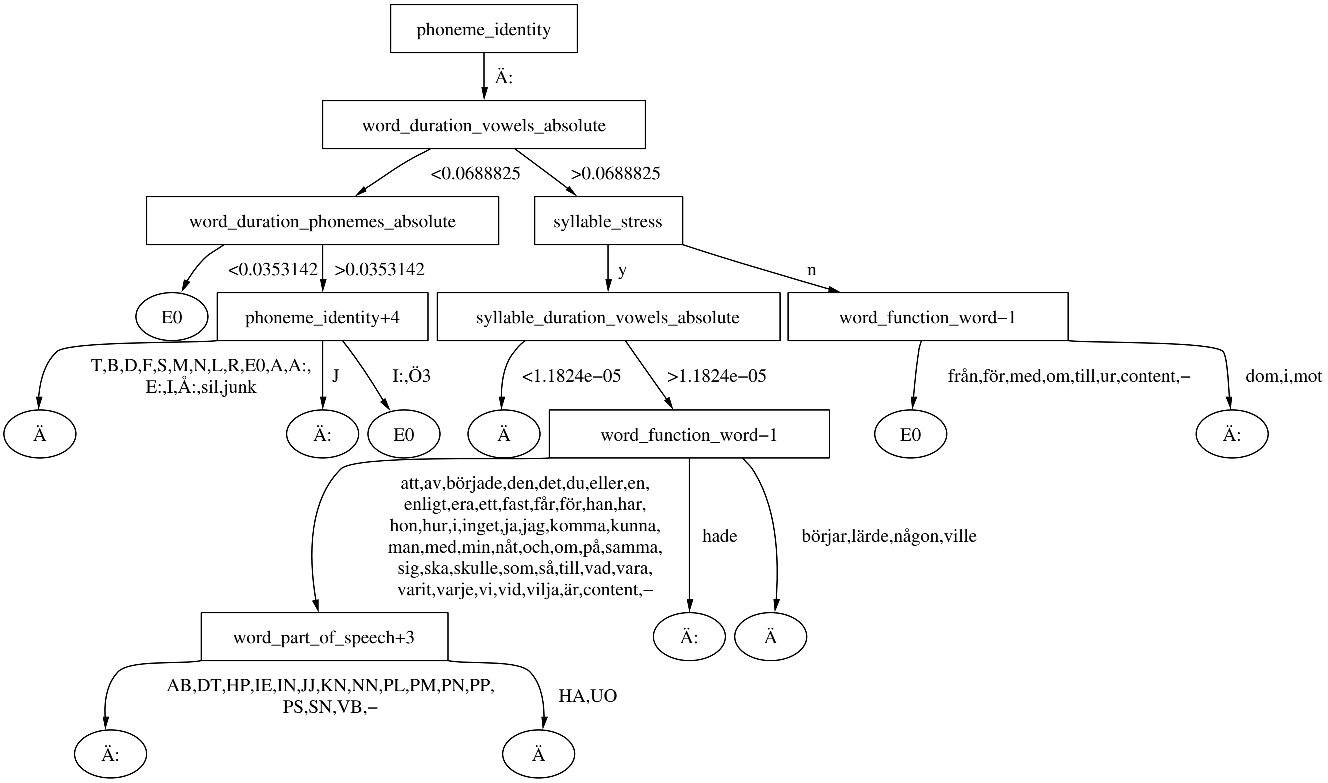

phoneme_identity
Ä:
word_duration_vowels_absolute
<0.0688825
>0.0688825
word_duration_phonemes_absolute
syllable_stress
<0.0353142
>0.0353142
y
n
E0
phoneme_identity+4
syllable_duration_vowels_absolute
word_function_word−1
T,B,D,F,S,M,N,L,R,E0,A,A:,
E:,I,Å:,sil,junk
J
I:,Ö3
<1.1824e−05
>1.1824e−05
från,för,med,om,till,ur,content,−
dom,i,mot
Ä
Ä:
E0
Ä
word_function_word−1
E0
Ä:
att,av,började,den,det,du,eller,en,
enligt,era,ett,fast,får,för,han,har,
hon,hur,i,inget,ja,jag,komma,kunna,
man,med,min,nåt,och,om,på,samma,
sig,ska,skulle,som,så,till,vad,vara,
varit,varje,vi,vid,vilja,är,content,−
hade
börjar,lärde,någon,ville
word_part_of_speech+3
Ä:
Ä
AB,DT,HP,IE,IN,JJ,KN,NN,PL,PM,PN,PP,
PS,SN,VB,−
HA,UO
Ä:
Ä

Figure J.15: *The realisations of the phoneme /ɛː/ (phoneme representations in the figure are in STA format).*

The Realisation of /æ/ and /æː/

The phonemes /æ/ and /æː/ share an arc from the top node of the decision tree model. Because of this, it is possible for the model to realise /æ/ as [æː], as shown in Table J.18, although this realisation of /æ/ never occurs in the key transcript.

Table J.18: *The realisations of the phoneme /æ/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
ə	72	28.69%	47	18.73%	44	61.11%
ɛ	65	25.90%	4	1.59%	4	6.15%
æ	114	45.42%	195	77.69%	108	94.74%
æː	0	0.00%	5	1.99%	0	0.00%
$\sum$	251	100.00%	251	100.00%	156	62.15%

The phonemes were clustered at creation of the first level of the decision tree, since the gain in information was maximised through clustering. The realisations of both /æ/ and /æː/ are distributed over several realisations with no realisation standing out in any extreme way in terms of frequency. The realisation distributions of /æ/ and /æː/, respectively, are too similar to be clearly separable.

One of the reasons for the similarity in realisation distributions is that /æː/ occurs in the very frequent copula verb *är* *'is'*, which is often reduced. The canonical pronunciation of *är*, /æːr/, is thus seldom used and the word is instead realised as [ær], [æː], [æ], [ɛ] or [ə].

Since a greedy algorithm is used for decision tree induction, no consideration is taken to how the creation of a higher level affects the possibility to create descriptive lower levels. In most cases, this is not a problem in practice, given the current training data. However, for the /æ/ and /æː/ phonemes, it would probably have been beneficial for the tree as a whole if the phonemes had not been clustered in the creation of the first tree level, but separated at an early stage of the tree induction process.

Table J.19: *The realisations of the phoneme /æː/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	1	0.07%	2	0.14%	1	100.00%
ə	456	31.91%	427	29.88%	425	93.20%
ɛ	79	5.53%	24	1.68%	23	29.11%
æ	269	18.82%	381	26.66%	244	90.71%
æː	624	43.67%	595	41.64%	562	90.06%
$\sum$	1,429	100.00%	1,429	100.00%	1,255	87.82%

Since the /æː/ phoneme is much more frequent than the /æ/ phoneme, as can be seen from tables J.18 and J.19, the distribution of realisations of /æː/ dominates the distribution of realisations for the /æ/-/æː/ cluster.

The consequence of this is that the share of correct realisation for the less frequent phoneme, /æ/, is very low. In total, the share of correct decisions made for this phoneme is 62.2%, but for the [ɛ] realisation, the share is as low as 6.2%. The high degree of noise in the data given the clustering of /æ/ and /æː/ is also reflected in the complexity of the decision tree model. Figure J.16 illustrates the complex model for /æ/ and /æː/ realisations.

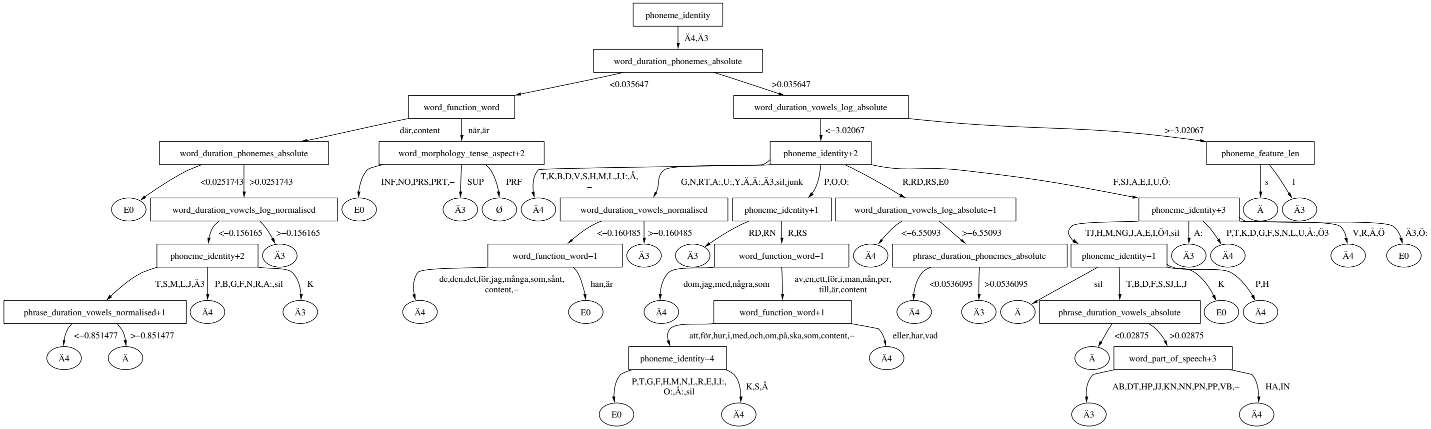


Figure J.16: *The realisations of the phonemes /æ/ and /æː/ (phoneme representations in the figure are in STA format).*

The Realisation of /œ/

The realisation rule for /œ/ is very simple: in stressed syllables, /œ/ is realised as [œ] and in unstressed syllables, the phoneme is elided. The rule is illustrated in Figure J.17. This simple rule seems to work well for the data used for training and evaluating the model, as seen in Table J.20. However, the rule may work less well when applied to new data.

Table J.20: *The realisations of the phoneme /œ/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	11	7.24%	11	7.24%	9	81.82%
ə	1	0.66%	0	0.00%	0	0.00%
œ	140	92.11%	141	92.76%	139	99.29%
$\sum$	152	100.00%	152	100.00%	148	97.37%

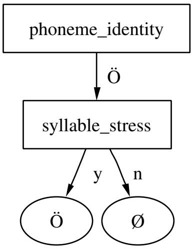


Figure J.17: *The realisations of the phoneme /œ/ (phoneme representations in the figure are in STA format).*

The Realisation of /øː/

An /øː/ can be realised as [øː], [œ], [ə] or ∅, as can be seen in Table J.21. The model depends on duration-based measures and phoneme context to determine the realisation of /øː/ in context, as shown in Figure J.18.

Table J.21: *The realisations of the phoneme /øː/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	1	0.38%	1	0.38%	1	100.00%
ə	3	1.15%	4	1.53%	3	100.00%
œ	29	11.07%	21	8.02%	18	62.07%
øː	229	87.40%	236	90.08%	225	98.25%
$\sum$	262	100.00%	262	100.00%	247	94.27%

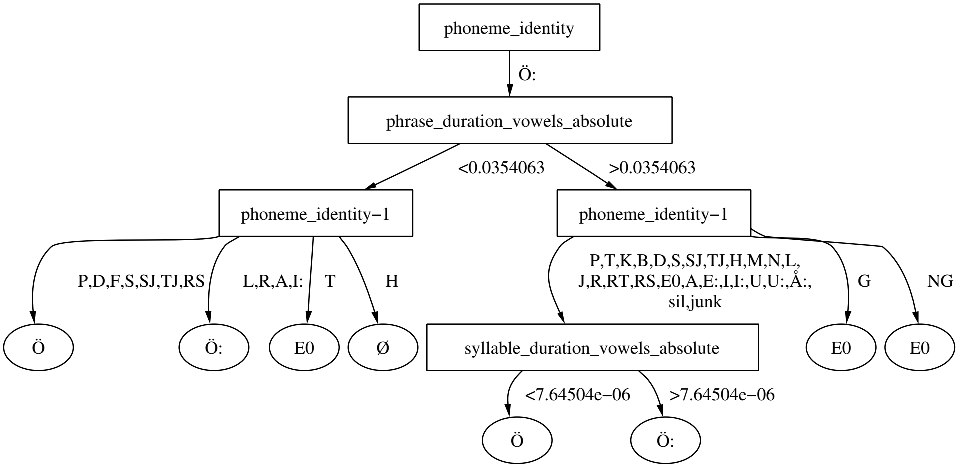


Figure J.18: *The realisations of the phoneme /øː/ (phoneme representations in the figure are in STA format).*

The Realisation of /œ̝/ and /œ̝ː/

Like the /æ/ and /æː/ phonemes, /œ̝/ and /œ̝ː/ share an arc from the top node of the decision tree pronunciation model. The part of the tree model responsible for predicting the realisations of /œ̝/ and /œ̝ː/ show the same kinds of problems as the part of the model responsible for /æ/ and /æː/ realisations, and for similar reasons.

Table J.22: *The realisations of the phoneme /œ̝/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	85	23.88%	0	0.00%	0	0.00%
ə	36	10.11%	4	1.12%	2	5.56%
œ	29	8.15%	0	0.00%	0	0.00%
œ̝	206	57.87%	352	98.88%	204	99.03%
∑	356	100.00%	356	100.00%	206	57.87%

The distributions are hard to separate and the more frequent class dominates the common tree structure. The /œ̝ː/ phoneme occurs in the canonical pronunciation representation of frequent preposition *för 'for'*, which has many realisations, one of the most common being /fœ̝/.

Table J.23: *The realisations of the phoneme /œ̝ː/ in the key transcript and in the decision tree model output, and the correct decisions made by the model.*

	Key transcript		Decision tree model output		Correct model decisions	
Phone	Instances	Share	Instances	Share	Instances	Share
∅	8	1.47%	0	0.00%	0	0.00%
ə	3	0.55%	1	0.18%	1	33.33%
œ	14	2.58%	2	0.37%	1	7.14%
œ̝	243	44.75%	241	44.38%	222	91.36%
œ̝ː	275	50.64%	299	55.06%	268	97.45%
∑	543	100.00%	543	100.00%	492	90.61%

Table J.22 shows the distribution of realisations and the share of correct classifications made by the model for the phoneme /œ̝/ and Table J.23 shows the corresponding statistics for the phoneme /œ̝ː/. In Table J.22, it can be seen that /œ̝/ has the lowest share of correct model decisions of any phoneme. Figure J.19 shows the part of the decision tree responsible for predicting the realisations of /œ̝/ and /œ̝ː/.

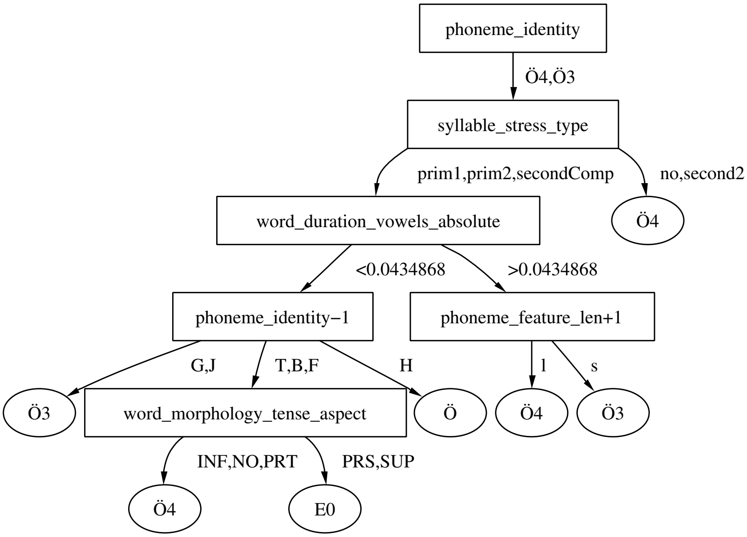


Figure J.19: *The realisations of the phonemes /œ̝/ and /œ̝ː/ (phoneme representations in the figure are in STA format).*