Reports in Phonetics, University of the Saarland
Berichte zur Phonetik, Universität des Saarlandes

PHONUS

Herausgegeben von: W. J. Barry & J. Trouvain

No. 10, Februar 2006

Institut für Phonetik
Universität des Saarlandes
D-66041 Saarbrücken, Februar 2006

ISSN 0949-1791

Foreword

The tenth volume of PHONUS presents a combined Phonetics and Computational Linguistics thesis which was accepted by the Philosophical Faculties of the Universität des Saarlandes in 2005. It demonstrates how speech-technology methods and linguistic-phonetic knowledge can be fruitfully combined in order both to increase our understanding of the prosodic structuring of speech and to improve the performance of speech-technology applications. The author applies machine-learning procedures to the prosodically labelled data of the 'Kiel Corpus of Read Speech' in order to predict the prosodic properties of German read texts. Two different prediction procedures are developed and evaluated in a systematic comparison with the corpus data. In a second evaluation step, 20 sentences are synthesized using both procedures and, together with a third version synthesized using a standard synthesis system, are measured in terms of their acceptability against a copy-synthesis version.

William Barry & Jürgen Trouvain, Saarbrücken, February 2006

Vorwort

Der zehnte Band in der Reihe PHONUS präsentiert eine kombinierte phonetische und computerlinguistische Arbeit, die im Jahre 2005 von den Philosophischen Fakultäten der Universität des Saarlandes angenommen wurde. Die Arbeit zeigt, wie sprachtechnologische Verfahren und linguistisch-phonetische Kenntnisse fruchtbar zusammenwirken können, um sowohl unser Verständnis der prosodischen Strukturierung von gesprochener Sprache voranzubringen als auch die Qualität sprachtechnologischer Anwendungen zu verbessern. Die Verfasserin setzt eine Reihe von maschinellen Lernverfahren ein, um aus den etikettierten Daten des 'Kiel Corpus of Read Speech' prosodische Eigenschaften für gelesene deutsche Texte in zwei unterschiedlichen Verfahren vorherzusagen. Die Vorhersagen werden in einem systematischen Vergleich mit den Korpusdaten getestet. In einem zweiten Schritt werden 20 Sätze gemäß der beiden Verfahren synthetisiert und in einem Perzeptionstest zusammen mit der Ausgabe eines Standardsynthesesystems gegen eine Copy-Synthese-Version gemessen. Die Perzeptionsergebnisse bestätigen die größere Akzeptabilität der von der Kandidatin erarbeiteten Verfahren.

William Barry & Jürgen Trouvain, Saarbrücken, Februar 2006

Improving Prosody Prediction for Speech Synthesis

With and Without Symbolic Prosody Features

Caren Brinckmann

vorgelegt als
Magisterarbeit im Fach Phonetik und Phonologie
Diplomarbeit im Fach Computerlinguistik

zum Thema
"The 'Kiel Corpus of Read Speech' as a Resource for Speech Synthesis"

Dezember 2004

Fachrichtung 4.7 Allgemeine Linguistik
Universität des Saarlandes

Verfasserin:	Caren Brinckmann Försterstr. 50 66111 Saarbrücken caren@brinckmann.de
Betreuer:	Prof. Dr. William J. Barry
Zweitgutachter Magisterarbeit:	PD Dr. Henning Reetz
Zweitgutachterin Diplomarbeit:	Dr. Sabine Schulte im Walde
Erstellungszeitraum:	23. März – 23. Dezember 2004
Webseite:	http://www.brinckmann.de/KaRS/

Abstract

The naturalness of synthetic speech produced by a text-to-speech (TTS) system depends strongly on the prediction of appropriate prosody, i.e. speech rhythm and melody. In many TTS systems the following prediction tasks contribute to the prosodic structure of the generated output: prediction of symbolic prosody features (such as accents and prosodic phrase boundaries), postlexical phonological processes, and acoustic parameters (duration and fundamental frequency F0). This thesis shows how to improve the prosody prediction of the German TTS system MARY, using the German speech database "Kiel Corpus of Read Speech" (KCoRS) comprehensively for all prosody prediction tasks.

The KCoRS comprises over four hours of labelled read speech. The original annotation includes sentence and word boundaries, realised and underlying (lexical) phonemes, orthography, and punctuation marks. The prosodic annotation incorporates the following domains: lexical stress, sentence accent, intonation contour, prosodic phrase boundaries, and pauses.

The original annotation of the KCoRS was extended automatically with the following additional features: sentence type, syntactic phrases, grammatical functions, part-of-speech, word frequency, and syllable boundaries. On this extended database, a set of classification and regression trees (CART) were trained for all prosody prediction tasks.

For the perceptual evaluation of the prediction models, 20 German utterances were each synthesised with MARY using four different prosody prediction methods:

- *copy synthesis*: phoneme, duration and F0 values were extracted from

the KCoRS and copy-synthesised with MARY

- *MARY*: existing MARY system without any modification
- *symbolic*: all trained prosody predictions models were used, *including* prediction of symbolic prosody features (accents, prosodic phrase boundaries, and phrase-final intonation contours)
- *direct*: direct prediction of postlexical processes, duration, and F0 values *without* using symbolic prosody features.

The perceptual evaluation showed that the overall perceptual quality of MARY can be significantly improved by training all models that contribute to prosody prediction on the same database. As expected, the copy-synthesis was perceived as best version. More importantly, it showed that the error introduced by *symbolic* prosody prediction perceptually equals the error produced by the *direct* method that does not exploit any symbolic prosody features. Thus, it can be concluded that the symbolic level of prosody prediction can be safely skipped, and the decision whether or not to include the symbolic prediction can be based entirely on the purpose of the TTS system (e.g. research tool vs. end user software).

Zusammenfassung

Die Prosodiemodellierung, d.h. die Vorhersage von Sprechrhythmus und -melodie, ist ein entscheidender Einflussfaktor für die Natürlichkeit synthetischer Sprache. Die vorliegende Arbeit untersucht die Einsatzmöglichkeiten des 'Kiel Corpus of Read Speech' (KCoRS) für die Prosodiemodellierung in der Sprachsynthese und zeigt, wie die Prosodievorhersage des deutschen Sprachsynthesesystems MARY verbessert werden kann. Dabei wird der Begriff der Prosodiemodellierung weit gefasst und beinhaltet sowohl die Vorhersage symbolischer Prosodiekategorien (Akzente und prosodische Phrasengrenzen), als auch die Modellierung postlexikalischer phonologischer Prozesse und die Vorhersage der akustischen Parameter Lautdauer und Grundfrequenz (F0).

Das KCoRS besteht aus mehr als vier Stunden Lesesprache. Es ist annotiert mit Laut-, Wort- und Satzgrenzen, zugrundeliegenden und tatsächlich realisierten Lauten, Orthographie und Interpunktion. Die prosodische Annotation umfasst lexikalischen Wortakzent, Satzakzent, Intonationskonturen, prosodische Phrasengrenzen und Pausen.

Die bestehende Annotation des KCoRS wurde automatisch mit folgenden Informationen ergänzt: Satztyp, syntaktische Phrasen und grammatische Funktionen, Wortart, Worthäufigkeit und Silbengrenzen. Auf dieser erweiterten Datenbasis wurden mit dem maschinellen Lernalgorithmus CART Klassifikations- und Regressionsbäume für alle Teilaufgaben der Prosodiemodellierung trainiert.

Für die perzeptuelle Evaluation der Prosodievorhersagemodelle wurden mit Hilfe des deutschen Sprachsynthesesystems MARY und den trainierten Klassifikations- und Regressionsbäumen 20 Äußerungen synthetisiert. Jede

Äußerung wurde mit vier verschiedenen Methoden erzeugt, wobei jeweils dieselben diphonbasierte MBROLA-Stimmen verwendet wurden:

- *Copy-Synthese*: Phonemsymbol, Dauer und F0-Werte wurden aus dem KCoRS extrahiert und mit MBROLA in MARY synthetisiert.
- *MARY*: Verwendung des bestehenden MARY Systems ohne Modifikation.
- *Symbolisch*: Verwendung aller trainierten Modelle, *inklusive* der symbolischen Prosodievorhersage von Akzenten, prosodischen Phrasengrenzen und phrasenfinalen Intonationskonturen.
- *Direkt*: Direkte Modellierung der postlexikalischen Prozesse, Lautdauern und F0-Werte *ohne* Verwendung symbolischer Prosodievorhersagemodelle.

Die perzeptuelle Evaluation ergab, dass die Sprachausgabe von MARY durch den Einsatz der automatisch trainierten Modelle signifikant verbessert werden kann. Die Copy-Synthese wurde wie erwartet als die beste Version rezipiert. Ein weiteres wichtiges Ergebnis war, dass sich die Methoden *Symbolisch* und *Direkt* perzeptuell nicht unterscheiden. Je nach Anwendungszweck des Synthesesystems (z.B. als Forschungsinstrument vs. Anwendersoftware) kann also auf die symbolische Prosodievorhersage verzichtet werden.

Acknowledgements

First of all I thank Bill Barry for always supporting me and giving me the freedom to follow my ideas.

I also thank Nick Campbell, who introduced me to the 'Kiel Corpus of Read Speech' as a resource for German speech synthesis, Eric Fosler-Lussier, who first showed me how to train a classification tree myself, and Chris Brew, who suggested using wagon when Weka ran out of memory. I am grateful to Benno Peters for patiently answering all my questions regarding the Kiel Corpus of Read Speech, especially the prosodic annotation.

For helpful discussions about various aspects of my research I thank Jacques Koreman, Marc Schröder, Sabine Schulte im Walde, Bistra Andreeva, and Stefan Baumann. Special thanks to Dominika Oliver, the best office mate ever, with whom I could discuss all the nitty-gritty details of prosodic modelling. I am also deeply indebted to Cordula Klein, Jürgen Trouvain, Anja Moos, and Silke Jarmut for their comments on earlier drafts of this thesis and extensive last minute proofreading. Of course, all remaining errors are mine.

I am very grateful to Gudrun Schuchmann for the reliable organisation of the perception experiment and the friendly supervision of the subjects. I also thank all participants the perception experiment for generously offering their time and energy for a mere bar of chocolate.

I thank all my friends and family for their constant encouragement and emotional support. Special thanks go to my parents, who patiently kept their promise and did not ask me about the state of my thesis until it was finished.

Most importantly, I give my warmest thanks to Martin Kempkes – thank you for always being there when I need you! ♡

Contents

List of Figures

List of Tables

Abbreviations

ACR	Absolute Category Rating
ANOVA	analysis of variance
ASCII	American Standard Code for Information Interchange
ASR	automatic speech recognition
ATR	Advanced Telecommunications Research Institute International
BN	Bayesian Network
CART	classification and regression trees (ML algorithm)
cc	correlation coefficient
CCR	Comparison Category Rating
CELEX	lexical database for Dutch, English and German
CHATR	unit selection speech synthesis system developed by ATR
CMOS	comparison mean opinion score
COS	comparison opinion score
CTS	concept-to-speech
DFKI	Deutsches Forschungszentrum für künstliche Intelligenz
F0	fundamental frequency, usually measured in Hz
GToBI	German Tones and Break Indices
HMM	Hidden Markov Model
HSD	Honestly Significant Difference
IPDS	Institute of Phonetics and Digital Speech Processing, Kiel
ITU-T	International Telecommunication Union – Telecommunication Standardization Sector
KCoRS	Kiel Corpus of Read Speech
KCoSS	Kiel Corpus of Spontaneous Speech
KIM	Kiel Intonation Model
MARS	Multivariate Adaptive Regression Splines

MARY	Modular Architecture for Research on speech sYnthesis
MBROLA	speech synthesiser based on diphone concatenation
ML	machine learning
MOS	mean opinion score
MULI	Multilingual Information Structure
NEGRA	Nebenläufige Grammatische Verarbeitung
NN	neural network
PaIntE	Parametric Intonation Events
PCFG	probabilistic context-free grammar
POS	part of speech
PROLAB	prosodic labelling system based on KIM
RIPPER	Repeated Incremental Pruning to Produce Error Reduction
RMSE	root mean squared error
SAMPA	Speech Assessment Methods Phonetic Alphabet
SCAPE	System for Computer-Aided Perception Experiments
SCHUG	shallow and chunk based unification grammar
SSML	Speech Synthesis Markup Language
stddev	standard deviation
STTS	Stuttgart Tübingen Tag Set
TBL	Transformation Based Learning
TnT	Trigrams'n'Tags
ToBI	Tones and Break Indices
TTS	text-to-speech

Introduction

The first text-to-speech (TTS) systems relied mostly on rules that were hand-crafted by human experts. The construction of these rules was based on introspection, carefully controlled production experiments, and manual inspection of speech corpora. The parameters of these rules were often adjusted through a trial-and-error procedure by listening to synthesised utterances. Some of the first TTS systems were barely intelligible, but even if they generated clearly understandable utterances, they sounded quite monotonous compared to human speech.

For more than a decade, these hand-crafted rules have been successively replaced by models that are automatically trained on annotated corpora with machine learning (ML) methods. For example, a speech corpus that is annotated with information about accent placement can be used to train a model that predicts which words in an utterance carry an accent. These models are usually more complex than the hand-crafted rules, resulting in the output of more varied speech.

The creation of suitable databases has become very important. These databases can be exploited for training models that solve specific prediction tasks. Large annotated speech databases can also be used for non-uniform unit selection synthesis, in which speech segments of different sizes are concatenated to generate natural sounding speech.

The German speech database "Kiel Corpus of Read Speech" (henceforth KCoRS) was chosen for the present study. With only half an hour of speech per speaker, the KCoRS is too small to serve as a reliable speech database for unit-selection synthesis (cf. Brinckmann, 1997). Nevertheless, it can be used for the training of the following TTS modules contributing to prosody predic-

tion: symbolic prediction of accents and prosodic boundaries, prediction of postlexical phonological processes (i.e. pronunciation modelling), and prediction of acoustic parameters (duration and F0 values). For the present study, two diphone-based voices of the German TTS system MARY (Schröder & Trouvain, 2003) were used to generate synthetic speech with the values predicted by the trained models.

An impressively large number of previous studies focussed on the improvement of models for one particular prediction task, e.g. symbolic prosody prediction, duration prediction, or prediction of F0 values. Pronunciation modelling has been almost entirely neglected for speech synthesis applications. Only very few studies use one database comprehensively for all prosody prediction tasks. The evaluation of the automatically trained models was mostly corpus-based, i.e. the predictions of the respective model were compared with the actual realisations in a database. However, formal perceptual evaluation is needed to determine whether the corpus-based improvements are perceptually relevant in a complete TTS system. For example, Brinckmann & Trouvain (2003) showed that the corpus-based differences of two duration prediction models could not be discerned by listeners as soon as the symbolic input to the duration models was not flawless, because it had been generated by a TTS system. Since the ultimate goal in TTS is to improve the overall quality, "TTS quality is still assessed best by human listeners" (Strom, 2002).

Goals

The major goals of this thesis are to show the following:

- The KCoRS can be used for machine learning-based training of prosody prediction models by expanding its original annotation with features that can be derived with pre-existing tools in a reasonable amount of time.
- The overall perceptual quality of the German TTS system MARY can be significantly improved by training all models that contribute to prosody prediction on the same database, namely the KCoRS.
- The error introduced by symbolic prosody prediction perceptually equals the amount of error produced by a direct method which does not exploit any symbolic prosody features.

Outline

In Chapter 1, this thesis starts with a brief introduction to the general architecture of a TTS system, focussing on the German TTS system MARY. The next section describes the core concepts and methods in machine learning, explaining a particular machine learning algorithm, CART, which was used to train classification and regression trees for prosody prediction. The selective summary of previous studies illustrates the diversity of machine learning methods that have been applied to prosody prediction tasks. Finally, the first chapter concludes with remarks on error accumulation within a TTS system and outlines two approaches to reduce it.

Chapter 2 motivates the choice of the KCoRS as a database for prosody prediction. It gives a detailed description of the original annotation in the KCoRS and explains the features that were added semi-automatically with pre-existing tools and tailored Perl programs. It concludes with some remarks on the limitations of the KCoRS and further possibilities.

Chapter 3 describes the methods that were applied to train classification and regression trees for the following prosody prediction tasks: prediction of prosodic boundaries, accent location and type, phrase-final intonation contours, postlexical phonological processes, duration, and F0 values. Two types of prediction models were trained: The first one, called *Symbolic*, uses symbolic prosody features for the prediction of segmental features (i.e. realised phoneme, duration, and F0). The second one, called *Direct*, is a method which predicts the segmental features directly without using any symbolic prosody features. The predictions of all models were evaluated by comparing

them to the actual realisations in the KCoRS.

Chapter 4 explains the perception experiment that was carried out to evaluate the predictions of the automatically trained models perceptually. The results of the perceptual evaluation show that the output of the German TTS system MARY can be significantly improved by training all models that contribute to prosody prediction on the KCoRS. More importantly, they show that the error introduced by the level of symbolic prosody prediction perceptually equals the amount of error produced by the direct method that does not exploit any symbolic prosody features.

This thesis concludes with an outlook on future directions in speech synthesis.

1. Fundamentals

1.1. Text-to-Speech Synthesis

Speech synthesis can be defined as the automatic transformation of a symbolic representation into an acoustic signal that sounds similar to human speech (Zboril, 1997). Two concepts have to be distinguished:

1. A **speech synthesis system** produces speech from written text (text-to-speech: TTS) or a conceptual representation (concept-to-speech: CTS).

2. A **speech synthesiser** produces speech from a representation of control parameters. The speech synthesiser is usually the last module of a speech synthesis system.

MARY (Schröder & Trouvain, 2003), the TTS system utilised for this study, uses the speech synthesiser MBROLA (Dutoit et al., 1996). The architecture of the German MARY system, which is shown in Figure 1.1, can be regarded as a typical TTS architecture (cf. Dutoit, 1997). MARY accepts plain text as input and is also able to parse speech synthesis markup such as SABLE (Sproat et al., 1998) and SSML[1].

Due to the modular architecture, single modules can be replaced easily. An interface[2] allows the user to control each processing step and to change the input to each module manually. All MARY modules are described in the following sections (see Schröder, 2004, for further details). Except for the

[1] http://www.w3.org/TR/speech-synthesis/
[2] http://mary.dfki.de

part-of-speech tagger and the chunk tagger, all modules within MARY are realised with hand-crafted rules.

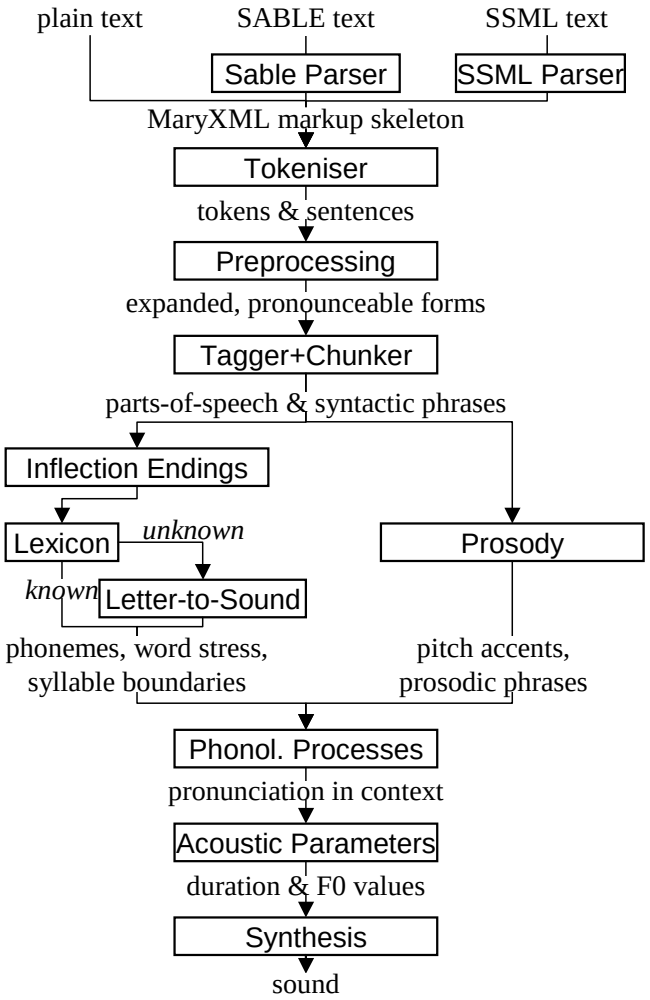

Figure 1.1.: Architecture of the MARY TTS system (from Schröder, 2004).

1.1.1. Preprocessing

Tokeniser

As a first step, the text is cut into separate tokens, namely words, numbers, special characters, and punctuation marks. MARY uses a set of hand-crafted rules to disambiguate periods into sentence-final periods, decimal number delimiters, and parts of ordinal numbers or abbreviations.

Text normalisation

The text normalisation module (termed "Preprocessing" in Figure 1.1) converts numbers and abbreviations into pronounceable forms.

1.1.2. Natural Language Processing

Part-of-Speech Tagging

Part-of-speech (POS) tagging is carried out with the statistical tagger TnT (Brants, 2000). The German language model of TnT was trained on the annotated NEGRA corpus (Brants et al., 1999) using the Stuttgart-Tübingen tag set (STTS, see Appendix B.1; Schiller et al., 1995). TnT uses second order Markov models, where the states represent tags and the outputs represent words. Smoothing is carried out with context-independent linear interpolation of unigrams, bigrams, and trigrams. Unknown words are handled by suffix analysis, where tag probabilities are set according to the word's final sequence of characters, with different estimates for uppercase and lowercase words.

Chunk Tagging

The chunk tagger described by Skut & Brants (1998) is used to recognise syntactic structures of limited depth ("chunk phrases"), namely the phrasal categories used in the NEGRA corpus. The chunk tagger uses a generalised

Markov Model-based tagging method based on the part-of-speech information provided by TnT and simple morphological information.

Grapheme-to-Phoneme Conversion

MARY uses the phonetic alphabet SAMPA[3] for German (Wells, 2004) for the phonemic transcription, adding also lexical stress and syllable boundaries. First, inflection endings are added to ordinals and abbreviations by a unification-based module. Second, the word is looked up in a lexicon derived from CELEX (Baayen et al., 1995). If needed, a simple compound treatment is performed. Unknown words, which cannot be phonemised by lexical lookup or compound treatment, are analysed by grapheme-to-phoneme rules, using a statistical morphological parser, syllabification rules, and lexical stress assignment rules. The resulting transcription represents the canonic pronunciation, i.e. it does not contain any segmental reductions.

Symbolic Prosody Prediction

The "Prosody" module assigns symbolic GToBI labels (Grice et al., 2005). GToBI[4] is an adaptation of ToBI (Tones and Break Indices; Silverman et al., 1992) for German, which describes the perceived intonation contour in terms of high and low tonal targets. Break indices are used to mark prosodic boundaries of intermediate phrases (break index 3) and intonation phrases (break index 4). All tonal targets must be related to either an accented syllable (accents) or the edge of a prosodic phrase (edge or boundary tones). GToBI accents are either simple tonal targets (`H*` and `L*`) or complex accents (`L+H*`, `L*+H`, `H+L*`, and `H+!H*`; `H` and `L` relate to high and low targets, and `*` is used to mark the tone of the accented syllable). GToBI boundary tones also include complex tones.

[3] Throughout this text, all transcriptions are given in SAMPA notation. For ease of reading all pronunciations are given between slashes (e.g. `/a:/`), irrespective of phonemic or phonetic status.

[4] `http://www.coli.uni-sb.de/phonetik/projects/Tobi/gtobi.php3`

MARY's hand-crafted prosody rules were derived through manual corpus analysis and are mostly based on part-of-speech and punctuation information. Intermediate and intonation phrase breaks are inserted at punctuation marks and at certain chunk phrase boundaries. Some parts-of-speech (e.g. nouns and adjectives) always receive an accent, others are only accented if the respective intermediate phrase contains no noun or adjective. The actual GToBI accents and boundary tones are assigned according to sentence type (statement, wh-question, yes/no-question, and exclamation) and position of the accent within the prosodic phrase.

Postlexical Phonological Processes

Once the prosodic boundaries, accents, and boundary tones are determined, the canonic pronunciation can be changed by postlexical phonological processes (cf. Kohler, 1990). These processes restructure the utterance on the segmental level as well as on the prosodic level. Examples of postlexical processes include

- segmental deletions and replacements, e.g. *haben* is pronounced as `/ha:b=m/`
- vowel reductions, e.g. *der* is pronounced as `/d@/`
- reducing the number of accents and phrase boundaries for fast speech.

Currently, MARY applies no postlexical rules. The models trained for the prediction of postlexical processes (see Section 3.3.2) deal with segmental changes only.

1.1.3. Calculation of Acoustic Parameters

MARY uses the MBROLA diphone synthesiser for synthesising the utterances. MBROLA processes a list containing the following information:

- phoneme in SAMPA

- duration in ms
- fundamental frequency (F0) targets in Hz.

An example of the MBROLA input format within MARY is given in Figure 1.2. After each phoneme, its duration is listed. The F0 values are given as pairs (*relative time in %, F0 in Hz*). For example, the first phoneme `/h/` in Figure 1.2, has a duration of 72 ms, and an F0 target value of 189 Hz at the very beginning of the phoneme. Phoneme `/E/` in the example even carries two F0 target values: The first one (204 Hz) is reached in the middle of the phoneme (50%), the second one (150 Hz) is reached at its end (100%). Intensity and spectral quality of the phonemes cannot be controlled with MBROLA.

```
h 72 (0,189)
a 72 (87,167)
l 63
o: 121 (50,205)
v 67
E 162 (50,204) (100,150)
l 55
t 66
_ 410
#
```

Figure 1.2.: Example of the MBROLA input format.

Duration and F0 values are predicted by the module "Acoustic Parameters" from the symbolic output of the preceding modules.

Duration Prediction

The duration of a sound segment depends on a variety of linguistic, pragmatic and phonetic factors (cf. Kohler, 1992b), e.g.:

- global speech tempo
- semantically important parts of an utterance are produced more slowly

- stress and accentuation: stressed syllables are longer than unstressed ones
- final lengthening at the end of a prosodic phrase
- a stressed syllable is shorter if it is followed by one or more unstressed syllables within the same word
- phonological quantity: phonologically long segments (tense) are longer than phonologically short segments (lax)
- phonetic context, e.g. segmental duration before fortis/lenis
- intrinsic segmental duration: high vs. low vowels, plosives vs. fricatives, fortis vs. lenis obstruents.

The duration rules currently implemented in MARY are a version of the Klatt rules (Klatt, 1979) adapted to German (Brinckmann & Trouvain, 2003). Klatt rules predict the segmental duration by multiplying the intrinsic duration of a given phoneme with a context-dependent factor. The result is then added to a phoneme-specific minimal duration, which can also be multiplied by a context-dependent factor. The adaptation of the context-dependent factor values to German was achieved by a manual trial-and-error procedure.

F0 Prediction

Rules to transform abstract ToBI labels into fundamental frequency (F0) values were described by Anderson et al. (1984) for English. For each prosodic phrase an F0 topline and an F0 baseline are assumed, both descending over the course of the utterance. `H` targets lie on the topline, whereas `L` targets are positioned on the baseline. Topline and baseline can be varied, e.g. according to the sex of the speaker or the sentence type (cf. Brinckmann & Benzmüller, 1999). Because of the declination of both lines, the F0 value of a phoneme in an accented syllable depends on the position of the syllable in the prosodic phrase.

1.1.4. Synthesis

MBROLA (Dutoit et al., 1996) is a speech synthesiser based on the concatenation of diphones. It takes a list of phonemes as input, together with prosodic information (duration of phonemes and F0 values), and produces speech samples at the sampling frequency of the diphone database used. The original F0 values of the diphones in the database are transformed by a time-domain algorithm with diphone smoothing capabilities. In this study, MARY's MBROLA diphone databases `de6` (male) and `de7` (female) are used for synthesis (see Section 4.1.2).

1.2. Machine Learning

Machine learning (ML) is an area of artificial intelligence concerned with the development of techniques which allow computers to "learn" through experience by finding and describing structural patterns in data. Machine learning methods take training data and form hypotheses or models that can be used to make predictions about novel data.

A training dataset consists of several instances, i.e. representations of objects. Instances are described by feature vectors. Features can be categorical (having a finite number of discrete values) or continuous (numeric).

Machine learning methods can be applied to the following tasks:

- classification: learn to put instances into pre-defined classes
- numeric prediction: learn to predict a numeric quantity instead of a class
- association: learn relationships between features
- clustering: discover classes of instances that belong together.

The TTS modules described in Section 1.1 solve classification tasks (part-of-speech tagging, chunking, symbolic prosody prediction, postlexical phonological processes) and numeric prediction tasks (calculation of acoustic parameters). The machine learning algorithm CART (Classification and Regression Trees; Breiman et al., 1984) can be applied to classification tasks (training of classification trees) as well as numeric prediction (training of regression trees). CART was used for all tasks relating to prosody prediction described in Chapter 3.

Machine learning algorithms can be divided into supervised and unsupervised methods. Supervised methods are used to learn the relationship between independent features and a designated dependent feature. Classification and numeric prediction algorithms are supervised methods. Unsupervised learning techniques group the instances of the training data without a pre-specified dependent feature. Clustering algorithms are usually unsupervised. Nevertheless, even for unsupervised methods human intuition cannot be entirely eliminated, because the designer of the task must specify how the data are to be represented and what mechanisms will be used to search for a characterization of the data.

1.2.1. Evaluation

When evaluating machine learning models there are some basic procedures to follow.

1. The dataset is divided into a (bigger) training set and a (smaller) test set. The training set is used to train the model, whereas the test set is used for evaluation only.

2. If the ML algorithm needs an additional dataset for a procedure against overfitting (e.g. pruning in CART, see Section 1.2.2), a three-fold division into training set, validation set, and test set is needed. The validation set is used (for pruning) during the training process.

3. Since annotated databases are very time-consuming to produce, one does not want to "waste" precious data for testing. The solution to this

dilemma is *k-fold cross-validation*: The corpus is divided into k mutually exclusive subsets (the "folds") of approximately equal size. The model is trained and tested k times. Each time it is trained on the dataset minus a fold and tested on that fold. The accuracy estimate is the average accuracy for the k folds. *Stratified* cross-validation ensures that each class is properly represented in the respective training and test sets. After evaluation, the final model for implementation is trained on the complete dataset.

Different performance metrics that can be used for evaluation are described in the following section.

Performance Metrics

Classification and numeric prediction are evaluated with different performance metrics. Confusion matrix, accuracy, recall, precision, and F-measure are used for the evaluation of classification models. Root mean squared error (RMSE) and correlation coefficient (cc) are used for evaluating numeric prediction models.

Confusion matrix A confusion matrix is a matrix showing the predicted and actual classifications. A confusion matrix is of size $L \times L$, where L is the number of different class values. The confusion matrix in Table 1.1 is for $L = 2$.

	predicted	
actual	positive	negative
positive	a	b
negative	c	d

Table 1.1.: Two-by-two confusion matrix for a class with 2 different values (*positive* and *negative*).

Accuracy Accuracy is defined as the rate of correct predictions made by the model on a test set (usually given in %). Using the variable names from Table 1.1, the formula for accuracy is: $(a + d)/(a + b + c + d)$.

Precision and recall If the values of the predicted class are not evenly distributed, precision and recall of each class value are more informative than overall accuracy:

- precision of class value "positive" $= a/(a + c)$
- recall of class value "positive"$= a/(a + b)$
- precision of class value "negative" $= d/(b + d)$
- recall of class value "negative"$= d/(c + d)$

If just one precision value is reported, it is usually the precision of the "positive" value (e.g. "boundary" in case of prosodic boundary prediction).

F-measure Precision and recall are combined in the F-measure:
F-measure $= (2 \times recall \times precision)/(recall + precision)$.

RMSE The root mean squared error is used for the evaluation of numeric predictions: $RMSE = \sqrt{\frac{\Sigma(predicted-actual)^2}{n}}$
RMSE is similar to the mean absolute error, but tends to exaggerate the effect of outliers.

Correlation coefficient Correlation determines the extent to which the actual and the predicted values are linearly related to each other. The value of correlation, the correlation coefficient, does not depend on the specific measurement units used. For example, if the predicted values are all multiplied with 100, the correlation with the actual values remains the same. Therefore, RMSE is usually reported in addition to the correlation coefficient.

1.2.2. CART

CART (Breiman et al., 1984) is a machine learning algorithm for automatically building classification and regression trees. Classification trees predict categorical features, while regression trees are used to predict numeric features.

Classification and regression trees contain a question about some feature at each node in the tree. The leaves of the tree contain the best prediction based on the training data, usually a single member of the predicted categorical feature (classification) or a predicted mean value (numeric prediction).

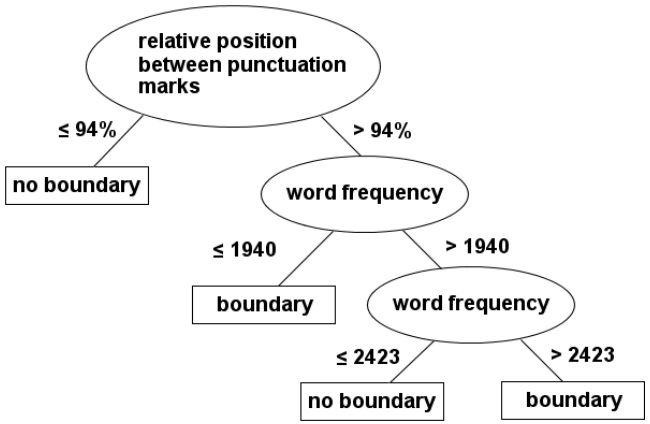

Figure 1.3.: Classification tree example (simplified from Figure 3.1). Nodes are marked with ellipses, leafs are presented in rectangles.

For example, the classification tree in Figure 1.3 can be used for the prediction of prosodic boundaries, i.e. it predicts whether a word is followed by a prosodic boundary or not. The root node (the topmost node) partitions the data according to the feature "relative position between punctuation marks". If an instance has a value $\leq 94\%$ for that feature, i.e. if it is not directly followed by a punctuation mark (that would result in a value of 100%), a leaf is reached, and the classification tree predicts that the respective word is not followed by a prosodic boundary. If the relative position between punctuation marks is $> 94\%$, the next node further down the tree concerns the feature "word frequency". If the word frequency of an instance is ≤ 1940 (the word frequency feature is explained in Section 2.3.1), another leaf is reached, and the tree predicts that the respective word is followed by a prosodic boundary. If the word frequency is > 1940, another question concerning word frequency has to be answered. The next node partitions the data into instances with a word frequency value ≤ 2423 and those with a value > 2423. The former instances are predicted to be followed by no prosodic boundary, whereas the latter ones receive the predicted value "boundary".

CART is a powerful machine learning algorithm because it

- permits both categorical and continuous features (as input features and predicted features)
- automatically selects the most significant features (but see Section 1.2.3)
- allows human interpretation of the result (up to a certain extent).

The basic CART building algorithm starts with the complete training set and determines the feature that splits the data minimising the mean "impurity" of the partitions. This splitting procedure is applied recursively on each partition of the data until some stop criterion is reached (e.g. a minimal number of instances in the partition). Since it chooses the locally best discriminatory feature at each stage in the process, CART is a greedy algorithm. This is suboptimal but a full search for a fully optimised set of questions would result in a very high computational cost. Because of the stepwise partitioning of the data, the size of the dataset that is considered at each node becomes smaller and smaller down the tree. Therefore, data sparsity can be a serious problem for CART, if the gaps in the training data are accidental rather than systematic.

Standard impurity measures are

- for categorical features: *entropy* $\times$ number of instances

- for continuous features: *variance* × number of instances.[5]

A very basic form of the tree building algorithm would lead to a fully exhaustive classification of all instances in the training set, and the resulting tree would *overfit* the data. A method to build trees that are more suitable to make the right predictions for new, unseen data is called *pruning*. This method holds out a portion of the training data (the *validation* set). The trained tree is pruned back until evaluation on the validation set does not improve any further.

Tools

The following software tools were used for the training of classification and regression trees for prosody prediction (as described in Chapter 3):

- **Weka** (Witten & Frank, 2000), version 3.4.2 with Java SDK 1.5.0. Weka is a collection of machine learning algorithms and contains tools for data pre-processing, classification, regression, clustering, association rules, and visualisation. Weka is open source software implemented in Java.
- **wagon** (King et al., 2003), version 1.2.3. wagon is an executable C/C++ program, part of the Edinburgh Speech Tools Library.

The CART algorithm implemented in Weka allows multiply branching nodes, whereas wagon trains only binary branching trees. Weka has been developed as an instructional tool for machine learning algorithms. Therefore, and because of the implementation in Java, the CART algorithm in Weka is comparatively slow and very memory-intensive. For the prosody prediction models described in Chapter 3, Weka was used for the training of classification trees, whereas regression trees were trained with wagon.

The tailored program that was used to extract the information from the database in the necessary format was written in Perl (Wall et al., 2000). Perl was also used to implement a prototype that incorporates the trained

[5] Entropy or variance alone would favour overly small partitions.

classification and regression trees for prosody prediction.

1.2.3. Feature Selection

In theory, most machine learning algorithms learn automatically which are the most appropriate features to make their predictions. For example, CART should never select irrelevant features, so that adding more features should only lead to better classification performance, never to worse results. However, John (1997) reported that classification accuracy of the CART algorithm deteriorates (typically by 5% to 10%) when a random binary feature is added to standard datasets. Even more surprisingly, sometimes the inclusion of highly relevant features can also diminish the classification accuracy (by 1% to 5% in the situations tested). Naive Bayes, another classification algorithm, assumes that all features are independent of each other. Therefore it robustly ignores irrelevant features, but its classification accuracy is damaged heavily when redundant features are added.

Since most machine learning algorithms are negatively affected by irrelevant or redundant features, it is important to precede training with a feature selection stage that selects only the most relevant features for the prediction task. "The best way to select relevant attributes is manually, based on a deep understanding of the learning problem and what the attributes actually mean. However, automatic methods can also be useful." (Witten & Frank, 2000).

Filters and Wrappers Automatic feature selection methods can be divided into filter methods and wrapper methods. Filter methods select the best features according to a reasonable criterion that is independent of the task. For example, a filter can select those features that are most linearly correlated to the target class. Wrapper methods apply a chosen machine learning algorithm (e.g. CART) to every subset of features. The best subset is the one with the best evaluation measures.

Greedy Search Since the number of possible feature subsets increases exponentially with the number of features, exhaustive search is impractical in most cases. Therefore the feature space is searched greedily, either starting with an empty feature set and adding one feature at a time (*forward selection*), or starting with the complete feature set and deleting features one at a time (*backward elimination*). The greedy search stops if the performance of the trained model does not increase anymore (or some other stopping criterion is reached). Forward selection usually results in smaller feature subsets than backward elimination.

Complexity The CART algorithms implemented in Weka and wagon both allow using a feature selection wrapper. Wrappers are potentially very time consuming, because the machine learning algorithm is carried out numerous times. The number of classification or regression trees that are trained during feature selection depends on the number of features in the original feature set (m) and the number of selected features (k). The forward selection wrapper starts out with testing each feature, thus building m trees. The feature that was used for building the best tree is retained, so that in the next step $m-1$ trees are built, and so on until the feature selection stops, because the performance of the trees does not increase anymore. At that point, $(2m-k)(k+1)/2$ trees have been built. In the worst case ($k=m$), the number of trees to be built during feature selection is quadratic to the size of the original feature set ($\mathrm{O}(m^2)$).

The time needed to build a single tree depends on the number of instances in the dataset (n) and the size of the feature set (m). The computational cost of the CART tree induction algorithm (including pruning) is $\mathrm{O}(mn \log n) + \mathrm{O}(n (\log n)^2)$ (Witten & Frank, 2000). The smallest dataset used for the training of prosody prediction models (see Section 3.2.1) consisted of 4750 instances with 52 features (word-level prosodic boundary prediction), whereas the largest dataset consisted of 22094 instances with 83 features (phoneme-level duration prediction). Thus, wrapper-based automatic feature selection was only feasible in a reasonable amount of time for prediction tasks on word or syllable level (i.e. symbolic prosody prediction). The phoneme-level classification and regression trees were trained without prior automatic feature selection.

1.3. Prosody Prediction with Machine Learning Methods

In this thesis, the term "prosody prediction" is defined rather broadly as the group of all prediction models that contribute to the rhythm and the melody of a synthesised utterance. More precisely, it includes all prediction models from symbolic prosody prediction, over the prediction of postlexical phonological processes to the prediction of acoustic parameters. The prediction of acoustic parameters is limited to the prediction of duration and F0 values, because these are the only two parameters that can be controlled for each phoneme using the MBROLA synthesiser. Other acoustic parameters that contribute to the perception of rhythm are intensity and spectral characteristics (e.g. steeper spectral tilt for reduced vowels).

Various machine learning algorithms have been applied to different prosody prediction tasks. Unless two algorithms are applied to the same dataset, the reported results are hard to compare because of the idiosyncrasies of the different datasets used for training. Nevertheless, the reported evaluation measures illustrate the difficulty of the respective task.

1.3.1. Prediction of Prosodic Boundaries

Fordyce & Ostendorf (1998) used transformation-based learning (TBL) and classification trees (CART) for the prediction of prosodic boundary locations. TBL is a supervised machine learning formalism introduced by Brill (1995) for part-of-speech tagging. It finds an ordered sequence of rules which successively change an initial classification of the data. These rules are chosen by

a greedy search over the entire corpus to minimise the overall classification error. Both TBL and CART were trained on the Boston University Radio News Corpus (Ostendorf et al., 1995). In terms of accuracy, the classification tree slightly outperformed TBL (84.1% vs. 82.6%).

Atterer & Schulte im Walde (2004) developed a relatively simple probabilistic context-free grammar (PCFG, cf. Manning & Schütze, 2001, ch. 11) for assigning intonation phrase boundaries to German text using STTS part-of-speech tags. To determine the probabilities of the grammar rules, the PCFG was trained in four iterations on 6,000 words (380 sentences) of the IMS Radio News Corpus (Rapp, 1998). The PCFG was compared with an approach based on Hidden Markov Models (HMMs; similar to Taylor & Black, 1998) using a window of POS-bigrams and a context length of 6. Evaluation showed that the PCFG was inferior to the HMMs (F-measure: 0.741 vs. 0.843).

Fackrell et al. (1999, 2001) used classification trees (CART) and two-layer neural networks (NN) to predict prosodic phrase boundary strength between words, values ranging from 0 to 3. Both were trained on databases of six different languages (Dutch, English, French, German, Italian, and Spanish). The evaluation measures over all languages showed that both methods performed equally well. The accuracy rates for the German database were 74.8% (NN) and 72.7% (CART).

Zervas et al. (2003) used CART, Naive Bayes and a Bayesian Network to predict prosodic boundary locations in a corpus of Modern Greek. CART (F-measure 0.608) and Naive Bayes (0.629) were outperformed by the Bayesian Network (0.704).

1.3.2. Accent Prediction

Fordyce & Ostendorf (1998) also used transformation-based learning (TBL) and classification trees (CART) for the prediction of pitch accent locations in the Boston University Radio News Corpus. For accent prediction, TBL

outperformed CART (accuracy: 86.8% vs. 85.6%).

Fackrell et al. (1999, 2001) used regression trees (CART) and two-layer neural networks (NN) to predict word prominence, values ranging from 0 to 9. Evaluation on databases of six different languages showed that CART performs slightly better than NN. The accuracy rates (exact classification +/-1) for German are 74.5% (NN) and 74.8% (CART).

Hirschberg & Rambow (2001) used a propositional rule learner, RIPPER (Cohen, 1995) to predict pitch accent locations (i.e. whether a word carried an accent or not). The model is expressed as an ordered set of if-then-rules (i.e. each rule only applies if the preceding ones do not) which contain each a conjunction of conditions and a consequent classification. RIPPER was trained on a corpus of read Wall Street Journal texts, which were transcribed and annotated with ToBI labels. The best feature set used for training led to an F-measure of 0.903.

1.3.3. Postlexical Phonological Processes

Most studies dealing with pronunciation variation are concerned with automatic speech recognition (ASR). Some synthesis-related studies used pronunciation modelling for improved labelling of large databases for unit-selection speech synthesis (Bennett & Black, 2003; Jilka & Syrdal, 2002; Breuer, 2000). Whenever possible, these databases are labelled automatically, so that an accurate pronunciation prediction is important. Otherwise the realised phonemes are always labelled with their canonic counterparts, not taking into account any reductions.

Hoste et al. (2000) used TBL and CART to extract phonemic knowledge and rules from pairs of pronunciation lexicons for Northern Dutch and Flemish. The motivation was to adapt speech synthesis systems to regional variants. The overall accuracy in predicting the pronunciation of a Flemish word pronunciation from the Dutch pronunciation was 89% for TBL and 92% for CART.

Miller (1998) inferred individual postlexical phonologies from labelled corpora of read American English using a recurrent neural network. The main postlexical phonological processes to be modelled were glottalisation, vowel reductions and the reduced realisation of /t/ (e.g. as flap). The highest accuracy reached was 89.6%.

1.3.4. Duration Prediction

One of the first machine learning techniques that was applied to duration prediction is CART (Riley, 1992). The regression trees trained by Brinckmann & Trouvain (2003) reached an RMSE of 22.46 ms (male voice) and 21.40 ms (female voice), performing significantly better than the tested Klatt rules. Nonetheless, this difference was not perceptible once the duration prediction models were implemented in MARY.

Since data sparsity can pose a problem for CART, other machine learning techniques have been suggested for duration prediction. Möbius & van Santen (1996) applied a sums-of-products model (a supervised, data-driven approach) to the Kiel Corpus of Read Speech. The overall correlation between observed and predicted durations is 0.896. Riedi (1997) used Multivariate Adaptive Regression Splines (MARS) to predict segmental durations from a corpus of German read speech. The resulting model has a correlation coefficient of 0.90.

Goubanova & Taylor (2000) compared a Bayesian Network (BN) to CART and to a sums-of-products model. All three models were trained on a database of American English read speech. BN achieved a RMSE of 5 ms, outperforming both CART (20 ms) and the sums-of-products model (9 ms).

1.3.5. F0 Prediction

Black & Hunt (1996) predicted three F0 values for every syllable with linear regression models, using features representing ToBI labels, lexical stress and

syllable position. The linear regression models were trained on the Boston University Radio News Corpus (Ostendorf et al., 1995). The F0 contours generated by this method have a correlation coefficient of 0.62 and 34.8 Hz RMSE when compared with the original realisations, whereas a previous rule-driven method (Anderson et al., 1984) resulted in a correlation coefficient of 0.40 and 44.7 Hz RMSE.

The F0 prediction described by Dusterhoff & Black (1997) used CART to predict parameterised descriptions of the F0 contour using the Tilt intonation model (Taylor & Black, 1994). Evaluation on the Boston University Radio News Corpus resulted in a correlation coefficient of 0.60 and 32.5 Hz RMSE.

Syrdal et al. (1998) compared three different F0 prediction methods, namely one primarily rule-based approach and two data-driven approaches, on a corpus of read prompts and Wall Street Journal texts. The rule-based approach was based on manually corrected ToBI labels, the two data-driven approaches used parameterised descriptions of the F0 contour with Tilt or PaIntE parameters (Parametric Intonation Events; Möhler & Conkie, 1998). All methods were compared in a formal listening test. PaIntE received the highest mean opinion scores (on a 5-point scale), followed by the rule-based approach and the Tilt method, which received the lowest scores.

1.4. Error Accumulation

As can be seen in Figure 1.1, a TTS system consists of several modules. All those modules make predictions that are not 100% perfect. Whenever one module makes an error, the modules that follow further down the processing chain "inherit" this error. If their predictions depend on a feature that was predicted incorrectly, they are likely to produce a follow-up error. For example, if the part-of-speech tagger predicts that a word is a content word rather than a function word, the symbolic prosody prediction will probably put an accent on that word, even though it should not be accented.

Automatic training of statistical models is usually carried out on corpora

that have been labelled semi-automatically, i.e. where the annotations were checked manually. Thus, the annotations are near-perfect. Therefore, the statistical models that were trained on perfect data make their predictions based on the assumption that their input is perfect. When these models are then implemented into a TTS system, they will most certainly get input that contains some errors. Some of these errors will have no further effect, some will lead to follow-up errors. Two methods aiming at reducing error accumulation in a TTS system are explained in the following sections: The first one uses only automatically predicted features during training, the second one predicts the acoustic parameters directly without using any intermediate symbolic prosody features.

1.4.1. Training on Automatically Predicted Features

The first method uses the same tools and models that are implemented in the respective TTS system to label the training data. For example, for the training of the symbolic prosody prediction model, the automatic predictions of the POS tagger and the chunk tagger are used without any manual corrections. In addition, the newly-trained prosodic boundary prediction model is used to relabel the training data, i.e. whenever the model predicts a prosodic boundary, this is annotated in the database. The accent prediction model is then trained on this partly erroneously labelled database. The predictions of the accent model are in turn used to re-label the database with accent information for the training of duration and F0 prediction models.

Training the models on automatically predicted features has the following advantage: Since the models are trained on erroneous data, they can "learn" to make right predictions from erroneous input (as long as the errors are not random). When implemented in a complete TTS system, the predictions for duration and F0 might be better than those from models that were trained on perfect data.

Fordyce & Ostendorf (1998) compared two models for accent prediction:

The first model was trained on a database containing manually corrected prosodic boundaries, the second model was trained on automatically predicted boundaries. The accuracy of the first model deteriorated from 86.8% to 86.3% when it received automatically predicted features as input, whereas the second model reached an accuracy of 86.7% on automatically predicted features. They concluded that most of the loss in accuracy can be regained by retraining the accent prediction model on automatically predicted features.

In a forced-preference comparison listening test, Fackrell et al. (1999) compared the following two methods for the prediction of duration and F0: The first method (called MAN) used models that were trained on a manually corrected database, whereas the second method (called AUT) used only automatically labelled training data. Both methods were compared with each other, as well as to copy-synthesised utterances (i.e. duration and F0 values were copied from a recording) and a pre-existing TTS system. Fackrell et al. (1999) found that the difference between MAN and AUT is not significant, and that the copy-synthesised originals are significantly better than MAN and the pre-existing TTS system.

Both studies suggest that prediction models can be trained on automatically predicted features without resulting in a deteriorated performance. However, there are two potential problems to be addressed:

1. The models are trained on data containing very system-specific errors. Whenever a model further up the processing chain is changed, *all* models that depend on its output have to be retrained. In contrast, a model that is trained on manually corrected data can be applied more generally.

2. If the TTS system is not used as a "black box", but rather as an instructional or research tool (such as MARY), the user is able to manually change intermediate representations. This can lead to rather strange behaviour of models that have been trained on automatically predicted data. Consider the following example: For some reason the symbolic

accent prediction always wrongly predicts a peak (high) accent instead of a valley (low) accent under certain conditions. Imagine that the F0 prediction has learned to correct this error by assigning a low F0 value under these conditions, even though the symbolic accent prediction predicts a peak accent. If a user now explicitly assigns a peak accent, it might happen that the produced output will have a low F0 value for the phonemes in the affected syllables.

As described in Section 3.2.2, the importance of using manually corrected features was tested in a preliminary experiment. As shown in Table 3.3, the differences in accuracy between prediction tasks using only automatically predicted features vs. using manually corrected features were rather small.

Therefore, and as a solution to the problems described above, the final symbolic prediction models described in Section 3.2 were trained on automatically predicted features, whereas the so-called *Symbolic* duration and F0 prediction models described in Section 3.3 were trained on *correct* symbolic prosody features.

1.4.2. Direct Prediction

The second method aims at reducing error accumulation simply by predicting duration and F0 values directly without intermediate symbolic prosody prediction. So-called *Direct* prediction models, which do not use any symbolic prosody features, are described in Section 3.3. Both the Symbolic and the Direct models are included in the perceptual evaluation (Chapter 4), showing that they do not differ significantly.

Nevertheless, the Direct prediction method is not a viable solution for TTS systems that are to be used as instructional or research tools, because it does not offer an intermediate symbolic prosody representation that could be manipulated by the user.

2. Database

As illustrated in Section 1.3, different machine learning algorithms often lead to similar results as long as the chosen database (corpus) contains the information needed for training. The choice of a suitable corpus and the representation of the data is of utmost importance. In order to train models for prosody prediction we need a speech corpus that is annotated with information about:

- word boundaries
- syllable boundaries
- phonemic (even better: phonetic) segmental labels
- pauses
- prosodic phrase boundaries
- accents (location and type)
- boundary tones or phrase-final intonation contours
- lexical stress.

Unfortunately, corpora of read speech with these levels of annotation do not abound for German. Apart from the "Kiel Corpus of Read Speech"[1], which is described in detail in the following sections, I know of only two other German annotated speech corpora: the "IMS German Radio News Corpus" (Rapp, 1998) and the "Siemens Synthesis Corpus (SI1000P)"[2]. Both contain read speech of professional broadcasting announcers. The former consists of

[1] http://www.ipds.uni-kiel.de/publikationen/kcrsp.en.html
[2] http://www.phonetik.uni-muenchen.de/Bas/BasSI1000Peng.html

radio news items and is available upon personal request. The latter contains 1000 newspaper sentences, and the license is rather expensive. Both corpora are only partly annotated with the required information, and the automatic segmental annotations were not manually verified for the whole material.

The KCoRS has several advantages: It is publicly available at a low price, it is almost completely annotated with the information needed for prosody prediction, and the annotations are manually verified. Nevertheless, it has one drawback: It consists mostly of isolated sentences (there are just two complete texts). Prosodic phenomena that depend on paragraph or information structure cannot be modelled with the KCoRS. Pause modelling is also practically impossible (cf. Section 3.1). However, those shortcomings are outweighed by the very comprehensive and consistent annotation.

Another question that arises when choosing a suitable corpus is whether one should base the prosodic models on read or rather on spontaneous speech. The Institute of Phonetics and Digital Speech Processing (IPDS) at the University of Kiel also offers the "Kiel Corpus of *Spontaneous* Speech" (KCoSS), which is annotated in the same way as the KCoRS. So, why not use the KCoSS, since its contents are much closer to speech occurring in real life than the ones of the KCoRS? One of my goals was to improve MARY, a German *text*-to-speech system, which is a reading machine, rather than a communication machine. Of course, MARY could be used as the output device of a dialogue system. However, an important prerequisite would be that the generated utterances are also "spontaneous". To my knowledge, breathing, back-channel utterances, grunts, hesitations and similar characteristics of conversational speech are not implemented in current dialogue systems. So, for the time being, it makes more sense to train the statistical models on read speech rather than on spontaneous speech.

It is very important to get to know the details of a database before starting to train any models. For example, the first German synthesiser that was built for the unit-selection synthesis system CHATR (Black & Taylor, 1994) was very unsatisfactory (sometimes even unintelligible) mainly for two

reasons (Brinckmann, 1997):

1. We had not realised that the two speakers who had read the complete textual material of the KCoRS were each named with two different IDs in different parts of the corpus (`kko` and `k61` for the male speaker, and `rtd` and `k62` for the female speaker). Thus, less than half of the available speech material was used at first.

2. We did not know that the segmental labelling in the KCoRS is mostly phonemic (with only a few phonetic additions). For example, we believed that a segment labelled with /`i:`/ is always a tense long vowel, when in fact it is often realised as a short schwa-like vowel in function words. So, when such a reduced variant was used by CHATR within an accented, unreduced syllable, the resulting synthesised speech became almost unintelligible.

In Section 2.1 and 2.2 the material and the original annotation of the KCoRS are described in detail. These two sections are mainly written for those who would like to use the KCoRS themselves but are daunted by the labelling format, which can be rather confusing for first-time users. In Section 2.3, I describe the features that I added to the KCoRS and the tools I used for these additions. Finally, I conclude with some remarks on the limitations of the KCoRS, and why some features were not added.

2.1. The Kiel Corpus of Read Speech

The KCoRS is a corpus of read German, which was collected and annotated at the IPDS. It comprises over four hours of labelled read speech and is available on CD-ROM (IPDS, 1994).

The KCoRS originates from the PHONDAT project, preparatory works starting in 1989. The aim of the project was to build a phonetic database of spoken German as a resource for automatic speech recognition and general

linguistic, phonological and phonetic questions (Kohler, 1992d). Within the PHONDAT project, the same textual material (described in Section 2.1.1) was used for recordings at four different universities in Germany – Bochum, Bonn, Kiel and München. Only the speech material recorded at the University of Kiel constitutes the KCoRS.

2.1.1. Textual Material

The textual material[3] consists mostly of isolated sentences taken from a variety of contexts.

- **Phonetically balanced material** (398 sentences[4]): The starting point for the compilation of phonetically balanced material were the 'Berlin and Marburg sentences' (Sotscheck, 1984). These are short sentences with high-frequency vocabulary, which contain all German phonemes and many of the phoneme pairs that are allowed according to the phonotactic restrictions of German (Kohler, 1992c). The other sentences of the phonetically balanced material were chosen so that all possible German phoneme pairs are covered.

- Two **short stories** (22 sentences): "Die Buttergeschichte" and "Nordwind und Sonne" (German version of "The Northwind and the Sun").

- **Train timetable queries** (204 sentences):

  – "Siemens sentences": invented, grammatically correct sentences, e.g. *Ich brauche für übernächsten Montag nachmittag eine Zugverbindung von Baden-Baden nach Oldenburg.*

  – "Erlangen sentences": selected transliterations of recorded spontaneous dialogues (not always grammatically correct), e.g. *Grüß Gott, ich bräuchte eine Fahrkarte nach Hamburg und wollte fragen, also wann der Zug abgeht dann.*

[3] The complete textual material of the KCoRS is listed on the following web pages:
`http://www.phonetik.uni-muenchen.de/Bas/BasPD1Contents`
`http://www.phonetik.uni-muenchen.de/Bas/BasPD2Contents`

[4] In the KCoRS, everything that ends either in a full stop, a question mark, or an exclamation mark counts as a sentence.

In total, these are 624 sentences, containing 4932 word tokens and 1673 word types (i.e. orthographically different words). The main textual characteristics are summarised in Table 2.1. With a mean value of 7.9 words, the sentences are relatively short. The shortest sentences consist of only one word, and all of them are "Erlangen sentences" (e.g. *nein* or *danke*). This illustrates that one-word utterances are quite possible in spontaneous speech. The longest sentence contains 29 words and is part of the short story "Die Buttergeschichte".

	complete material	phonetically balanced	rest
mean sentence length (in words)	7.9	6.2	11.0
frequency of sentences with at least one comma	22.3%	10.1%	43.8%
frequency of interrogative sentences	16.3%	5.8%	35.0%
frequency of exclamations	4.6%	6.5%	1.8%

Table 2.1.: Characteristics of the KCoRS textual material. Figures are given for the complete textual material and two subsets: the phonetically balanced material and the rest of the corpus (i.e. short stories and train timetable queries).

The histogram of sentence lengths in Figure 2.1 shows that the single most frequent sentence length is 5 words (sentences with a length of 5 words make up more than a quarter of the whole corpus), but this peak is almost entirely caused by the phonetically balanced material. Within the rest of the textual material, the sentence lengths are much more evenly distributed. Only 22.3% of the 624 sentences contain a comma, which is mainly due to the general shortness of the sentences. The KCoRS includes 102 interrogative sentences (sentences ending with a question mark) and 29 exclamations (sentences ending with an exclamation mark).

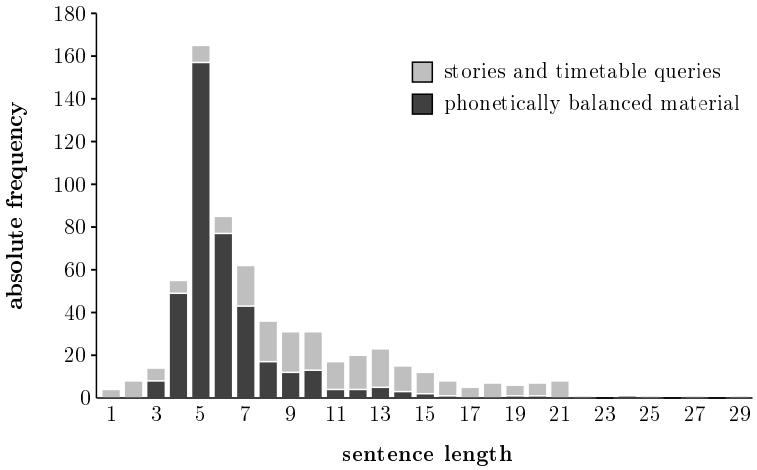

Figure 2.1.: Histogram of sentence lengths (counted in number of words) in the textual material of the KCoRS. Notice the difference between the phonetically balanced material (mostly short sentences) and the short stories and train timetable queries (longer sentences, more evenly distributed sentence lengths).

All of these textual characteristics have to be kept in mind as possible influencing factors for the performance of the statistical models that were trained on the database.

2.1.2. Recordings

The PHONDAT project was carried out in two phases. For PHONDAT1 the phonetically balanced material and the short stories were recorded. PHONDAT2 covered the train timetable queries.

At the IPDS, 53 speakers (26 female, 27 male, all older than 20 years) were recorded in a sound-treated room. One female and one male speaker read the whole textual material, each of the remaining 51 speakers read a subcorpus of the 624 different sentences. Every speaker was advised to read carefully but fluently. If an error occurred, the recording was interrupted by the supervisor and the sentence was repeated.

The signals were digitized at 16kHz sampling frequency with 16-bit resolution. They were stored in separate files for each sentence and associated with exactly one label file with the following file naming conventions: `xxxyyyyy.16` for the signal files and `xxxyyyyy.s1h` for the label files, where `xxx` is the speaker ID and `yyyyy` the sentence ID[5]. For the PHONDAT1 material, the speaker ID follows the format `k`<*number*>, and information about the speaker is coded as follows:

- even *number* → female speaker
- odd *number* → male speaker
- *number* <= 30 → speaker is not older than 30
- *number* > 60 → speaker is older than 30.

This coding convention was abandoned in PHONDAT2, so that the two speakers who read the whole corpus each have two different speaker IDs, depending on the part of the corpus: the female speaker is named `k62` and `rtd`, the male speaker has the speaker IDs `k61` and `kko`. For the training of the prosody models (cf. Chapter 3), only the data of those two speakers is used. The complete speech material of kko/k61 is 43.5 minutes long, rtd/k62's material amounts to 41 minutes. Deducting all pauses (most of them are at the beginning and at the end of a file), this leads to 29 minutes (kko/k61) and 26 minutes (rtd/k62) of 'pauseless' speech material.

2.2. Original Annotation

The character set used in the label files is 7-bit ASCII, where German umlauts are represented with special characters (e.g. "}" for "ü"). As can be seen in the example in Figure 2.2, the label files have the following syntax:

<name of label file>
<orthography>
`oend`

[5]For the PHONDAT2 material, the sentence ID consists only of four digits (`yyyy`).

<canonical transcription>
kend
<realised form>
hend
<start sample> <label> <start time>
<start sample> <label> <start time>
...

The canonical transcription was derived semi-automatically from the orthography by manually correcting the output of the grapheme-to-phoneme conversion module of the German text-to-speech synthesis system RULSYS (Kohler, 1992a).

The field <realised form> contains the sequence of all labels from the <label> section without any position markers. Aiming for brevity, the labels are very compact and rather hard to decipher for first-time users of the KCoRS. For example, `#&1(` labels an early peak accent with the accentuation level 1, whereas `#&1.` denotes the intonation contour "mid fall". In the following sections, all annotation symbols for the field <label> are described in detail.

2.2.1. Orthography

The orthographical representation of the words is given at the very beginning of the label file in the field <orthography>. Within the <label> section, the following symbols relating to the orthography are used:

- Word boundaries: The symbol of the first phoneme of a word is marked with a prefixed `##`. All labels *within* a word are prefixed with `$`, all others start with `#`.

- Sentence boundaries are labelled with `#c:` with the same sample number as the first phoneme of the sentence.

- Punctuation marks are always preceded by `#.` `!` `?` `.` `,` are annotated as they appear in the <orthography>, other punctuations marks (e.g. the colon) are labelled with `,` .

```
k61be022.s1h
Achte auf die Autos!
oend
Q 'a x t @  Q aU f+  d i:+  Q 'aU t o: s  !
kend
c:  &2(  Q- 'a x t -h @  &0  Q- aU f+  &0  d -h i:+  &1.  &2)
  Q- -q 'aU t -h o: s  !  &2.  &PGn
hend
     8455 #c:           0.5283750
     8455 #&2(          0.5283750
     8455 ##Q-          0.5283750
     8455 $'a           0.5283750
     9853 $x            0.6157500
    11308 $t            0.7066875
    12181 $-h           0.7612500
    12431 $@            0.7768750
    13378 #&0           0.8360625
    13378 ##Q-          0.8360625
    13378 $aU           0.8360625
    14881 $f+           0.9300000
    15839 #&0           0.9898750
    15839 ##d           0.9898750
    16445 $-h           1.0277500
    16632 $i:+          1.0394375
    18001 #&1.          1.1250000
    18001 #&2)          1.1250000
    18001 ##Q-          1.1250000
    18001 $-q           1.1250000
    18001 $'aU          1.1250000
    20987 $t            1.3116250
    21886 $-h           1.3678125
    22588 $o:           1.4116875
    25240 $s            1.5774375
    29161 #!            1.8225000
    29161 #&2.          1.8225000
    29161 #&PGn         1.8225000
```

Figure 2.2.: KCoRS label file k61be022.s1h

2.2.2. Morpheme Boundaries and Parts-of-Speech

Only those morpheme boundaries that are connected with particular phonetic characteristics (e.g. lengthening or aspiration) are marked using `$#` before the phoneme symbol.

Function words are marked by placing the symbol `+` after the symbol of the last phoneme of the word (e.g. `$i:+` at sample `16632` in Figure 2.2).

2.2.3. Phonemes

The segmental labelling of the KCoRS is "broad phonetic" (Barry & Fourcin, 1992), i.e. the segmental label inventory is "essentially phonological with a small number of phonetic additions" (Kohler et al., 1995). It is based on the canonical transcription, and the phonemes[6] are transcribed with a modified version of SAMPA (Speech Assessment Methods Phonetic Alphabet; Wells, 2004):

- 21 vowels (7 short vowels, 8 long vowels, 3 diphthongs, 2 schwas, 1 nasal vowel[7]): `I`, `Y`, `E`, `9`, `a`, `O`, `U`, `i:`, `y:`, `e:`, `2:`, `E:`, `a:`, `o:`, `u:`, `aI`, `OY`, `aU`, `@`, `6`, `a~`
- 15 `/6/`-diphthongs (short or long vowels followed by the vocalised *r* `/6/`,
  e.g. `/i:6/` in *Bier*): `I6`, `Y6`, `E6`, `96`, `a6`, `O6`, `U6`, `i:6`, `y:6`, `e:6`, `2:6`, `E:6`, `a:6`, `o:6`, `u:6`
- 22 consonants, including the glottal stop `/Q/`: `p`, `b`, `t`, `d`, `k`, `g`, `Q`, `m`, `n`, `N`, `f`, `v`, `s`, `z`, `S`, `Z`, `C`, `x`, `r`, `h`, `j`, `l`.

Whenever the realised form deviates from the canonic transcription, the following symbols are added:

[6] Since the transcription in the KCoRS is mostly phonemic, I refer to the labelled sounds as "phonemes" throughout this text and refrain from distinguishing between phones and phonemes.

[7] German SAMPA has symbols for four different nasal vowels, but only one of them appears in the KCoRS.

- Deletions are marked with a hyphen after the symbol of the deleted phoneme, e.g. `Q-`.
- Insertions are marked with a hyphen before the symbol of the inserted segment, e.g. `-t`.
- Replacements are marked with a hyphen after the symbol of the canonic form, followed by the realised form, e.g. `n-m` (where `n` is realised as `m`). Only *phonemic* changes (e.g. reduction from a full vowel to schwa) are labelled this way, phonetic variations in vowel quality or quantity are not marked.

Table 2.2 lists the percentage of deletions, replacements and insertions of all canonic phonemes. The most commonly deleted canonic phonemes are `/Q/` (kko/k61: 64% vs. rtd/k62: 54%), `/@/` (38% vs. 47%), and plosive releases (36% vs. 39%).

	deletions	replacements	insertions
kko/k61	12.2%	2.0%	0.17%
rtd/k62	13.7%	2.1%	0.05%

Table 2.2.: Percentage of deletions, replacements and insertions of all canonic phonemes for speakers kko/k61 and rtd/k62.

In addition to the canonic labels, the following labels are used to mark phonetic aspects of the realised segments:

- Glottalisation / creaky voice is labelled with `-q`.
- Nasalisation is labelled with `-~`, only if a nasal has been deleted and the neighbouring realised phonemes are nasalised.
- Hesitational lengthening: If a segment is hesitationally lengthened, the label `z:` is placed at the sample number of the following phoneme (i.e. *after* the lengthened phoneme).

- Plosive release: The closure and the release phase of a plosive are labelled separately. The release is always transcribed with `-h`, regardless of the respective plosive. If the plosive is followed by a fricative, the plosive release phase is usually not labelled separately but assigned to the duration of the fricative.

- Uncertainty: If the beginning of a phoneme cannot be determined with certainty, the corresponding label is prefixed with `%`.

2.2.4. Prosody

The KCoRS is annotated with the prosodic labelling system PROLAB (Kohler, 1995; Peters & Kohler, 2004), which is based on the pitch contour-based Kiel Intonation Model (KIM; Kohler, 1997). It incorporates the following domains: lexical stress, accent, intonation contour, prosodic boundaries, and pauses. Labels for accent, intonation contour and prosodic boundaries always contain `&` in order to separate them from the segmental labels.

Lexical Stress

There are no syllable boundaries marked in the KCoRS. Therefore, primary and secondary lexical stress is indicated by prefixing the symbol of the vowel of the stressed syllable with `'` or `"` respectively (e.g. `$'a` at sample `8455` in Figure 2.2).

Function words receive no lexical stress marking, even though there are several multi-syllabic function words in German (e.g. *warum*, *desto*, *wegen*). If a function word carries a sentence accent (see below), the label `$''` is inserted before the vowel of the stressed syllable.

If the realised lexical stress position in a word deviates from the canonical transcription, this is marked the same way as phonemic changes (see Section 2.2.3, e.g. `a-'a:`).

Accent

Sentence accent is usually an attribute of the whole word. Therefore, the accent labels are placed before the respective word and prefixed with `#&`. If a word carries more than one accent and one accent label must be placed *within* a word, it is prefixed with `$&`. Usually, the accent falls on the syllable with primary stress. If a vowel is preceded by the label `$''`, the accent falls on the syllable containing that vowel. If a word carries more than one accent, a sentence accent marker is provided for each accent, either before the respective morpheme boundary (if present), or directly before the accented phoneme.

Within one accent label, the following information is coded: accentuation level, accent type, alignment, and upstep. A complete list of all PROLAB accent labels that occur in the KCoRS is given in Appendix A.1.

Accentuation level Four levels of accentuation are distinguished:

`0` unaccented
`1` partially accented
`2` accented
`3` reinforced.

As shown in Figure 2.3, the most frequent accentuation level is `0`, closely followed by `2`. Speaker kko/k61 produced only 39 reinforced accents compared to 136 reinforced accents for speaker rtd/k62.

Accent type and alignment Any syllable that is not unaccented (i.e. is labelled with an accentuation $>$ `0`), carries one of three possible accent types: flat, peak, or valley. In addition, peak and valley labels carry information about their alignment, i.e. the position of the maximum or minimum in the F0 contour with respect to the accented syllable.

Flat accents show very little change in F0 across several phonemes or syllables, even though an accent can be perceived. Kohler (2003) calls this

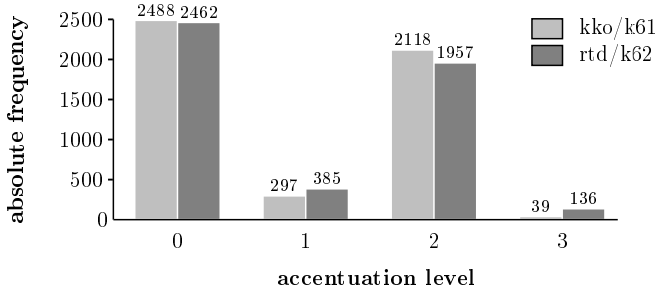

Figure 2.3.: Absolute frequency of accentuation levels for speaker kko/k61 and rtd/k62.

type of accent *force accent* in order to distinguish it from *pitch accents*, which are always associated with an F0 movement. In PROLAB, flat accents are labelled with `-`.

Peak accents have a local maximum in the F0 contour in the neighbourhood of the accented syllable. Three values for alignment are available for peak accents: early, mid, and late, with the respective F0 maximum before, within, and after the nucleus of the accented syllable. The PROLAB labels are: `)` (early peak), `^` (mid peak), and `(` (late peak). Figure 2.4 shows that peak accents are the most frequent accent types for both speakers (kko/k61: 85.8%, rtd/k62: 85.2%).

Valley accents have a local minimum in the F0 contour in the neighbourhood of the accented syllable. Only two types of alignment are distinguished for valley accents: `]` (early valley: F0 minimum before the nucleus of the accented syllable) and `[` (non-early valley: F0 minimum within or after the nucleus of the accented syllable).

Upstep As a default, the F0 minima and maxima of the accents are expected to decline over the course of an utterance, so that the first peak accent in an utterance is higher than the second one and so on (for a detailed discussion of declination cf. Cohen et al., 1982). Therefore, this regular 'downstep'

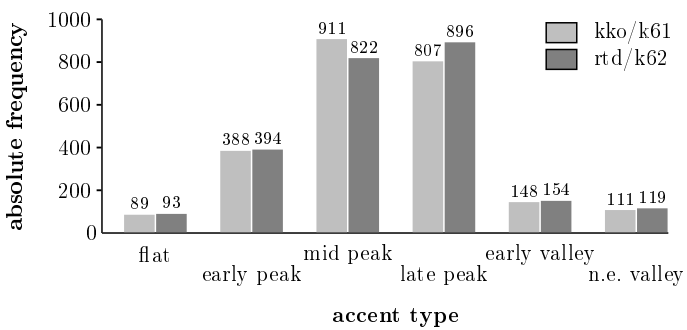

Figure 2.4.: Absolute frequency of accent types for speaker kko/k61 and rtd/k62 (n.e. valley = non-early valley).

of accents is not labelled. However, when an accent's minimum or maximum is higher than, or as high as the preceding accent, it is labelled with upstep: `!`. All accents with an accentuation level greater than `0` can be upstepped. Only 5.2% of kko/k61's accents and 5.8% of rtd/k62's accents are upstepped.

Intonation Contours

Concatenation and phrase-final contours PROLAB labels for intonation contours between accented words (so-called "concatenation contours") and at the end of a prosodic phrase (phrase-final contours) always end with a punctuation mark: `;` is used to label a minimal rise ("pseudo-terminal contour"; Peters, 1999), `,` denotes a low rise, `?` marks a high rise and `.` is used for several types of falls. The `.` is preceded by a digit to denote the strength of the fall: `0` (level), `1` (mid fall), and `2` (terminal fall). All *fall* categories can be combined with all *rise* categories resulting in 9 additional, complex intonation contours. All intonation contour labels are listed in Appendix A.2. As shown in Figure 2.5, falls form the most frequent class of intonation contours, whereas high rises are very infrequent.

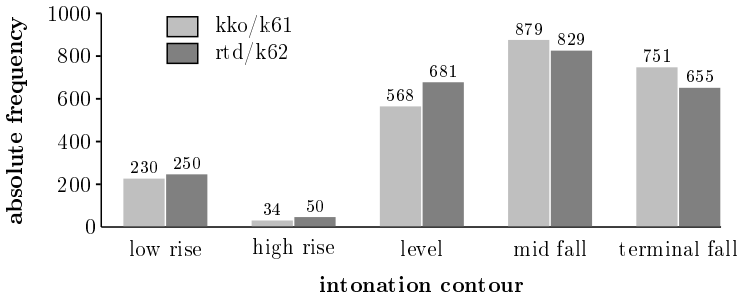

Figure 2.5.: Absolute frequency of simplified concatenation and phrase-final contours (simplification as described in Section 3.2.2).

Phrase-initial contours Most prosodic phrases begin with several unaccented syllables, the "pre-head". As a default, the F0 contour of this pre-head is lower than the F0 maximum of the first accented syllable. Two "high pre-head" categories are labelled in PROLAB: `HP1` marks a pre-head with a F0 contour that is as high as the following accent, whereas `HP2` is used for a pre-head starting with a high F0 contour that falls steadily until the first accented syllable is reached. If the first accent is a valley, it is not possible to distinguish between low and high pre-head. In these cases, the default (low pre-head) is assumed.

Prosodic boundaries, register, and speech rate

Prosodic phrase boundaries are marked with `PGn`. They are phonetically signalled by phrase-final segmental lengthening and usually by F0 resetting after them. They often coincide with pauses (see below). Phrase boundaries are not further divided into subclasses with differing boundary strengths (a division into `PG1` and `PG2` was planned, but has not been carried out to date).

Usually, the declination of the accents is reset at the beginning of a prosodic phrase and the downstep starts anew. If there is no reset after a prosodic phrase boundary, the boundary is labelled with `=PGn`.

If a speaker deviates from his or her normal F0 range, this is labelled

with `HR` (high register) or `LR` (low register). Similarly, deviations from the normal speaking rate of the speaker are marked with `RP` (rate plus) or `RM` (rate minus). Since register and speech rate labels are very rare in the KCoRS (they were introduced to PROLAB mainly for spontaneous speech), they were not used for statistical modelling.

Pauses

The following types of pauses are labelled in the KCoRS:

- silent pause (`p:`)
- pause filled with
  - breathing (`h:`)
  - clicking or lip-smacking (`s:`)
  - segmental material because the speaker stumbled or misread a word (`v:`).

The vast majority of the pauses produced by the selected speakers kko/k61 and rtd/k62 are silent pauses (95% and 97% respectively), which is not surprising given the fact that most sentences are rather short and produced in isolation. Speaker kko/k61 produces more pauses than speaker rtd/k62, which is in line with his slower speech rate.

	prosodic boundary type			
	no boundary	**reset**	**no reset**	total
	kko / rtd	kko / rtd	kko / rtd	kko / rtd
pause	0 / 1	673 / 653	2 / 4	675 / 658
no pause	3967 / 3905	236 / 311	54 / 58	4257 / 4274
total	3967 / 3906	909 / 964	56 / 62	4932

Table 2.3.: Co-occurrences of following prosodic boundaries and pauses per word for speaker kko/k61 and rtd/k62.

Summarising the co-occurrences of pauses and prosodic boundaries in Table 2.3, we can formulate the following sets of simple rules:

1. prediction of pauses from boundaries (accuracy: 94.4%)
   - no boundary ⇒ no pause
   - "no reset" boundary ⇒ no pause
   - reset boundary ⇒ pause
2. prediction of boundaries from pauses (accuracy: 93.2%)
   - pause ⇒ reset boundary
   - no pause ⇒ no boundary

Since most pauses occur at the beginning or at the end of the speech files, we have very little data about the duration of pauses. Therefore, the KCoRS is not a suitable training database for pause modelling (cf. Section 3.1).

2.3. Added Features and Changes

Although the KCoRS already contains a lot of information and is annotated very consistently, I added several features[8] that are important for prosody prediction. Concerning the textual data these are: sentence type, part of speech, syntactic phrases, grammatical functions, and word frequencies. The annotation of the speech signal was enriched with information about syllable boundaries and F0 median values. In addition to some minor changes to the orthography, I changed the annotation of lexical stress and some phoneme labels. All these additions and changes are described in detail in the following sections.

Since the models that are trained on these features shall eventually be implemented in MARY as an alternative prosody prediction, the tools I chose for automatically adding features to the KCoRS had to satisfy one of the following conditions: They had to be either:

[8] All files containing the added features are available from `http://www.brinckmann.de/KaRS/`.

- already implemented in MARY (part-of-speech tagger, syntactic chunk tagger)
- easily implementable with a small algorithm (sentence type, syllable boundaries)
- publicly available (word frequencies from CELEX)
- available within the DFKI (SCHUG parser for grammatical functions).

Some of the automatically added features were corrected manually (part-of-speech, syntactic phrases, grammatical functions, syllable boundaries), others were not corrected, either because this would have been too time-consuming (F0 values) or because it is unnecessary (sentence type, word frequencies). Both automatically derived and manually corrected feature sets were tried out for symbolic prosody prediction (cf. Section 3.2) in order to estimate the amount of error introduced to the models by erroneous feature values.

2.3.1. Textual Data

Orthography

In order to facilitate textual processing, the orthography was changed in the following cases:

- spelling mistakes were corrected, e.g. *Jung's* in sentence `mr006` was changed to *Jungs*
- numbers were expanded, e.g. *11.* in sentence `cn020` was converted into *elften*
- spellings of denominations for the time of day were harmonised following §55(6) of the new regulations of German orthography (IDS, 1996): denominations for the time of day are capitalised when they follow *heute*, *(vor)gestern* or *(über)morgen*, e.g. *heute Abend*.

Sentence Type

Brinckmann & Benzmüller (1999) showed that in German scripted speech the four utterance types statement, wh-question, yes/no-question, and declarative question differ significantly concerning final boundary tone, F0 range, and F0 slope. Therefore, every sentence in the textual material of the KCoRS was automatically labelled with one of the following sentence types: *statement* (ends with a full stop), *exclamation* (ends with an exclamation mark), or *question* (ends with a question mark). The questions were further subdivided into the types listed in Table 2.4.

type	description	example
wh-question	contains an interrogative pro-form	*Wann geht der nächste Zug nach Mannheim?*
yes-no question	inflected verb at the beginning of the sentence	*Steigt Dein Drachen sehr hoch?*
negative yes-no q.	contains *nicht* or *kein*	*Muß der Zucker nicht dort drüben stehen?*
alternative question	presents two possible answers connected with *oder*	*Wünschen Sie Raucher oder Nichtraucher?*
declarative question	same word order as in a statement	*Und später fährt keiner mehr?*
polite request	starts with *Könnten Sie ...* or *Können Sie ...*	*Könnten Sie mir bitte Züge von Regensburg nach Frankfurt heute abend sagen?*

Table 2.4.: Question types in the textual material of the KCoRS.

As can be seen in Figure 2.6, the most frequent sentence type in the KCoRS is the statement (78.7%). 104 sentences (16.7%) were classified as questions[9], and only 29 (4.6%) as exclamations.

[9] Two sentences ending with a full stop were classified as polite requests, thus falling into the category "question".

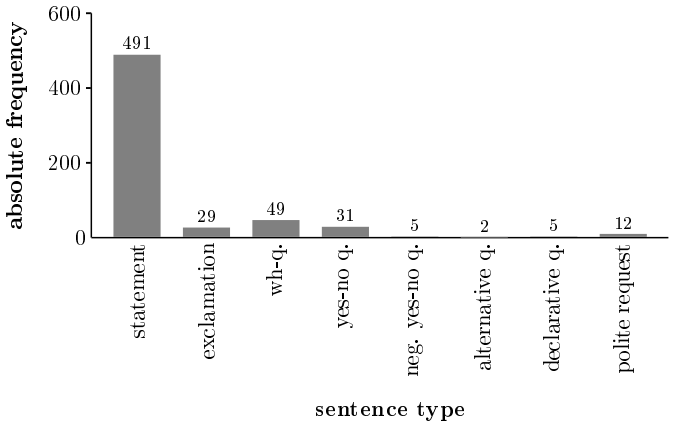

Figure 2.6.: Histogram of sentence types in the textual material of the KCoRS (q. = question, neg. =negative).

Part-of-Speech Tags

The original annotation of the KCoRS distinguishes between function and content words, reflecting the assumption that function words are usually unaccented. A more refined part-of-speech classification could be helpful for the prediction of accentuation. For example, separated verbal particles and attributive indefinite pronouns (*keine*, *beide*) are often accented, even though they are usually classified as function words.

Part-of-speech tagging was carried out in two steps. First, the statistical tagger TnT (Brants, 2000) was applied to the textual data. The German language model of TnT had been trained on the annotated NEGRA corpus (Brants et al., 1999) using the Stuttgart-Tübingen tag set (STTS). Second, the tags were manually corrected following the guidelines for STTS (Schiller et al., 1995).

A comparison between the statistically tagged data and the manually corrected version revealed that only 3.4% of the tags had to be corrected. Table 2.5 shows that TnT performs significantly better on known tokens (i.e. tokens that are part of the lexicon generated from the NEGRA corpus)

than on unknown tokens. Even though the KCoRS textual data is rather unlike the NEGRA corpus (which is a collection of newspaper texts), the accuracy figures are very similar.

	percentage	tagging accuracy		
	unknown tokens	known tokens	unknown tokens	overall
KCoRS	10.9%	97.8%	86.9%	96.6%
NEGRA	11.9%	97.7%	89.0%	96.7%

Table 2.5.: TnT's part-of-speech tagging accuracy for the KCoRS textual data and the NEGRA corpus (figures for NEGRA from Brants, 2000). Unknown tokens are tokens that are not in the lexicon generated from the NEGRA training corpus.

Out of 54 possible STTS tags, 47 are present in the KCoRS (see Appendix B.1 for a complete list of STTS tags with examples and information about their absolute frequency in the KCoRS). Only the following seven tags are missing: APPO, FM, PPOSS, PRELAT, TRUNC, VMPP, and XY.

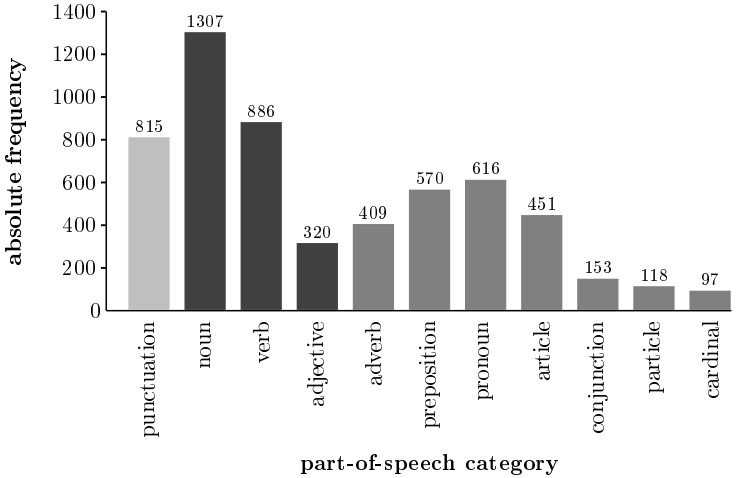

Figure 2.7.: Histogram of simplified part-of-speech categories in the textual material of the KCoRS.

Syntactic Chunks

Words that belong to the same syntactic phrase are usually not separated by a prosodic phrase break, at least in read speech. MARY uses the chunk tagger (Skut & Brants, 1998) to recognise syntactic structures of limited depth. The chunk tagger was applied to the textual material of the KCoRS, and the output was corrected manually.

The chunk tagger assigns the phrasal categories used in the NEGRA corpus (Brants et al., 1999), but only multi-word phrases receive such a phrasal chunk tag (44.5% of all word tokens in the KCoRS are not part of a multi-word phrase). For example, if a noun phrase consist only of one pronoun, it keeps the POS tag assigned by TnT.

Out of 20 possible phrasal chunk tags, 14 are present in the KCoRS (see Table B.3 in Appendix B.2 for a detailed list). By far the most frequent phrasal chunk tags are NP (noun phrase) and PP (adpositional phrase) – together with their respective coordinated variants (CNP and CPP) they make up 92% of the labelled multi-word phrases (see Figure 2.8). Top-level chunk phrases, i.e. chunk phrases that are not embedded in any other phrase, make up 81.9% of all multi-word phrases.

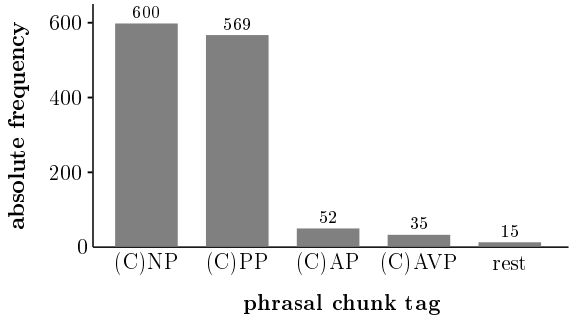

Figure 2.8.: Histogram of multi-word phrasal chunk tags in the textual material of the KCoRS. Figures for NP, PP, AP and AVP are given together with their respective coordinated variants. All other chunk tags are collapsed into the category "rest".

A comparison of the automatically derived top-level categories with the manual corrections revealed a word-level accuracy of 85.1% (i.e. the top-level phrasal category or POS of 85.1% of all word tokens was not changed manually). Regarding the absolute position of each word within the top-level chunks (i.e. whether it is the first, second, third etc. word within the chunk), the chunk tagger reached an accuracy of 87.4%.

Grammatical Functions

Wolters & Mixdorff (2000) reported that the grammatical function of a phrase has an influence on the accentability of the words it contains, e.g. nouns in genitive adjuncts are less likely to be accented than nouns in subjects. This could be explained by the fact that genitive adjuncts are frequently used to link new discourse entities to discourse-old entities or world knowledge.

MARY contains no grammatical function tagger yet, but the SCHUG parser developed at the DFKI (Declerck, 2002) is readily available for this purpose. Therefore, SCHUG was used to assign phrasal categories and grammatical functions to the textual material of the KCoRS. The SCHUG parser is a rule-based system using morphological and part-of-speech information. In contrast to the chunk tagger, it assigns phrasal categories also to phrases consisting of only one word. Table 2.6 lists all SCHUG categories and their possible grammatical functions.

SCHUG was applied to the complete textual material of the KCoRS, and its output was corrected manually. As shown in Figure 2.9, the most frequent SCHUG categories in the KCoRS are NP, VG and PP. If the grammatical function of a noun phrase is ambiguous (according to SCHUG's rules), SCHUG assigns a set of all grammatical functions that are deemed possible for that phrase. Since this is the case for 49.7% of the automatically derived noun phrases (even for some pronouns with overt case marking), some improvement is necessary here. Another field for future improvements of SCHUG is the recognition of embedded phrases. An inspection of the man-

category	description	possible grammatical functions
AP	adjective phrase	PREDICATIVE_AP
AdvP	adverbial phrase	PREDICATIVE_ADVP
NP	noun phrase	SUBJ, SUBJ/DEEP_OBJ, AKK_OBJ, DAT_OBJ, GEN_OBJ, NP_ADJUNCT_GEN, PREDICATIVE_NP
PP	prepositional phrase	PP_ADJUNCT, PP_OBJ
SUBORD_CLAUSE	subordinated clause	XADJUNCT, XCOMP
VG	verb group	–
W	word (conjunctions)	–

Table 2.6.: SCHUG categories and possible grammatical functions.

ually corrected SCHUG phrases showed that 16.0% of all SCHUG phrases in the KCoRS are embedded phrases (see Table B.2 in Appendix B.2). Currently, SCHUG is only capable of recognising top-level phrases.

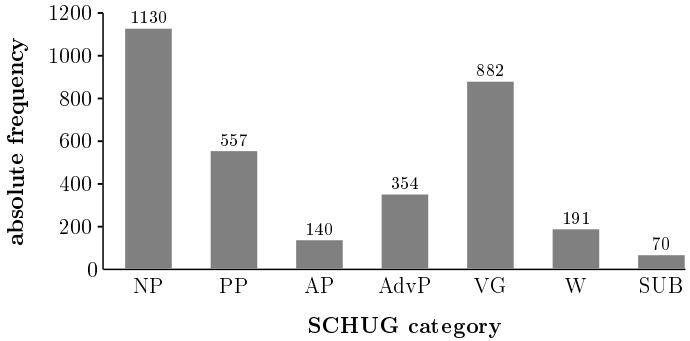

Figure 2.9.: Histogram of SCHUG categories in the textual material of the KCoRS (SUB = SUBORD_CLAUSE).

A comparison of the automatically derived top-level categories with the manual corrections revealed a word-level accuracy of 76.3% (i.e. the top-level SCHUG category of 76.3% of all word tokens was not changed manually), whereas the grammatical functions were correct only for 51.0% of all words. Regarding the absolute position of each word within the respective top-level

phrase (i.e. whether it is the first, second, third etc. word within the phrase), the SCHUG parser reached an accuracy of 78.3%. At first glance these accuracy figures seem to suggest that the SCHUG parser performs worse than the chunk tagger. However, almost half of the words do not receive a phrasal chunk tag from the chunk tagger, instead they keep their original part-of-speech tag. So the accuracy of the chunk tagger benefits very much from the reliability of TnT. Nonetheless, both the SCHUG parser and the chunk tagger need further improvement.

Word Frequencies

Fidelholtz (1975) showed that the frequency of a word has a significant effect on the reduction of its vowels (the higher the word frequency, the more probable a vowel reduction). Word frequency also correlates with the content/function word distinction: Function words usually have a higher frequency than content words. Thus, the accentability of a word might be rather a consequence of its frequency than of its part-of-speech.

Even if homographic wordforms are distinguished regarding their part-of-speech, the textual material of the KCoRS contains only 1733 different wordforms. Since a TTS system has to rely on a much bigger lexicon, the frequency information that was added for each wordform was not computed directly from the KCoRS. Instead, it was taken from the lexical database CELEX (Baayen et al., 1995). The frequency information in CELEX is based on the "Mannheim" corpus (1984 version) of the "Institut für Deutsche Sprache", which contains about 6.0 million words from mostly written and some spoken sources.

CELEX offers a variety of frequency figures, both for lemmas and for wordforms. I chose *MannMln*, i.e. the wordform frequency scaled down to a range of 1 to 1.0 million (instead of the original 1 to 6.0 million). The minimal value of *MannMln* in CELEX is 0, the maximum is 25287 (for the word *und*). Those 227 wordforms of the KCoRS that are not present in CELEX (mostly nouns), also received the frequency value 0 (totalling in 352 zero-frequency wordform types).

Of course there are wordforms that are very frequent in the KCoRS, but not that frequent in the Mannheim corpus. For example, the most frequent wordform in the KCoRS is *nach* (142 tokens), which is due to the large number of train timetable queries (such as `s008`: *Ich möchte morgen abend nach Köln fahren*). In the Mannheim corpus, *nach* receives the frequency figure of 1738 (when scaled down to the size of the KCoRS, this is the equivalent of 9 tokens). Nevertheless, both the frequency figures based on the KCoRS itself as well as the ones from CELEX behave very similarly when it comes to their distribution within the KCoRS. As can be seen in Figure 2.10, there are many wordforms in the KCoRS with a low frequency figure (e.g. 1 or –only in the case of CELEX frequency figures– 0), some with a medium frequency figure, and only very few wordforms with a high frequency figure.

2.3.2. Speech Data

Phonemes

In the original annotation, all plosive releases are labelled with `-h`, suggesting that the release phase is not canonic, but rather an insertion. Since release phases of fortis plosives are generally longer than lenis releases, their labels were changed, marking them separately with the additional symbols `p_h`, `t_h`, `k_h`, `b_h`, `d_h`, `g_h`. Furthermore, the plosive releases were regarded as canonic.

Lexical Stress

Lexical stress information was added for all function words, so that all words received one primary stress location.

In the original annotation of the KCoRS, two words carry two primary stress locations: *B'aden-B'aden* and *sp'ät'abends*. After listening to the

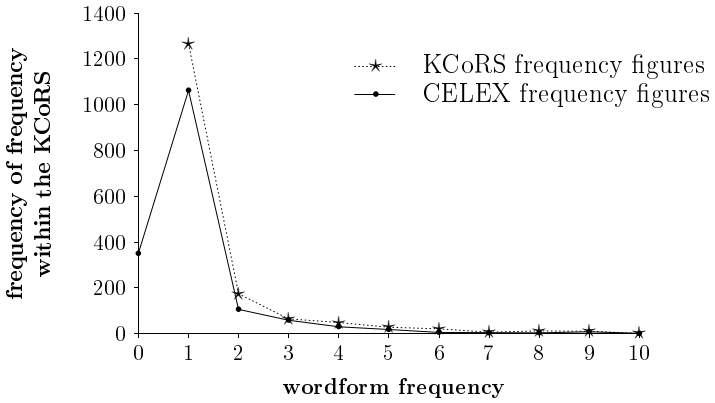

Figure 2.10.: Frequency of CELEX and KCoRS frequency figures of wordform types in the KCoRS textual material. X-axis: The frequency figure for each wordform is either computed directly from the KCoRS (KCoRS frequency figures) or taken from the *MannMln* figure of CELEX and scaled down to the size of the KCoRS (CELEX frequency figures). Y-axis: The frequency of frequency of wordform types is based on the KCoRS textual material. Note that only the CELEX frequency figure can have a value of 0.

realisations of the speakers, one primary stress location was changed to a secondary stress: *B"aden-B'aden* and *sp"ät'abends*.

Syllable Boundaries

Automatic syllabification was carried out with a simple algorithm which defined every vowel as syllable nucleus and every sonorant `/m,n,N,l/` that is preceded by a consonant as potential syllable nucleus. The syllabification of the segments between the established nuclei was based on the following rules and standard phonological principles:

- Every word boundary and every labelled morpheme boundary is a syllable boundary.

- A glottal stop /`Q`/ is always the onset of a syllable.
- Plosive closure and following plosive release or fricative (in case of affricates) are not separated by a syllable boundary.
- Ambisyllabicity: a consonant following a lax vowel in a VCV pattern is marked as ambisyllabic.
- Obligatory Coda: a syllable must be closed (or followed by an ambisyllabic consonant) after a short, lax vowel (except /`@, 6`/).
- Maximal Onset Principle: make the syllable onset as long as it legitimately can be according to the phonotactic restrictions of German.

Two types of syllable boundaries were distinguished: "`_`" marks a syllable boundary which is followed by an ambisyllabic consonant, while normal syllable boundaries are marked with "`-`".

The syllabification algorithm was applied to two datasets: The first one (the 'lexicon') contained all wordforms of the KCoRS with their respective canonic phoneme sequence, the second one ('connected speech'), consisted of all realised phone sequences of the two speakers kko/k61 and rtd/k62. Both syllabified datasets were corrected manually, the second one by listening to all utterances of the two speakers. Compared to the manual corrections, for the 'lexicon' 99.1% of the automatically derived syllable boundaries are correct, while for 'connected speech' the accuracy dropped to 97.2%. This is mainly due to the following phenomena:

- Postlexical resyllabification across word boundaries, e.g. in `k61be031`: *gibt es* realised as /`g g_h I p - t t_h E s`/.
- Glottal stop /`Q`/ is possible at the end of a syllable when it replaces a plosive, e.g. in `k61mr069`: *Zentner* realised as /`t s E n Q - n 6`/.
- Potentially syllabic sonorants following a vowel or /`l`/ are problematic, e.g. *einen* (realised as monosyllabic /`Q aI n`/ or disyllabic /`Q aI -`

n/ ?) and *rollen* (monosyllabic /r O l n/ or disyllabic /r O _ l n/ ?). Each decision was based on my auditory impression, e.g. monosyllabic /r O l n/ in k62be087 and disyllabic /r O _ l n/ in k61be087 (the disyllabic impression seems to be due to the slower speech rate of speaker k61).

F0 Values

The acoustic parameters to predict are duration and F0. The duration of each phoneme can be computed using the labelled phoneme boundaries. F0 values were estimated with ESPS's get_f0 algorithm (Talkin, 1995), which uses the normalised cross correlation function and dynamic programming. As frame step, the default of 10ms was chosen; for the female speaker rtd/k62 the minimum F0 value was set to 120Hz, the maximum to 400Hz, whereas for the male speaker kko/k61 the minimum and maximum were set to 50Hz and 250Hz respectively[10].

Median F0 For every vowel and sonorant (m, n, N, l), the median of these raw F0 values was calculated. The median was chosen instead of the mean, because it is more robust to outliers. Nevertheless, there were still some erroneous median F0 values, especially within portions spoken with a creaky voice, because of doubling or halving errors.

Last F0 Thus, we have one F0 value for every vowel and sonorant. If there is a pitch accent on the last syllable of a prosodic phrase, and this syllable only contains one vowel or sonorant, one F0 value is not sufficient to capture a valley accent followed by a rising intonation contour. Therefore, for every prosodic phrase a final F0 value was stored by computing the median of the last three F0 values of the last vowel or sonorant of that prosodic phrase. If get_f0 cannot estimate any F0 value, median and last F0 are set to 0.

[10]Informal inspection revealed that these values were adequate for those two voices.

2.3.3. Further Possibilities and Limitations

Other features that could be added to an annotated speech corpus include word predictability, discourse features, GToBI labels, intensity and spectral tilt.

Pan & Hirschberg (2000) showed that word predictability, measured in terms of bigram word predictability $log(Prob(w_i|w_i-1))$, is a useful predictor of pitch accent placement for nouns. In order to compute this measure we need a suitable textual corpus. Aiming for a rather neutral prosody of sentences that can occur in any context, we would have to use a very big textual corpus – otherwise the measures would be very domain-specific. This is in line with using the frequency numbers from CELEX, which were calculated from the 6 mio. token Mannheim Corpus rather than directly from the KCoRS. Since bigram word predictability can be helpful mostly for limited domain synthesis, I decided not to add this feature to the KCoRS.

Discourse features like the givenness of a referring expression have an influence on the pitch accent and phrasing (cf. Wolters & Mixdorff, 2000), but since the KCoRS consists mostly of isolated sentences and not of complete texts (except for the two short stories), this kind of information cannot be added. For information structural features, a corpus of read newspaper texts such as the one built in the MULI project (Baumann et al., 2004) and the "IMS German Radio News Corpus" (Rapp, 1998) should be investigated instead.

MARY uses GToBI labels for the symbolic prosody prediction. GToBI labelling was not carried out for the two selected speakers of the KCoRS mainly because of two reasons:

1. Even though Braunschweiler (2003) described an approach to predict GToBI labels automatically from the F0 curve and intensity measures, these automatically predicted labels still have to be corrected manually, which is very time-consuming.

2. The prosody prediction described in Chapter 3 consists of symbolic

prosody prediction and prediction of acoustic parameters. Since the only module of MARY needed for this approach is the MBROLA synthesis, GToBI labels are not necessary as intermediate symbolic representation. The PROLAB labels can be used instead.

Intensity and spectral tilt of realised phonemes influence the perception of rhythm, but since they cannot be modelled by MBROLA, those measurements were not included as features to predict.

3. Prosody Prediction with CART

As described in Section 1.3, in this thesis prosody prediction is defined as containing all prediction tasks that contribute to the prediction of the realised phoneme, its duration, and its F0 values. The following separate prediction tasks are described in the subsequent sections:

- pause prediction
- symbolic prosody prediction:
  - ⋆ prosodic boundaries
  - ⋆ accentuation level
  - ⋆ accents: location and type
  - ⋆ phrase-final intonation contours
- prediction of postlexical phonological processes:
  - ⋆ type of change: none, deletion, replacement
  - ⋆ in case of replacement: replacement rule
- prediction of acoustic parameters:
  - ⋆ duration
  - ⋆ median F0
  - ⋆ last F0.

One major goal of this thesis is to show that the output of a text-to-speech system can be significantly improved by training all models that contribute to prosody prediction on the same database. As described in Section 1.3, many different machine learning algorithms have been applied for the different prediction tasks. It was not my aim to find the best feature

set, the best algorithm, and the best model for each prediction task. Instead I applied the same machine learning algorithm (CART; Breiman et al., 1984) to train classification and regression trees for all prediction tasks.

Because of reasons related to the implementation of the machine learning software tools (see Section 1.2.2), all classification trees were trained with Weka (version 3.4.2; Witten & Frank, 2000), whereas all regression trees were trained with wagon (version 1.2.3, King et al., 2003). All classification trees were evaluated with stratified 10-fold cross-validation. Since wagon does not offer stratified cross-validation, the performance of the regression trees was estimated on a randomly selected separate test set.

Automatic feature selection (greedy forward selection wrapper) was only performed for word-level and syllable-level prediction tasks (i.e. symbolic prosody prediction). The phonemic datasets were too large to make automatic feature selection computationally feasible in a reasonable amount of time (see Section 1.2.3).

Datasets 20 sentences from the KCoRS were randomly selected for the perceptual evaluation (see Table 4.1 in Section 4.1.2). These 20 sentences were *not* included for training, validation and corpus-based testing of the classification and regression trees. Apart from these 20 sentences, the complete KCoRS and all added features (as described in Section 2.3) are used as database to produce the input datasets for Weka and wagon.

3.1. Pause Prediction

As mentioned in Section 2.2.4, only very few pauses occur within a sentence (or rather: between two words), so that information about their duration is available only for 62 and 52 pauses respectively for kko/k61 and rtd/k62 in the training data. Because of the extreme data sparsity, it is impossible to model pause duration with a regression tree. Therefore, two very simple rules based on a trial-and-error procedure with MARY were applied instead:

1. Pause location: A word is followed by a pause, only if it is followed by a punctuation mark.

2. Pause duration: If the word is followed by a comma or a dash, the pause duration is 100ms, if it is followed by another punctuation mark, the pause duration is 300ms.

Of course, this is not a very satisfying solution, but for a successful training we would need a database that consists of complete texts.

3.2. Symbolic Prosody Prediction

3.2.1. Prosodic Boundary Prediction

The classification task for prosodic boundary prediction is to predict for each word whether it is followed by a prosodic boundary or not. Originally, it was planned to predict also the type of the boundary (reset vs. no reset), but since the "no reset" boundaries make up only 6% of all boundaries in the KCoRS, they proved to be impossible to predict with reasonable precision and recall. So I decided to predict only the classes "boundary" and "no boundary".

Features

For each word the following features were extracted from the database:

- **word level features**, for a window of 5 words (the respective word and 2 neighbouring words to the left and to the right):
  - ⋆ part-of-speech: STTS and simplified (simplifications as in Figure 2.7)
  - ⋆ word frequency (CELEX)
- **punctuation features**:
  - ⋆ preceding punctuation

  - ⋆ following punctuation: original and simplified (*none, comma, other*)
  - ⋆ absolute and relative position[1]: distance to preceding and following punctuation (in words), and relative position between punctuation marks

- **sentence features**:
  - ⋆ sentence length (in words)
  - ⋆ sentence type (as defined in Section 2.3.1)
  - ⋆ absolute and relative position of the word in sentence

- **SCHUG features**:
  - ⋆ for a window of 3 SCHUG phrases (the respective phrase and 1 neighbouring phrase to the left and to the right): category, grammatical function (as in Table 2.6), and length (in words) of topmost encompassing SCHUG phrase (depth=0)
  - ⋆ absolute and relative position of the word within the topmost SCHUG phrase

- **chunk phrase features**:
  - ⋆ for a window of 3 chunk phrases (the respective phrase and 1 neighbouring phrase to the left and to the right): category and length (in words) of
    - ◇ topmost phrase (depth=0)
    - ◇ second-level phrase (depth=1)
  - ⋆ absolute and relative position of the word within the topmost and second-level chunk phrase.

[1] Calculation of all relative position features:
relative position = 100 × absolute position / (length of the stretch −1), so that the first and the last segment of a stretch receive relative position values of 0% and 100% respectively.

All features relating to part-of-speech, SCHUG and chunk phrases were automatically predicted (cf. discussion in Section 3.2.2). Whenever a feature was missing (e.g. because the first word of a sentence does not have a left neighbour), it received the value −100, which never occurred as regular value of any feature. Thus, it was not missing for CART, but contained usable information (e.g. about the position of a word).

Feature Selection and Classification Trees

For speaker kko/k61, the automatic feature selection resulted in a feature set consisting of only two features: *relative position between punctuation marks* and *word frequency*. For speaker rtd/k62, the selected feature set was even more reduced and consisted only of the feature *distance to the following punctuation in words*. This illustrates that prosodic phrasing in read speech depends mostly on punctuation.

The two trained classification trees are very simple (see Weka output in Figure 3.1 and 3.2), e.g. for speaker rtd/k62: Only if the word is followed by a punctuation mark is it followed by a boundary. The numbers given in parentheses after each leaf of the classification tree (first/second) indicate the total number of instances from the training set at the respective leaf (first) and the number of incorrectly classified instances at that leaf (second).

```
betweenpunctposition_rel <= 94: none (3967.0/174.0)
betweenpunctposition_rel > 94
|   CELEXfreq <= 1940: boundary (743.0/14.0)
|   CELEXfreq > 1940
|   |   CELEXfreq <= 2423: none (10.0/2.0)
|   |   CELEXfreq > 2423: boundary (30.0/2.0)
```

Figure 3.1.: Prosodic boundary classification tree for speaker kko/k61.

```
distancefollowingpunct <= 0: boundary (782.0/22.0)
distancefollowingpunct > 0: none (3968.0/235.0)
```

Figure 3.2.: Prosodic boundary classification tree for speaker rtd/k62.

Evaluation

Even though the trees are so simple, they have a fairly high accuracy of 95.96% (kko/k61) and 94.74% (rtd/k62), illustrating that prosodic phrase boundaries can be predicted fairly easily for read speech.

	F-measure		accuracy
	boundary	no boundary	
kko/k61	0.887	0.975	95.96%
rtd/k62	0.866	0.967	94.74%

Table 3.1.: 10-fold cross-validated performance measures for the prosodic boundary classification trees.

3.2.2. Accent and Intonation Contour Prediction

In the KCoRS, for each word its accentuation level is annotated, ranging from 0 to 3. If the word carries two accents, the accentuation level is specified separately for each accent, assuming that the accentuation level spreads to all following syllables in that word. Each accent is labelled in terms of location, type, alignment and upstep (see Section 2.2.4).

Only 5.2% of kko/k61's accents and 5.8% of rtd/k62's accents are upstepped. Preliminary tests showed that upstep could not be predicted from the available features (the trained classification trees were merely decision stumps that predicted "no upstep").

Accent type and alignment were treated as one by combining them to the following six complex accent types: flat, early peak, mid peak, late peak, early valley, and non-early valley.

In the KCoRS, three types of intonation contours are labelled: phrase-initial contours, concatenation contours, and phrase-final contours (see Section 2.2.4). Preliminary tests showed that phrase-initial contours depend very much on the type of accent they precede and the length of the pre-head, whereas concatenation contours depend on the types of the accents they concatenate. In order not to introduce too many extra errors in the symbolic prosody prediction, I decided not to train any models for phrase-initial and concatenation contours. In contrast, the last intonation contour of a prosodic phrase can be modelled without knowing the type of its preceding accent.

Therefore, accent and intonation contour prediction consists of four separate tasks:

- for each syllable: prediction of the accentuation level
- for each syllable: prediction whether it carries an accent or not (i.e. accent location)
- for each syllable carrying an accent: complex accent type
- for each syllable carrying the last accent of the prosodic phrase: phrase-final intonation contour.

All trained classification trees are far too big to be presented on paper (e.g. the classification tree for the accentuation level prediction of rtd/k62 has 1025 leaves), but they can be downloaded from my thesis web page[2].

Features

For each canonic syllable, the same features as for prosodic boundary prediction were used (see Section 3.2.1). In addition, the following features were extracted from the database:

- **syllable level features**:
  - ⋆ lexical stress
  - ⋆ syllable length (in canonic phonemes)
- **positional features**:
  - ⋆ absolute and relative position of the syllable in the word
  - ⋆ absolute and relative position of the syllable in the sentence

[2] http://www.brinckmann.de/KaRS/

  - ⋆ distance preceding and following prosodic boundary (in words and syllables)
  - ⋆ distance preceding and following pause (in words and syllables)
  - ⋆ relative position in prosodic phrase (in words and syllables)
  - ⋆ relative position in inter-pause stretch (in words and syllables)

- **sentence feature**: sentence length (in syllables).

All features relating to pauses and prosodic phrase boundaries are predicted by the respective pause and prosodic boundary models.

Feature Selection

Greedy forward feature selection was carried out for all four prediction tasks, separately for each speaker. Table 3.2 shows which features were selected automatically for the respective prediction task using only automatically predicted features (≈) vs. using manually corrected features (√) (cf. Section 1.4 for a general discussion about the use of automatically predicted vs. manually corrected features). As a general tendency concerning the use of syntactic phrase features, the prediction tasks with manually corrected features used a greater number of SCHUG features, whereas the prediction tasks with automatically predicted features used more chunk phrase features. For example, for the prediction of accentuation level, the grammatical function of a phrase was used only if it was manually corrected. This seems to support the statement in Section 2.3.1, namely that SCHUG needs further improvement before it can be successfully integrated into MARY.

In order to determine whether it is important to use features that are as correct as possible, the accuracy values of trained classification trees using only automatically predicted features vs. using manually corrected features were compared in a preliminary experiment. As shown in Table 3.3, in nearly all cases the classification trees trained with manually corrected features have a higher accuracy (determined by 10-fold cross-validation). For the prediction of accentuation level and accent location the differences in accuracy are only minor. However, it could be argued that an improvement of the pre-existing tools (TnT, SCHUG, and chunk tagger) and an improvement in the

	prediction task			
feature type	accentuation level	accent location	accent type	final contour
part of speech	√ ≈	√ ≈	√	
neighbour POS			≈	
word frequency	√ ≈	√ ≈		
neighbour word freq.	√ ≈	√	≈	√
following punctuation	√ ≈	√	√ ≈	√ ≈
sentence length	√ ≈	≈		√
sentence type		√ ≈	√	√ ≈
SCHUG category	√	√	√	
neighbour SCHUG cat.		≈		√
SCHUG grammatical function	√			
neighbour SCHUG gram.funct.			√	≈
SCHUG phrase length	√	√		
neighbour SCHUG length	√			
chunk phrase cat.	≈			
neighbour chunk phrase cat.	≈	≈	√ ≈	√ ≈
chunk phrase length		≈		√
neighbour chunk phrase length		≈		
lexical stress		√ ≈		
syllable length		≈	≈	
position in sentence	≈	≈		√
position between punctuation	≈			
position in inter-pause stretch	≈	≈	√ ≈	≈
position in prosodic phrase	√ ≈	√ ≈	√	√ ≈
position in SCHUG phrase	√	√ ≈		
position in chunk phrase				√ ≈
position in word		√ ≈	√ ≈	

Table 3.2.: Automatically selected feature types for the prediction of accentuation level, accent location, accent type, and phrase-final intonation contour. The features marked with √ are used for the prediction with manually corrected features. The features marked with ≈ are used for the prediction with automatically predicted features.

prediction of pauses and prosodic boundaries would have a positive effect on the accuracy of the prediction of accent types and phrase-final intonation contours. Nonetheless, for the training of the classification trees described in the following sections only automatically predicted features were used.

	prediction task			
	accentuation level	accent location	accent type	final contour
kko/k61	−0.2	0.2	1.4	0.7
rtd/k62	0.4	0.3	1.6	1.5

Table 3.3.: Differences in accuracy (in percentage points) between prediction tasks using manually corrected features vs. only automatically predicted features. A negative value means that the accuracy is higher if automatically predicted features are used.

Accentuation Level

Classification Trees The root node of both classification trees predicting accentuation level partitions the data according to word frequency (≤ 549 vs. > 549), followed by nodes concerning part of speech. This illustrates the fact that accentuation level is mainly determined by frequency factors: The more frequent a word, the lower its probability of containing accented syllables. The classification tree for kko/k61 ends with a leaf that assigns accentuation level 0 (unaccented) to all syllables in words with a frequency higher than 1253.

Evaluation 10-fold cross-validation led to accuracy values of 90.3% (kko/k61) and 86.6% (rtd/k62). Detailed confusion matrices are shown in Table 3.4. We can assign a cost matrix, so that the cost of a classification error is computed by the distance between the actual accentuation level and the predicted one. This reflects the amount of damage done by a wrong classification. For example, if the actual accentuation level is 1 (partially accented), but the model predicts 3, the cost is 2. The average cost is 0.142 for kko/k61 and 0.187 for rtd/k62. We can conclude that accentuation is easier to model for kko/k61 than for rtd/k62.

	kko/k61				**rtd/k62**			
actual	classified as							
accentuation	0	1	2	3	0	1	2	3
0	2718	70	168	3	2690	87	183	7
1	122	248	129	2	118	368	171	3
2	158	55	4071	6	182	107	3598	50
3	8	1	37	34	13	6	121	129

Table 3.4.: Confusion matrices for accentuation level prediction.

Accent Location

Classification Trees The root node in both classification trees for accent location prediction partitions the data according to lexical stress, so that only syllables with primary stress receive an accent. Closely following nodes concern part-of-speech, word frequency (only kko/k61), and relative position of the syllable within the word. For example, in rtd/k62's classification tree, most syllables with a relative position smaller than 80% within the word carry an accent, the others do not. This captures the fact that most function words are monosyllabic, the only syllable receiving a position of 100%. Finer distinctions are made by part-of-speech nodes further down the tree.

Evaluation Again, the classification tree for kko/k61 (accuracy 93.6%) performs slightly better than the tree for rtd/k62 (accuracy 92.1%, see Table 3.5).

Accent Type

Classification Trees The higher nodes in both classification trees predicting accent type partition the data according to following punctuation and distance to the following pause. This illustrates that accent type depends largely on positional features. Further down the tree, rtd/k62 relies mostly

	F-measure		accuracy
	accent	no accent	
kko/k61	0.893	0.954	93.6%
rtd/k62	0.873	0.943	92.1%

Table 3.5.: 10-fold cross-validated performance measures for the accent location classification trees.

on the feature "word frequency of the left neighbour", whereas kko/k61's tree uses the feature "part-of-speech of the right neighbour".

Evaluation Accuracy figures for the classification trees predicting accent type are rather low (kko/k61: 54.3%, rtd/k62: 58.1%), reflecting the difficulty of the task: six different accent types are to be predicted. However, for this prediction task the classification tree for rtd/k62 performs better than the tree for kko/k61.

F-measures are extremely low for "flat" (< 0.06), "early valley" (< 0.1), and (only for kko/k61) "non-early valley" (0.09). As shown in Table 3.6, the most common misclassification for flat accents and valleys are mid peaks and late peaks. If an early valley is misclassified as late peak, this error could be regarded as not so severe, e.g. an inexperienced human labeller could also make this mistake. Peaks are mostly misclassified as other peaks.

	kko/k61						**rtd/k62**					
actual	classified as											
accent type	fl	ep	mp	lp	ev	nev	fl	ep	mp	lp	ev	nev
flat	2	5	52	28	0	0	3	7	32	47	1	2
early peak	1	276	85	11	2	1	2	301	30	44	1	1
mid peak	4	122	488	250	10	7	4	75	345	343	7	18
late peak	5	6	241	506	10	8	6	1	165	687	4	2
early valley	0	2	62	58	8	13	2	1	46	80	8	9
non-early valley	1	2	44	39	13	6	2	2	26	37	3	44

Table 3.6.: Confusion matrices for the prediction of accent types (fl = flat, ep = early peak, mp = mid peak, lp = late peak, ev = early valley, nev = non-early valley).

Phrase-Final Intonation Contour

After manual inspection of the data, the phrase-final intonation contour classes were simplified by forming the following groups:

- lowrise: low rise, level-low rise, mid fall-low rise, and terminal fall-low rise
- highrise: high rise, level-high rise, mid fall-high rise, and terminal fall-high rise
- level: level and level-minimal rise
- midfall: mid fall and mid fall-minimal rise
- termfall: terminal fall and terminal fall-minimal rise.

Classification Trees The main features used in the classification tree for speaker kko/k61 are distance to the following prosodic boundary and simplified part-of-speech of the second right neighbour. This suggests that for speaker kko/k61 the position of the last accented syllable in the prosodic phrase is the most important factor to determine the phrase-final intonation contour.

For speaker rtd/k62 the main features are distance to the following pause, sentence type, and simplified following punctuation. The beginning of the classification tree shown in Figure 3.3 reveals that speaker rtd/k62 uses mostly terminal falls in statements, exclamations, and alternative questions, low rises in wh-questions and polite questions, and high rises in yes-no questions, declarative questions and negative questions. Therefore, it is worthwhile to distinguish between different question types, which is in line with the findings reported by Brinckmann & Benzmüller (1999).

Evaluation Overall accuracy of both trees is 81.5% (kko/k61) and 74.1% (rtd/k62). However, the F-measures for rtd/k62 are all above 0.5, except for mid fall, whereas for kko/k61 the F-measures of high rise, level, and mid fall

```
distancefollowingpause_words_auto <= 1
|   sentencetype = st: termfall (471.0/6.0)
|   sentencetype = ex: termfall (24.0/1.0)
|   sentencetype = wh: lowrise (46.0/22.0)
|   sentencetype = yn: highrise (30.0/10.0)
|   sentencetype = dq: highrise (4.0)
|   sentencetype = neg: highrise (5.0/1.0)
|   sentencetype = alt: termfall (2.0)
|   sentencetype = pol: lowrise (12.0/6.0)
distancefollowingpause_words_auto > 1
|   followingpunct_simple_word = comma
    [...]
```

Figure 3.3.: Beginning of rtd/k62's classification tree for predicting phrase-final intonation contour classes.

are all below 0.2. As can be seen in the confusion matrix for speaker kko/k61 in Table 3.7, the recall of highrise is 0 (i.e the classification tree never predicts a high rise) – most high rises are even misclassified as terminal falls.

	kko/k61					**rtd/k62**				
actual	classified as									
accentuation	lr	hr	lev	mf	tf	lr	hr	lev	mf	tf
lowrise	175	0	8	4	14	104	13	41	7	19
highrise	4	0	0	0	28	18	27	1	0	2
level	52	0	6	0	5	49	0	75	5	10
midfall	22	0	1	3	1	21	0	15	8	7
termfall	33	0	1	0	576	32	1	15	2	525

Table 3.7.: Confusion matrices for the prediction of phrase-final intonation contours.

3.3. Segmental Predictions

On the segmental (phonemic) level, we predict the features that are needed to generate input for MBROLA, namely

- realised phoneme (i.e prediction of postlexical phonological processes)

- duration
- median F0
- last F0.

3.3.1. Features

For the prediction of the segmental features, two different feature sets are used. The first one, called *Symbolic*, contains features relating to prosodic boundaries, accents, and phrase-final intonation contours. The second one, called *Direct*, does not contain any of those symbolic prosody features. As shown in Figure 4.1 (Section 4.1.2), the Direct prediction method leaves out the symbolic prosody prediction completely. This way, it loses some information, but it also reduces error accumulation. The two feature sets for Symbolic and Direct prediction are described in the following sections. No automatic feature selection was performed, because the datasets were too large, making automatic feature selection unfeasible in a reasonable amount of time.

Symbolic Feature Set

For each phoneme, the same features as for syllable-level symbolic prosody prediction (see Section 3.2.2) are used, except for SCHUG and chunk phrase features. In addition, the following features were extracted from the database:

- **phoneme level features**, for a window of 5 phonemes (the respective phoneme and 2 neighbouring phonemes to the left and to the right):
  - ⋆ phoneme identity
  - ⋆ phoneme type (vowel, consonant)
  - ⋆ consonant fortis/lenis (undefined, fortis, lenis)
  - ⋆ structural position in syllable (onset, nucleus, coda, ambisyllabic)

  - ⋆ number of phonemes in the same syllable structure position (not for neighbouring phonemes)

- **syllable level features**:
  - ⋆ accentuation level
  - ⋆ accent location (none, accent)
  - ⋆ distance to preceding and following accented syllable

- **accent group level features**:
  - ⋆ accent type
  - ⋆ accent type of following accent group
  - ⋆ phrase-final intonation contour ("none" for non-phrase-final accent groups).

An accent group was defined as a group of syllables consisting of one accented syllable and all following syllables up to, but not including, the next accented syllable. Syllables in pre-heads were defined as belonging to the following accent group.

The Symbolic feature set exists in two variants: The first one is used for the prediction of postlexical phonological processes and uses features of *canonic* phonemes and syllables. The second one is used for the prediction of duration, median F0, and last F0, and uses features of *realised* phonemes and syllables.

Direct Feature Set

For each phoneme, the same features as for syllable-level symbolic prosody prediction (see Section 3.2.2) are used, except for those features relating to prosodic boundaries. In addition, the following features were extracted from the database:

canonic phoneme level features, for a window of 5 canonic phonemes (the respective phoneme and 2 neighbouring phonemes to the left and to the right):

- ⋆ phoneme identity
- ⋆ phoneme type (vowel, consonant)
- ⋆ consonant fortis/lenis (undefined, fortis, lenis)
- ⋆ structural position in syllable (onset, nucleus, coda, ambisyllabic)
- ⋆ number of canonic phonemes in the same syllable structure position (not for neighbouring phonemes).

3.3.2. Prediction of Postlexical Phonological Processes

Glottalisation cannot be synthesised by the MBROLA synthesiser. Therefore, whenever a glottal stop was deleted, but left glottalisation behind, this deletion counted as replacement (marked with the glottalisation symbol `/q/`). This way it was possible to insert a glottal stop of 10ms during synthesis to mimic glottalisation (cf. Section 4.1.2).

The prediction of postlexical phonological processes (i.e. prediction of the realised phoneme) was carried out in two steps.

1. *Change*: In the first step, it was predicted whether the canonic phoneme was deleted, replaced, or left unchanged.

2. *Replacement*: In the second step, for all replaced phonemes a replacement rule was predicted.

Only certain "replacement rules" are possible, a canonic phoneme cannot be replaced by any other phoneme. All [canonic → realised] pairs that occur in the KCoRS were allowed as replacement rules, e.g. the replacement of `/E:/` with `/e:/` was accepted as replacement rule [`E:` → `e:`]. By predicting these rules instead of the realised phonemes, the prediction of impossible replacements because of data sparsity was prevented. For example, speaker kko/k61 always leaves a canonic `/Y/` unchanged. Therefore, CART had no

information regarding replacements rules for /Y/ (data sparsity). As a result, the trained classification tree assigned the replacement rule [Q → q], which is the most frequent replacement rule. Whenever necessary, these "impossible" replacements were ignored during prediction.

The prevalent features used in the trees predicting *change* and *replacement* are phoneme identity, features of phoneme neighbours, syllable length, structural position in the syllable, lexical stress, and accentuation level (for the Symbolic method only). The accuracy figures that are listed in Table 3.8 show that the Symbolic prediction is not always better than the Direct prediction.

	change		*replacement*	
	Symbolic	Direct	Symbolic	Direct
kko/k61	94.6%	93.2%	92.3%	92.0%
rtd/k62	92.8%	92.9%	94.0%	94.8%

Table 3.8.: Accuracy of the two tasks for the prediction of postlexical phonological changes.

3.3.3. Prediction of Acoustic Parameters

Wagon was used to train the regression trees for the prediction of the acoustic parameters duration, median F0, and last F0. Stop values and the size of the held-out validation set were determined in a trial-and-error procedure by comparing the evaluation measures RMSE and correlation coefficient (cc) on a separate test set. When the best settings had been determined, the whole dataset was used for the training of the final regression trees.

Duration Prediction

z-scores The only feature from the feature set that was not used for duration prediction is phoneme identity. The reason behind this is that every phoneme has a certain intrinsic duration which has a strong influence on the duration of the phoneme, e.g. tense vowels are longer than lax vowels, and fortis plosives are longer than lenis plosives. In order to factor out the influence of intrinsic duration, the absolute duration values were converted into z-scores, and the mean duration and standard deviation of each phoneme were stored in a separate file. The z-scores that are predicted by the regression trees can be converted back into absolute duration values by applying the following formula:

$$\text{absolute duration} = (z\text{-score} \times \text{stddev}) + \text{mean duration}$$

For the Symbolic prediction, the z-scores were computed on the realised phonemes, for the Direct prediction they were computed on the canonic phonemes.

Regression trees The Symbolic regression trees for duration prediction use the following features near the roots of the trees: positional features (position in prosodic phrase, neighbouring phonemes), accent location and type, lexical stress, syllable structure, and phoneme type. The Direct regression trees also rely heavily on positional features (neighbouring phonemes, following punctuation); in addition they use part-of-speech, word frequency, syllable length and structure, as well as lexical stress and phoneme type.

F0 Prediction

z-scores The raw F0 values were also transformed into z-scores, but not by using separate mean and stddev values for each phoneme. Instead, for each speaker the mean and stddev of the median F0 values was calculated. By predicting F0 values in terms of z-scores it is possible to use one regression tree for several voices. For Symbolic prediction, last F0 is the last F0 before a prosodic boundary. Since the Direct prediction uses no features about prosodic boundaries, in this case last F0 is the last F0 before a pause.

Regression trees The Symbolic regression trees for the prediction of median F0 use phrase-final intonation contour, positional features, accent type

and location, lexical stress, and syllable structure as topmost features. The Direct regression trees for median F0 prediction rely heavily on positional features; in addition they use lexical stress, part-of-speech, and word frequency.

The regression trees for the prediction of last F0 are compact enough, so that one of them is shown in Figure 3.4. It can be read as follows: The root node asks whether the word is followed by a question mark. If yes, the next question is about the distance of the preceding pause in words. If this distance is smaller than 8, the predicted z-score is 0.456 (the first value given at each leaf denotes the stddev of all instances of the training set at that leaf). If the distance is at least 8, then the next question is whether the sentence type is a wh-question. If yes, the predicted z-score is 0.495; all other questions have a last F0 z-score of 1.578 (i.e they end with a high intonation). Words that are not followed by a question mark follow the other branch of the root node. All predicted z-scores in this branch are negative, thus predicting a low F0 value.

Evaluation

Since wagon does not offer stratified cross-validation, the evaluation was carried out by dividing the dataset into a training set (90%) and a test set (10%). The evaluation measures listed in Table 3.9 show the performance of the regression trees on the test set. In terms of RMSE and cc, the Symbolic prediction is always better than its respective Direct counterpart. In the case of the Symbolic prediction, the evaluation on the test set uses the correct symbolic prosody features from the database (prosodic boundaries, accents, phrase-final intonation contours). Therefore, it is quite possible that the Symbolic prediction performs worse as soon as it is implemented in a TTS system, where it is faced with incorrectly predicted symbolic prosody features. Since the Direct method does not rely on any correct symbolic prosody features (it uses only automatically predicted features), the evaluation measures can be seen as fairly accurate predictors of its performance in a complete TTS system.


```
((followingpunct_word is quest)
 ((distanceprecedingpause_words_auto < 8)
  ((1.8746 0.456342))
  ((sentencetype is wh)
   ((1.4928 0.495193))
   ((1.42952 1.57833))))
 ((rightneighbour1_POS_auto is -100)
  ((lexicalstress is none)
   ((leftneighbour1_POS_auto is NN)
    ((syllpositioninword_abs < 2.3)
     ((0.207099 -1.83956))
     ((0.19063 -1.94317)))
    ((leftneighbour1_simplePOS_auto is pronoun)
     ((0.178767 -1.88098))
     ((toplevelSCHUGchunkcategory_auto is PP)
      ((0.160586 -2.01656))
      ((leftneighbour2_simplePOS_auto is verb)
       ((0.166037 -2.00491))
       ((simplePOS_auto is verb)
        ((0.217362 -1.91042))
        ((distanceprecedingpunct_inwords < 5.2)
         ((0.191442 -1.98462))
         ((0.177953 -1.93929)))))))))
   ((distanceprecedingpause_words_auto < 4)
    ((0.216729 -1.96712))
    ((syllablelengthinphonemes < 5.6)
     ((leftneighbour1_simplePOS_auto is adj)
      ((0.171697 -1.92763))
      ((leftneighbour1_CELEXfreq < 2528.7)
       ((syllablelengthinphonemes < 3.2)
        ((0.641384 -1.75926))
        ((0.39531 -1.69561)))
       ((0.190776 -1.90698))))
     ((0.934826 -1.61646)))))
  ((0.576984 -1.48369))))
```

Figure 3.4.: Regression tree (wagon output format) predicting last F0 z-score for speaker rtd/k62. At branching nodes the "yes"-branch is given first, followed by the "no"-branch. The first value at a leaf denotes the standard deviation, the second value is the mean (i.e. the predicted last F0 z-score). Negative z-scores denote F0 values below the speaker's mean, positive z-scores imply a high F0 value.

There is also an interesting difference between the two speakers: Duration prediction is better for kko/k61, whereas F0 prediction is better for speaker rtd/k62.

prediction task	evaluation measure	**kko/k61** Symbolic	**kko/k61** Direct	**rtd/k62** Symbolic	**rtd/k62** Direct
duration	RMSE	0.773	0.791	0.8441	0.882
	cc	0.612	0.594	0.572	0.528
median F0	RMSE	0.708	0.783	0.653	0.744
	cc	0.698	0.609	0.762	0.677
last F0	RMSE	1.307	1.487	0.666	1.010
	cc	0.543	0.350	0.895	0.712

Table 3.9.: Evaluation of the prediction of duration and F0 z-scores with regression trees trained on the Symbolic and the Direct datasets. The evaluation measures are root mean squared error (RMSE) and correlation coefficient (cc).

4. Perceptual Evaluation

In Chapter 3, the trained classification and regression trees were evaluated by comparing their predictions with the actual realisations in the KCoRS. The corpus-based evaluation measures RMSE and correlation coefficient allow us to compare different machine learning schemes or different datasets. For example, the F0 values of the female speaker rtd/k62 seem to be easier to predict than the F0 values of the male speaker kko/k61 (see Section 3.3.3). On the other hand, kko/k61's models for duration prediction are better than the ones for rtd/k62.

Especially for speech synthesis it is advisable to test the predictions of a model not only by comparing it to the realisations in a corpus, but also by measuring subjective listener preferences with perception experiments, for the following reasons:

1. It is unknown which of the following three is more/most important: a good F0 prediction, a good duration prediction, or a good prediction of postlexical phonological processes? And even if one synthesis system is superior to another one in all three respects, it is still possible that this difference cannot be perceived by listeners.

2. The corpus-based evaluation measures implicitly assume the realisations of one particular speaker as gold standard. However, usually there are several acceptable ways to produce an utterance. If the model commits an error in the prediction compared to the corpus, this "error" might be just as acceptable as the corpus realisation.

3. Listeners may have differing idiosyncratic preferences. For example, Portele (1997) and Brinckmann & Trouvain (2003) showed that one

group of listeners prefers the text-to-speech system to speak as "correctly" as possible, with no deviations from the canonic pronunciation, while the other group prefers the inclusion of some common segmental postlexical processes, such as schwa-deletion and assimilation of nasals.

4. Some listeners might even prefer a machine to sound unnatural, because they feel uncomfortable if they cannot tell whether they are communicating with a machine or with a human being.

In order to avoid implementing "improvements" to the TTS system that are not accepted by the listeners, one should therefore conduct a perception experiment.

The first perceptual evaluations of speech synthesis systems were intelligibility tests, e.g. by using semantically unpredictable sentences (SUS test; Benoît et al., 1996). This was worthwhile for the TTS systems at that time, because some were barely intelligible. Nowadays nearly all systems are clearly intelligible, so most perception experiments focus on naturalness, acceptance, or preference by asking the subjects to rate the synthesised stimuli on some scale or to compare two (or more) stimuli with each other.

Some experiments try to compare the systems more indirectly by giving the subjects a task (e.g. to follow the instructions produced by a TTS system) and measuring their reaction time or recording their gaze with an eye-tracker (Swift et al., 2002). If the subjects generally react faster when listening to the stimuli generated by one system, it is argued that this system is better than the others, which is certainly true for the respective task.

In my perception experiment, I followed the recommendations P.85 and P.800 by ITU-T[1] (International Telecommunication Union – Telecommunication Standardization Sector; ITU-T, 1994, 1996). These recommendations describe procedures for the perceptual evaluation of speech signals that have been agreed upon by the members of ITU-T (currently 359 institutions worldwide). They have been tested thoroughly and can be can be viewed as a

[1] `http://www.itu.int/ITU-T/`

standard, even though they are not used very often in the speech synthesis community.

4.1. Materials and Methods

4.1.1. General Procedure

Two of the methods described by the ITU-T recommendation P.800 (ITU-T, 1996) are Absolute Category Rating (ACR) and Comparison Category Rating (CCR). In the ACR procedure, the subjects are asked to judge the quality of each synthesised stimulus they hear using the following five-point scale:

5 excellent
4 good
3 fair
2 poor
1 bad

The mean of all scores (MOS = mean opinion score) is then calculated for each stimulus type.

According to ITU-T (1996), the ACR method tends to lead to low sensitivity in distinguishing among good quality TTS systems. A modified version of the ACR procedure, the CCR procedure, affords higher sensitivity. In the CCR procedure, the stimuli are presented to listeners by pairs (A-B) where A is a copy-synthesised original and B is synthesised by the systems to be compared. Some "null pairs" (A-A) are included to check the quality of anchoring. According to recommendation P.800, samples A and B should be separated by a pause of 500 to 1000ms duration. Since we cannot assume that A is always more acceptable than B, the order of the samples is chosen at random for each trial. On half of the trials, A is followed by B. On the remaining trials, the order is reversed. This way, it is also possible to examine the ratings of each subject for consistency. The subjects use the following

scale to judge the quality of the second sample relative to the quality of the first sample:

3	much better
2	better
1	slightly better
0	about the same
−1	slightly worse
−2	worse
−3	much worse

In effect, the subjects provide two judgements with one response: "Which sample has better quality?" and "By how much?". Simple averaging of the numerical scores should yield a mean score of approximately 0 for all conditions. It is necessary to recode the raw data: In those cases where the order of presentation is B-A, the sign of the numerical score must be reversed (i.e. $-1 \rightarrow 1$, $1 \rightarrow -1$). These recoded scores are used to compute CMOS (comparison mean opinion score). Thus, the results are presented in terms of the A-B order. Appropriate analyses of variance (ANOVA) and a posteriori Tukey HSD (Honestly Significant Difference) multiple comparison tests can be performed on the recoded scores. Because of the higher sensitivity, I chose the CCR method for my perception experiment. The specific set-up of the experiment (generation and presentation of stimuli, rating procedure, and group of subjects) is described in the following sections.

4.1.2. Stimuli

The 20 sentences listed in Table 4.1 were randomly selected from the KCoRS as synthesis sentences for the perception experiment. They had not been used as training, validation or test items for the classification and regression trees described in Chapter 3. The mean sentence length (in canonic syllables) is 14.5 (minimum: 5, maximum: 34).

All 20 test sentences were processed by MARY with no manual modifications (the phonemic pronunciation was examined for errors, but none were detected). MARY offers three female and four male MBROLA voices. For the perception experiment, I chose the two voices that were recorded for MARY's emotional synthesis (Schröder, 2004), named `de6` (male voice) and `de7` (female voice). In addition to those versions produced by the original

`be006`	Montag war es uns zu regnerisch.
`be038`	Die Ärzte sind damit gar nicht einverstanden.
`be074`	Vater mischt gleich die Karten.
`cn015`	Der gesuchte Weg erscheint auf dem Stadtplan in roten Leuchtpunkten, indem Sie auf die Taste mit dem entsprechenden Namen drücken.
`e026`	Gibt es eine Zugverbindung heute abend nach Frankfurt, und wenn ja, auf welchem Gleis fährt der Zug ab?
`e040`	Ich möchte am dreiundzwanzigsten zwölften nach Oldenburg fahren, und zwar möchte ich in Oldenburg früh sein, wenn möglich vor neun Uhr.
`e042`	Ja das ist zu früh.
`ko029`	Sie döst müde vor sich hin.
`ko039`	Das Kamel hat zwei Höcker.
`ko049`	Die Bejahung dieser Frage ist meine Bedingung für einen Neuanfang.
`mr007`	Wer weiß dort genau Bescheid?
`mr016`	Iss dein Essen nie hastig!
`mr018`	Bist Du sehr kalt geworden?
`mr040`	Sechs Mädchen wollen Schwester werden.
`mr088`	Einige Busse fahren heute später.
`s041`	Ich möchte in vierzehn Tagen von München über Hannover nach Hamburg fahren.
`s072`	Welchen Zug muß ich nehmen, um gegen zehn Uhr in Würzburg zu sein?
`s1017`	Achtlos wirft der Knirps Matsch durchs Eckfenster.
`s1040`	Nicht alle Menschen verkraften den Linksverkehr sofort.
`tk010`	Bei dieser Sachlage müssen wir die Hirschjagd aufschieben und uns kurz nach neun Uhr zurückmelden.

Table 4.1.: List of the 20 test sentences (with their respective ID in the KCoRS) for the perception experiment.

MARY system, three other generation methods were applied to each sentence for each voice: Copy-synthesised Originals, Direct Prediction, and Symbolic Prediction, which are all described in the following sections.[2] All stimuli were stored as 16-bit, 22050 Hz wav files.

Copy-synthesised Originals

In order to produce the copy-synthesised originals, the following features were extracted from the KCoRS and printed in the MBROLA format (cf. Section 1.1.3):

- realised phoneme
- for each realised phoneme: duration in ms
- for each realised sonorant and vowel: median F0 in Hz, placed at duration 50% in the phoneme
- for each last realised phoneme before a pause: last F0 in Hz.

Because of some MARY/MBROLA characteristics, the extracted features had to be changed in the following cases:

- MBROLA cannot synthesise glottalisation. So, whenever a glottal stop had been deleted in the original realisation and the following realised phoneme was glottalised, a glottal stop of 10ms was inserted in order to mimic glottalisation (or at least to make sure that some sort of juncture was audible).
- Neither of the chosen MBROLA voices distinguishes between plosive closure and release (there is only one symbol for each plosive). Therefore, all neighbouring plosive closures and releases were combined into one phoneme. Also, if the plosive closure had been deleted, but the release was still present, the symbol was changed into the MBROLA plosive symbol.

[2] All stimuli are available as sound files from `http://www.brinckmann.de/KaRS/`.

- In the KCoRS there are no phonemic labels for affricates; closure and release are labelled separately. After listening to some trial stimuli, I decided to combine all neighbouring `/t/` and `/s/` to the affricate `/ts/`.
- Since MBROLA voices do not offer `/6/`-diphthongs, these were divided up into the vowel (receiving 2/3 of the diphthong's duration, and placing the median F0 value at 75% of the vowel's duration) and `/6/`.

Based on the assumption that we are aiming for natural-sounding speech synthesis, these copy-synthesised originals constitute the upper limit of MBROLA, i.e. one cannot get any closer to natural read speech with the MBROLA diphone synthesis method. Phonetic vowel reductions and nasalisation cannot be captured at all, and glottalisation can only be mimicked very crudely. The plosive release cannot be modelled separately from the plosive closure, even though the plosive releases are deleted much more often than the closures (cf. Section 3.3.2), especially in consonant clusters. As informal inspection revealed, plosive release deletion is sometimes successfully captured by the respective diphone (especially by the diphones of the female voice `de7`).

Symbolic and Direct Prediction

Both Symbolic and Direct prediction are methods that use the classification and regression trees that were trained on the KCoRS database (as described in Chapter 3). Both methods use only automatically derived features as input. As shown in Figure 4.1, the Symbolic method predicts symbolic prosody features (prosodic boundaries, accentuation level, accents, and intonation contours) before predicting the MBROLA input features realised phoneme string, duration, F0 median, and last F0. The Direct method uses predicted pauses as the only additional feature for the prediction of the MBROLA features. In order to generate proper MBROLA input, the predicted features had to be changed in the same way as the ones of the copy-synthesised originals.

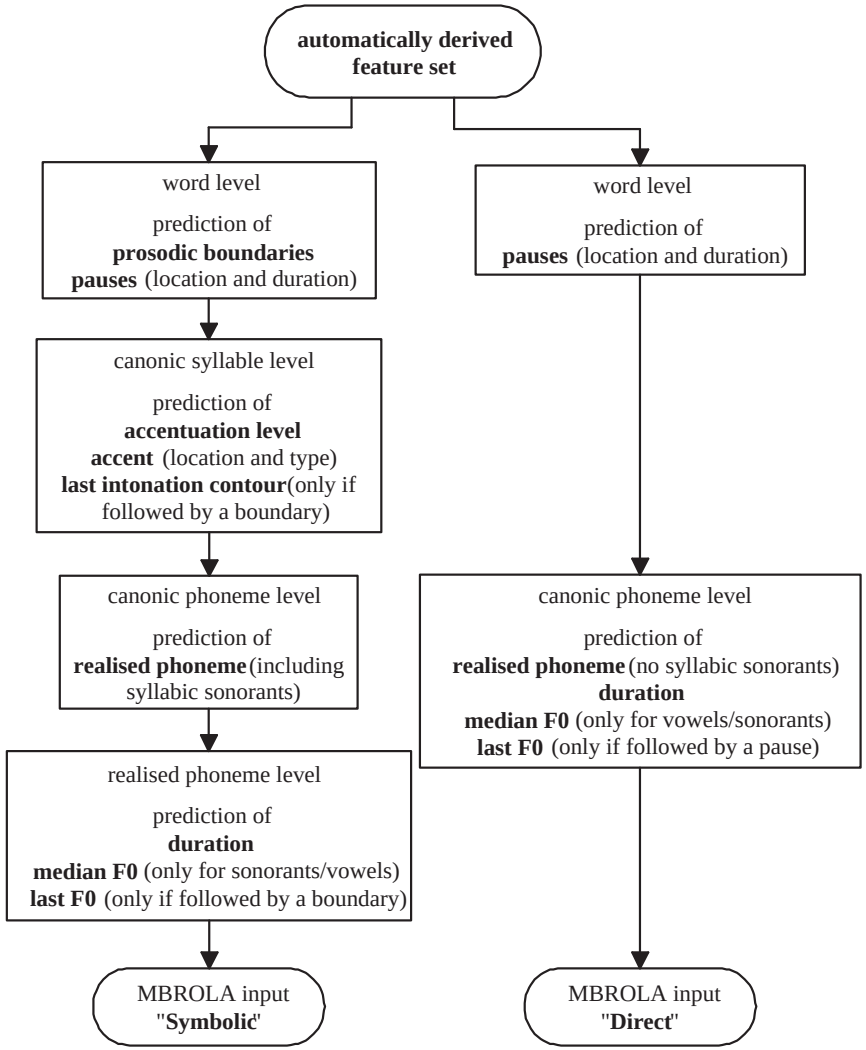

Figure 4.1.: Generation of MBROLA input for "Symbolic" and "Direct" stimuli.

Stimulus Pairs

For every sentence, the copy-synthesised sample (A) was paired with each of the automatically predicted samples (B), namely MARY, Direct, and Symbolic. A and B were separated by a pause of 800ms. In order to be able to examine the consistency of the subjects' ratings, both orders (A-B) and (B-A) were included in the experiment, resulting to a total of 120 ($20 \times 3 \times 2$) stimulus pairs.

In addition to these stimulus pairs, the sentence *Heute ist schönes Frühlingswetter.* was used to generate four pairs for the training section at the beginning of the experiment.

Four identical (A-A) and (B-B) pairs, where both samples were *exactly* the same, were also included. These identical pairs were used to examine whether the subjects were listening carefully.

All stimulus pairs were generated with the male voice `de6` and with the female voice `de7`. Since each stimulus pair had a mean duration of 7s, the experiment would have been longer than 30 minutes if each subject had to rate both voices. I regard 30 minutes as the maximum length for a perception experiment where the subjects have to listen carefully and remain very focussed on the task. Therefore, two separate experiments were set up: one with the female voice and one with the male voice.

4.1.3. Presentation

Before starting the perception experiment, the subjects were asked to fill in a questionnaire which asked for information regarding their age, sex, and the region of Germany they grew up in (dialectal background), as well as professional background and prior experience with speech synthesis (choosing "none", "little", "regular user", or "expert"). After the experiment, the subjects were asked for any comments.

The perception experiment itself was conducted with SCAPE (System for Computer-Aided Perception Experiments; Grabowski & Bauer, 2004),

a small, flexible program written in Java. The instructions for the subjects (see Table C.1 in Appendix C) were presented on screen, and the stimuli were presented via headphones. The subjects were instructed to listen carefully to both samples of each pair and to rate the overall quality of the second sample compared to the first one using the seven-point CCR scale by clicking on the respective radio button (see Figure 4.2). The subjects could listen to each stimulus pair only once, and as soon as the radio button was clicked, the next stimulus pair was presented. After rating the four training pairs, the subjects were prompted to ask any questions regarding the procedure of the experiment. After the training pairs and the prompt, all stimuli (including the identical pairs) were presented in a randomised order (with a different order for every subject).

Figure 4.2.: Screenshot of perception experiment with SCAPE.

SCAPE stores the following information for each presented stimulus pair:

- subject ID
- presentation number of the stimulus pair
- filename of the stimulus pair

- duration of the stimulus pair
- reaction time, measured from the beginning of the stimulus pair
- rating.

4.1.4. Subjects

32 subjects took part in the perception experiment. All are native German speakers, 26 of them being students or staff members of the Department of General Linguistics. Both synthesis voices were rated by an equal number of female and male listeners.

The ratings of each subject were screened for reaction time and consistency. A reaction time that is smaller than the duration of the stimulus pair means that the subject gave his or her rating before hearing the complete stimulus pair. Since every stimulus pair was presented twice in the experiment (A-B vs. B-A), the percentage of stimuli pairs that were rated similarly (both negative, both positive, or both 0) was taken as consistency measure.

Two subjects (neither had any prior experience with speech synthesis) had given more than 10 of their ratings before completely hearing the stimulus pair. Their consistency scores were also rather low (33.3% and 40%). I concluded that those two subjects had been unable to cope with the task and excluded their ratings from further analysis. Since these consistency analyses were conducted directly after each subject had completed the task, we were able to reassign the following subjects to new groups, ensuring that both synthesis voices were rated by an equal number of female and male listeners. The remaining 30 subjects were aged between 20 and 40 years (mean: 28 years).

The dialectal background of the subjects might have an influence on their preference of certain intonational patterns and segmental postlexical processes (e.g. concerning the replacement of `/E:/` by `/e:/`). Since the statistical models were trained on two speakers from Schleswig Holstein (Northern Germany), the subjects were grouped into "northern" (grown up in Schleswig-

Holstein, Hamburg, or Lower Saxony) and "other" (grown up in any other federal state).

Each subject is characterised by the following four features (number of subjects with that feature in parentheses):

- sex: male (14) vs. female (16)
- prior experience with speech synthesis: none or little experience (16) vs. regular user or expert (14)
- dialectal background: northern (8) vs. other (22)
- synthesis voice the subject had to rate: male (15) vs. female (15).

The distribution of all pairwise feature combinations among the subjects is listed in Table 4.2. A chi-square test revealed that unfortunately the features *dialectal background* and *prior experience* are not independently distributed among the subjects ($\chi^2 = 4.045$, $p < 0.05$). Only one of the subjects who grew up in Northern Germany is a regular user or expert, the other 7 have no or little experience with speech synthesis. In contrast, 59% of the subjects who grew up in another part of Germany are regular users or experts.

		sex		experience		synthesis voice	
		male	female	none/little	reg/exp	male	female
backgr.	northern	3	5	7	1	4	4
	other	11	11	9	13	11	11
sex	male			6	8	7	7
	female			10	6	8	8
exper.ce	none/little					9	7
	reg/exp					6	8

Table 4.2.: Absolute frequencies of pairwise feature combinations among the subjects of the perception experiment (reg/exp = regular user or expert of speech synthesis).

4.2. Results and Discussion

The significant differences and interactions described in the following sections were found by performing univariate analyses of variance (ANOVA) and post-hoc Tukey-HSD multiple comparisons with the statistical software SPSS 10. Correlations and their significance were analysed using Pearson correlation.

4.2.1. Subjects' Comments

The comments of the subjects are not only helpful for improving the procedure of the experiment, they also shed light on the reasons behind some of the ratings:

Scale One subject (with little experience) commented that the seven-point scale was too fine-grained for him, he would have preferred a three-point scale (better vs. equal vs. worse). On the other hand, another subject (expert) commented that she was very happy with the seven-point scale, which allowed her to make fine distinctions.

Pauses Some subjects found the pause between the two samples too short. One of these subjects found it rather stressful that the next stimulus was played automatically after he had placed his rating. Another subject found it hard to stay concentrated throughout the whole experiment and would have preferred an explicit pause after a block of 60 stimuli. Especially for naive subjects one should consider introducing longer pauses or allowing repeated playback.

Randomisation One subject complained that despite randomisation, sometimes the same sentence was repeated several times. Another subject even suspected that the order of stimuli depended on his ratings. If possible, the "randomisation" should be controlled, so that two neighbouring stimuli pairs always consist of different sentences.

Sentence length One subject (with little experience) commented that it was much easier for him to make a decision if the sentences were longer. This is in line with the generally lower scores for longer sentences (see Section 4.2.3) and the correlation between absolute scores and consistency (see Section 4.2.2): If a subject is unsure about his rating, he tends to give a rating that is close to 0.

Reaction time Two subjects confessed that they had placed their rating before listening to the end of the second sample whenever the samples differed so greatly that they had a very strong preference.

Sentence choice One subject complained that sentence `mr018` (*Bist Du sehr kalt geworden?*) was ungrammatical for her (she would have preferred *Ist Dir sehr kalt geworden?*).

MBROLA One female subject complained that the fundamental frequency of the male voice was sometimes too high, whereas one male subject found the low F0 of the female voice too low. This illustrates the limitations of MBROLA (and idiosyncratic preferences).

Dialectal preferences Several subjects with a Southern German dialectal background (raised in Saarland, Hessen, or Baden-Württemberg) complained that the female copy-synthesised sample of sentence `e042` (*Ja das ist zu früh.*) sounded perfectly natural, but very arrogant. Most of them said they had voted for the less natural sample, which sounded more friendly to them. In fact, as can be seen in Figure 4.9, the Direct and Symbolic samples of `e042` received even a positive CMOS (i.e. better than the copy-synthesised original). This illustrates the fact that a natural-sounding synthesis is not always the most accepted one.

4.2.2. Consistency

The mean percentage of similar ratings across all subjects is 61.1%, showing the difficulty of the task. One subject achieved only 31.7% similar ratings, whereas the most "consistent" subject had 85.0% similar ratings. 79.2% of all identical pairs were recognised (i.e. they were rated with 0), but only 46.7% of the subjects recognised all four identical pairs. The percentage of similar ratings of a subject and his or her recognition rate of identical pairs do not correlate significantly (correlation coefficient: 0.233).

Consistency (1=similar rating, 0=different rating) and absolute COS have a significant correlation coefficient of 0.335 [$p \leq 0.01$] over all stimuli, i.e. the more extreme the rating, the more consistent (see Figure 4.3). For example, if an item is rated with -3, it is very likely that the second presentation of the item is rated with a negative score as well.

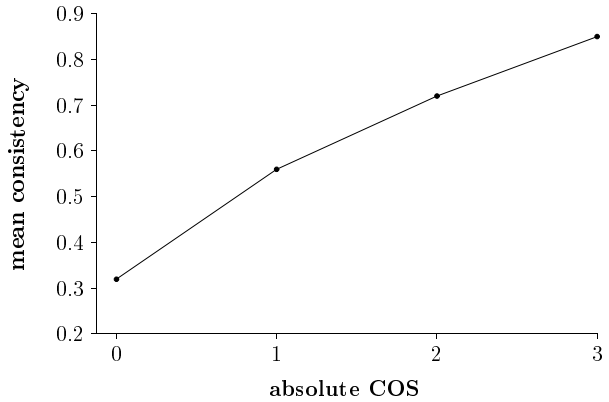

Figure 4.3.: Correlation between absolute COS and consistency of ratings.

An ANOVA revealed that the mean consistency (proportion of similar ratings across all (A-B)/(B-A) pairs) is significantly higher for MARY (0.75) than for Direct (0.54) and Symbolic (0.55) [$p \leq 0.005$]. This illustrates that MARY receives more extreme ratings and also suggest that subjects are rather unsure about their ratings of Direct and Symbolic.

4.2.3. Comparison Mean Opinion Score (CMOS)

Main Effects and Interactions

The mean overall CMOS (over both voices and all three synthesis methods) is −1.04. The following significant CMOS differences were found (by ANOVA and Tukey HSD):

- **synthesis method**: Symbolic (−0.76) ≈ Direct (−0.80) > MARY (−1.55) [$p < 0.001$]
- **synthesis voice**: female voice (−0.93) > male voice (−1.15) [$p < 0.001$]
- **prior experience with speech synthesis**: none/little (−0.98) > regular/expert (−1.11) [$p < 0.01$]
- **sex of listener**: male listener (−0.96) > female listener (−1.11) [$p \leq 0.001$]
- **dialectal background**: northern (−0.85) > other (−1.11) [$p < 0.001$]

Regarding CMOS, significant interactions were found for:

- synthesis voice and method [$p < 0.001$]
- synthesis voice, experience, and sex of listener (three-way interaction) [$p < 0.001$].

All main effects and interactions are described in detail in the following sections.

Synthesis Method Over all subjects and both synthesis voices, MARY receives significantly lower ratings than both Symbolic and Direct (which do not differ significantly). As shown in Figure 4.4, 24.6% of all MARY stimuli receive a COS (comparison opinion score) of −3, in contrast to only 9.3% Direct and 8.1% Symbolic stimuli. 15.4% of all MARY stimuli have a COS

of 0 or better, whereas 38.9% Direct and 39.4% Symbolic stimuli are rated having a similar or better quality than the copy-synthesised original.

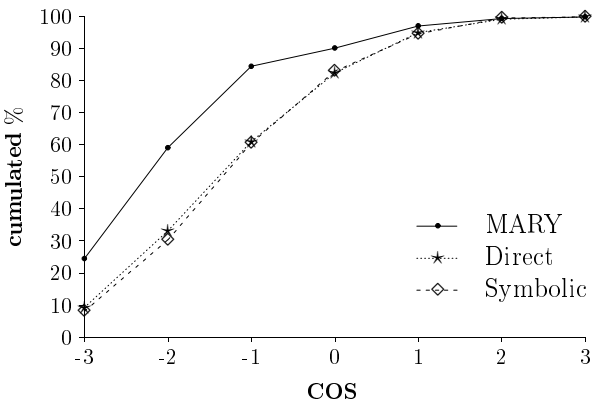

Figure 4.4.: COS cumulative distributions over both synthesis voices for the three synthesis methods MARY, Direct, and Symbolic. COS = 0 means that the stimulus was rated having the same overall quality as the copy-synthesised original, stimuli with a positive COS were rated having a better quality than the copy-synthesised original.

Synthesis Voice If the two synthesis voices are analysed separately, the same significant difference is observed for each voice: MARY receives significantly lower ratings than both Symbolic and Direct (which do not differ significantly). In addition, there is an interesting interaction between synthesis voice and method. As shown in Figure 4.5, both synthesis voices receive the same low CMOS for MARY (−1.55). For the Direct synthesis method, the male voice gets a lower CMOS (−0.88) than the female voice (−0.72), but this difference is not significant. For the Symbolic method, the CMOS of the male voice is significantly lower (−1.00) than the CMOS of the female voice (0.53) [$p \leq 0.005$]. The additional layer of symbolic prosody prediction seems to be slightly helpful only for the female voice.

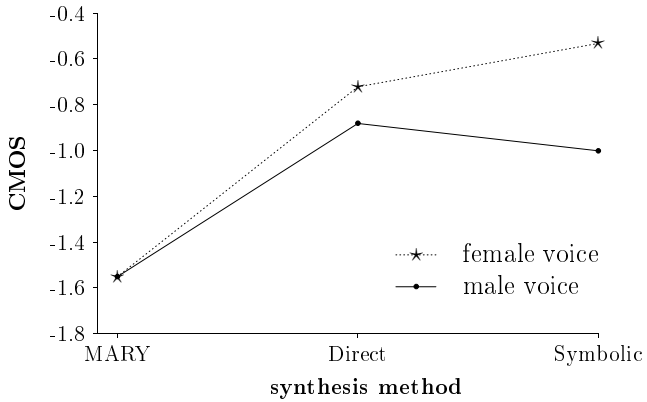

Figure 4.5.: Interaction between synthesis method and synthesis voice.

Prior experience with speech synthesis Subjects with regular/expert experience generally give lower ratings than subjects with no or little prior experience. This can be explained by the fact that through their prior experience with speech synthesis, regular/experts have a clear preference of what a TTS system should sound like, and they are able to hear finer differences. There is also an interesting interaction between synthesis method and prior experience [$p < 0.05$]: As shown in Figure 4.6, the CMOS of regular users and experts is especially low for MARY (-1.71) – more experienced TTS users expect the synthesis to sound more natural.

Sex of listener Looking at the CMOS of male and female listeners, we find that male listeners give significantly higher ratings (-0.96) than female listeners (-1.11) (this is true for all synthesis methods). However, there is a significant interaction between synthesis voice, experience, and sex of listener. As can be seen in Figure 4.7, the lowest ratings are given by "naive" female listeners (with no or little experience with speech synthesis) listening to the male voice. Naive female listeners and all male listeners prefer the female voice, whereas expert female listeners prefer the male voice. But since there are only two expert female listeners who listened to the male voice, these

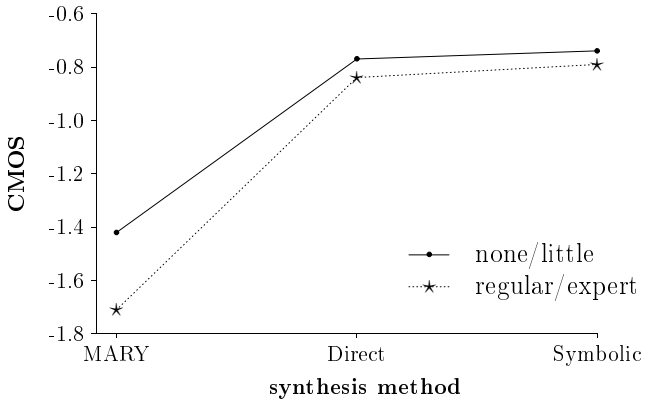

Figure 4.6.: Interaction between experience of the listener and synthesis method.

group results have to be treated with caution. In order to decide whether these interactions really reflect differences between groups, or whether they simply show idiosyncrasies of the subjects who just happen to belong to those groups, we need more subjects per group.

Dialectal background Subjects with a Northern German background give significantly higher ratings than subjects with a non-northern background. As mentioned in Section 4.1.4, the features *dialectal background* and *prior experience* are not independently distributed among the subjects. Therefore, we need further analyses to determine the cause of the higher CMOS of the Northern German subjects: Is it higher because they prefer the characteristics of their home dialect in a synthetic voice, or is it higher because they are more "naive" subjects, who generally give higher ratings? If the dialectal background of a subject has an influence on the ratings, this effect should only occur for those stimuli that were not generated with original MARY, because MARY was not trained on any corpus and produces Standard German output without any reductions. Figure 4.8 shows that this is not the case: the subjects with a Northern German background generally give more

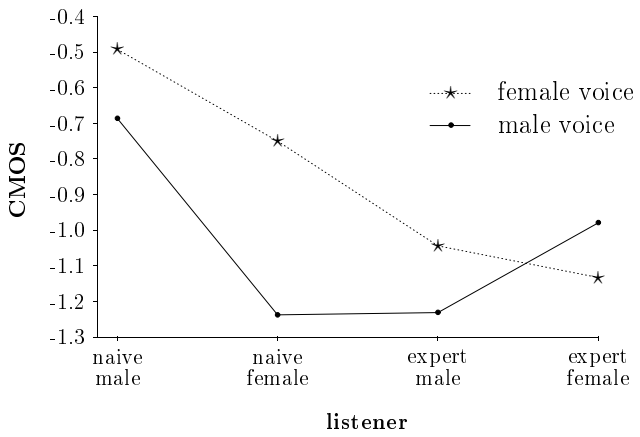

Figure 4.7.: Interactions between synthesis voice, experience, and sex of listener. "Naive" listeners are those with no or little experience with speech synthesis. "Expert" listeners are regular users or experts of speech synthesis.

positive ratings, no matter which synthesis method they are listening to. Therefore, the cause of the higher CMOS must be their inexperience with speech synthesis.

Single Sentences

A possible argument against using machine learning (ML) methods for prosody prediction is that even though the overall quality of ML-based synthesis systems is better than the quality of rule-based systems, ML-based systems show a greater variance, i.e. some sentences of ML-based systems sound excellent, whereas others sound very bad. It could be argued that rule-based systems might sound worse, but because they do so consistently, the user is not surprised by any sudden quality changes, leading to a higher acceptance.

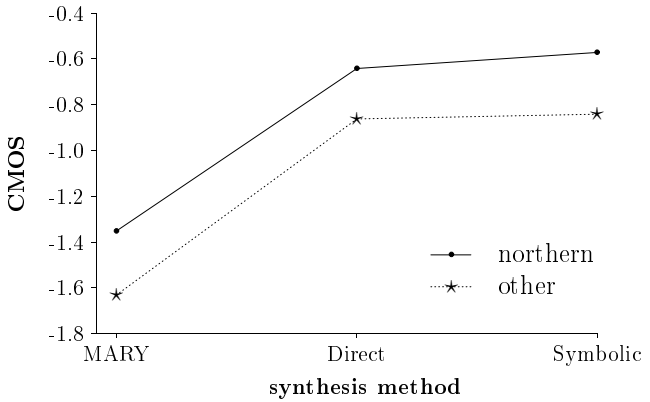

Figure 4.8.: Influence of the subjects' dialectal background on CMOS.

Sentence Length Across all stimuli, COS correlates negatively with sentence length, i.e. the longer the sentence, the lower the rating (correlation coefficient -0.149, $p \leq 0.01$), suggesting that listeners need longer sentences to make consistent decisions (cf. Section 4.2.1). The absolute value of the correlation coefficient is significantly lower for MARY (-0.098) than for Direct (-0.174) and Symbolic (-0.191). This could be explained by the fact that the KCoRS consists mostly of short sentences, so that both ML-based methods perform worse for longer sentences than for shorter ones, whereas MARY uses the same set of rules for every sentence.

Variance Figures 4.9 and 4.10 show the CMOS of each sentence separately for the female and the male voice. For the female voice, the variance of CMOS is lowest for MARY (MARY: 1.48, Direct: 1.81, Symbolic: 1.60). Nonetheless, the Symbolic method *always* receives higher ratings than MARY, suggesting that the Symbolic method should be the chosen for the female voice.

For the male voice, the ratings of MARY even have the highest variance of all three methods (MARY: 1.81, Direct: 1.74, Symbolic: 1.68). Compared to the Direct synthesis method, MARY is only better for sentence `cn015`, so that I would recommend using the Direct method for the male voice.

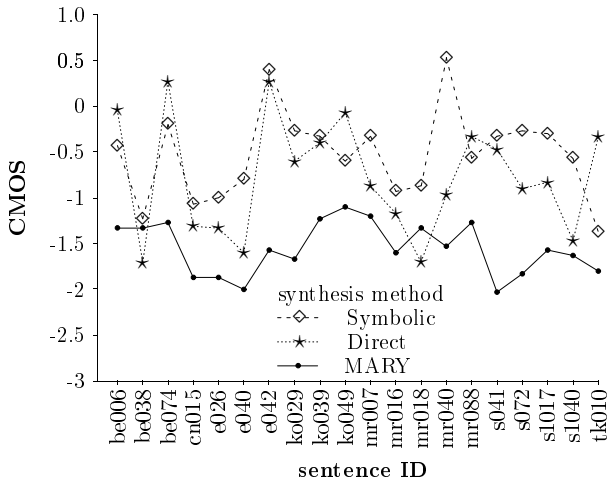

Figure 4.9.: CMOS for each sentence (female voice).

4.3. Conclusions

The perceptual evaluation showed that all three synthesis methods mostly receive negative scores. Even though there are some exceptions, one can generally assume the copy-synthesised originals as gold standard. It also showed that both ML-based methods (Symbolic and Direct) are superior to the original rule-based MARY method.

Comparing the two ML-based methods, I conclude that the symbolic level of prosody prediction can be safely skipped without obtaining a significantly lower CMOS. On the other hand, the inclusion of symbolic prosody prediction is not detrimental either. Therefore, the decision whether or not to include the symbolic level can be based entirely on the purpose of the synthesis system. If it is an instructional or research tool (such as MARY), one should include the symbolic prediction level, if it is just a "black box" for the user, one can use the Direct prediction method. If only one of the voices used in the present study was to be chosen, it should be the female voice `de7`, which generally received higher ratings.

As a general rule, the more experienced a TTS user, the higher his

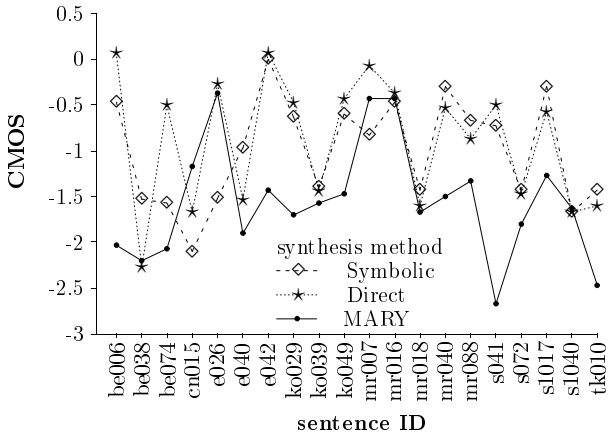

Figure 4.10.: CMOS for each sentence (male voice).

expectations regarding naturalness. If we aim for a wider usage of speech synthesis, it is necessary to improve it.

Finally, for a small follow-up study, the following procedure could be carried out to find out whether corpus-based and perceptual evaluation measures correlate: By comparing the synthesised stimuli with the original realisations, for each stimulus we could measure

- RMSE and correlation coefficient of duration values
- RMSE and correlation coefficient of median F0 values
- accuracy of predicted segmental changes.

These corpus-based evaluation measures could then be directly compared with the perceptual ratings. The results might also shed light on the question which of the three parameters – duration, F0, or postlexical phonological processes – is most important.

Conclusion and Outlook

The perceptual evaluation shows that the output of a text-to-speech system can be significantly improved by training all models that contribute to prosody prediction on the same database, namely the 'Kiel Corpus of Read Speech', which was enriched with additional features. More importantly, it shows that the error introduced by symbolic prosody prediction perceptually equals the amount of error produced by the direct method that does not exploit any symbolic prosody features.

More time and effort could be spent introducing other features and trying out different machine learning and feature selection methods. However, I doubt whether the resulting models would lead to a perceptually improved output. I think that the limitations of the KCoRS and MBROLA have been reached with the presented approach.

One major drawback of the KCoRS is its textual material consisting almost entirely of isolated sentences. In order to model prosodic properties of longer texts, we need a corpus of read newspaper texts or radio news. The available speech corpora in that domain (IMS German Radio News Corpus, S1000P, MULI; cf. Section 2.3.3) are not completely labelled with segmental and prosodic information. Therefore, a possible approach would be to extend the annotations of these corpora.

Instead of using the MBROLA diphone synthesiser, an even more promising approach is to try a different synthesis method, namely non-uniform unit selection, which generally produces more natural sounding output. The speech material in the KCoRS, which is not more than half an hour of speech per speaker, is not sufficient for a reliable non-uniform unit selection speech synthesiser (cf. Brinckmann, 1997). To my knowledge, there

exists no publicly available German database with two or more hours of labelled speech per speaker so far. Therefore, it would be worthwhile producing such a large labelled speech corpus. With this corpus of read speech, one could also include breathing pauses occurring in read speech, making the generated output sound more natural.

Breathing is not the only "noise" in natural speech. Campbell (2004) reported that in a large database of daily conversational speech (the 'Expressive Speech Processing' corpus) grunts and other noises are remarkably frequent. Instead of clear emotional states (such as happiness, sadness, anger, and fear), a great variety of different speaking styles is present, which express attitudes and interpersonal relationships.

I think that the challenge for the next years is to move onward from "reading machines" to truly conversational speech synthesis, which could be used in a dialogue system or as an aid for vocally disabled persons. As Campbell (2004) argues very convincingly, in order to achieve this long-term goal, we will have to move away from text-based synthesis by using a large database of naturally occurring conversational speech, which remains to be built for German.

A. PROLAB

In Table A.1, Table A.2, and Table A.3 all PROLAB labels used in the KCoRS are listed and described. Additionally, the absolute frequency of each label is given for the complete KCoRS, speaker kko/k61, and speaker rtd/k62.

A.1. Accent and alignment labels

		absolute frequencies		
PROLAB	**description**	**KCoRS**	**kko/k61**	**rtd/k62**
accentuation level: unaccented				
#&0	unaccented	15775	2484	2455
#&%0	uncertain accentuation level	106	4	7
accentuation level: partially accented				
#&1-	flat	390	51	65
$&1-	flat within a word	2	0	0
#&%1-	flat, uncertain accentuation level	6	0	2
#&1^	mid peak	780	164	180
$&1^	mid peak within a word	6	2	0
#&|1^	mid peak with upstep	6	1	0
#&%1^	mid peak, uncertain accentuation level	8	0	2

continued on next page

PROLAB	description	absolute frequencies KCoRS	kko/k61	rtd/k62
#&1^%	mid peak, uncertain alignment	5	0	0
#&1)	early peak	167	23	15
$&1)	early peak within a word	3	0	1
#&%1)	early peak, uncertain accentuation level	6	0	0
#&1)%	early peak, uncertain alignment	1	0	0
#&1(	late peak	411	36	73
$&1(	late peak within a word	7	1	1
#&|1(	late peak with upstep	4	0	2
#&%1(	late peak, uncertain accentuation level	7	2	0
#&1(%	late peak, uncertain alignment	8	0	0
#&1]	early valley	78	11	7
$&1]	early valley within a word	1	0	0
#&%1]	early valley, uncertain accentuation level	1	0	0
#&1]%	early valley, uncertain alignment	1	0	0
#&1[	non-early valley	129	5	36
$&1[	non-early valley within a word	2	0	0
#&1[%	non-early valley, uncertain alignment	4	1	1
$&1[%	non-early valley within a word, uncertain alignment	1	0	0
accentuation level: accented				
#&2-	flat	253	37	26
#&|2-	flat with upstep	3	0	0
#&%2-	flat, uncertain accentuation level	4	1	0
#&2-%	uncertain flat	1	0	0

continued on next page

PROLAB	description	absolute frequencies KCoRS	kko/k61	rtd/k62
#&2^	mid peak	3539	623	453
$&2^	mid peak within a word	6	0	0
#&|2^	mid peak with upstep	373	75	67
#&%2^	mid peak, uncertain accentuation level	18	1	3
#&%|2^	mid peak with uncertain upstep	1	0	1
#&|%2^	mid peak with upstep, uncertain accentuation level	1	0	0
#&2^%	mid peak, uncertain alignment	65	11	8
#&|2^%	mid peak with upstep, uncertain alignment	2	1	0
#&2)	early peak	2503	357	357
$&2)	early peak within a word	2	1	1
#&|2)	early peak with upstep	25	1	7
#&%2)	early peak, uncertain accentuation level	1	0	0
#&2)%	early peak, uncertain alignment	44	6	9
#&2(	late peak	4294	709	724
$&2(	late peak within a word	9	3	2
#&|2(	late peak with upstep	302	45	63
#&%2(	late peak, uncertain accentuation level	10	1	2
#&2(%	late peak, uncertain alignment	30	3	7
#&|2(%	late peak with upstep, uncertain alignment	2	1	0
#&2]	early valley	857	133	141
$&2]	early valley within a word	4	1	1
#&|2]	early valley with upstep	9	1	0
#&%2]	early valley, uncertain accentuation level	3	0	2
#&2]%	early valley, uncertain alignment	16	2	2

continued on next page

PROLAB	description	absolute frequencies KCoRS	kko/k61	rtd/k62
#&|2]%	early valley with upstep, uncertain alignment	1	0	0
#&2[	non-early valley	606	98	79
$&2[	non-early valley within a word	4	2	1
#&|2[	non-early valley with upstep	12	2	1
#&%2[	non-early valley, uncertain accentuation level	3	2	0
#&2[%	non-early valley, uncertain alignment	8	1	0
accentuation level: reinforced				
#&3^	mid peak	369	32	105
$&3^	mid peak within a word	1	0	1
#&|3^	mid peak with upstep	5	1	2
#&3^%	mid peak, uncertain alignment	2	0	0
#&3)	early peak	11	0	4
#&3(	late peak	107	6	20
#&|3(	late peak with upstep	1	0	0
#&3(%	late peak, uncertain alignment	3	0	2
#&3]	early valley	3	0	1
#&3[	non-early valley	3	0	1

Table A.1.: PROLAB pitch accent and alignment labels used in the KCoRS with the absolute frequency of occurrence for the complete KCoRS, and the speakers kko/k61 and rtd/k62.

A.2. Intonation contour labels

PROLAB	description	absolute frequencies KCoRS	kko/k61	rtd/k62
concatenation and phrase-final contours				
#&,	low rise	1428	217	222
$&,	low rise within a word	1	1	0
#&?	high rise	257	34	42
#&0.	level	3810	567	678
$&0.	level within a word	7	1	2
#&%0.	uncertain level	5	0	1
#&0;	level - minimal rise	1	0	0
#&0.,	level - low rise	1	0	1
#&0.?	level - high rise	2	0	2
#&1.	mid fall	5218	874	824
$&1.	mid fall within a word	30	5	5
#&%1.	uncertain mid fall	8	0	0
#&1;	mid fall - minimal rise	20	0	4
#&1.,	mid fall - low rise	54	8	5
#&1.?	mid fall - high rise	1	0	0
#&2.	terminal fall	4335	748	654
$&2.	terminal fall within a word	8	3	1
#&%2.	uncertain terminal fall	13	0	0
#&2;	terminal fall - minimal rise	284	0	20
#&2.,	terminal fall - low rise	50	4	22
#&2.?	terminal fall - high rise	6	0	6
phrase-initial contours				
#&HP2	high-falling pre-head	26	3	1
#&HP1	high-level pre-head	466	36	49
#&HP2%	uncertain high-falling pre-head	2	1	0
#&HP1%	uncertain high-level pre-head	1	0	0

continued on next page

PROLAB	description	absolute frequencies KCoRS	kko/k61	rtd/k62
#&HP1^	upstepped high-level prehead	1	1	0

Table A.2.: PROLAB intonation contour labels used in the KCoRS with the absolute frequency of occurrence for the complete KCoRS, and the speakers kko/k61 and rtd/k62.

A.3. Prosodic phrase boundaries, register, and speech rate labels

PROLAB	description	absolute frequencies KCoRS	kko/k61	rtd/k62
prosodic phrase boundaries				
#&PGn	with reset	6038	908	954
#&=PGn	without reset	423	56	62
#&%PGn	uncertain boundary with reset	20	1	7
#&%=PGn	uncertain boundary without reset	5	0	0
register				
#&HR	high register	28	2	0
#&LR	low register	21	2	0
speech rate				
#&RP	increased speech rate	1	0	0
#&RM	decreased speech rate	1	0	0

Table A.3.: PROLAB prosodic phrase boundary, register and speech rate labels used in the KCoRS with the absolute frequency of occurrence for the complete KCoRS, and the speakers kko/k61 and rtd/k62.

B. Syntactic Features

B.1. STTS part-of-speech tagset

In Table B.1 all part-of-speech tags of the Stuttgart-Tübingen Tag Set (STTS) are described. Additionally, the absolute frequency of each tag (i.e. number of words with that tag) in the KCoRS is given.

POS	freq	description	example
ADJA	176	attributive adjective	schönes [Frühlingswetter], [den] elften [Dezember]
ADJD	144	predicative or adverbial adjective	[es war] regnerisch, länger [schlafen]
ADV	409	adverb	gestern, jetzt
APPR	490	preposition or left part of circumposition	in, durch, auf
APPRART	75	preposition with article	im, am, zum
APPO	0	postposition	[ihm] zufolge
APZR	5	right part of circumposition	[von dort] aus
ART	451	article	den, einen
CARD	97	cardinal number	zehn, siebzehn
FM	0	material of a foreign language	a big fish
ITJ	5	interjection	naja, na
KOKOM	5	comparative conjunction	wie, als

continued on next page

continued from previous page

POS	freq	description	example
KON	108	coordinating conjunction	und, oder, aber
KOUI	4	subordinating conjunction with *zu* and infinitive	ohne [sich zu schämen], um [noch etwas zu erhalten]
KOUS	36	subordinating conjunction with a sentence	
NE	288	proper noun	Berlin, Erna
NN	1019	common noun	Kuchen, Hunger, Vater
PDAT	28	attributive demonstrative pronoun	diese [Drängelei]
PDS	15	substituting demonstrative pronoun	das [paßt]
PIAT	16	attributive indefinite pronoun that cannnot be preceded by a determiner	keine [Scheu], mehrere [Tage]
PIDAT	32	attributive indefinite pronoun that can be preceded or followed by a determiner	[von] beiden [Zügen]
PIS	29	substituting indefinite pronoun	man, keiner
PPER	322	irreflexive personal pronoun	ich, es, ihr
PPOSAT	47	attributive possessive pronoun	seine [zweite Chinareise]
PPOSS	0	substituting possessive pronoun	meins, deiner
PRELAT	0	attributive relative pronoun	[der Mann,] dessen [Hund]
PRELS	16	substituting relative pronoun	[ein Wanderer,] der [in einen warmen Mantel gehüllt war]

continued on next page

continued from previous page

POS	freq	description	example
PRF	25	reflexive personal pronoun	[Du bewirbst] dich
PROAV	10	pronominal adverb	danach, trotzdem, deshalb, demgemäß
PTKA	11	particle with adjective or adverb	am [schnellsten], zu [regnerisch]
PTKANT	19	answer particle	ja, nein, danke
PTKNEG	33	negation particle	nicht
PTKVZ	50	separated verbal particle	[auf welchem Gleis fahren die Züge] ab
PTKZU	5	*zu* before an infinitive	[ohne sich] zu [schämen]
PWAT	18	attributive interrogative pronoun	welche [Züge]
PWAV	47	adverbial interrogative or relative pronoun	wann, wie, wo, wobei
PWS	11	substituting interrogative pronoun	wer, was
TRUNC	0	first (separated) part of composition	An- [und Abreise]
VAFIN	130	finite auxiliary	ist, habe, hätte
VMFIN	152	finite modal	[dann] kann [ich], [wir] wollen
VVFIN	380	finite content verb	[alle] eilen, [Zug] endet [hier]
VAIMP	1	auxiliary imperative	sei [gewarnt]
VVIMP	24	content verb imperative	achte [auf die Autos]
VAINF	18	auxiliary infinitive	sein, haben, werden
VMINF	1	modal infinitive	[man hatte lesen] können
VVINF	146	content verb infinitive	[Mutter konnte länger] schlafen
VAPP	4	auxiliary past participle	geworden
VMPP	0	modal past participle	[er hat es] gekonnt
VVPP	27	content verb past participle	[wurde] eröffnet, [hat] angetreten

continued on next page

continued from previous page

POS	freq	description	example
VVIZU	3	content verb infinitive with incorporated *zu*	anzustellen, abzunehmen
XY	0	non-word, containing special characters	D2XW3
$,	174	comma	,
$.	633	sentence final punctuation mark	. ? ! :
$(	8	other punctuation mark	- “

Table B.1.: STTS part-of-speech tagset with the absolute frequency (i.e. number of tokens) of each part-of-speech tag in the KCoRS

B.2. Syntactic Chunk Phrases

	absolute frequency					
category	d=0	d=1	d=2	d=3	d=4	Σ
AP	127	9	4	–	–	140
AdvP	308	36	10	–	–	354
NP	992	109	26	2	1	1130
PP	374	139	39	2	3	557
SUBORD_CLAUSE	49	18	2	1	–	70
VG	809	51	19	2	1	882
W	135	48	8	–	–	191
Σ	2794	410	98	7	5	3324

Table B.2.: Frequency of SCHUG categories in the textual material of the KCoRS. d gives the level of embedding, i.e. a syntactic phrase (or word) with d=0 is a top-level phrase, which is not embedded in any other phrase.

		absolute frequency				
category	description	d=0	d=1	d=2	d=3	Σ
AA	superlative phrase with *am*	3	–	–	–	3
AP	adjective phrase	27	12	1	2	42
AVP	adverbial phrase	23	5	3	–	31
CAC	coordinated adpositions	–	–	–	–	–
CAP	coordinated adjective phrase	2	6	1	1	10
CAVP	coordinated adverbial phrase	2	2	–	–	4
CCP	coordinated complementiser	–	–	–	–	–
CNP	coordinated noun phrase	14	15	2	1	32
CO	coordinated different categories	2	–	–	–	2
CPP	coordinated adpositional phrase	1	–	–	–	1
CVP	coordinated verb phrase	1	–	–	–	1
CVZ	coordinated *zu*-infinitive	–	–	–	–	–
ISU	idiosyncratic unit	–	–	–	–	–
MPN	multi-word proper noun	3	–	–	–	3
MTA	multi-token adjective	–	–	–	–	–
NM	multi-token number	1	–	–	1	2
NP	noun phrase	517	41	10	–	568
PP	adpositional phrase	441	105	20	2	568
QL	quasi-language	–	–	–	–	–
VZ	*zu*-marked infinitive	5	–	–	–	5
	Σ	1041	186	37	7	1271

Table B.3.: Frequency of phrasal chunk tags assigned with the chunk tagger to the textual material of the KCoRS. d gives the level of embedding, i.e. a phrase with d=0 is a top-level phrase, which is not embedded in any other phrase.

C. Perception Experiment

The instructions for the subjects of the perception experiment were presented on screen and read as follows:

Du nimmst an einem Experiment zur subjektiven Bewertung von **Sprachsynthesemethoden** teil.
In diesem Experiment wirst Du **paarweise Varianten von Äußerungen** hören, die mit verschiedenen Sprachsynthesemethoden erzeugt wurden.
Du hörst jeweils eine Variante, gefolgt von einer kurzen Pause und einer zweiten Variante. Bitte höre Dir beide Varianten sorgfältig an und **beurteile die zweite Variante** im Vergleich zur ersten Variante mit Hilfe der folgenden Skala:
Die **zweite** Variante ist, verglichen mit der ersten Variante, viel besser besser etwas besser ungefähr gleich etwas schlechter schlechter viel schlechter.
Bei der Bewertung geht es um Deinen **persönlichen Gesamteindruck**.
Wir werden mit 4 Übungsbeispielen beginnen, damit Du Dich an die Testprozedur gewöhnen und die Lautstärke so einstellen kannst, wie sie Dir angenehm ist. Nach den Übungsbeispielen kannst Du eine Pause einlegen, um Fragen zum Ablauf des Experiments zu stellen, falls Du irgendwelche Probleme hast.
Das Experiment dauert ungefähr 30 Minuten.
Vielen Dank für Deine Teilnahme! :–)

Table C.1.: Instructions for the subjects of the perception experiment.

Bibliography

Anderson, M., Pierrehumbert, J. & Liberman, M. (1984). Synthesis by rule of English intonation patterns. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 84)*, vol. 1. San Diego, USA, 2.8.1–2.8.4.

Atterer, M. & Schulte im Walde, S. (2004). A PCFG for Prosodic Structure. Experiments on German. In: *Proceedings of Speech Prosody 2004*. Nara, Japan, 505–508.

Baayen, R.H., Piepenbrock, R. & Gulikers, L. (1995). The CELEX lexical database (release 2). CD-ROM, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, USA [Distributor].
URL `http://www.kun.nl/celex/`

Barry, W.J. & Fourcin, A.J. (1992). Levels of labelling. *Computer Speech and Language* **6**, 1–14.

Baumann, S., Brinckmann, C., Hansen-Schirra, S., Kruijff, G.J., Kruijff-Korbayová, I., Neumann, S., Steiner, E., Teich, E. & Uszkoreit, H. (2004). The MULI project: Annotation and analysis of information structure in German and English. In: *Proceedings 4th International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, Portugal, 1489–1492.

Bennett, C.L. & Black, A.W. (2003). Using acoustic models to choose pronunciation variations for synthetic voices. In: *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*. Geneva, Switzerland, 2937–2940.

Benoît, C., Grice, M. & Hazan, V. (1996). The SUS test: a method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. *Speech Communication* **18**(4), 381–392.

Black, A.W. & Hunt, A.J. (1996). Generating F0 contours from ToBI labels using linear regression. In: *Proceedings of the International Conference*

on Spoken Language Processing (ICSLP 96), vol. 3. Philadelphia, USA, 1385–1388.

Black, A.W. & Taylor, P. (1994). CHATR: a generic speech synthesis system. In: *Proceedings of COLING-94*, vol. 2. Kyoto, Japan, 983–986.

Brants, T. (2000). TnT – a statistical part-of-speech tagger. In: *Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLP-2000)*. Seattle, USA.
URL `http://www.coli.uni-sb.de/~thorsten/tnt/`

Brants, T., Skut, W. & Uszkoreit, H. (1999). Syntactic annotation of a German newspaper corpus. In: *Proceedings of the ATALA Treebank Workshop*. Paris, France, 69–76.
URL `http://www.coli.uni-sb.de/sfb378/negra-corpus/`

Braunschweiler, N. (2003). ProsAlign – the automatic prosodic aligner. In: *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 2003)*. Barcelona, Spain, 3093–3096.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). *Classification and Regression Trees*. Pacific Grove, USA: Wadsworth & Brooks.

Breuer, S. (2000). Reduktionsanalyse mit CART. In: *Tagungsband 11. Konferenz Elektronische Sprachsignalverarbeitung (ESSV 2000)*. Cottbus, Germany.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics* **21**(4), 543–565.

Brinckmann, C. (1997). German in eight weeks – a crash course for CHATR. Tech. Rep. TR-IT-0236, ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan.

Brinckmann, C. & Benzmüller, R. (1999). The relationship between utterance type and F0 contour in German. In: *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH '99)*, vol. 1. Budapest, Hungary, 21–24.

Brinckmann, C. & Trouvain, J. (2003). The role of duration models and symbolic representation for timing in synthetic speech. *International Journal of Speech Technology* **6**(1), 21–31.

Campbell, N. (2004). Getting to the heart of the matter; speech is more than just the expression of text or language. In: *Proceedings 4th International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, Portugal.

Cohen, A., Collier, R. & 't Hart, J. (1982). Declination: Construct or intrinsic feature of speech pitch? *Phonetica* **39**, 254–273.

Cohen, W.W. (1995). Fast effective rule induction. In: *Proceedings of the Twelfth International Conference on Machine Learning (ICML 1995)*. Tahoe City, California, USA, 115–123.

Declerck, T. (2002). A set of tools for integrating linguistic and non-linguistic information. In: *Proceedings of Semantic Authoring, Annotation & Knowledge Markup (SAAKM 2002)*. Lyon, France.

Dusterhoff, K. & Black, A.W. (1997). Generating F0 contours for speech synthesis using the Tilt intonation theory. In: *Proceedings of ESCA Tutorial and Research Workshop on Intonation: Theory, Models and Applications*. Athens, Greece, 107–110.

Dutoit, T. (1997). *An Introduction to Text-to-Speech Synthesis*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F. & van der Vrecken, O. (1996). The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In: *Proceedings of the Fourth International Conference on Spoken Language Processing (ICSLP 96)*, vol. 3. Philadelphia, USA, 1393–1396.

Fackrell, J., Vereecken, H., Grover, C., Martens, J.P. & Van Coile, B. (2001). Corpus-based development of prosodic models across six languages. In: E. Keller, G. Bailly, A. Monaghan, J. Terken & M. Huckvale (eds.) *Improvements in Speech Synthesis*, chap. 17. Chichester, UK: Wiley & Sons, 176–185.

Fackrell, J., Vereecken, H., Martens, J.P. & Van Coile, B. (1999). Multilingual prosody modelling using cascades of regression trees and neural networks. In: *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH '99)*, vol. 4. Budapest, Hungary, 1835–1838.

Fidelholtz, J.L. (1975). Word frequency and vowel reduction in English. In: *Proceedings of the 11th Meeting of the Chicago Linguistic Society (CLS 11)*. Chicago, USA, 200–213.

Fordyce, C.S. & Ostendorf, M. (1998). Prosody prediction for speech synthesis using transformational rule-based learning. In: *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 98)*, vol. 3. Sidney, Australia, 843–846.

Goubanova, O. & Taylor, P. (2000). Using Bayesian Belief Networks for model duration in text-to-speech systems. In: *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*. Beijing, China.

Grabowski, R. & Bauer, D. (2004). System for Computer-Aided Perception Experiments (SCAPE). Version 1.0 Beta 2.
URL `http://www.coli.uni-sb.de/~doba/scape/`

Grice, M., Baumann, S. & Benzmüller, R. (2005). German intonation in autosegmental-metrical phonology. In: S.A. Jun (ed.) *Prosodic Typology – The Phonology of Intonation and Phrasing*. Oxford: OUP, 55–83.

Hirschberg, J. & Rambow, O. (2001). Learning prosodic features using a tree representation. In: *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH 2001)*, vol. 2. Aalborg, Denmark, 1175–1178.

Hoste, V., Gillis, S. & Daelemans, W. (2000). Machine learning for modeling Dutch pronunciation variation. In: P. Monachesi (ed.) *Computational Linguistics in the Netherlands 1999. Selected Papers from the Tenth CLIN Meeting*. Utrecht, Netherlands: UiL OTS, 73–83.

IDS (1996). Deutsche Rechtschreibung: Regeln und Wörterverzeichnis. Amtliche Regelung. Mannheim, Germany.
URL `http://www.ids-mannheim.de/reform/`

IPDS (1994). The Kiel Corpus of Read Speech. Volume I. CD-ROM, Universität Kiel, Germany.

ITU-T (1994). A method for subjective performance assessment of the quality of speech voice output devices. ITU-T Recommendation P.85, International Telecommunication Union – Telecommunication Standardization Sector, Geneva, Switzerland.

ITU-T (1996). Methods for subjective determination of transmission quality. ITU-T Recommendation P.800, International Telecommunication Union – Telecommunication Standardization Sector, Geneva, Switzerland.

Jilka, M. & Syrdal, A.K. (2002). The AT&T German text-to-speech system: Realistic linguistic description. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2002)*. Denver, USA.

John, G.H. (1997). *Enhancements to the data mining process*. Ph.D. thesis, Computer Science Department, Stanford University.

King, S., Black, A.W., Taylor, P., Caley, R. & Clark, R. (2003). Edinburgh speech tools library. System documentation edition 1.2, for 1.2.3 24th Jan 2003.
URL `http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/`

Klatt, D.H. (1979). Synthesis by rule of segmental durations in English sentences. In: B. Lindblom & S. Öhmann (eds.) *Frontiers of Speech Communication Research*. London, New York, San Francisco: Academic Press, 287–299.

Kohler, K.J. (1990). Segmental reduction in connected speech in German: Phonological facts and phonetic explanations. In: W.J. Hardcastle & A. Marchal (eds.) *Speech Production and Speech Modelling*. Dordrecht, Netherlands: Kluwer Academic Publishers, 69–92.

Kohler, K.J. (1992a). Automatische Generierung der kanonischen Transkription und des Aussprachelexikons. *Arbeitsberichte des Instituts für Phonetik der Universität Kiel (AIPUK)* **26**, 175–196.

Kohler, K.J. (1992b). Dauerstrukturen in der Lesesprache. Erste Untersuchungen am PHONDAT-Korpus. *Arbeitsberichte des Instituts für Phonetik der Universität Kiel (AIPUK)* **26**, 225–238.

Kohler, K.J. (1992c). Erstellung eines Textkorpus für eine phonetische Datenbank des Deutschen. *Arbeitsberichte des Instituts für Phonetik der Universität Kiel (AIPUK)* **26**, 11–39.

Kohler, K.J. (1992d). Vorwort. *Arbeitsberichte des Instituts für Phonetik der Universität Kiel (AIPUK)* **26**, 7–9.

Kohler, K.J. (1995). PROLAB – the Kiel system of prosodic labelling. In: *Proceedings of the XIIIth International Congress of Phonetic Sciences (ICPhS 95)*, vol. 3. Stockholm, Sweden, 162–165.

Kohler, K.J. (1997). Modelling prosody in spontaneous speech. In: Y. Sagisaka, N. Campbell & N. Higuchi (eds.) *Computing Prosody: Computational Models for Processing Spontaneous Speech*, chap. 13. New York, USA: Springer, 187–210.

Kohler, K.J. (2003). Neglected categories in the modelling of prosody – pitch timing and non-pitch accents. In: *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 2003)*. Barcelona, 2925–2928.

Kohler, K.J., Pätzold, M. & Simpson, A. (1995). From scenario to segment. The controlled elicitation, transcription, segmentation and labelling of spontaneous speech. *Arbeitsberichte des Instituts für Phonetik der Universität Kiel (AIPUK)* **29**.

Manning, C.D. & Schütze, H. (2001). *Foundations of statistical natural language processing*. Cambridge, USA: MIT Press.

Miller, C. (1998). Individuation of postlexical phonology for speech synthesis. In: *Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis*. Jenolan Caves, Australia.

Möbius, B. & van Santen, J. (1996). Modeling segmental duration in German text-to-speech synthesis. In: *Proceedings of the Fourth International Conference on Spoken Language Processing (ICSLP 96)*, vol. 4. Philadelphia, USA, 2395–2398.

Möhler, G. & Conkie, A. (1998). Parametric modeling of intonation using vector quantization. In: *Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis*. Jenolan Caves, Australia, 311–316.

Ostendorf, M., Price, P.J. & Shattuck-Hufnagel, S. (1995). The Boston University Radio News Corpus. Tech. Rep. ECS-95-001, Boston University.

Pan, S. & Hirschberg, J. (2000). Modeling local context for pitch accent prediction. In: *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*. Hong Kong, 233–240.

Peters, B. (1999). Prototypische Intonationsmuster in deutscher Lese- und Spontansprache. *Arbeitsberichte des Instituts für Phonetik der Universität Kiel (AIPUK)* **34**, 1–177.

Peters, B. & Kohler, K.J. (2004). Trainingsmaterialien zur prosodischen Etikettierung mit dem Kieler Intonationsmodell KIM. Manuscript, IPDS, Universität Kiel, Germany.
URL `http://www.ipds.uni-kiel.de/pub_exx/bpkk2004_1/TrainerA4.pdf`

Portele, T. (1997). Reduktionen in der einheitenbasierten Sprachsynthese. In: *Fortschritte der Akustik (DAGA 97)*. Oldenburg, Germany: DEGA, 386–387.

Rapp, S. (1998). *Automatisierte Erstellung von Korpora für die Prosodieforschung*. Ph.D. thesis, Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung.

Riedi, M. (1997). Modeling segmental duration with multivariate adaptive regression splines. In: *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH '97)*, vol. 5. Rhodes, Greece, 2627–2630.

Riley, M.D. (1992). Tree-based modelling of segmental durations. In: G. Bailly, C. Benoît & T.R. Sawallis (eds.) *Talking Machines: Theories, Models, and Designs*. Elsevier Science Publishers, 265–273.

Schiller, A., Teufel, S. & Thielen, C. (1995). Guidelines für das Tagging deutscher Textcorpora mit STTS. Tech. rep., Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, and Seminar für Sprachwissenschaft, Universität Tübingen.
URL `http://www.sfs.nphil.uni-tuebingen.de/Elwis/stts/stts.html`

Schröder, M. (2004). *Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis*. Ph.D. thesis, PHONUS 7, Research Report of the Institute of Phonetics, Saarland University.

Schröder, M. & Trouvain, J. (2003). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology* **6**, 365–377.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. & Hirschberg, J. (1992). ToBI: a standard

for labelling English prosody. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP 92)*, vol. 2. Banff, Canada, 867–870.

Skut, W. & Brants, T. (1998). Chunk tagger – statistical recognition of noun phrases. In: *Proceedings of the ESSLLI-98 Workshop on Automated Acquisition of Syntax and Parsing*. Saarbrücken, Germany.

Sotscheck, J. (1984). Sätze für Sprachgütemessungen und ihre phonologische Anpassung an die deutsche Sprache. In: *Fortschritte der Akustik (DAGA 84)*. Bad Honnef, Germany: DPG, 873–876.

Sproat, R., Hunt, A., Ostendorf, M., Taylor, P., Black, A., Lenzo, K. & Edgington, M. (1998). SABLE: A standard for TTS markup. In: *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 98)*, vol. 5. Sidney, Australia, 1719–1722.

Strom, V. (2002). From text to prosody without ToBI. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2002)*. Denver, USA.

Swift, M.D., Campana, E., Allen, J.F. & Tanenhaus, M.K. (2002). Monitoring eye movements as an evaluation of synthesized speech. In: *Proceedings of the IEEE 2002 Workshop on Speech Synthesis*. Santa Monica, USA.

Syrdal, A.K., Möhler, G., Dusterhoff, K., Conkie, A. & Black, A.W. (1998). Three methods of intonation modeling. In: *Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis*. Jenolan Caves, Australia.

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In: W.B. Kleijn & K.K. Paliwal (eds.) *Speech Coding and Synthesis*, chap. 14. New York, USA: Elsevier, 495–518.

Taylor, P. & Black, A.W. (1994). Synthesizing conversational intonation from a linguistically rich input. In: *Proceedings of the 2nd ESCA/IEEE Workshop on Speech Synthesis*. Mohonk, USA, 175–178.

Taylor, P. & Black, A.W. (1998). Assigning phrase breaks from part-of-speech sequences. *Computer Speech and Language* **12**, 99–117.

Wall, L., Christiansen, T. & Orwant, J. (2000). *Programming Perl*. Cambridge, USA: O'Reilly.

Wells, J.C. (2004). SAMPA computer readable phonetic alphabet. URL `http://www.phon.ucl.ac.uk/home/sampa/`

Witten, I.H. & Frank, E. (2000). *Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco, USA: Morgan Kaufmann.

Wolters, M. & Mixdorff, H. (2000). Evaluating radio news intonation – autosegmental versus superpositional modelling. In: *Proceedings Sixth International Conference on Spoken Language Processing (ICSLP 2000)*, vol. 1. Beijing, China, 581–584.

Zboril, D. (1997). Einführung in die Sprachsynthese. URL `http://www.phonetik.uni-muenchen.de/Lehre/Skripten/Seminare/HS_WS1997/Synthese.html`

Zervas, P., Maragoudakis, M., Fakotakis, N. & Kokkinakis, G. (2003). Bayesian induction of intonational phrase breaks. In: *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*. Geneva, Switzerland, 113–116.