Diss. ETH No. 18210

Polyglot Text-to-Speech Synthesis

Text Analysis & Prosody Control

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY
ZURICH

for the degree of
DOCTOR OF TECHNICAL SCIENCES

presented by
HARALD ROMSDORFER
Dipl. Ing.
born July 27, 1973
citizen of Gmunden, Austria

accepted on the recommendation of
Prof. Dr. Lothar Thiele, examiner
Prof. Dr. Jan P. H. van Santen, co-examiner

2009

Meinen Eltern, Peter, Regina, Antonia und Jacob Felix gewidmet

Acknowledgments

In particular, I would like to thank Prof. Dr. Lothar Thiele and Dr. Beat Pfister for supervising and guiding my research, and Prof. Dr. Jan van Santen for his support as a co-examiner of my thesis.

I would like to thank Dr. Christof Traber, Dr. Marcel Riedi, and Volker Jantzen for designing and implementing the SVOX system and for recording the polyglot speech databases. Although the time spent together was very short, I drew a lot of inspiration from their work on the SVOX system.

I wish to thank the linguistic experts who supported me in lexicon construction. Especially, I want to thank Alexis Wilpert for his efforts in constructing the English, French, and Spanish linguistic resources.

I am indebted to numerous students who have directly or indirectly contributed to the linguistic resources and the design of the polySVOX system.

Contents

List of Abbreviations

ANN	Artificial neural network
ASCII	American standard code for information interchange
CART	Classification and regression tree
ETH	Eidgenössische Technische Hochschule (Swiss Federal Institute of Technology)
ETHPA	ETH computer phonetic alphabet
DCG	Definite clause grammar
FST	Finite-state transducer
GLM	Generalized linear model
GMM	Gaussian mixture model
IPA	International phonetic association
HMM	Hidden Markov model
LNRE	Large number of rare events
LPC	Linear predictive coding
MARS	Multivariate adaptive regression splines
MFCC	Mel-frequency cepstral coefficient
MLP	Multi-layer perceptron
MSE	Mean squared error
NMSE	Normalized mean squared error
OOV	Out-of-vocabulary
RNN	Recurrent neural network
RMS	Root of the mean squared error
SAMPA	Speech assessment methods phonetic alphabet
TD-PSOLA	Time-domain pitch-synchronous overlap-add
TTS	Text-to-speech

Abstract

The conversion of mixed-lingual texts, i.e., texts that contain inclusions of multiple other languages in form of phrases, words, or even parts of words, into artificial speech signals poses several problems to today's text-to-speech synthesis systems.

This thesis describes a new polyglot text-to-speech synthesis system that presents a solution to these problems. Concretely, new approaches to mixed-lingual text analysis, to mixed-lingual phonological processing, and to polyglot prosody control are presented in detail.

This system, as intended for the multilingual setting of Switzerland, is currently able to analyze text of any mixture of English, French, German, and Italian, and to generate artificial, polyglot speech signals in these languages with polyglot prosody consisting of French and German parts. In order not to restrict this polyglot text-to-speech synthesis system to a certain multilingual setting, strong emphasis is put on a modular system architecture. This architecture allows the construction of a polyglot synthesis system by combining independent, monolingual resources.

The first part of this thesis describes the implementation of the morphological and syntactic analysis of mixed-lingual text, in which an extended definite clause grammar formalism is applied. This mixed-lingual analyzer achieves both, precise language identification and accurate morphological and syntactic structure determination. A special architecture of this analyzer makes it possible to construct the mixed-lingual analyzer by combining several independent, monolingual analyzers. This approach has been the worldwide first mixed-lingual analyzer applied in text-to-speech synthesis.

A mixed-lingual phonological processing component, that is also constructed from independent monolingual resources, provides mixed-lingual prosodic phrasing, sentence accentuation, and phonological transformations.

In the second part of this thesis, a new approach to prosody modeling is presented. This approach applies weighted ensembles of neural networks with optimized input factor sets. A perceptual evaluation verifies that these ensembles of neural networks are able to generate natural sounding fundamental frequency contours and segment duration sequences, even when trained on automatically segmented prosody corpora. About 90% of 80 different test sentences having synthetic prosody were judged indistinguishable from the corresponding original recordings with human prosody.

Finally, a new approach to polyglot prosody control is presented. This approach allows switching between monolingual prosody models without audible rhythmic or melodic discontinuities. To the author's knowledge, this is the worldwide first polyglot prosody control applied in speech synthesis. A perceptual evaluation provides evidence that the synthetic polyglot prosody sounds for about 82% of the test sentences as natural as human prosody.

Kurzfassung

Die Umwandlung von gemischtsprachigen Texten, damit sind Texte gemeint, die Einschlüsse von mehreren anderen Sprachen, in Form von Gliedsätzen, Wörtern oder sogar Wortteilen beinhalten, in ein künstliches Sprachsignal bereitet heutigen Sprachsynthesesysteme verschiedene Probleme.

Diese Arbeit beschreibt ein neuartiges, polyglottes Sprachsynthesesystem, das eine Lösung für diese Probleme bietet. Konkret werden neue Ansätze für eine gemischtsprachige Textanalyse, eine gemischtsprachige phonologische Verarbeitung und eine polyglotte Prosodiesteuerung im Einzelnen vorgestellt.

Da dieses System für die mehrsprachige Umgebung der Schweiz gedacht ist, ist es zur Zeit in der Lage, Texte mit einer beliebigen Mischung aus Englisch, Französisch, Deutsch und Italienisch zu analysieren und künstliche, polyglotte Sprachsignale in diesen Sprachen mit polyglotter Prosodie, bestehend aus französischen und deutschen Teilen, zu erzeugen. Um dieses polyglotte Sprachsynthesesystem nicht auf eine bestimmte, mehrsprachige Umgebung zu beschränken, wurde besonderes Augenmerk auf eine modulare Systemarchitektur gelegt. Diese Architektur erlaubt die Konstruktion eines polyglotten Sprachsynthesesystems durch die Kombination von unabhängigen, einsprachigen Ressourcen.

Der erste Teil der Arbeit beschreibt die Implementation der morphologischen und syntaktischen Analyse von gemischtsprachigem Text, wobei ein erweiterter Definite-clause-grammar-Formalismus eingesetzt wird. Diese gemischtsprachige Analyseroutine erzielt sowohl eine präzise Identifikation der Sprache, wie auch eine genaue Bestimmung der morphologischen und syntaktischen Struktur. Die besondere

Architektur dieses Analyseverfahrens erlaubt die Konstruktion eines gemischtsprachigen Analysemoduls durch Kombination von mehreren unabhängigen, einsprachigen Analysemoduln. Dieser Ansatz war das weltweit erste, gemischtsprachige Analyseverfahren für Sprachsynthesesysteme.

Eine gemischtsprachige, phonologische Verarbeitungskomponente, die ebenfalls aus unabhängigen, einsprachigen Ressourcen zusammengestellt wird, ermöglicht gemischtsprachige Phrasierung, Akzentverteilung und die Anwendung von phonologischen Transformationen.

Im zweiten Teil der Arbeit wird ein neuer Ansatz zur Prosodiemodellierung vorgestellt. Dieser Ansatz verwendet gewichtete Ensembles aus neuronalen Netzen mit einer optimalen Eingangsfaktorenauswahl. Ein Hörtest bestätigt, dass diese Ensembles aus neuronalen Netzen in der Lage sind, natürlich klingende Grundfrequenzkonturen und Lautdauersequenzen zu erzeugen. Und das sogar, wenn sie auf automatisch segmentierten Prosodiekorpora trainiert wurden. Etwa 90% von 80 verschiedenen Testsätzen mit künstlich erzeugter Prosodie können nicht von den entsprechenden Originalaufnahmen mit natürlicher Prosodie unterschieden werden.

Ein neuer Ansatz zur polyglotten Prosodiesteuerung wird am Ende der Arbeit vorgestellt. Dieser Ansatz erlaubt es, ohne hörbare, rhythmische oder melodische Störungen zwischen mehreren einsprachigen Prosodiemodellen zu wechseln. Soweit der Autor informiert ist, ist dies die weltweit erste, polyglotte Prosodiesteuerung in einem Sprachsynthesesystem. Ein Hörtest zeigt, dass die so erzeugte polyglotte Prosodie für etwa 82% der Testsätze so natürlich wie menschliche Prosodie klingt.

Chapter 1

Introduction

1.1 Purpose and Scope of this Thesis

A text-to-speech (TTS) synthesis system converts written orthographic text (in computer readable form) into corresponding artificial speech signals. Research in TTS synthesis has concentrated in the last decade on two main fields: improving the speech quality of the artificial speech signal in terms of segmental quality and speech prosody, and making TTS synthesis systems multilingual, i.e., processing texts in one of multiple possible languages using the same TTS synthesis system, cf. [vSSM$^+$97, Spr97, SK06].

In multilingual countries, however, texts become more and more *mixed-lingual*, i.e., texts that contain inclusions of multiple other languages in form of phrases, words, or even parts of words. In such multilingual cultural settings, listeners expect a high-quality TTS synthesis system to read such texts in a *polyglot* manner, i.e., in such a way that the origin of the inclusions is heard, by using correct language specific pronunciation and prosody. Multilingual TTS synthesis systems, however, are unable to correctly convert such mixed-lingual texts into polyglot speech signals.

This thesis presents a report on the work done by the author to construct a polyglot TTS synthesis system that is able to convert such mixed-lingual texts into artificial, polyglot speech signals. This system

is able to analyze mixed-lingual texts having English, French, German, or Italian words or parts of words, and it is able to generate polyglot prosody consisting of French or German parts. The primary goal of this work was to explore and prove new concepts for

- construction of a mixed-lingual text analyzer from a combination of monolingual text analyzers, and
- construction of a polyglot prosody control from monolingual prosody controls, that are trained on automatically segmented natural speech data of the same speaker.

A second goal of this thesis was to reuse as much as possible of the formalisms and algorithms of the existing, monolingual German TTS system SVOX that has been developed at ETH Zürich. And a third goal consisted in designing the architecture of the new polyglot system in a way that allows an easy integration of new, additional languages.

As the polyglot TTS system reuses a lot of the formalisms for lexicon and grammar entries, for two-level rules, and for the annotation of speech corpora and some of the basic algorithms for text analysis and diphone concatenation of the original SVOX system, the new, polyglot system was termed "polySVOX". The original formalisms and algorithms are very well described in [Tra95]. Therefore, this thesis concentrates on the description of the new system architecture and the new algorithms for mixed-lingual text analysis, for mixed-lingual phonological processing, and for polyglot prosody control.

1.2 Scientific Contribution

A new algorithm for morpho-syntactic analysis of mixed-lingual texts is presented, that achieves both, precise language identification and accurate morphological and syntactic structure determination, while maintaining a strict separation of the linguistic databases for each language. An experiment gives evidence of the high language identification performance of the algorithm. This approach, first presented in [PR03], was the worldwide first mixed-lingual text analysis applied in TTS synthesis, see, e.g., [BL04].

A new formalism and new algorithms for mixed-lingual phonological processing are described that allow sentence accentuation, prosodic phrasing, and the application of phonological transformation on polyglot utterances.

A new approach to prosody modeling is investigated. This approach applies weighted ensembles of neural networks with optimized input factor sets for fundamental frequency control and segment duration control. A perceptual evaluation verifies that this approach enables networks trained on automatically segmented prosody corpora to generate natural sounding speech prosody.

Finally, a new approach to polyglot prosody modeling is presented, that allows switching between monolingual prosody models without audible rhythmic or melodic discontinuities. To the author's knowledge this is also the worldwide first approach to polyglot prosody modeling. A perceptual evaluation provides evidence that the synthetic polyglot prosody sounds as natural as the prosody of the original recordings.

1.3 TTS System Contribution

It is necessary here to clearly state what the author's own contributions were to the realization of the polyglot TTS system described in this thesis, as several people have contributed to the design and realization of the monolingual SVOX system.

The author's own contributions include a complete redesign and re-implementation of the overall TTS system architecture (cf. Section 1.5) that was originally designed and implemented by Christof Traber [Tra95]. The new architecture also includes the addition of a component for phonological processing (cf. Chapter 3).

The analysis of mixed-lingual texts required also a complete redesign and re-implementation of the syntactic and morphological analysis (cf. Chapter 2) that was originally realized by Thomas Russi [Rus90] and later extended for the use in the SVOX system by Christof Traber [Tra95]. However, the basic formalisms and algorithms for DCG-based bottom-up parsing and for the application of two-level rules were reused from [Tra95].

The author reused and extended the existing German lexicon of the SVOX system. However, he had to rewrite German word and sentence

grammar from scratch in order to meet the requirements of mixed-lingual text analysis. Word and sentence grammars of English, French, and Italian, as well as the inclusion grammars were finally all written by the author. The construction of the English and French lexica was accomplished with the help of linguist experts. The still rather small Italian lexicon was set up by the author.

Monolingual prosody control is completely replaced by the polyglot prosody control described in the Chapters 6 and 7. However, the author drew a lot of inspiration from the fundamental frequency control realized by Christof Traber [Tra95] and from both versions of duration controls implemented by Marcel Riedi [Rie98] and by Karl Huber [Hub91].

The recording, segmentation, and labeling of the quadrilingual diphone corpus and the recording of the bilingual prosody corpus was done in the POSSY project, cf. [HPT98, THN+99]. The segmentation and labeling of the bilingual prosody corpus used in this thesis, however, was done by the author.

Diphone-based speech signal generation of the SVOX system was extended by the author to enable polyglot diphone synthesis and to support multilingual diphone corpora. This TD-PSOLA based speech signal generation is not presented in this thesis. [Tra95] presents a good description of this concatenation algorithm.

1.4 Linguistic Terminology

This section introduces in alphabetic order the most important linguistic terms used throughout this thesis. More specific linguistic terms will be introduced when needed in later chapters.

Accent, stress: Accent or stress denotes prominence of an uttered syllable over other syllables. Several levels of prominence can be differentiated.

Foot: The term "foot" is mainly used in the context of rhythm and timing of speech. A foot basically consists of one accented syllable and all unaccented syllables to the right (*left-headed foot*) or to the left (*right-headed foot*) until the next accented syllable or until a sentence or phrase boundary.

Grapheme, graph: A grapheme is the smallest unit of written text that is capable of distinguishing different words. A graph is the individual realization of a grapheme. Graphemes and grapheme sequences are written between ⟨ ⟩. Graphs and graph sequences, or text, is denoted between " ".

Morpheme, morph: The morpheme is the smallest unit of a language that is capable of carrying a meaning. A morpheme is either a word or part of a word. The actual realization of a morpheme is called a morph.

Morphology: The description of the structure of morphemes and how words can be built from morphemes.

Phoneme, phone: A phoneme is the smallest unit of the spoken language that distinguishes different meanings of utterances. The set of possible phonemes is language-dependent. In the speech signal, a phone is one particular realization of a phoneme and forms the minimal segment of an utterance. Phones and phone sequences are denoted by phonetic symbols between [ ]. Phonemes are represented by phonemic symbols between / /. To illustrate incorrect phone sequences, they are written between ∗[ ].

Phonetic transcription: A phonetic transcription is a standardized, canonical description of the pronunciation of words as found in phonetic dictionaries. The phonetic dictionaries used in this thesis were for English [JRHS03], for French [War96], for German [Dud05], and for Italian [Pon95].

Phonetics: The investigation of the production and perception of phones in spoken utterances in terms of acoustic, articulatory, and perceptual parameters.

Phonological representation: In this thesis, a phonological representation denotes a minimal, voice-independent abstract description of an utterance, that includes the phonetic symbols and an abstract description of the prosody of the utterance. The phonological representation will be described in detail in Section 3.2.1.

Phonology: The investigation and treatment of minimal necessary distinctions that separate two spoken utterances with different meaning from one another.

Phrase, prosodic phrase: A phrase is a part of an utterance which in general consists of one or more words that are uttered as rhythmic and/or melodic unit. The boundaries between phrases may be indicated by pauses, by lengthening of the final syllable(s) before the boundary, or by certain melodic patterns. The term "prosodic phrase" is used to distinguish the phrase in the prosodic sense from the phrase in the syntactic sense (as, e.g., in "noun phrase").

Pragmatics: The investigation of the intentions and meanings that speech units (words, sentences, and texts) have in a particular context in which they are uttered. Of special interests are the intentions and communicative functions of these units in a discourse.

Prosody: The manner of articulation of a speech sound sequence in terms on intonation, rhythm, and loudness.

Semantics: The semantic level of description deals with the meaning of words and sentences.

Sentence: A sentence is a sequence of words forming a syntactically and semantically closed statement or question. In this thesis, the term "sentence" refers to a sentence in its orthographic form, which is usually terminated by a punctuation symbol, like ".", "!", or "?".

Syllable: A syllable is a unit of the spoken language that contains a voiced center, the so-called syllabic nucleus (a vowel, a diphthong, or a syllabic consonant), and one or more optional consonants preceding or following the nucleus. The preceding consonants are called the onset, the following consonants the coda of the syllable.

Syntax: The description of well-formed sentence structures of a language.

Word: A morphologically and semantically closed unit that denotes a specific item (object, action, property) of the world. Distinctions are often drawn between an orthographic word, a grammatical word, and a phonological word: an *orthographic word* is a graph sequence in written language delimited by white spaces. A *grammatical word* (or *syntactic word*) is the terminal element of syntax

analysis and forms the interface between morphology and syntax. A *phonological word* is a unit of spoken language defined by language specific, phonological criteria. In English, e.g., such a criterion is that a phonological word contains only one main stress. For example, the English sequence "the people'll have" consists of three orthographic, four grammatical, and two phonological words.

1.5 The polySVOX Architecture

A polyglot TTS system transforms a mixed-lingual text given as a sequence of graphemes into a speech signal. In order to cope with the high complexity of polyglot TTS synthesis, the polySVOX system was constructed from independent monolingual systems. This approach is feasible provided the architecture of the monolingual systems has been chosen suitably.

The polySVOX architecture strictly separates language-independent algorithms from language-dependent linguistic and acoustic data. Furthermore, following the linguistic view adopted as a basis for the ETH TTS project, a voice-independent part is separated from a voice-dependent part. The voice-independent part, termed *transcription*, includes text analysis and phonological processing. It maps (or transcribes) the input text onto an abstract intermediate representation, the *phonological representation*. The voice-dependent part, the so-called *phono-acoustical model*, comprises prosody control and speech signal generation. It produces from the phonological representation the speech signal.

The polySVOX system transforms a text paragraph by paragraph in four steps into a speech signal. Figure 1.1 illustrates these steps. The applied methods of the four corresponding system components are as follows:

- **Text analysis** derives the morphological structure of the words and the syntactic structure of the sentences, and delivers the phonetic transcription and the language identification of each morpheme. For text analysis, strictly rule-based processing is applied, i.e., a chart parser, which uses word, sentence, and paragraph

grammars and two-level rules for lexicon-to-surface mapping implemented as finite state transducers.

- **Phonological processing** applies phonological transformations, like schwa elision, French liaison, or English linking-r, which are formulated using so-called multi-context rules. It also assigns sentence accentuation and prosodic phrase boundaries based on the syntactic structure of a sentence. This abstract prosodic description, together with the phonetic transcription of each word, constitutes the phonological representation.

- **Prosody control** generates from the phonological representation the physical prosodic parameters. These are the duration values

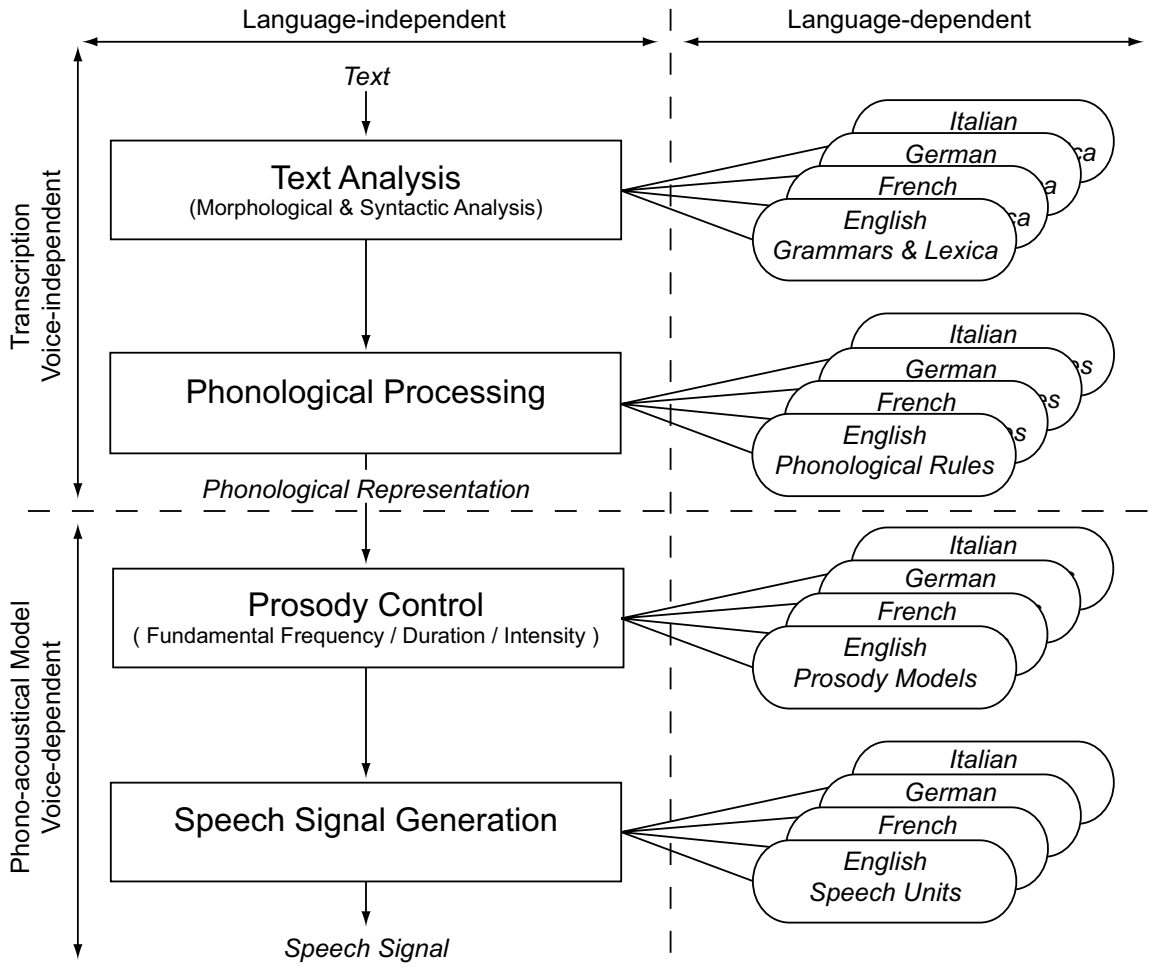

Figure 1.1: *The overall structure of the polySVOX TTS system.*

of all phones and pauses and the fundamental frequency contour of an utterance. Phone duration and fundamental frequency control are realized by means of trainable statistical models (artificial neural networks), which directly map the symbols of the phonological representation onto phone duration and fundamental frequency values.

- **Speech signal generation** is based on concatenation of diphone units extracted from natural speech. Prior to concatenation the diphones have to be prosodically modified such that they match the specified phone duration and fundamental frequency values.

This architecture suits monolingual as well as polyglot TTS synthesis. Basically, the linguistic and acoustic data define the set of languages that can be processed. Thus, the current set of languages of the polySVOX system can easily be expanded to new languages.

1.6 Outline of this Thesis

Chapter 2 identifies requirements and consequences for mixed-lingual text analysis and presents the polySVOX approach to mixed-lingual text analysis using several examples of mixed-lingual texts. An experiment measures the performance on language identification on a publicly available text corpus. Chapter 3 introduces a new rule formalism for describing phonological transformations and presents mixed-lingual approaches to sentence accentuation, prosodic phrasing, and phonological transformations. This chapter closes with a definition of the phonological representation. Chapter 4 surveys different existing approaches to speech prosody modeling and gives a definition of multilingual and polyglot prosody modeling. Chapter 5 describes setup, automatic segmentation, and labeling of the natural speech data used for the construction of the prosody models. The weighted neural network ensembles applied in Chapter 6 allow to generate natural sounding speech prosody using models trained on these automatically segmented prosody corpora. The ensemble-based prosody models are applied in the approach to polyglot prosody modeling presented in Chapter 7. A comparison to the prosody models applied in the SVOX system and a perceptual evaluation of the polyglot prosody control closes this chapter. Chapter 8 concludes the

thesis by reviewing its goals and by summarizing the major results. The appendix contains the ASCII representation of the phonetic symbols for all languages used in polySVOX, the grammars and lexica applied for the description of mixed-lingual text analysis, a list of all input factors for duration control and fundamental frequency control, and all test sentences that were used for perceptual evaluation.

Chapter 2

Mixed-lingual Text Analysis

2.1 Introduction

Text analysis in a TTS synthesis system is a combination of the analysis of the morphological structure of the words of a sentence and of the analysis of the syntactic structure of the sentence:

- Morphological analysis is used to obtain the correct pronunciation of the whole word from the pronunciation of the morphemes of which the word is composed, and to extract the structure of compounds and inflected forms in order to derive the correct word category.
- Syntax analysis serves several purposes. Beside of the determination of the syntactic structure of a sentence, which is highly relevant for the derivation of accentuation and prosodic phrasing of the sentence, it resolves ambiguities between homographs to derive their correct pronunciation, and it may serve as basis for a future semantic analysis of sentences.

The task of mixed-lingual text analysis in a polyglot TTS synthesis system additionally includes language identification on syntactic and

morphological level in order to apply the correct language specific pronunciation, accentuation, and prosodic phrasing. As this task is much more complex than monolingual text analysis, a detailed review of the requirements and the consequences for mixed-lingual text analysis is given.

2.1.1 Requirements for Mixed-lingual Text Analysis

The requirements for text analysis in a polyglot TTS synthesis arise from the texts, which have to be converted into speech, and from phonological and prosodic requirements of the subsequent synthesis steps. To illustrate these requirements, Table 2.1 lists some mixed-lingual example sentences with various foreign inclusions, as they can be found in Swiss newspapers or on Swiss web pages. The inclusions are put in parentheses and are indexed according to their language either as ($_{\mathrm{E}}$English), ($_{\mathrm{F}}$French), ($_{\mathrm{G}}$German), or ($_{\mathrm{I}}$Italian). The sentences themselves are also put in parentheses to indicate the sentence's base language.

In the following, language mixing phenomena typically encountered in published texts are illustrated first, then follow phonological and prosodic requirements:

Language Mixing Phenomena

The mixed-lingual sentences in Table 2.1 illustrate basically three major types of foreign inclusions:

Mixed-lingual words that are produced from a foreign stem by means of base language declension or conjugation or by means of compound word formation together with a base language word. Examples for such mixed-lingual words in Table 2.1 are

- "($_{\mathrm{G}}$($_{\mathrm{E}}$up)ge($_{\mathrm{E}}$dat)et)" in sentences 11 and 15: some German past participle construction of the English verb "to update".
- "($_{\mathrm{G}}$($_{\mathrm{E}}$Musical)programm)" in sentence 11: a German compound noun construction of the English noun "musical" and the German noun "Programm".

- "($_{E}$($_{F}$cuisine)'s)" in sentence 2: an English s-genitive construction of a French noun.

1. ($_{E}$Asia welcomes ($_{F}$bon ami Chirac).)
2. ($_{E}$One of French ($_{F}$nouvelle cuisine)'s objectives is to cook foods lightly.)
3. ($_{E}$She's not really ($_{F}$au fait) with my ideas.)
4. ($_{F}$À la mi-mars, le ($_{E}$Tokyo Game Show) sera l'occasion de nouvelles annonces pour ces "($_{E}$world game companies)".)
5. ($_{F}$Comment avez-vous osé vous attaquer à l'Adagio d'($_{G}$Hammerklavier)!)
6. ($_{G}$Wird das ($_{F}$Café) nicht von Ihren ($_{E}$Fans) belagert?)
7. ($_{G}$In 50 m nach links in die ($_{F}$Avenue de l'église) abbiegen!)
8. ($_{G}$Die ($_{F}$Femme fatale) ist die zentrale Frauenfigur des ($_{F}$Film noir).)
9. ($_{G}$Tessiner Städte im ($_{I}$Italianità) ($_{E}$Rating).)
10. ($_{G}$Die ($_{E}$Greatest Nation) hat die ($_{F}$Grande Nation) als tonangebende Nation abgelöst.)
11. ($_{G}$Das ($_{E}$Musical)programm ($_{E}$New York's) wurde ($_{F}$en passant) ($_{E}$up)ge($_{E}$dat)et.)
12. ($_{G}$($_{E}$Lobbying) ($_{F}$à discrétion) vor der Vergabe der Olympischen Spiele von 2012 in Singapur.)
13. ($_{G}$Geniessen Sie einen ($_{I}$Caffè Latte) oder eine feine italienische Spezialität im ($_{F}$Salon Rouge) des Landesmuseums.)
14. ($_{G}$Bis Ende März will sich der Kölner Konzern entscheiden, wie es mit dem ($_{E}$Discounter), der Bestandteil der Schweizer Tochter ($_{F}$Bon appétit) ($_{E}$Group) ist, weitergehen soll.)
15. ($_{G}$($_{F}$Peu à peu) wird der ($_{E}$High Performance) ($_{F}$Fonds) vom ($_{F}$Fonds)($_{E}$manager) ($_{E}$up)ge($_{E}$dat)et.)
16. ($_{G}$Der ($_{E}$Teammanager) ($_{I}$Luigi Riva) sieht im höheren Altersdurchschnitt der ($_{F}$Equipe Tricolore) keinen Vorteil für die ($_{I}$Squadra Azzurra).)
17. ($_{G}$Die ($_{E}$Air Force 1) landet in Frankfurt.)
18. ($_{G}$Kunstvolles Dekor im ($_{F}$Louis XIV) Stil.)
19. ($_{G}$Ich ($_{E}$dat)e das System ($_{E}$up).)
20. ($_{I}$($_{E}$General Motors) pagherà a Fiat 1,55 miliardi di euro per risolvere il ($_{E}$Master Agreement), inclusa la cancellazione della ($_{E}$put option).)

Table 2.1: *Examples of mixed-lingual sentences with various foreign inclusions. The inclusions and the sentences are put in parentheses and are indexed according to their language.*

Full foreign words that are embedded in a base language context. These word forms follow foreign morphology, but possibly disagree syntactically with the base language context. Examples in Table 2.1 are

- "($_{\mathtt{G}}$das ($_{\mathtt{F}}$Café))" in sentence 6: a French masculine noun embedded as a German neuter noun.
- "($_{\mathtt{I}}$della ($_{\mathtt{E}}$put option))" in sentence 20: an English neuter compound noun embedded as an Italian feminine noun.

Foreign multi-word inclusions, which are syntactically correct foreign constituents. These foreign constituents are embedded within the base language context according to the base language's syntax. Table 2.1 also contains examples of this inclusion type, like:

- "($_{\mathtt{E}}$New York's)" in sentence 11: an English s-genitive construction, which is embedded in the German sentence, according to the German syntax, after the referent.
- "($_{\mathtt{F}}$Avenue de l'église)" in sentence 7: a French noun phrase embedded in the German sentence in place of a German noun.
- "($_{\mathtt{E}}$Lobbying) ($_{\mathtt{F}}$à discrétion)" in sentence 12: a mixed-lingual multi-word inclusion, which consists of an English noun and a French prepositional phrase, embedded as a German noun phrase.

Phonologic and Prosodic Requirements

Multilingual listeners expect mixed-lingual sentences to be read in a way that the origin of foreign inclusions is heard. Particularly, a polyglot TTS synthesis system must generate the correct language specific sequence of phones with appropriate prosody. This means that polyglot TTS synthesis must comply with the following phonologic and prosodic requirements:

- **Language specific pronunciation**: foreign inclusions must be pronounced in a language specific manner. E.g., in Switzerland the French noun "Avenue" in the German sentence 7 of Table 2.1

must be pronounced [av(ə)ny] using French phones (and not in a German fashion ∗[aveːnuə]).

- **Language specific word stress**: the word stress of foreign inclusions must follow language specific rules; thus, French nouns in German sentences, like "Avenue" [av(ə)ˈny], are end-stressed, even if German nouns are generally front-stressed. Applying a German word stress pattern, e.g., ∗[ˈav(ə)ny], makes the word difficult to understand.

- **Language specific phonological phenomena**: phonological transformations within longer foreign inclusions follow the phonological rules of the inclusion language. E.g., the application of German phonological rules onto the French noun phrase "Bon appétit" in the German sentence 14 of Table 2.1 produces the incorrect transcription ∗[bõ.ʔa.pe.ti]. This pronunciation sounds strange to Swiss listeners, as it lacks the French liaison consonant and has a German glottal stop inserted (as it is usually done in German sentences before a vowel starting a word). In contrast, the application of French phonological rules, like denasalization and liaison, results in the correct transcription [bɔ.na.pe.ti].

- **Language specific sentence accentuation**: the intonation of larger, multi-word foreign inclusions, like "world game companies" in sentence 4 of Table 2.1, follows the foreign sentence accentuation patterns. Thus, "world game companies" is accented [[2]wɜːld.[1]ge͡ɪm.[3]kʌm.pə.niz][1] according to English accentuation, and not according to the accentuation of the sentence's base language, French: ∗[[2]wɜːld.ge͡ɪm.[3]kʌm.pə.[1]niz].

- **Language specific phrasing**: placement of phrase boundaries within foreign inclusions disobeys in general the base language's phrasing rules; the phrasing rules of the inclusion language specify their correct placement. E.g., in German and English sentences nouns are followed by potential phrase boundaries; in the French inclusion "Salon Rouge" in the German sentence 13 of Table 2.1, however, no phrase boundary may be placed after the noun "Salon", as the subsequent adjective "Rouge" is part of the French noun phrase.

[1] [1] denotes the main phrase accent; [2] and [3] denote weaker accents.

2.1.2 Consequences for Mixed-lingual Text Analysis

Text analysis of a TTS synthesis system that has to pronounce sentences like the ones of Table 2.1 and thereby meet the requirements given in Section 2.1.1 must fulfill three basic tasks:

- language identification,
- language-dependent phonetic transcription, and
- language-dependent syntactic structure analysis.

The following sections describe these tasks in more detail.

Language Identification and Language-dependent Transcription

First of all, mixed-lingual text analysis must be capable to identify the correct language of each portion of the input text. This is necessary in order to transcribe these text portions according to their languages and in order to apply appropriate word stress. The size of such portions, as shown in Table 2.1, varies from single morphemes to complete sentences.

Interlingual homographs, i.e., words of different languages with the same graphemic sequence but with different pronunciations, make the task of language identification of a given portion of text even more complex. For example, "hat" is an English noun as well as a German verb, or "die", which is an English verb as well as a German determiner. Certain types of interlingual homographs, like loanwords, logograms, abbreviations, or acronyms, are especially difficult to disambiguate:

- **Loanwords** are strongly assimilated to the base language, not only in morpho-syntactic terms, but also with respect to the pronunciation. Loanwords in mixed-lingual text may, however, raise an additional issue concerning homographs in places where their pronunciation depends on the language context. Consider, e.g., the word "Nation" in sentence 10 of Table 2.1, which is first pronounced in English as [ˈne͡ɪʃən], then in French as [nɑˈsjɔ̃] and finally in German as [naˈt͡si̯oːn].

- **Logograms** like numbers, Roman numerals, currency units, or special symbols ("%", "&", etc.) are also a form of interlingual homographs. The correct pronunciation of these logograms depends on the language context. E.g., the two sentences, "(G Die (E Air Force 1) landet in Frankfurt.)" and "(G Kunstvolles Dekor im (F Louis XIV) Stil.)", contain examples of logograms that are part of foreign inclusions and that are therefore pronounced according to their inclusion languages.

- **Abbreviations** are short forms of words with or without final period, which are pronounced as the full form they represent. Common abbreviations are often graphemically identical across multiple languages, but are pronounced differently. For example, the abbreviation "dr" is pronounced in English as [ˈdɒktə(r)], in French as [dɔkˈtœːʁ], in German as [ˈdɔktoːɐ̯] and in Italian as [dotˈtoːre].

- **Acronyms** are short forms of words or phrases, which are either spelled or pronounced. Acronyms are usually composed of the initial letters of the words they symbolize. There exist a lot of graphemically identical acronyms across multiple languages. The rules of pronunciation, however, vary a lot depending on the language: in Italian most acronyms are read, in German and English they are normally spelled, and in French, according to [BdMF01], approximately half of the acronyms are read and half are spelled. As an example consider "IRA", which is read in French [iˈʁa] and Italian [ˈiːra], but spelt in English [ˌa͡ɪɑːrˈe͡ɪ] and German [iːɛrˈaː].

Language-dependent Syntactic Structure Analysis

The analysis of the mixed-lingual syntactic structure of the input sentence forms the basis for subsequent language-dependent phonological processing. A prerequisite for such a mixed-lingual syntactic analysis is the correct identification of syntactic word and sentence boundaries. Syntactic words are the terminal elements of syntactic analysis. In contrast to orthographic words, which are delimited by blank characters and which are therefore easily identified by text preprocessing, syntactic words are more difficult to identify and do not always correspond to orthographic words due to different graphemic phenomena:

- **Word contractions**, e.g., English "he's", "Mary's", German "das ist's" (that's it), or Italian "po'd'acqua" (some water).
- **Multi-word lexemes**, i.e., word forms spanning multiple orthographic words, like English "in fine" (adverb) or French "est-ce que" (interrogative particle).
- **Ambiguous punctuation symbols**, e.g., a period at the end of an abbreviation may at the same time be a full stop to indicate the end of the sentence.
- **Cross-line hyphenation of words** at line breaks, e.g., consider English "in-<LF>put" vs. "in-<LF>and output" ('<LF>' is the line end symbol).
- **Missing designated word separation symbols** in languages like Chinese or Japanese. E.g., [SCGC96] give a good overview of the problems text analysis for Chinese is confronted with.

Section 2.2 describes the overall architecture and key aspects of the approach to mixed-lingual morpho-syntactic analysis as implemented in the polySVOX TTS synthesis system. Section 2.3 explains the text analysis procedure in detail and illustrates syntactic word and sentence boundary identification. The subsequent Sections 2.4 and 2.5 present solutions for all types of language mixing phenomena listed above by applying the polySVOX mixed-lingual text analysis. This text analysis also allows the disambiguation of interlingual homographs, as shown in Section 2.6, and a grapheme-to-phoneme conversion of unknown words in mixed-lingual sentences, illustrated in Section 2.7. Finally, Section 2.8 presents the results of a language identification experiment. Appendix B provides a format specification of lexicon entries and grammar rules, and lists excerpts of the lexica and grammars for English, French, German, and Italian.

2.2 Mixed-lingual Morpho-Syntactic Analysis

The polySVOX approach to mixed-lingual text analysis simultaneously solves all three tasks given in Section 2.1.2: language identification,

language-dependent phonetic transcription, and language-dependent syntactic structure analysis. In order to compare this approach qualitatively with other approaches to language identification, a survey of other algorithms for language identification is given first.

2.2.1 Language Identification

Language identification from written text is important in different application areas, including mixed-lingual TTS synthesis, multilingual speech recognition, e.g., [THRJ02], and document classification, e.g., [CT94]. Therefore, numerous approaches towards this task exist. The majority of them is based on statistical information about word and character sequences of the languages in question. These approaches usually apply a character context window of a fixed length onto the input character sequence; an often used window contains three characters in front of and three after the character in question. The most likely language for a given word or a sentence is then calculated employing methods like n-grams [Sch91, Gre95], neural networks [TS04], decision trees [HT01], or some combination of them. Some approaches also make use of basic linguistic knowledge, e.g., in form of a heuristic method using language specific character frequencies plus language specific lists of function words and word endings [Gig95]. Common to all of these approaches is that the granularity of language identification is either a sentence or at most a word.

As can be verified by the examples of Table 2.1, language identification on word level is not accurate enough for a polyglot TTS system. Foreign inclusions in mixed-lingual words can be very short, as the English inclusions in the German verb "($_{G}$($_{E}$up)ge($_{E}$dat)et)" demonstrate. Such mixed-lingual words require language identification to be applied on morpheme level. As typical character context windows are larger than these inclusions, it is difficult for statistical approaches based on character context windows to correctly identify the language of foreign inclusions in mixed-lingual words.

Additionally, a mixed-lingual sentence, like "($_{G}$($_{F}$Peu à peu) wird der ($_{E}$High Performance) ($_{F}$Fonds) vom ($_{F}$Fonds)($_{E}$manager) ($_{E}$up)ge($_{E}$dat)et.)", demonstrates that there exist words with a majority of characters not belonging to the word's base language, as well as sentences, in which the sentence's base language is not the language

with the maximum number of words of this sentence. Considering this, simple statistical approaches for identifying the base language of a word or a sentence are too unreliable for language identification in mixed-lingual sentences.

2.2.2 The polySVOX Approach

Investigations of mixed-lingual texts showed two central findings:

1. *Inclusions of foreign constituents into a context of another language can be described by specific bilingual morphological and syntactic rules.* As an example, consider the following mixed-lingual German verbs containing English verb stems:

"($_{G}$($_{E}$updat)en)"	[ʌpˈdeɪtn̩]
"($_{G}$($_{E}$brows)en)"	[ˈbraʊzn̩]
"($_{G}$($_{E}$scann)en)"	[ˈskænən]

   These verbs contain the present tense form of English verb stems (without silent "e", but with optional consonant doubling) and follow weak German conjugation. English verb prefixes are bound to the English verb stem.

   The mixed-lingual German past participle form consists of the German past participle prefix "ge" followed either by the present tense form of the English verb stem plus a German past participle ending or by the complete English past participle. A possible English verb prefix may optionally be separated from the English verb stem and be included in front of the German past participle prefix "ge". As examples consider the following, equally frequently used mixed-lingual German forms of "updated":

"($_{G}$($_{E}$up)ge($_{E}$dat)et)"	[ˈʔʌpgəˌdeɪtət]
"($_{G}$($_{E}$up)ge($_{E}$dated))"	[ˈʔʌpgəˌdeɪtɪd]
"($_{G}$ge($_{E}$updat)et)"	[gəʔʌpˈdeɪtət]
"($_{G}$ge($_{E}$updated))"	[gəʔʌpˈdeɪtɪd]

2. *Within foreign constituents only foreign monolingual morphological and syntactic rules are relevant.* This can be illustrated by the mixed-lingual German sentence

> "($_{G}$Die ($_{F}$Femme fatale) ist die zentrale
> Frauenfigur des ($_{F}$Film noir).)"

As German syntax does not allow adjectives to be placed after the corresponding noun, the syntactic structure of this sentence can only be correctly analyzed if French syntactic rules are applied within the two French noun phrase inclusions. The French noun phrases are then included as noun phrase constituents in the German sentence.

2.2.3 Architecture Overview

The polySVOX system follows a modular approach to mixed-lingual text analysis and strictly separates monolingual analyzers from bilingual inclusion grammars, cf. [RP07]. Each *monolingual analyzer* contains a monolingual morpheme lexicon as well as a word, a sentence, and a paragraph grammar. A *bilingual inclusion grammar* contains bilingual grammar rules that describe, which foreign constituents can be mapped as foreign inclusions onto corresponding constituents of the base language. Thus, monolingual grammars need not be modified at all when including new languages. Only small bilingual inclusion grammars are necessary, which are loaded together with the corresponding monolingual grammars. The size of such an inclusion grammar is normally less than five percent of the size of monolingual grammars (e.g., 18 inclusion grammar rules specifying English inclusions in German compared to 797 monolingual German grammar rules).

Figure 2.1 illustrates the modular architecture of the polySVOX morphological and syntactic text analysis. It is realized as a bottom-up chart parser for penalty-extended definite clause grammars (DCGs). An input scanner normalizes the graphemic input text character by character in a stream-like fashion, cf. [RP06]. For this normalized character stream, a contiguous sequence of matching lexemes is looked up in all monolingual morpheme lexica. The chart parser itself operates on three different levels: a word, a sentence, and a paragraph level. Each level is provided with a set of separate monolingual grammars, which are joined by a set of bilingual inclusion grammars. The input scanner triggers word analysis at unambiguous word boundaries in the input stream. Likewise, word analysis starts sentence analysis at un-

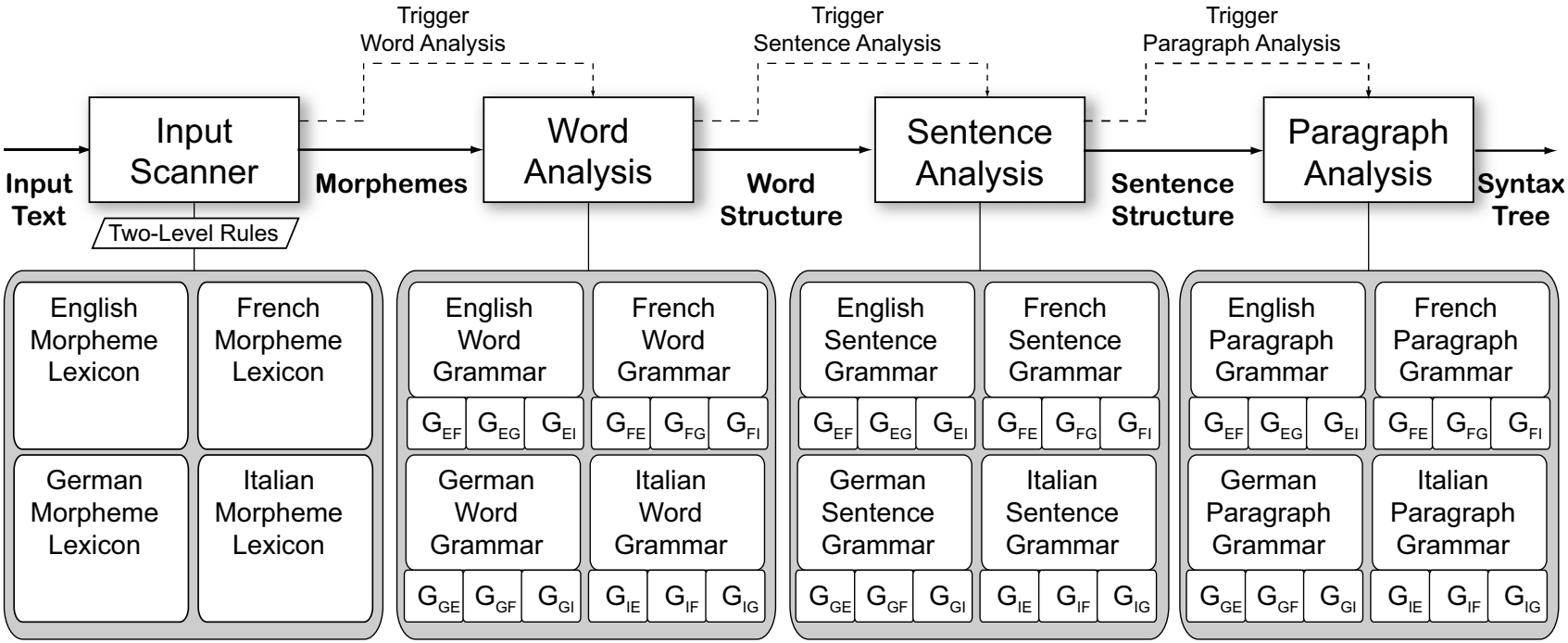

Figure 2.1: *Architecture of morphological and syntactic analysis of the polySVOX TTS synthesis system. The notation G_{ij} specifies an inclusion grammar that describes inclusions of language j in language i. E, F, G, and I are abbreviations of English, French, German, and Italian, respectively.*

ambiguous sentence boundaries. And sentence analysis finally triggers paragraph analysis at unambiguous paragraph boundaries.

The DCG formalism is a natural extension of context-free grammars. This extension is done by augmenting the context-free production rule skeleton with feature terms and the term unification operation. Theoretically, DCGs have the power of a Turing machine, and in that sense are as general as they can be, cf. [PW80]. In the polySVOX system, DCG rules have additionally got a penalty value in order to select the optimal solution among several ambiguous solutions, and they have got optional keywords for controlling the building of parse trees, cf. [Tra95], and for identifying syntactic word and sentence boundaries.

2.2.4 Inclusion Grammars

An inclusion grammar consists of bilingual grammar rules that specify mappings from foreign constituents and their feature terms to corresponding constituents of the base language. Examples of such inclusion grammar rules for English, French, German and Italian are given in Appendix B.

Inclusion grammar rules allow to formulate constraints on including foreign constituents using two basic concepts: constituent mapping restrictions and inclusion penalties. Both are specified using the extended DCG formalism.

Constituent mapping restrictions allow to map a specific foreign constituent to a base language constituent. E.g., the verb stem inclusion rule R84 in Appendix B.3 specifies that only the present tense form of English verb stems `VS_E`, indicated by the feature value '`pres`', may be included as German verb stem `VS_G` that must additionally follow weak German conjugation.

Inclusion penalties allow to disambiguate interlingual ambiguities. These penalty values are set manually by a linguistic expert according to following guidelines:

- The penalty values of inclusion rules for a given constituent must be generally higher than the overall penalty values of any monolingual analysis of this constituent.

- The penalty values of larger inclusions, e.g., noun phrases, shall be typically lower than the penalty values of smaller inclusions, e.g., nouns or adjectives.
- The penalty values of all inclusion rules for the same constituent must be harmonized across all inclusion languages.

These inclusion grammar rules provide a very accurate description of how foreign inclusions are analyzed in the base language context. The strictly bilingual definition of these rules makes it possible to construct a modular and flexible morphological and syntactic grammar for any desired language combination.

2.2.5 Language Switching Flag

Inclusion grammars are loaded together with their monolingual grammars. If inclusion grammars of several languages are loaded, cyclic dependencies in bottom-up chart parsing and incorrect analysis results are inevitable. Cyclic dependencies arise from loading inclusion grammars of two languages specifying inclusions of each other. E.g., simultaneous loading of the following two inclusion grammar rules (rules R47 and R87 of Appendix B) will obviously result in a cyclic dependency when parsing English or German nouns:

```
N_E (?NR,?,?) ==> N_G (?NR,?,?) * 100
N_G (?NR,?,?) ==> N_E (?NR,?,?) * 100
```

Incorrect analysis results emerge, if certain morphological or syntactic structures that are valid for one language are forbidden in another language, and if inclusion grammar rules exist that specify appropriate mappings between these languages. E.g., the positioning of an adjective after an noun in a French noun phrase, as for example in “le film noir”, is forbidden in German noun phrases. The sequence ∗“der Film schwarze” must therefore not be analyzed as a German noun phrase. However, applying French-German inclusion grammar rules R68 and R69 of Appendix B, and using French sentence grammar rule R64, ∗“Film schwarze” is analyzed as a French noun phrase that contains two German inclusions. The German-French inclusion grammar rule R101 maps this French noun phrase back to a German noun phrase

nucleus. Applying the German sentence grammar rule R79 would result in the incorrect analysis of ∗"der Film schwarze" as a German noun phrase.

In order to prevent cyclic dependencies in bottom-up chart parsing and incorrect analysis results, all but the first application of inclusion grammar rules for every single lexeme must be inhibited. This is achieved using a so-called *language switching flag* that prevents inclusions of foreign constituents that already contain a foreign inclusion themselves. The flag is basically a Boolean feature term implemented using the DCG formalism. It is therefore completely transparent to the parsing algorithm.

Table 2.2 shows the application of the language switching flag to the grammar rules which are necessary to analyze the above example noun phrase. Each constituent obtains an additional feature term, which represents the language switching flag. This feature term either has the value `true` or `false`, or is a variable named `LSF1`, `LSF2`, and so forth. Each monolingual grammar rule evaluates this feature by ap-

```
       BOOL_OR (false,false,false)      ==> * 0 :INV
       BOOL_OR (true, true, false)      ==> * 0 :INV
       BOOL_OR (true, false,true)       ==> * 0 :INV
       BOOL_OR (true, true, true)       ==> * 0 :INV
[R64]  NP_F (?N,?P,?G,?LSF3)            ==> N_F (?N,?P,?G,?LSF1)
                                            ADJ_F (n,?N,?G,?LSF2)
                                            BOOL_OR (?LSF3,?LSF1,?LSF2) *
[R68]  N_F (?NR,?,true)                 ==> N_G (?NR,?,?,false) * 100
[R69]  ADJ_F (?,?N,?,true)              ==> ADJ_G (?,?N,?,?,?,false) * 100
[R79]  NP_G (?C,?NR,?P,?G,?NT,?LSF3)    ==> DET_G (?C,?NR,?G,?F,?TYP,?LSF1)
                                            NPNUC_G (?C,?NR,?P,?G,?TYP,?NT,?LSF2)
                                            BOOL_OR (?LSF3,?LSF1,?LSF2) *
[R101] NPNUC_G (?,?NR,pers3,?,?,?,true) ==> NP_F (?NR,?,?,false) * 90
```

Table 2.2: *French and German sentence grammar rules showing feature variables* `LSF1`, `LSF2`, *and* `LSF3` *as language switching flags. Language switching flag features are written in bold. Boolean or is implemented by additional* `BOOL_OR` *rules. Bilingual inclusion grammar rules toggle the value of the language switching flag feature from* `false` *to* `true`*. Monolingual rules evaluate the language switching flags of their subconstituents using the* `BOOL_OR` *rules. The rule numbers indicate the equivalent grammar rules of Appendix B.*

plying Boolean OR to the language switching flags of the constituents of the rule body (cf. the `BOOL_OR` rules, rule R64, and rule R79 in Table 2.2). Each inclusion grammar rule requires the language switching flag of the foreign constituent to be `false`, and sets the language switching flag of the base language's constituent to `true` (cf. inclusion rules R68, R69, or R101). Thus, the language switching flags of a constituent that contains a foreign inclusion and all constituents derived from it are always set to `true`. Additionally, no inclusion grammar rule can be applied to such a constituent anymore, as this would require their language switching flags to be `false`.

The language switching flag prevents cyclic dependencies in bottom-up chart parsing due to inclusion grammar rules, as it stops possible cycles after the first application of an inclusion grammar rule. The flag also prevents incorrect analysis results like the one illustrated above, as it allows only direct foreign inclusions. The polySVOX system automatically extends grammar rules by the necessary `BOOL_OR` rules and language switching flag features when loading the grammars. Therefore, the grammar rules listed in Appendix B do not contain language switching flag features or grammar rule extensions.

2.3 Word and Sentence Boundary Identification

In order to correctly identify syntactic words within a graphemic input text, morphological and syntactic knowledge is necessary. Therefore, it is unreasonable to do this identification in some text preprocessing step. We better integrate identification of syntactic words into morphological and syntactic text analysis.

Figure 2.2 illustrates word and sentence boundary identification of polySVOX with a morpho-syntactic analysis of the English sentence: "It's in St. Mary's St." The correct pronunciation of this sentence [ɪts ɪn sənt me͡əriz stri:t] requires to identify the period in the first "St." as part of the abbreviation and the period in the second "St." as a period terminating the sentence. This can be achieved by syntactic means, which have to provide the correct analysis of "It's" as a personal pronoun followed by a contracted verb form and of "Mary's" as possessive form of a noun.

The application of the main processing steps of the polySVOX text analysis (cf. Figure 2.1) to this example sentence are described in the following:

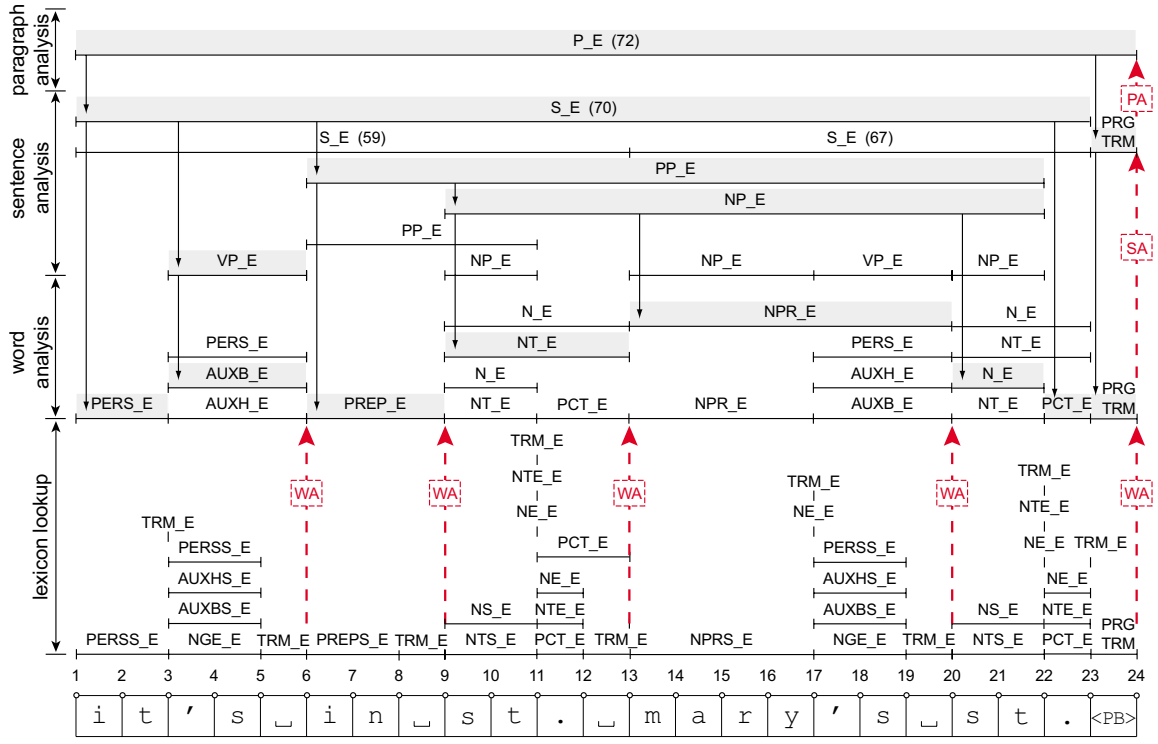

Figure 2.2: *Representation of the simplified chart resulting from morphological and syntactic analysis of the sentence "It's in St. Mary's St." At the bottom the normalized input character sequence is shown. Edges are drawn without constituent feature values. If a set of edges with the same associated constituent but different feature values span the same vertices, only one of these edges is shown here. Important penalty values of edges are shown in parentheses. The "lexicon lookup" section contains edges associated with the lexemes found during lexicon lookup. The "word analysis", "sentence analysis", and "paragraph analysis" sections contain edges associated with constituents resulting from the respective analysis levels. An arrow with a dashed line tagged with* `WA`*,* `SA`*, or* `PA` *indicates a word, sentence, or paragraph analysis trigger event, resp. The constituents of the final syntactic parse tree are shown with grey background.*

Text normalization generates out of the graphemic input sequence a well-defined character stream. Note that also punctuation characters, the blank character, carriage return, the newline character, and other special characters are included as separate tokens. Text normalization primarily takes care that all capital letters are converted to lowercase letters, that all sequences of contiguous space characters are reduced to one space character, and that all input characters not defined in one of the language specific sets of legal input characters are deleted from the character stream. Additionally, a paragraph boundary symbol `"<PB>"` is inserted at the end of a paragraph and at the end of the stream.

Lexicon lookup looks for all possible decompositions of the character stream into lexemes of the morpheme lexicon. For each matching lexeme, a corresponding edge is inserted into the chart. These edges are shown in the "lexicon lookup" section in Figure 2.2. In the morpheme lexicon the keyword `':WORD_END'` indicates a possible word boundary after the lexeme, as can be seen, e.g., in the lexicon entries L1, L2, L3, or L4 in Appendix B.1.

Word analysis is started only at unambiguous word boundaries in order to prevent incorrect analysis results. A chart vertex is an unambiguous word boundary if the associated lexemes of all edges ending in this vertex are tagged by the keyword `':WORD_END'`, and no edge is crossing this vertex. The character token sequence starting form the previous unambiguous word boundary up to the current one is then parsed for all contiguous sequences of words that are morphologically correct as defined by a word grammar, see, e.g., Appendix B.1. The word analysis results are inserted into the chart. These constituents are shown in the "word analysis" section in Figure 2.2.

Sentence analysis is designed similar to word analysis. Terminal elements are the word constituents of word analysis. Sentence analysis is started only at an unambiguous sentence boundary. This is at the next chart vertex where the associated word constituents of all edges ending in this vertex are tagged by the keyword `':SENT_END'` and no edge is crossing this vertex. This keyword is set by word grammar rules, as shown, e.g., in the grammar

rules R1 or R2 in Appendix B.1. Sentence analysis is needed to disambiguate morphologically ambiguous words. The results of sentence analysis are all possible syntactically correct sequences of sentences, as defined by a sentence grammar. These results are again inserted into the chart as shown in section "sentence analysis" in Figure 2.2.

Paragraph analysis is started at an unambiguous paragraph boundary. This is at the next chart vertex where the associated sentence constituents of all edges ending in this vertex are tagged by the keyword `':PARA_END'` and no edge is crossing this vertex. This keyword is set by sentence grammar rules, cf. grammar rule R25 in Appendix B.1. The sentence constituents serve as terminal elements for syntactic analysis of the paragraph. Out of the set of possible sentence sequences, paragraph analysis returns the sentence sequence with minimal total penalty.

2.3.1 Contracted Word Forms

The approach presented here allows to correctly analyze ambiguous contracted word forms. The basic idea is to include in morphological analysis beside of blank characters also empty characters as word delimiters. E.g., for English, these delimiters are listed as `TRM_E` in the morpheme lexicon in Appendix B.1 and are used in the word grammar rules in Appendix B.1 to terminate each word constituent. Thus, joint orthographic words can be split into a sequence of syntactic words. In order to prevent incorrect word splits, the empty word delimiter has a higher penalty, cf. lexicon entry L5. Additionally, specific word categories like abbreviations can use separate empty word delimiters with a lower penalty value, e.g., lexicon entry L6. These empty word delimiters are without a `':WORD_END'` tag, so word analysis is triggered only at the unambiguous ends of orthographic words.

The use of empty word delimiters for the analysis of contracted word forms is illustrated by the token sequence `"'s"` in the sentence in Figure 2.2. `"'s"` can be a contracted form of a verb, a contracted personal pronoun or the suffix of a noun in possessive form. As illustrated, four different lexemes, L10, L33, L39, and L40 match `"'s"` and are inserted into the chart. For `"it's␣"` word analysis returns only

three morphologically correct sequences of syntactic words: a personal pronoun `PERS_E` followed by either the contracted form of the personal pronoun "us" (`PERS_E`) or of the auxiliaries "be" (`AUXB_E`) or "have" (`AUXH_E`). For `"mary's␣"` the word grammar rule R5 additionally allows a morphological analysis of the complete orthographic word as possessive form of a proper noun `NPR_E`.

As can be verified in Figure 2.2, this input sequence can also be analyzed as a sequence of two English sentences. Doing so, the first `"st."` would be incorrectly analyzed as abbreviation of "street", and the second `"'s"`, also incorrectly, as an auxiliary "be".

Paragraph grammar rules, as shown for English in Appendix B.1, that define a paragraph as a sequence of sentences, prevent that incorrect analysis result. As the penalty values of grammar rule production and of the rule subconstituents are added up to form the penalty value of the rule head, the penalty value of a paragraph consisting of the two short sentences is higher (7 + 59 + 67) than the penalty value of a paragraph consisting only of the longer sentence (2 + 70).

2.3.2 Multi-word Lexemes

The approach presented here is also well-suited for multi-word lexemes. E.g., consider the preposition "in front of". As blank characters are processed like other characters, lexicon lookup treats multi-word lexemes like any other lexeme. Additionally, word analysis is started only at the end of such a multi-word lexeme, because the associated chart edge spans the whole multi-word lexeme including the blank characters. Thus, word analysis is not triggered after "in" and "front".

To describe "in front of" as a multi-word lexeme is convenient for syntactic analysis, whereas it is irrelevant for pronunciation. For other word forms, like the adverb "in fine", pronounced as [ɪn 'fa͡ɪni], multi-word analysis is a necessity to disambiguate it from the preposition "in" [ɪn] followed by the adjective "fine" [fa͡ɪn]. E.g., consider the sentence "He's in fine condition in fine.". Using multi-word lexemes, the final "in fine" is correctly analyzed as an adverb.

2.3.3 Sentence End Identification

The correct identification of the end of a sentence in case of ambiguous punctuation symbols is an important issue in TTS synthesis, as sentence end is a necessary feature in prosody generation. Therefore, numerous approaches have already been presented, including simple heuristics such as detecting capitalized words following periods, cf. [McA89, Dut93, LC91], probabilistic ones, e.g., [Ril89], and an elaborated, morphology based approach presented by [Tra95]. All of these approaches, however, are applied in a preprocessing step and therefore lack syntactic disambiguation capabilities.

Similar to the identification of syntactic words, the identification of sentence ends also requires morphological and syntactic knowledge. In polySVOX, sentence end identification is integrated into morphological and syntactic analysis and punctuation symbols are analyzed as a special form of syntactic words. The following points summarize the

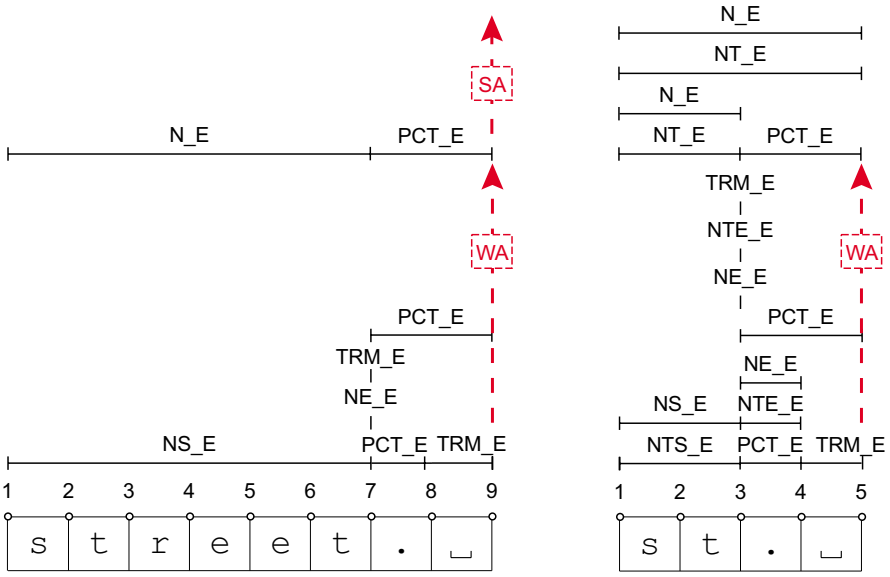

Figure 2.3: *For the input text "Street. " word analysis returns a noun* `N_E` *followed by an unambiguous sentence end* `PCT_E`*. Thus, sentence analysis is started at chart vertex 9. In case of the input text "St. " the period is ambiguous. It is either a punctuation symbol* `PCT_E`*, a part of a noun* `N_E`*, or a noun title* `NT_E`*. Therefore sentence analysis is not triggered at vertex 5.*

central ideas in sentence end identification:

- In case of *unambiguous sentence-final punctuation symbols*, sentence analysis is started immediately. This is done at chart vertices where all word category edges that end in this vertex are tagged with the keyword `':SENT_END'`.
- For *ambiguous punctuation symbols*, all alternative word categories containing the punctuation symbol are inserted into the chart. Sentence analysis is started only at the next unambiguous sentence end.
- At the *end of a paragraph*, indicated by the paragraph boundary symbol `"<PB>"`, sentence analysis is always started.

Figure 2.3 illustrates the first two situations: In case of `"street.␣"`, as presented on the left side, word analysis returns an English noun `N_E`, which contains an empty noun ending `NE_E` and an empty word delimiter `TRM_E`. This noun is followed by an unambiguous sentence end `PCT_E`, which spans the period and the blank character. The corresponding morpheme lexicon entries are listed in Appendix B.1.

In contrast to this, the right side of Figure 2.3 shows word analysis results in case of an ambiguous sentence end. The period in the input sequence `"st.␣"` may be a full stop indicating the sentence end as well as the termination of the abbreviation of "street" or "Saint". Word analysis therefore produces four different word sequences for this input: a noun `N_E` or a noun title `NT_E`, or a sequence of a noun or a noun title followed by a punctuation symbol `PCT_E`.

These alternative word sequences can be disambiguated by subsequent syntactic analysis. Figure 2.2 illustrates such a disambiguation. As sentence end decision in chart vertex 13 is ambiguous (two word category edges without `':SENT_END'` keyword end in this vertex), sentence analysis is started only after the final paragraph boundary symbol `"<PB>"` has been reached. Sentence analysis produces two different sentence sequences containing two different readings of the first period, i.e., a full stop or part of an abbreviation. Subsequent paragraph analysis finally disambiguates the category of this punctuation symbol by selecting the sentence sequence with minimal total penalty, as described in Section 2.3.1.

2.4 Mixed-lingual Morphological Analysis

Mixed-lingual word analysis applies separate monolingual lexica and separate monolingual word grammars plus additional bilingual word inclusion grammars in parallel to parse a given graphemic input sequence morphologically. Appendix B contains example lexica and grammars for the languages English, French, German, and Italian. The central idea when analyzing mixed-lingual input text is to favor always monolingual analysis results over mixed-lingual ones. This is achieved by setting the penalty values of inclusion rules for a given constituent higher than the overall penalty values of any monolingual analysis result of this constituent.

Figure 2.4 illustrates on the left side the morphological analysis of the mixed-lingual German word

"($_{\texttt{G}}$($_{\texttt{E}}$up)ge($_{\texttt{E}}$dat)et)" (updated)

and on the right side the morphological analysis of the monolingual German word

"($_{\texttt{G}}$datiert)" (dated).

The stem "dat" is highly ambiguous. It can be an English verb stem `VS_E` (cf. lexicon entry L34), an English noun stem `NS_E` (L19), a French verb stem `VS_F` (L86), or an Italian adjective stem `AS_I` (L168). Additionally, "datiert" can be analyzed using the German verb stem `VS_G` "datier" (L127), and "upgedatet" using the French noun stem `NS_F` "date" (L76).

"upgedatet" on the left side of Figure 2.4 demonstrates how language mapping restrictions using inclusion grammar rules are applied in practice. The English and French verb stems can be both analyzed as foreign German verb stems using the inclusion rules of Appendix B.3. Inclusion grammar rule R84 maps the English verb stem to a German verb stem with verb class feature value `v1`. Inclusion grammar rule R86 includes the French verb stem using feature value `v12`. Another inclusion grammar rule, R85, maps the English prefix "up" to a German prefix. As the verb class feature value of the German past participle ending `P2E_G` "et" is `v1`, only the embedded English verb stem can be unified with this ending using the German word grammar rule R75. Thus, the

only word constituent, that can be analyzed using word grammar rules, is a German past participle `P2_G` with two English inclusions. This is in fact the desired analysis of "($_{G}$($_{E}$up)ge($_{E}$dat)et)".

"datiert" on the right side of Figure 2.4 shows how inclusion rule penalty values are used to disambiguate the correct analysis result of multiple possible results. The German verb stem `VS_G` "datier" together with the German past participle ending `P2E_G` "t" forms a monolingual German past participle `P2_G` with an overall penalty value of 4. Also, all multilingual variants of the stem "dat" are inserted into the chart. The French verb stem `VS_F` is mapped to a German verb stem with the verb class feature value `v12`. Using German word grammar rule R76 this stem can be unified with the German past participle ending "iert" to a German past participle with an overall penalty value of 163. As both `P2_G` constituents are grammatically equal, subsequent sentence analysis will choose the one with the lower penalty value. So, "($_{G}$datiert)" is correctly analyzed as a monolingual German word.

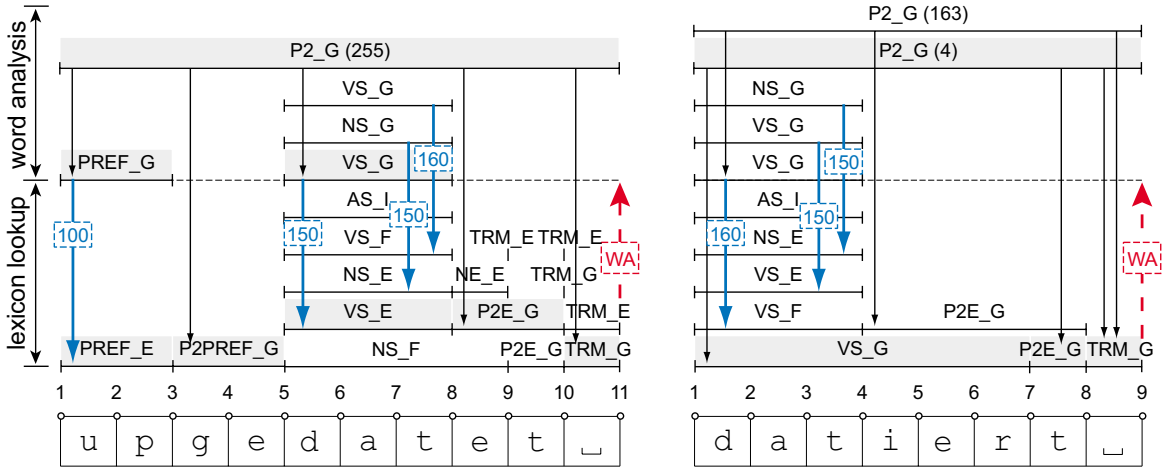

Figure 2.4: *Representations of the simplified charts resulting from morphological analysis of the mixed-lingual German word "upgedatet " (updated) and the monolingual German word "datiert " (dated). An arrow with bold line tagged with a penalty value denotes the application of an inclusion grammar rule. The constituents of the final morphological parse tree are shown with grey background.*

2.5 Mixed-lingual Syntactic Analysis

The polySVOX system accomplishes syntactic analysis in two steps: a sentence analysis step and a paragraph analysis step. Similar to mixed-lingual word analysis, separate monolingual sentence and paragraph grammars plus additional bilingual sentence and paragraph inclusion grammars are applied for mixed-lingual syntactic analysis. The result of syntactic analysis is a mixed-lingual morpho-syntactic parse tree, which describes the syntactic structure of the sentences and the morphological structure of the words. Foreign inclusions are easily identified within this parse tree, as each constituent is tagged by a suffix indicating the language.

Figure 2.5 demonstrates how polySVOX analyzes the mixed-lingual

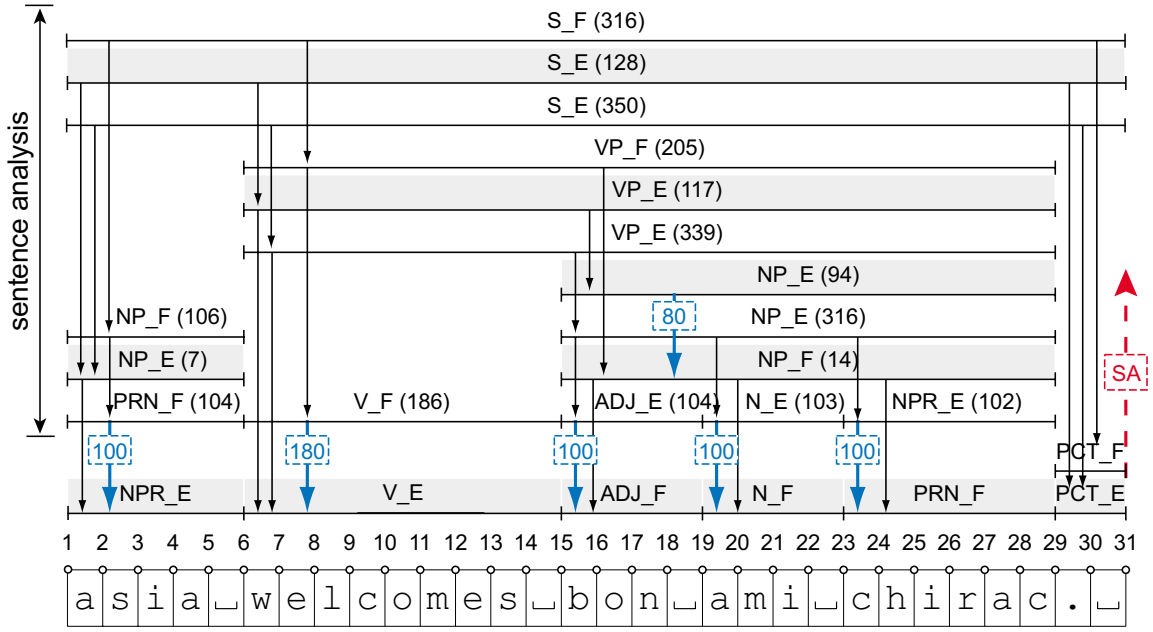

Figure 2.5: *Representation of the simplified chart resulting from morphological and syntactic analysis of the mixed-lingual English sentence "Asia welcomes bon ami Chirac." The bottom line of constituents comprises word constituents resulting from mixed-lingual word analysis. The "sentence analysis" section contains edges associated with constituents resulting from sentence analysis.*

English sentence

"($_{\texttt{E}}$Asia welcomes ($_{\texttt{F}}$bon ami Chirac.))"

that contains a majority of French words. In this sentence it is also important to correctly analyze the French noun phrase, as this information is necessary for subsequent phonological processing, like the application of French liaison rules, and the generation of a proper prosody. Three different syntactic analysis results are discussed, which are illustrated in Figure 2.5:

- `S_E (350)`: Word analysis returns "bon" as French adjective `ADJ_F`, "ami" as French noun `N_F`, and "Chirac" as French proper noun `PRN_F`. These French constituents are mapped onto corresponding English constituents using the inclusion grammar rules R40, R41, and R42 of Appendix B.1. Their inclusion penalties sum to 300. Applying monolingual English sentence grammar rules these embedded French inclusions can be analyzed as an English noun phrase `NP_E` with an overall penalty value of 316. With this English noun phrase the analysis as an English sentence gets an overall penalty value of 350.

- `S_E (128)`: The French adjective, noun, and proper noun can also be analyzed as a French noun phrase `NP_F`. This French noun phrase is then mapped to an English noun phrase with an overall penalty value of 94 using inclusion grammar rule R43. In this case, the inclusion penalty is only 80. The analysis as an English sentence with this embedded French noun phrase results in an overall sentence penalty value of only 128.

- `S_F (316)`: Also, it is possible to analyze the English proper noun "Asia" and the English verb "welcomes" as foreign inclusions within a French sentence. These English constituents are mapped to French constituents using French inclusion rules R66 and R67 of Appendix B.2. The summed inclusion penalty is 280. As the inclusion of an English verb within a French sentence is more unlikely, the inclusion grammar rule R67 has a higher penalty value. Overall sentence penalty value of the so analyzed French sentence `S_F` is 316.

Of these three results of sentence analysis, the one with the lowest overall penalty is finally chosen. This is the English sentence including the complete French noun phrase `NP_F` as a foreign inclusion. The resulting morpho-syntactic parse tree contains the correct identification of the sentence base language and of the language of the foreign multi-word inclusion, the correct mixed-lingual phone sequence, and the correct syntactic structure of the foreign multi-word inclusion.

Figure 2.5 shows how mixed-lingual syntactic analysis correctly analyzes the syntactic structure of foreign inclusions. This is achieved inherently by choosing the result with minimal overall penalty, as the inclusion of larger constituent structures adds up fewer inclusion penalties to the overall penalty. The constraints set by monolingual sentence grammars and bilingual sentence inclusion grammars additionally specify how larger foreign inclusions are syntactically analyzed, and restrict thereby the number of possible solutions. The specification of different inclusion penalty values allows to distinguish between common and uncommon foreign inclusions.

Treatment of Unparsable Sentences

If a *sentence cannot be analyzed* using the given sentence grammar rules, an artificial parse tree is created by finding the way through the chart with minimal total penalty, cf. [Tra95]. This is done by applying a dynamic programming technique. All constituents of this minimal path are then interpreted as direct descendants of an artificially generated, language-independent sentence constituent `NOSYNS`. An additional edge penalty leads to the preference of higher constituents over the combination of lower constituents. If, e.g., no sentence constituents `S_E` or `S_F` exist in the chart of Figure 2.5 after parsing, the constituent `NOSYNS` will be inserted into the chart having the constituents of the minimal path, i.e., the sequence of `NP_E`, `VP_E`, and `PCT_E`, as direct descendants.

2.6 Disambiguation of Interlingual Homographs

Interlingual homographs, as outlined in Section 2.1.2, are a major problem to language identification. Figure 2.6 illustrates by means of the

mixed-lingual German sentence

> "($_{\texttt{G}}$Die ($_{\texttt{E}}$Greatest Nation) hat die ($_{\texttt{F}}$Grande Nation) als tonangebende Nation abgelöst.)"
> (The Greatest Nation replaced the Grande Nation as leading nation),

how mixed-lingual syntactic analysis of the polySVOX system disambiguates such interlingual homographs.

In this sentence the first instance of "nation" is English, the second one is French, and the final one is German. The disambiguation of these homographs alone is impossible, as they all have the same syntactic function (all are singular nouns). For correct disambiguation the language of the neighboring words of these nouns must be considered additionally. The English adjective "greatest" produces an English noun phrase `NP_E` with the English variant of "nation". Likewise, the French adjective "grande" forms a French noun phrase `NP_F` with the French variant of "nation", and the German adjective "tonangebende" a German noun phrase `NP_G` with the German variant of "nation". The English and French noun phrases are finally included as foreign noun phrase inclusions within the German sentence.

An alternative analysis of this sentence includes the English and French adjectives as foreign adjective inclusions within German noun phrases. But, as the inclusion of a complete noun phrase is less penalized than the inclusion of a separate adjective (see, e.g., inclusion rules R91 and R92 in Appendix B.3), the analysis with complete foreign noun phrase inclusions is preferred. E.g., by including "greatest nation" as a foreign noun phrase the first German noun phrase (`NP_G`) in Figure 2.6 gets an overall penalty of 111 compared to 169 if only the English adjective "greatest" would have been included as foreign inclusion. Likewise, the final German noun phrase gets an overall penalty of 132 when including the complete French noun phrase "grande nation" versus an overall penalty of 143 in case of including only the French adjective "grande".

Figure 2.6 also shows two other interlingual homographs that do not arise from loanwords: one is "hat", which is an English noun as well as a German verb. The other one is "die", which is an English verb as well as a German determiner. These interlingual homographs are disambiguated by syntactic means: Using the German variants of

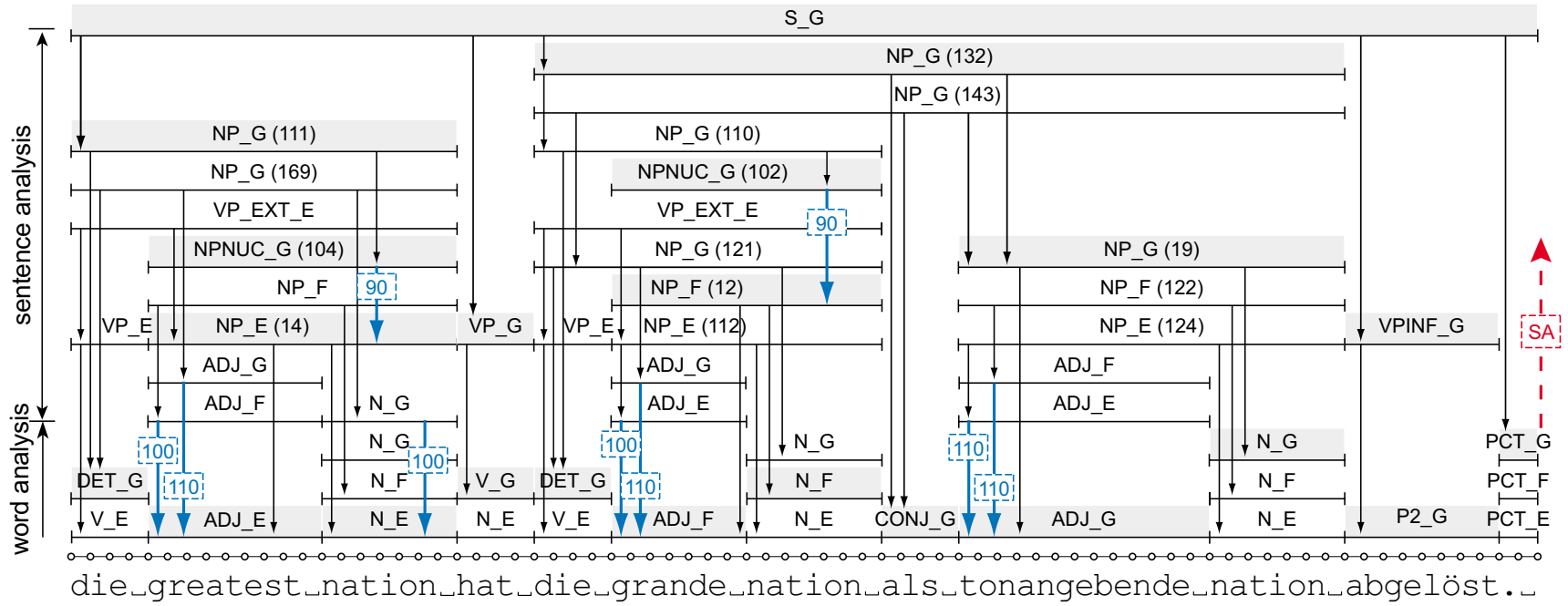

Figure 2.6: *Representation of the simplified chart resulting from morphological and syntactic analysis of the mixed-lingual German sentence "Die Greatest Nation hat die Grande Nation als tonangebende Nation abgelöst."*

these homographs a correct German sentence can be analyzed. With the English variants no syntactically correct sentence is possible.

2.7 Unknown Words in Mixed-lingual Sentences

TTS synthesis systems are expected to read any text, even if these texts contain words that are not parseable by word analysis, either as one or more morphemes are missing in the lexicon, or as they are simply misspelled. Such words are often referred to as "unknown" words. Concerning missing lexicon entries, their number can be reduced by utilization of large lexica. Nevertheless, there will always be a remainder of words (especially proper nouns) that are not covered even by very large lexica, cf. [CCL90]. Concerning misspelled words, listing them in lexica is obviously impossible.

Monolingual TTS synthesis systems usually incorporate some form of a rule-based grapheme-to-phoneme mapping by which a monolingual pronunciation of any unknown word is derived in the system's language.

In mixed-lingual sentences unknown words may additionally be full foreign words or even mixed-lingual words. However, unknown mixed-lingual words occur very seldom, as mixed-lingual words are typically built using common foreign stems, which are anyway included in a reasonably sized lexicon. In order to derive the correct pronunciation of unknown foreign words, polyglot TTS synthesis systems need to comprise a mixed-lingual text analysis component that is able to identify the correct language of unknown words and provide language-dependent grapheme-to-phoneme mappings. The following mixed-lingual sentence illustrates an example of an unknown English proper noun, which is embedded in a German sentence. The unknown word is noted in italics:

"($_{\texttt{G}}$Er lebt in ($_{\texttt{E}}$British *Columbia*).)"

The correct pronunciation of this sentence requires the unknown word to be analyzed by the English grapheme-to-phoneme conversion and not, e.g., the one of the sentence's base language.

The following assumptions served as a basis for the design of the polySVOX mixed-lingual grapheme-to-phoneme conversion:

- Closed word categories, like prepositions, conjunctions, determiners, or pronouns, of each language of the system are supposed to be completely listed in the respective monolingual lexica. Thus, the only possible word categories for an unknown word are the open word categories, i.e., nouns, verbs, adjectives, and adverbs.
- An unknown word is analyzed as monolingual word in each of the system's languages in parallel using the respective monolingual grapheme-to-phoneme conversion algorithms.
- Mixed-lingual syntactic analysis finally disambiguates all available multilingual pronunciations of an unknown word by the word categories.

The polySVOX system applies for all four languages the same algorithm to unknown word analysis. This algorithm is based on chart parsing using penalty extended DCG rules. It implements for each language a separate, monolingual *word stem analysis*, which decomposes unknown words into unknown word stems and known inflectional endings, prefixes, and suffixes, which are part of the monolingual morpheme lexica. Special word grammar rules describe the syllabic structure, word stress assignment, and pronunciation of unknown word stems. For unknown English stems, e.g., these rules define stressed versus unstressed and long versus short pronunciation of vowels, and word initial, word final, and word central pronunciation of consonant clusters. Appendix B.1 lists some of the English unknown stem analysis rules (rules R12 to R24).

In addition to the morpheme lexica a separate, monolingual lexicon of grapheme clusters, the so-called submorphemic lexicon, is loaded for each monolingual unknown word analysis. Each submorphemic lexicon contains all possible syllable onsets, codas, and nuclei together with their pronunciation variants of the respective language. Appendix B.1 shows some entries of the English submorphemic lexicon. Given specific inflectional endings and suffixes it is possible to make assumptions about the syntactic categories of unknown words and to provide appropriate pronunciations and word stress assignments. [Tra95] describes in detail this stem analysis for unknown German words.

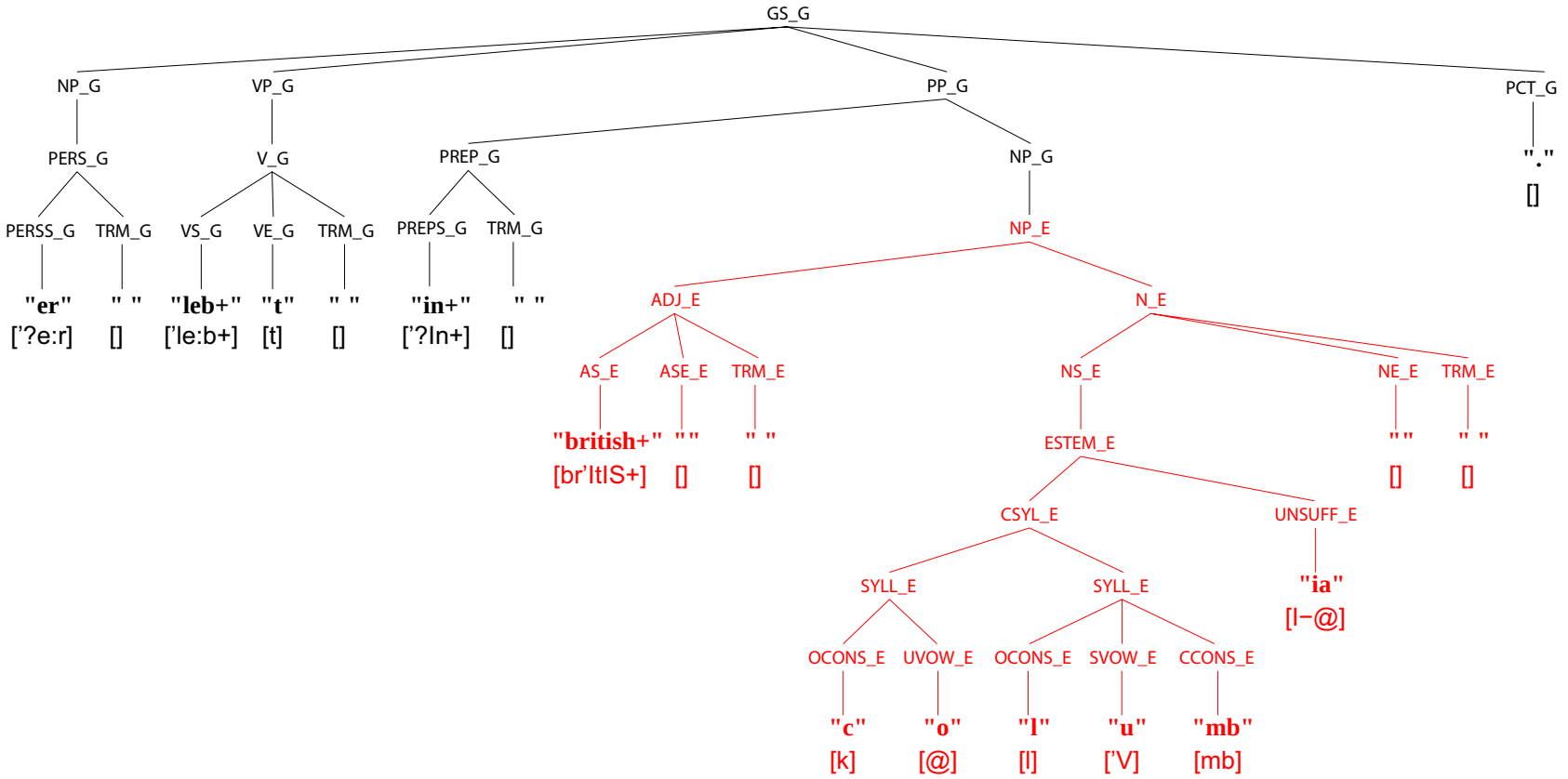

Figure 2.7: *Morpho-syntactic tree resulting from polySVOX morpho-syntactic analysis of the mixed-lingual German sentence "Er lebt in British* Columbia.*" The English proper noun "Columbia" is unknown to the system. Phonetic symbols are denoted using the ASCII-representation of Appendix A.*

Figure 2.7 shows the morpho-syntactic tree resulting from polySVOX morpho-syntactic analysis of the mixed-lingual example sentence "($_{\texttt{G}}$Er lebt in ($_{\texttt{E}}$British Columbia).)". As the English proper noun "Columbia" is unknown, grapheme-to-phoneme mappings in English, French, German, and Italian are applied to this word. Each of these monolingual grapheme-to-phoneme conversions derives pronunciations of "Columbia" for one or more word categories. Of these alternatives mixed-lingual sentence analysis selects the English noun category `N_E`, as this word category fits best the syntactic structure of the remaining sentence.

2.8 Language Identification Experiments and Discussion

2.8.1 Mixed-lingual Sentence Corpus

For the evaluation of language identification accuracy of the polySVOX system, the author collected and manually tagged a test corpus of 612 mixed-lingual sentences. The majority of these sentences comes from Swiss newspapers in German ("Neue Zürcher Zeitung", "Blick", and "20 Minuten") and in French ("Le Matin" and "Tribune de Genève"). Additionally, the example sentences of Table 2.1 were included in the corpus. Some of the English sentences, finally, were taken from articles found on the internet.

This test corpus contains 36 English, 35 French, and 541 German mixed-lingual sentences, which together comprise 8511 words. Punctuation symbols are not counted. Table 2.3 shows detailed word number statistics of the corpus. 1903 (22.4%) of all words are foreign inclusions. These inclusions consist of 1593 full foreign words and 310 mixed-lingual words.

2.8.2 Sentence Base Language Identification

All sentences of the mixed-lingual corpus have been separately analyzed without any context sentences. Still, the base language of all 612 sentences but one was correctly identified. This one German mixed-lingual sentence, "Weltcup-Leader Simon Schoch out" - a heading in a

Sent.	Word	Base Lang.		Full Incl.		Mixed Incl.		Sum	
ENG	ENG	410	(30)	-		3	(0)	413	(33)
	FRE	-		58	(3)	0		58	(3)
	GER	-		12	(0)	0		12	(0)
	ITA	-		28	(0)	0		28	(0)
	Sum	410	(30)	98	(3)	3	(0)	511	(33)
FRE	ENG	-		109	(13)	0		109	(13)
	FRE	412	(34)	-		0		412	(34)
	GER	-		16	(7)	0		16	(7)
	ITA	-		13	(1)	0		13	(1)
	Sum	412	(34)	138	(21)	0	(0)	550	(55)
GER	ENG	-		861	(117)	0		861	(117)
	FRE	-		420	(53)	0		420	(53)
	GER	5786	(327)	-		307	(28)	6093	(355)
	ITA	-		76	(17)	0		76	(17)
	Sum	5786	(327)	1357	(187)	307	(28)	7450	(542)
Sum		6608	(391)	1593	(211)	310	(28)	8511	(630)

Table 2.3: *Number of words of the mixed-lingual sentence corpus. This corpus contains 36 English, 35 French, and 541 German sentences with English, French, German, and Italian inclusions. The numbers are grouped row-wise according to the base language of the sentence (Sent.) and the language of the words (Word). The columns contain the numbers of monolingual words of the sentence base language (Base Lang.), of full foreign inclusions (Full Incl.), and of mixed-lingual words (Mixed Incl.). For every category, the number of unknown words is given in parentheses beside the number of all words.*

Swiss German newspaper, was analyzed as an English mixed-lingual sentence.

As "Schoch" is an unknown word in the TTS system and "Simon" is an English as well as a German forename, the analysis as an English sentence, "($_{\texttt{E}}$($_{\texttt{G}}$Welt)cup-Leader Simon Schoch out)", with only one foreign inclusion gets lower penalty than the analysis as a German sentence, "($_{\texttt{G}}$Welt($_{\texttt{E}}$cup)-($_{\texttt{E}}$Leader) Simon Schoch ($_{\texttt{E}}$out))", containing three foreign inclusions. However, as this sentence is the heading of a German article, paragraph analysis would finally chose the correct German reading of this sentence.

2.8.3 Language Identification of Words

Table 2.4 shows the word language confusion matrix grouped by the sentence base languages. From this table one can easily verify that the language of 8314 of all 8511 words (97.7%) was correctly identified. However, this number is not really representative, as it largely depends on the rate of sentence base language words. Therefore, separate numbers for word language identification of sentence base language words and of foreign inclusions are presented.

Table 2.5 gives detailed word language identification results in terms of precision and recall for sentence base language words and full foreign inclusions grouped by the sentence base languages and in total for the whole mixed-lingual corpus. In total, the language of full foreign inclusions was identified with a balanced *F-score of 95.5%*. The language of sentence base language words was identified with an *F-score of 98.7%*.

F-scores for language identification of foreign inclusions within the English and French sentences sets separately are even higher, i.e., 98.0%

Sent.	Word	ENG	FRE	GER	ITA
ENG	ENG	412	1	0	0
	FRE	3	55	0	0
	GER	0	0	12	0
	ITA	0	0	0	28
FRE	ENG	105	3	1	0
	FRE	2	409	1	0
	GER	0	0	16	0
	ITA	0	0	0	13
GER	ENG	801	2	58	0
	FRE	17	379	24	0
	GER	54	19	6017	3
	ITA	0	0	9	67

Table 2.4: *Word language confusion matrix of English, French, and German mixed-lingual sentences. The rows show the reference language of a word, the columns show the language assigned to a word by polySVOX.*

and 97.1% resp. The language of words of the sentence base language was identified with F-scores of 99.6% for English and 99.3% for French mixed-lingual sentences. These results verify the authors' experience that foreign inclusions can easier be identified in English or French sentences than in German ones. However, the results for English and French sentences alone must be read with some care as the number of English and French sentences is rather limited (i.e., 36 English and 35 French sentences).

2.8.4 Inclusions in Mixed-lingual Words

Mixed-lingual words are mainly found in German sentences. Table 2.3 shows that 307 of 1664 foreign inclusion (18.4%) in the German sentences are mixed-lingual words. In the English test sentences only 3 mixed-lingual words were found. These are actually full foreign inclusions with English plural or s-Genitive suffixes, i.e., "($_{\texttt{I}}$cappucino)s", "($_{\texttt{I}}$lasagna)s", and "($_{\texttt{F}}$cuisine)'s". In the French test sentences no mixed-lingual words exist.

All English mixed-lingual words and 260 of the 307 German mixed-

		Word Language				Base	Foreign
		ENG	FRE	GER	ITA	Language	
ENG	Precision	99.3	98.2	100	100	99.3	99.0
	Recall	99.8	94.8	100	100	99.8	96.9
FRE	Precision	98.1	99.3	88.9	100	99.3	97.1
	Recall	96.3	99.3	100	100	99.3	97.1
GER	Precision	91.9	94.8	98.5	95.7	98.5	95.6
	Recall	93.0	90.2	98.8	88.2	98.8	94.9
Total	Precision	95.3	97.1	98.5	97.3	*98.6*	*95.8*
	Recall	94.6	94.7	98.8	92.3	*98.8*	*95.1*

Table 2.5: *Precision and recall results of word language identification given in percent for English, French, and German mixed-lingual sentences separately and for the mixed-lingual corpus in total. To the right, precision and recall results for word language identification of full foreign inclusions and of sentence base language words are shown.*

lingual words, *84.8%* of the mixed-lingual words in total, are correctly analyzed. The erroneous analyses of German mixed-lingual words originate from four main sources:

1. *Analysis as a full foreign word:* the polySVOX analysis prefers full, foreign monolingual noun inclusions that syntactically agree over mixed-lingual inclusions. Therefore, a mixed-lingual word that contains a German word part which is ambiguous to a word of the inclusion language is analyzed as a monolingual foreign word (cf. the examples in Table 2.6). This is the most common error of mixed-lingual word analysis.

2. *Analysis as monolingual base language word:* as our analysis prefers monolingual words of the sentence base language over any foreign inclusion, every mixed-lingual word containing a foreign inclusion that is ambiguous to a base language morpheme is always analyzed as monolingual word in the sentence base language.

3. *Ambiguous base language and foreign morphemes:* the polySVOX analysis prefers ambiguous base language morphemes that accord with the morphological rules over foreign morphemes. Exceptions are incorrectly analyzed, as shown in Table 2.6.

	Reference analysis	Analysis by polySVOX
1	($_{G}$($_{F}$Gourmet)-($_{E}$Festival)) ($_{G}$($_{E}$Bluetooth)-System) ($_{G}$Index($_{F}$fonds)) ($_{G}$Auto($_{E}$rowdy)) ($_{G}$($_{E}$Lifestyle)-Hotel)	($_{F}$Gourmet-Festival) ($_{E}$Bluetooth-System) ($_{F}$Indexfonds) ($_{E}$Autorowdy) ($_{E}$Lifestyle-Hotel)
2	($_{G}$($_{E}$Hacker)attacken)	($_{G}$Hackerattacken)
3	($_{G}$Firmen-($_{E}$Website)) ($_{G}$($_{E}$All-In-One)-Navigationsgerät)	($_{G}$Firmen-Web($_{E}$site)) ($_{G}$All-In-($_{E}$One)-Navigationsgerät)
4	($_{G}$Feinschmecker-($_{F}$Restaurants)) ($_{G}$($_{F}$Amateur)truppe) ($_{G}$Viertel($_{F}$final))	($_{G}$Feinschmecker-($_{E}$Restaurants)) ($_{G}$($_{E}$Amateur)truppe) ($_{G}$Viertel($_{E}$final))

Table 2.6: *Examples of incorrectly analyzed mixed-lingual words of the test corpus. The number in the left column indicates the description number in Section 2.8.4.*

4. *Ambiguous foreign morphemes of multiple languages:* ambiguous English and French morphemes are very common in German mixed-lingual words. As English morphemes are more often used in the German speaking part of Switzerland, the polySVOX text analysis prefers English morphemes over ambiguous French morphemes. Table 2.6 shows exceptions, in which the French pronunciation is more common. These are incorrectly analyzed using the English pronunciation.

2.8.5 Language Identification of Unknown Words

239 (12.6%) of the foreign inclusions and 391 (5,9%) of the sentence base language words are unknown (cf. Table 2.3). This means, they must be analyzed using unknown word analysis described in Section 2.7. Note, that the strict morphological analysis of words reduces the number of unknown words already considerably when compared to full form lexicon lookup.

Table 2.7 shows the word language identification results of unknown words and the inclusion language identification results of unknown mixed-lingual words. The polySVOX system assigns the correct language to 95.1% of the unknown words in the sentence base language, and to *72.5% of the unknown foreign inclusions.*

As the polySVOX system analyzes unknown words in a monolingual fashion (cf. Section 2.7), unknown mixed-lingual words are analyzed in-

	Unknown Words	Correct	Correct (%)
Base Lang.	391	372	95.1
Full Incl.	211	153	72.5
Mixed Incl.	28	11	39.3
Total	630	536	85.1

Table 2.7: *Results of language identification of unknown words. The rows show the total number of words, and the number and percentage of correctly identified words of unknown words of the sentence base language (*`Base Lang.`*), of unknown full foreign inclusions (*`Full Incl.`*), and of unknown mixed-lingual words (*`Mixed Incl.`*).*

correctly by default. However, unknown mixed-lingual hyphenated compounds are analyzed as a sequence of hyphenated monolingual words. Thus, polySVOX is still able to identify the correct inclusion languages within 11 of 28 (39.3%) unknown mixed-lingual words. In total, the language of *85.1% of 630 unknown words* was correctly identified.

Chapter 3

Mixed-lingual Phonological Processing

3.1 Introduction

In the polySVOX system, syllabification, accentuation, prosodic phrasing and various phonological transformations are done in the so-called phonological processing component. This component processes the syntax tree of the morpho-syntactic analysis and generates the phonological representation. The sequence of processing steps is as follows:

1. Word syllabification
2. Word accentuation
3. Prosodic phrasing
4. Sentence accentuation
5. Phonological transformations
6. Sentence re-syllabification

The following sections explain, after a description of the phonological representation and the multi-context rule formalism, each processing step in detail.

3.2 Formalism of Phonological Processing

3.2.1 Phonological Representation

The phonological representation contains the phonetic transcription of the words to be uttered, the position of language switches, the accentuation level of syllables, the position and type of phrase boundaries, the indicator of phrase types, and the indicator of the base language of the utterance. For the sentence

> "Le monde à l'envers, ein französischer Film, war ein sehr guter, bekannter Film."
> (Le monde à l'envers, a French movie, was a very good, well known movie.)

the following phonological representation is obtained in the current polySVOX system:

> #{P:G:0} \F\lə- m[2]õː-d a- l ɑ̃-v[1]ɛːʁ #{P:1} \G\ʔa͡in- fran-ts͡s[2]øː-zɪ-ʃɐ- f[1]ɪlm #{P:1} v[2]aːɐ̯- ʔa͡in- z[1]eːr- g[2]uː-tɐ #{T:1} bə-k[2]an-tɐ- f[1]ɪlm

In addition to the phone symbols, the following special symbols appear in the phonological representation:

\L\ indicates a language switch. All phones following this switch up to the next language switch are produced using language *L*. Currently, the languages English, French, German, and Italian are supported. These are denoted by \E\, \F\, \G\, and \I\, resp.

\- marks the boundary between two adjacent syllables. Also word final syllable boundaries are denoted, since in some languages, like French, word boundaries are no mandatory syllable boundaries. A word boundary can optionally be indicated by a blank character.

[A] denotes the accentuation level *A* of the syllable. It is placed before the syllable nucleus. The accentuation levels are interpreted as follows:

[1] denotes the main accent of an intonational phrase. This is the anchor point of the phrase intonation pattern.

`[2]` denotes a pitch accent, i.e., an accent with a major pitch movement and a lengthened syllable duration.

`[3]` denotes a non-pitch accent on the main stress syllable of a word, i.e., an accent with a lengthened syllable duration, but without a major pitch movement.

`[4]` denotes a syllable with secondary or tertiary word stress.

`[0]` denotes an unaccented syllable. This level is set for all syllables without explicit accent mark.

`[E]` denotes an emphatic accent. An emphatic accent is characterized by a major pitch movement that normally exceeds standard pitch accents, a lengthened syllable duration, and an increased syllable signal energy. No further distinction between different types of emphatic accents is made.

`#{T:L:B}` indicates a phrase boundary, where T indicates the type of the following phrase, L the base language of the following phrase, and B the type of the phrase boundary.

The classification of phrase types is based on the distinction made in [Dud84] and [vE56]. They classify phrases according to the final part of their intonation contour into phrases with

terminal intonation pattern ("Vollschluss" in German), i.e., a complete fall of the phrase final intonation contour,

semi-terminal intonation pattern ("Halbschluss"), having a non-complete fall of the phrase final intonation contour,

progredient intonation pattern ("Schwebekadenz", also referred to as *continuation rise*), with no or only a slight rise of the phrase final intonation contour, and

interrogative intonation pattern ("Steigkadenz"), i.e., a high rise of the phrase final intonation contour.

The following *phrase types* are currently defined:

`T` denotes a phrase with *terminal* intonation pattern and a low overall tone range.

`E` denotes a phrase with *terminal* or *semi-terminal* intonation pattern, a high overall tone range, and a high overall signal intensity.

`S` denotes a phrase with *terminal* or *semi-terminal* intonation pattern and an overall tone range that lies between the tone ranges of `T` and of `E` phrase types.

`P` denotes a phrase with *progredient* intonation pattern.

`Y` denotes a phrase with *interrogative* intonation pattern.

`YC` denotes a phrase with *interrogative* intonation pattern for confirmation. The intonation contour has a short final fall after the rise.

The *base language* indicator of a phrase is optional except for the first phrase of an utterance. If the base language indicator is not present, the language of the preceding phrase is taken. The same languages as for language switches are possible: an English base language is denoted by `E`, French by `F`, German by `G`, and Italian by `I`.

The following *phrase boundary types* are defined:

`0` indicates a sentence-final phrase boundary with a pause.

`1` indicates a sentence-internal phrase boundary with a pause.

`2` indicates a sentence-internal phrase boundary without a pause.

3.2.2 Multi-context Rules

A rule formalism that is flexible enough to describe all possible context restrictions of phonological transformations was introduced in [RP04, RPB05]. This so-called multi-context rule formalism allows to define phonological transformations which are restricted by specific syntactic, graphemic and/or phonological contexts. Formally, a multi-context rule consists of a subtree pattern, the separation symbol ':' and an associated phonological transformation:

SubtreePattern : Transformation ;

Transformations are specified in the form: $\sigma/\rho \Leftrightarrow L \_ R$. These context-dependent rewrite rules are similar to the well-known *two-level rules* (see e.g. [Kos83, Rus90, Rus92]), but have been extended here to operate on all types of symbols, i.e., graphemic and phonological ones at the same time. The polySVOX system accepts these two-level rules only in the form of *finite state transducers* (FSTs). The formalisms of two-level rules and of FSTs used in this thesis are presented in [Tra95]. The rule-to-FST conversion was done using software described in [Tra97].

The *subtree pattern* specifies the syntactic context and defines for each constituent whether the graphemic and/or phonological terminals are subject to the phonological transformation defined by the rule. These patterns are represented as strings and may be specified using constituent symbols plus additional wild-card symbols as listed in Table 3.1. The application of the associated transformation gets triggered whenever the subtree pattern can be matched with a part of the syntax tree. The following examples of such subtree patterns, also shown in Figure 3.1, are used in the current polySVOX system:

```
NP_F ( * ADJ_F {} [] N_F [] * )
NP_F ( * N_F <NR=PL> {} [] ADJ_F [] * )
```

The first pattern (pattern a) specifies the syntactic context of French mandatory liaison between a noun and a preceding adjective within a French noun phrase. The operator '`[]`' selects the phonological terminals of both constituents for application by the associated phonological

`*`	any sequence (0...n) of constituents including their (possibly empty) subtrees
`?`	any constituent (exactly one) including its (possibly empty) subtree
`(...)`	syntax hierarchy marker
`<...>`	feature specification
`[]`	phonological representation operator
`{}`	graphemic representation operator
`%id`	set identifier

Table 3.1: *Wild-card and special symbols used within multi-context rules.*

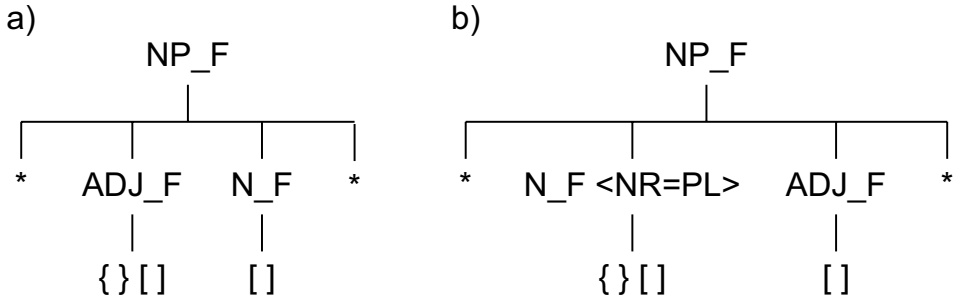

Figure 3.1: *Examples of subtree patterns for use in the polySVOX phonological transformation step.*

rule. Analogously, the operator '`{}`' selects the graphemic terminals of the first constituent. The second pattern (pattern b) specifies the syntactic context of French optional liaison between a plural noun and a subsequent adjective. The pattern specifies the noun with an additional feature-value pair `<NR=PL>`, i.e., to select only plural nouns.

3.3 Syllabification

The syllabification of the phonetic transcription, i.e., the assignment of a syllable boundary between each pair of neighboring syllable nuclei (vowels, diphthongs, or syllabic consonants), is an important preparatory step for all subsequent phonological and prosodic processing. Before any other phonological transformation, syllabification is applied on the phonetic sequence of each word. At the end of phonological processing, the phonetic sequence of the whole sentence is (re-)syllabified, in order to correct syllable boundaries in case of phone elisions or insertions generated by phonological transformations.

Some morphologically motivated syllable boundaries are already set by the morpho-syntactic analysis, e.g., obligatory syllable boundaries before stems or after prefixes. The remaining boundaries must be assigned according to language specific, phonetic criteria. As references for such phonetic criteria, the author used for English [JRHS03], for French [War96], for German [Dud05], and for Italian [Pon95].

Mixed-lingual syllabification in the polySVOX system applies DCG-based bottom-up parsing using language specific, phonetic lexica of on-

set and coda consonant clusters. An additional, language-independent lexicon contains all syllable nuclei, all individual consonants, and all symbols for syllable language, syllable boundary, and phrase boundary. Small, language specific grammars describe the syllable structure of each language. The language tag associated with every syllable nuclei constrains parsing to the grammar rules of the corresponding language. A language-independent grammar finally describes the phonetic sequence of a word or a sentence as a sequence of syllables. Consonant sequences, that do not conform with any consonant cluster of the language specific lexica, are analyzed using the individual consonants of the language-independent lexicon.

An appropriate setting of the grammar rule weights makes the result of parsing follow the "onset maximalization" principle, cf. [Kah76], constrained by language specific onset and coda consonant clusters, and information about the morphological structure of words.

For example, mixed-lingual morpho-syntactic analysis will analyze "downgeloadet" (downloaded) as a sequence of two verb prefixes, an English one and a German one, followed by an English verb stem and a German verb ending. To indicate morphological constraints to syllabification, both verb prefixes are already terminated by a syllable boundary marker:

[d\E\'a͡ʊn-] [g\G\ə-] [l\E\'ə͡ʊd] [\G\ət].

Mixed-lingual syllabification would then syllabify this input as [\E\dˈa͡ʊn-\G\gə-\E\lˈə͡ʊ-\G\dət] which is the correct syllabification in this case. Accent markers that occur in the phonetic transcription are ignored. The English [d] is moved to the German syllable [dət] and thereby transformed into a German [d]. This conforms well with the assumption that syllables are always uttered using only one language.

3.4 Word Accentuation

For each word in the syntax tree, language specific word accentuation rules are applied. While the positions of primary and secondary word accents of English, German, and Italian lexemes are already set in the lexicon, French lexicon entries do not contain stress markers. For French words, a multi-context rule is applied that sets the primary word accent

on the last syllable whose nucleus is not a schwa. French inclusions in English, German, or Italian words also receive a primary word accent on the last non-schwa syllable.

If there is more than one primary word accent in the word, all but one of the primary accents are reduced to secondary word accents. French and Italian words are generally *right-accented*, this means, the most right primary word accent remains whereas other primary accents are changed to secondary word accent. English and German words, in contrast, are normally *left-accented*, i.e., the most left primary word accent remains. These default word accent patterns are implemented using the following multi-context rules, that match all accentuable word categories of a language and apply left-accentuation or right-accentuation on the phonetic sequence of the word:

```
%AccCons_F [] : "'" / "," <=> _ ? "'" ;
%AccCons_I [] : "'" / "," <=> _ ? "'" ;
%AccCons_E [] : "'" / "," <=> "'" ? _ ;
%AccCons_G [] : "'" / "," <=> "'" ? _ ;
```

In these rules, `%AccCons_*` specifies for each language a set of all accentuable word categories. The phonological transformations reduce all primary accents except the last or the first one, resp., to secondary accents.

The following examples illustrate this default word accentuation on English, French, German, and Italian word, resp.:

"typewriter"	[tˈa͡ɪp] + [rˈa͡ɪ.tə(r)]	⇒ [tˈa͡ɪp.rˌa͡ɪ.tə(r)]
"homme-orchestre"	[ˈɔ.m(ə)] + [ɔʀ.kˈɛs.tʀ(ə)]	⇒ [ɔ.mɔʀ.kˈɛs.tʀ(ə)]
"Bürgermeister"	[bˈʏr.gɐ] + [mˈa͡i.stɐ]	⇒ [bˈʏr.gɐ.mˌa͡i.stɐ]
"millecento"	[mˈil.le] + [t͡ʃˈɛn.to]	⇒ [mˌil.le.t͡ʃˈɛn.to]

In contrast to the default word accentuation, certain English and German words may have quite complex word accent patterns that depend on the morphological structure of the words. In the following, some of the word accentuation phenomena, for which multi-context rules have been implemented in the polySVOX system, are presented:

German verb prefixes can be separated into three groups with respect to word accentuation:

P1 Unaccented verb prefix. E.g., the prefix "ge" in "geleiten" (to accompany) is pronounced as [gə.lˈa͡i.tən].

P2 Verb prefix with primary word accent, that is removed if a stressed syllable follows: e.g., the prefix "miss" is unstressed in "missleiten" (to mislead) [mɪs.lˈa͡i.tən], but it carries word primary stress in "missgeleitet" (misled) [mˈɪs.gə.lˌa͡i.tət].

P3 Verb prefix with primary word accent, that reduces the verb stem accent: e.g., the prefix "hinunter" in "hinunterleiten" (to lead downwards) is stressed as [hɪ.nˈʊn.tɐ.lˌa͡i.tən]. Also foreign verb prefixes in mixed-lingual German verbs belong to this category. For example, the word accent pattern of "downgeloadet" (downloaded) is correctly analyzed as [\E\dˈa͡ʊn.\G\gə.\E\lˌə͡ʊ.\G\dət].

Verb prefix entries in the morpheme lexicon have got an additional feature indicating the group they belong to: `P1`, `P2`, or `P3`. Verb prefixes of the first group (`P1`) do not require any modifications of word accent pattern. Word accentuation for verbs having a prefix of the third group (`P3`) follows the default left-accentuation of German words. Verb prefixes of the second group (`P2`), however, require a special multi-context rule:

```
V_G ( * PREF_G <TYPE=P2> [] ? [] * ) :
          "'" / @ <=> _ ? ']['  { %C } "'" ;
```

In this rule, `%C` specify the set of all consonants. `']['` indicates the boundary between the phonetic sequences of the prefix and any subsequent constituent. Figure 3.2 shows the subtree pattern of this rule.

Applying this multi-context rule on the morphological structure of "missleiten" [mˈɪs-] [lˈa͡it] [ən] gives the following transformation

$$[\text{mˈɪs-}]\ [\text{lˈa͡it}] \quad \Longrightarrow \quad [\text{mɪs-}]\ [\text{lˈa͡it}]$$

and results in [mɪs.lˈa͡i.tən].

Applying this rule on "missgeleitet" [mˈɪs-] [gə-] [lˈa͡it] [ət] gives

$$[\text{mˈɪs-}]\ [\text{gə-}] \quad \Longrightarrow \quad [\text{mˈɪs-}]\ [\text{gə-}]$$

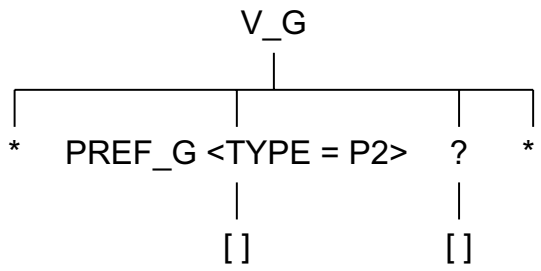

Figure 3.2: *Subtree pattern for German word accentuation: the subtree pattern matches a prefix of type* `P2` *and any subsequent constituent within a German verb. The operator '*`[]`*' selects the phonological terminals of both constituents for application by the associated phonological rule. The phonological rule will remove the primary word stress of the prefix, in case the subsequent morpheme has a primary word accent on the first syllable.*

and German left-accentuation results in [mˈɪs.gə.lˌa͡i.tət].

British English compound words are right-accented if the compound functions as verb, adjective, or adverb. This is accomplished using multi-context rules that apply right-accentuation on the corresponding English word constituents, i.e., `ADJ_E`, `V_E`, and `ADV_E`:

```
ADJ_E [] : "’" / "," <=> _ ? "’" ;
V_E [] : "’" / "," <=> _ ? "’" ;
ADV_E [] : "’" / "," <=> _ ? "’" ;
```

These rules, for example, generate the correct word accent pattern for the adjective "bad-tempered" as [bˌæd.tˈem.pəd], for the adverb "headfirst" as [hˌed.fˈɜːst], and for the verb "downgrade" as [dˌa͡ʊn.grˈe͡ɪd].

3.5 Prosodic Phrasing

Prosodic phrasing is the division of an utterance into several speech groups, or phrases. Depending on the language, prosodic phrases either depend more on syntactic constraints, as, e.g., in German utterances,

or more on rhythmic constraints, as, e.g., in French utterances. Mixed-lingual prosodic phrasing must therefore provide a language-dependent combination of both constraints.

Prosodic phrasing in the polySVOX system is based on the algorithm given by Bierwisch in [Bie66]. This algorithm was implemented in the German SVOX system by Traber, who describes the algorithm very detailed in [Tra95]. Therefore, only an overview of the phrasing algorithm itself is given here.

In contrast to the monolingual German phrasing of the SVOX system, the mixed-lingual prosodic phrasing is applied after word accentuation, but before sentence accentuation, as French sentence accentuation relies on the position of phrase boundaries. The phrasing procedure has the following steps:

- For each word in the syntax tree, a *numeric word accent pattern* is extracted from its phonetic representation. The primary word stress (') is therefore converted into a primary accent (1), and each secondary word stress (,) into a secondary word accent (4). Unstressed syllables are assigned 0. Some word categories (such as articles, prepositions, personal pronouns, coordination particles, and others) are declared to be unstressed. The primary word stress of these words is assigned a special, very weak accent level (99), and a secondary word stress becomes 102. For French words, all non-accentuable syllables, i.e., syllables containing schwa [ə] or optional schwa [(ə)], are indicated by a negative accent value (-1).

- *Initial phrase boundaries* are set between each pair of adjacent syntactic words according to the level of the closest common ancestor node of the two words in the syntax tree. The level of the root of the syntax tree is defined to be 3, and the level number is increased by 1 with each new tree level. Syntactically closely connected words are therefore separated by a weaker initial boundary than loosely connected words. A weaker boundary is indicated by a larger boundary value.

  Punctuation symbols, like periods or commas, are analyzed as word constituents by syntax analysis, and initial phrase boundaries are also set at their boundaries. To strengthen punctuation boundaries, their boundary values are reduced by 3. As punctuation symbols of clauses and subordinate clauses are in the current

syntax grammar near to the root level of the syntax tree, they get strong initial boundaries. Commas as part of a list, however, get weaker initial boundaries. This results in the desired effect, that clause boundaries are realized with strong phrase boundaries and short list items may have a weaker or even no phrase boundary at all between them.

- *Phonological word formation* combines all unaccented (clitic) words with the syntactically closer accented neighboring word to an initial phrase, that is often called a *phonological word*, cf. [BF90]. This initial phrase is thereby assigned the language of the first constituent of the syntax tree, form a bottom-up perspective, that spans all syllables of the phrase. These initial phrases also form the basis in the phrasing algorithm for English in [BF90] and for French in [Mer99].

- *Prosodic phrase formation* finds the final prosodic phrases by cyclic deletion of some of the intermediate phrase boundaries with a boundary value larger than some predefined threshold (currently 3). The criterion for deletion of intermediate boundaries is language-dependent. Two types of phrase boundaries are distinguished: Monolingual phrase boundaries separate phonological words of the same language. Mixed-lingual phrase boundaries separate phonological words of different languages.

  For monolingual English, German, or Italian phrase boundaries, the criterion described in [Tra95] is applied in cyclic fashion starting with the weakest boundary: a phrase boundary is deleted, if its boundary value is larger or equal to both neighboring boundary values, and if the number of accented syllables between this boundary and the boundary to the right or to the left is smaller than a threshold q, where

$$q = \begin{cases} p+1 & : \quad nsyl \leq 2 \\ p & : \quad 2 < nsyl < 5 \\ p-1 & : \quad nsyl \geq 5 \end{cases}$$

  with $nsyl$ specifying the number of syllables between this boundary and the boundary to the right or to the left, resp. p is a parameter defining the desired degree of phrasing. Usually, $p = 1$

or $p = 2$ is appropriate. Higher values of p delete more, lower values delete fewer boundaries.

Monolingual French phrase boundaries are deleted in such a way, that the number of syllables of the resulting prosodic phrases is as near as possible to a predefined parameter, that defines the desired degree of phrasing. This parameter is currently set to 6.

For the deletion of mixed-lingual phrase boundaries, the criterion according to the sentence's base language is applied.

The values of the resulting phrase boundaries are finally normalized to match the phrase boundary types defined in Section 3.2.1. In the current system, the phrase boundary value 0 is assigned the phrase boundary type `#{*:0}`, values of 1 or 2 are assigned `#{*:1}`, and any value larger than 2 the boundary type `#{*:2}`. The wildcard `*` denotes phrase type information.

Phrase Type Assignment

Phrases are assigned phrase types with respect to the modality of the sentence. Syntactic analysis of the polySVOX system is currently able to distinguish between statements, commands, exclamations, requests, wh-questions, yes/no-questions, and alternative questions. Table 3.2 shows the heuristic that is applied for phrase type assignment, given the sentence type, the number of phrases in the sentence, and the punctuation constituents.

For example, the phrase in a German statement having only one phrase receives the phrase type S, the phrases in a German statement with 5 phrases are assigned the phrase type sequence P P P P T, and the phrases in a French wh-question with 4 phrases the phrase types YC P P Y. The French alternative question

"Tu veux du café, de la bière, du vin ou du coca?"
(Do you want coffee, beer, wine, or coke?)

consists of 4 phrases which are assigned the phrase type sequence YC YC YC T.

sentence base language	English, German, or Italian		French	
number of phrases	1	≥ 2	1	≥ 2
statement	S	P^+ T	S	P^+ T
command, request, or exclamation	E	P^+ T	E	P^+ S
wh-question	Y	P^+ T	Y	YC P^* Y
yes/no-question	Y	(P^* YC , $)^+$ P^* Y	Y	(P^* YC , $)^+$ P^* Y
alternative question		(P^* YC ,& $)^+$ P^* T		(P^* YC ,& $)^+$ P^* T

Table 3.2: *Phrase type assignment with respect to sentence type, the number of phrases, and punctuation constituents. '*' denotes 0 to n repetitions. '+' indicates 1 to n repetitions. ',' represents a punctuation constituent at the phrase boundary. ',&' indicates a punctuation or a disjunctive conjunction constituent at the phrase boundary.*

Example of Mixed-lingual Phrasing

Figure 3.3 shows the syntax tree, that is generated by the current polySVOX system for the mixed-lingual German sentence "Le monde à l'envers, ein französischer Film, war ein sehr guter, bekannter Film".

The phrasing algorithm of polySVOX first extracts the numeric word accent patterns and sets initial phrase boundaries between words and punctuation symbols according to the syntactic connection strengths (Phrase boundaries are represented by `'#n'`. Words are denoted by word accent patterns in angular brackets):

```
#0 <-1> #6 <1 -1> #6 <99> #7 <> #8 <0 1> #2 <> #3 <99> #7
<0 1 0 0> #7 <1> #3 <> #1 <1> #4 <99> #5 <1> #6 <1 0> #3
<> #3 <0 1 0> #5 <1> #0 <> #0
```

Then, unaccented words are combined with the syntactically closest accented neighboring word to form phonological words:

```
#0 <-1>    <1 -1> #6 <99>    <>    <0 1> #2 <>    <99>
<0 1 0 0> #7 <1>    <> #1 <1> #4 <99>    <1> #6 <1 0>
<> #3 <0 1 0> #5 <1>    <> #0
```

French prosodic phrase formation removes the boundary between the

```
   3  4  5  6  7  8
+- GS_G
   |
   +- MCL_G
   |  |
   |  +- NP_G
   |  |  |
   |  |  +- NP_F
   |  |  |  |
   |  |  |  +- ARTDEF_F   "le"   "l@"
   |  |  |  |
   |  |  |  +- N_F   "monde"   "m~O:d(@)"
   |  |  |  |
   |  |  |  +- PP_F
   |  |  |     |
   |  |  |     +- PREP_F   "à"   "a"
   |  |  |     |
   |  |  |     +- NP_F
   |  |  |        |
   |  |  |        +- ARTDEF_F   "l'"   "l"
   |  |  |        |
   |  |  |        +- N_F   "envers"   "~AvE:R"
   |  |  |
   |  |  +- APPOS_G
   |  |     |
   |  |     +- PUNCT   ", "   ""
   |  |     |
   |  |     +- NP_G
   |  |     |  |
   |  |     |  +- ARTINDEF_G   "ein"   "'?a_in"
   |  |     |  |
   |  |     |  +- ADJ_G   "französischer"   "fran't_s2:zIS@r"
   |  |     |  |
   |  |     |  +- N_G   "film"   "'fIlm"
   |  |     |
   |  |     +- PUNCT   ", "   ""
   |  |
   |  +- VPFIN_G   "war"   "'va:r"
   |  |
   |  +- NP_G
   |     |
   |     +- ARTINDEF_G   "ein"   "'?a_in"
   |     |
   |     +- ADJP_G
   |     |  |
   |     |  +- GRADPART_G   "sehr"   "'ze:r"
   |     |  |
   |     |  +- ADJ_G   "guter"   "'gu:t@r"
   |     |  |
   |     |  +- PUNCT   ", "   ""
   |     |  |
   |     |  +- ADJ_G   "bekannter"   "b@'kant@r"
   |     |
   |     +- N_G   "film"   "'fIlm"
   |
   +- PUNCT   "."   ""
```

Figure 3.3: *Syntax tree generated by polySVOX for the sentence "Le monde à l'envers, ein französischer Film, war ein sehr guter, bekannter Film". The tree is shown in a simplified indentation form down to the full-word level and without features. The levels in this tree structure, which are indicated at the top, are the basis of the phrasing algorithm.*

two short French phrases:

```
#0 <-1>    <1 -1>    <99>    <>    <0 1> #2 <>    <99>
<0 1 0 0> #7 <1>    <> #1 <1> #4 <99>    <1> #6 <1 0>
<> #3 <0 1 0> #5 <1>    <> #0
```

German prosodic phrase formation results in:

```
#0 <-1>    <1 -1>    <99>    <>    <0 1> #2 <>    <99>
<0 1 0 0>    <1>    <> #1 <1>    <99>    <1>    <1 0>
<> #3 <0 1 0>    <1>    <> #0
```

After final normalization of phrase boundaries and phrase type assignment, the result is (the final sentence boundary is not shown):

```
#{P:0} Le monde à l'envers #{P:1} , ein französischer
Film, #{P:1} war ein sehr guter, #{T:2} bekannter Film.
```

Remarks

Prosodic phrasing, as described above, is still a rather ad-hoc solution. Correct phrasing is closely connected to a correct syntactic analysis and a well-formed syntax tree, which requires some "skill" of the grammar designer. In order to make phrasing more independent of syntactic analysis and thereby more robust, a hybrid, statistical and rule-based phrasing algorithm, as the *Hybrid ϕ-Model for Phrase Break Prediction* presented in [Att05], may be applied in future. However, such an algorithm requires for each language large, prosodically annotated sentence corpora, that were not available for the present work.

3.6 Sentence Accentuation

3.6.1 Accentuation Principles

Mixed-lingual sentence accentuation in the polySVOX system combines two different approaches:

- The accentuation of English, German, and Italian constituents follows the syntax-based algorithm of Kiparsky presented in [Kip66]. This algorithm determines in cyclic fashion, i.e., from

the leaves of the morpho-syntactic tree to the sentence node, accentuation patterns within a constituent according to two rules: the *nuclear stress rule*, due to which the nucleus of a constituent remains primary accented, whereas all other accents are reduced, and the *rhythmic stress shift rule*, which changes some accent patterns due to rhythmic constraints.

- The accentuation of French phrases bases on Selkirk's notion of the *prosodic structure* of an utterance, cf. [Sel81]. This prosodic structure consists of a hierarchical organization of prosodic categories. The number of layers and the names of the categories in the prosodic hierarchy vary throughout the literature, but there is little disagreement on the higher levels of this hierarchy, i.e., the levels of phonological words, of intonational phrases, and of utterances. French accentuation follows the algorithm described in, e.g., [HDC84, DCDCV97]. This algorithm also operates in cyclic fashion, i.e., from the leaves of the prosodic structure to the utterance node. It assigns according to the *accentual bipolarisation principle* an initial and a final accent to phonological words and to prosodic phrases, and according to a *dominance principle*, at each level of the prosodic structure a higher level of prominence to the final accent than to the initial one.

The prosodic structure of an utterance consists in the polySVOX system, from a bottom-up perspective, of phonological words, of prosodic phrases bounded by weak or strong phrase boundaries, and of the utterance. A prosodic constituent is assigned the language of the first syntactic constituent of the syntax tree, again form a bottom-up perspective, that spans all syllables of the phrase. Figure 3.4 shows the prosodic structure of the sentence presented in Figure 3.3.

3.6.2 Mixed-lingual Accentuation Algorithm

The mixed-lingual accentuation algorithm is based on the algorithm implemented for the monolingual German SVOX system in [Tra95]. This algorithm applies a set of *accentuation patterns* on a modified syntax tree, in which the leaves, i.e., the graphemic/phonemic representation of words, are initially replaced by the corresponding numeric word accentuation, that has been extracted in the first step of phrasing.

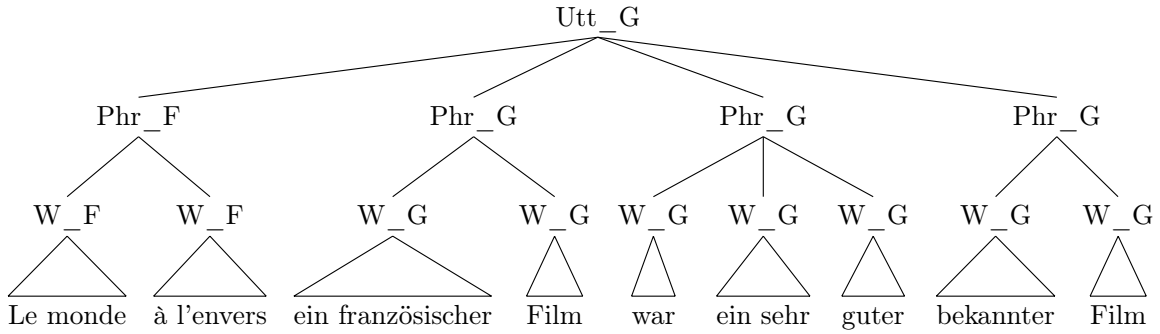

Figure 3.4: *The prosodic structure of the mixed-lingual German sentence "Le monde à l'envers, ein französischer Film, war ein sehr guter, bekannter Film" whose syntax tree is shown in Figure 3.3. The language of each prosodic constituent is indicated by language suffixes, e.g., a '_F' suffix for French and a '_G' suffix for German constituents. Phonological words are denoted by 'W', prosodic phrases are denoted by 'Phr'. 'Utt' specifies the utterance.*

The accentuation patterns all specify a possible subtree of the modified syntax tree and an action to be carried out if the pattern matches the given syntax tree. Due to the language suffix of the constituent types, it is possible to specify language specific and even mixed-lingual accentuation patterns. Accentuation patterns have the structure of ordinary syntax trees, but additionally, wildcard symbols may be used. Table 3.3 presents a list of all wildcard and special symbols that can currently be specified in accentuation patterns.

For each constituent type, a collection of accentuation patterns may

*	any sequence (0...n) of constituents including their (possibly empty) subtrees
**	like '*', but is stronger if it occurs together with '*', in that it matches as many constituents as possible.
?	any constituent (exactly one) including its (possibly empty) subtree
(...)	syntax hierarchy marker

Table 3.3: *Wildcard and special symbols used within accentuation patterns.*

be specified, which are compared with the syntax tree in order of declaration. If a pattern matches the syntax tree, the corresponding action is carried out, and no other patterns are applied any more for the same constituent.

The integration of prosodic structure-based accentuation into this algorithm is achieved by using a modified prosodic structure, in which the corresponding numeric accentuation sequence is assigned to each phonological word, as shown in Figure 3.5. Changes in this numeric accentuation sequence are synchronized with the numeric word accentuation of the modified syntax tree. Accent values in this modified prosodic structure are changed by traversing the structure from the leaves to the root and by applying a set of accentuation patterns. These accentuation patterns are similar to the ones used for syntax-based accentuation. All patterns specify a possible substructure of the prosodic structure and an action to be carried out if the pattern matches the given prosodic structure.

In the present accentuation algorithm, the following three actions can be specified:

acc1, m, acc2: matches an accent value 'acc1' and assigns it the new value 'acc2'.

acc1, nsr, acc2: matches an accent value 'acc1' and marks it with the new value 'acc2'. This action triggers the application of the stress reduction principle of the nuclear stress rule according to

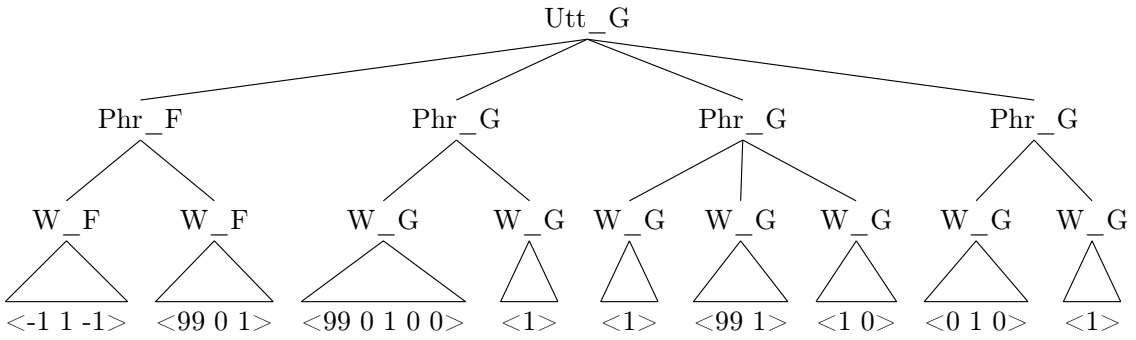

Figure 3.5: *Initial modified prosodic structure of the mixed-lingual German sentence "Le monde à l'envers, ein französischer Film, war ein sehr guter, bekannter Film".*

[Kip66]: If a marked accent is set to the new value 'acc2', all other accents of level 'acc2' within the same constituent are numerically increased by 1. This is repeated for all accents of level 'acc2' + 1, 'acc2' + 2, etc., until a gap in this sequence of accent levels is encountered.

acci, abp, accf: shifts the final accent within the matched constituent to the last accentuable syllable (having a non-negative accent value) and marks it with the value 'accf'. All other accent values within the same constituent are reduced according to the stress reduction principle of the nuclear stress rule (see above). If the constituent has more than one accent, the initial accent is shifted to the first syllable with a non-negative accent value smaller than the special level for unaccented words (99), and it is assigned the accent value 'acci'.

Examples of accentuation patterns are shown in Figure 3.6. The first pattern realizes left-accentuation within English compound nouns. The second and the third pattern specify accentuation within French phonological words ('`W_G`') and prosodic phrases ('`Phr_F`'), resp. The fourth pattern states that within a German noun phrase ('`NP_G`'), the primary accent remains on the rightmost noun ('`N_G`') and all other accents will be reduced. The fifth pattern states that within a German statement ('`GS_G`'), the rightmost noun phrase before the finite verb phrase will receive the sentence accent. The last pattern specifies default right-accentuation within German constituents.

At the end of each accentuation cycle, the *rhythmic stress shift rule*

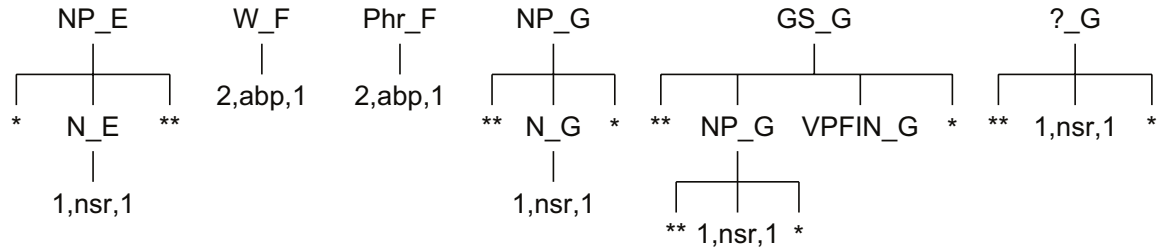

Figure 3.6: *Examples of subtree patterns used in the polySVOX sentence accentuation algorithm.*

is applied for all constituents except for some constituent types given in a special list. This list currently contains all main and subordinate clause constituents. The rhythmic stress shift rule changes the following accent constellations to new ones:

$$3\ 2\ 1 \rightarrow 2\ 3\ 1$$
$$2\ 2\ 1 \rightarrow 2\ 3\ 1$$
$$1\ 2\ 3 \rightarrow 1\ 3\ 2$$
$$1\ 2\ 2 \rightarrow 1\ 3\ 2$$

In these patterns, weaker accents (value 0 or greater than 3) may intervene, which remain unchanged.

Postcyclic Accentuation Rules

Some accent rules are executed after the traversal of the syntax tree. These *postcyclic rules* are again given as accentuation patterns.

In the polySVOX system, postcyclic patterns are currently used to correct the accent strength on finite verbs. These are usually too strongly accented by the application of the nuclear stress rule, because in the syntactic structure the finite verb phrase is a direct descendent of the sentence node. For example, the following postcyclic accentuation rules change accent values of 2 and 3 on all finite verbs below main clause constituents to a value of 4:

```
GS_G ( * VPFIN_G ( ** (2,m,4) * ) * )
GS_G ( * VPFIN_G ( ** (3,m,4) * ) * )
```

Accent Normalization

The accentuation algorithm described above may lead to a large number of different accent levels. The accentuation levels are normalized such, that within each phrase, the rightmost accent of the strongest level within the phrase is defined as *phrase accent* and therefore set to level 1. All other accents are strengthened as much as possible while maintaining their relative prominences within the phrase. The accent levels are then further restricted to the phonological interpretable levels 1 to 4 and 0. All remaining weaker accents in the range from 5 up to the special level for unaccented words (99) are also set to level 4. All levels larger or equal to 99 are set to 0.

Example of Mixed-lingual Accentuation

For the mixed-lingual German sentence "Le monde à l'envers, ein französischer Film, war ein sehr guter, bekannter Film", the word accentuation and phrase boundary sequence generated by polySVOX are (Phrase boundaries are represented by `'#n'`. Word accent patterns are enclosed in angular brackets):

```
#0 <-1> <1 -1> <99> <0 1> #1 <99> <0 1 0 0> <1> #1 <1>
<99> <1> <1 0> #2 <0 1 0> <1> #0
```

The application of prosodic-structure-based accentuation patterns for French phonological words generates no changes. The application of the patterns for French prosodic phrases changes the accentuation to

```
#0 <-1> < 2 -1> <99> <0 1> #1 <99> <0 1 0 0> <1> #1 <1>
<99> <1> <1 0> #2 <0 1 0> <1> #0
```

The application of the right-accented `NP_G` pattern on the first German noun phrase changes the accentuation to

```
#0 <-1> <2 -1> <99> <0 1> #1 <99> <0  2  0 0> <1> #1 <1>
<99> <1> <1 0> #2 <0 1 0> <1> #0
```

The default right-accentuation pattern applied on the German adjective phrase (`ADJP_G`) and subsequent rhythmic stress shift gives

```
#0 <-1> <2 -1> <99> <0 1> #1 <99> <0 2 0 0> <1> #1 <1>
<99> < 2 > < 3  0> #2 <0 1 0> <1> #0
```

The application of the right-accented `NP_G` pattern on the mixed-lingual German noun phrase changes the accentuation to

```
#0 <-1> < 3  -1> <99> <0  2 > #1 <99> <0  3  0 0> <1> #1 <1>
<99> <2> <3 0> #2 <0 1 0> <1> #0
```

The application of the right-accented `NP_G` pattern on the final German noun phrase gives

```
#0 <-1> <3 -1> <99> <0 2> #1 <99> <0 3 0 0> <1> #1 <1>
<99> < 3 > < 4  0> #2 <0  2  0> <1> #0
```

The application of default right-accentuation on the German sentence changes the accentuation to

```
#0 <-1> < 4  -1> <99> <0  3 > #1 <99> <0  4  0 0> < 2 > #1
< 2 > <99> < 4 > < 5  0> #2 <0  3  0> <1> #0
```

Postcyclic accent reduction on verb `VPFIN_G` changes

```
#0 <-1> <4 -1> <99> <0 3> #1 <99> <0 4 0 0> <2> #1 < 4 >
<99> <4> <5 0> #2 <0 3 0> <1> #0
```

Final accent normalization results in

```
#0 <0> <2 0> <0> <0 1> #1 <0> <0 2 0 0> <1> #1 <2> <0> <1>
<2 0> #2 <0 2 0> <1> #0
```

3.7 Phonological Transformations

The correct pronunciation of mixed-lingual text requires the application of a number of phonological transformations that comprise segmental assimilation, reduction, and insertion phenomena of each language involved. Some of these phenomena are mandatory, like French liaison or German terminal devoicing, others depend on the speaking style and speech rate. In addition to monolingual phenomena, also some mixed-lingual phenomena must be considered. The following examples present some of the monolingual and mixed-lingual phenomena currently comprised in the polySVOX system:

German aspiration: In word-initial position, the German unvoiced plosives [p], [t] and [k] preceding a vowel are aspirated, denoted as [p^h], [t^h] and [k^h], resp. They are also aspirated in word-final position before a break.

German terminal devoicing: All voiced plosives (obstruents) before a morpheme or word boundary are devoiced.

French liaison: In French noun groups, liaison is forbidden between a singular noun and the consecutive adjective, e.g., “un bruit effroyable” [œ̃.bʀɥi.e.fʀwa.jabl];
between a plural noun and the following adjective it is optional, e.g., “les amis agréables” [le.za.mi.(z)a.gʀe.abl];
liaison is mandatory between the preceding adjective and a noun, e.g., “un bon ami” [œ̃.bɔ.na.mi].

Liaison is generally avoided between a singular noun and the following verb, e.g., “l’étudiant entend” [le.ty.djã.ã.tã];
it is optional between a plural noun and the following verb, e.g., “les étudiants entendons” [le.ze.ty.djã.(z)ã.tã.dɔ̃];

but liaison is mandatory between a clitic personal pronoun and the following verb, e.g., "on entend" [ɔ̃.nã.tã].

French liaison consonant realization: The phonetic liaison consonant can be directly derived from the corresponding graphemic consonant: e.g., "s", "x" or "z" result in [z]; or "c", "q" or "g" in [k].

French optional schwa elision: Word-final optional schwa [(ə)] in French is not pronounced, if the following word begins with a vowel or an "h muet". In front of a word beginning with an "h aspiré", final [(ə)] is pronounced. E.g., "une bonne hache" [yn.bɔ.nə.ʔaʃ].

English linking "r": Word-final "r" is usually only pronounced, if the following word begins with a vowel, e.g., "four eggs" [fɔːr.egz] but "four pounds" [fɔː.pa͡ʊndz].

English plosive elision: In clusters of three plosives or two plosives and a fricative, the middle plosive may disappear, cf. [Roa91]. E.g., "act badly" [æk.bæd.li].

Italian raddoppiamento fonosintattico: Word-initial Italian consonants are lengthened when following a poly-syllabic Italian word with final stress. E.g., the initial consonant of the Italian word "latte" [l'at͡te] is pronounced in "caffè latte" as [kaf͡f'ɛ l͡l'at͡te].

Cross-lingual assimilations in German words: Foreign inclusions in mixed-lingual German words virtually keep the pronunciation prescribed by the originating language. The syllable onset and coda near the language switching position, however, may be weakly assimilated to the base language.

In a word like "Dufourstrasse", which is composed from the French proper name "Dufour" and the German noun "Strasse" (street), the French [ʀ] has to be replaced by the German [r]. It would sound rather affected to pronounce [dʏ.fuʀ.ʃtraː.sə] instead of [dʏ.fur.ʃtraː.sə].

These examples show that the phonological phenomena depend on various contexts, like phonetic, graphemic, syntactic, and language context as well as contextual information about accentuation and phrasing.

In the polySVOX system, all phenomena were implemented using the multi-context rule formalism presented in Section 3.2.2.

Example of Phonological Transformations

Within the example sentence “Le monde à l’envers, ein französischer Film, war ein sehr guter, bekannter Film.” the phonological transformations remove the optional schwa of “monde” [mɔ̃ːd(ə)]. After sentence re-syllabification, the following phonological representation is finally obtained:

#{P:G:0} \F\lə- m[2]ɔ̃ː-d a- l ɑ̃-v[1]ɛːʁ #{P:1} \G\ʔa͡in- fran-ts͡[2]øː-zɪ-ʃɐ- f[1]ɪlm #{P:1} v[2]aːɐ̯- ʔa͡in- z[1]eːr- g[2]uː-tɐ #{T:1} bə-k[2]an-tɐ- f[1]ɪlm

Chapter 4

Speech Prosody Modeling

This chapter describes speech prosody modeling in general. After an introduction to general prosodic phenomena, a review of state-of-the-art approaches to F_0 and segment duration modeling applied in TTS synthesis. This chapter concludes with a definition of multilingual and of polyglot prosody modeling.

4.1 Introduction

The prosody of a speech signal can be described at the perceptual level in terms of *pitch*, *(sentence) melody*, *(speech) rhythm*, and *loudness*. The physically measurable quantities by which speech segments can be modified are the acoustic parameters *fundamental frequency* (F_0), *segment duration*, and *signal intensity*. F_0 correlates with pitch and sentence melody, segment duration correlates with speech rhythm, and signal intensity correlates with loudness. *Intonation* refers to the rise and fall of F_0 of the voice in speech.

The prosody control component must generate these physical parameters to drive the speech generation component. The parameters are influenced by various factors, that depend on the roles prosody plays in human communication. The linguistic literature on these different roles of prosody is vast. E.g., [Cry69, Bol89, Lav94, Tra05] provide an overview. [Tra95, vSMK08] discuss the roles of prosody with respect to

TTS synthesis.

[Tra05] distinguishes *perspectival* (indicates distance and location of a speaker), *organic* (indicates age, sex, and health of a speaker), *expressive* (indicates emotions and attitudes of a speaker) and *linguistic aspects* of speech that influences speech prosody. The influences, or factors, of the first three aspects basically characterize a specific speaking environment, a particular speaker, and a certain speaking style. Most TTS systems keep perspectival, organic, and expressive factors constant by recording the voice of a given speaker with a "neutral" information communication speaking style in a well-defined studio environment. In the last decade, however, an own branch of TTS research, so-called "emotional TTS", started to explore expressive factors in speech prosody, see [Sch01, Sch08] for an overview.

In the polySVOX system presented in this thesis, the aim was to model linguistic factors of *polyglot speech prosody* and to keep perspectival, organic, and expressive factors as constant as possible. Therefore, the voice of a single, polyglot speaker with a "neutral" speaking style was recorded in several languages.

4.2 Linguistic Factors of Speech Prosody

The linguistic factors of speech prosody vary with each utterance and are furthermore language specific. From the linguistic perspective, speech prosody may be used for:

- *Distinguishing different meanings of a word*: In tone languages, such as Chinese, Vietnamese, Thai, or Swedish, the type of F_0 movement within certain syllables may distinguish different meanings of a word. In languages that have long and short phones, such as English, French, German, or Italian, different phone duration patterns are used to distinguish different meanings of a word. These tonal movement types and these temporal patterns are therefore of phonemic nature, but they must be treated in prosody control in close connection with other melodic and rhythmic contributions rather than with the sound segment production.

- *Semantic structuring of utterances*, that is used to group words, that belong together semantically, in one rhythmic and melodic

group, a so-called *(prosodic) phrase*, and to indicate relationships between these phrases. The linguistic notion therefore is *(prosodic) phrasing*. Prosodic phrasing may also be used to disambiguate different semantic contents of an utterance. In the utterance "Charles the first king of England", the semantic contents define whether the correct grouping is "(Charles the first) (king of England)" or "(Charles) (the first king of England)".

- *Emphasizing of words* for distinguishing semantically more important words from less important ones. The linguistic term therefore is *(sentence) accentuation*. In many languages, sentence accentuation is used to highlight so-called content words, like nouns, verbs, adjectives, or adverbs. Stronger emphasizing of words may also be used for rational highlighting and expression of contrast, or for amplifying their importance. [Koh06] refers to these phenomena as *emphasis for focus* and *emphasis for intensity*.

  The usage of emphasis to indicate semantic focus, however, varies among different languages. In German and in English utterances, semantic focus on certain words is expressed by emphasizing them, like in "Ihr wollt *morgen* abreisen?" and "You want to leave *tomorrow*?". In French, this semantic focus is expressed without any emphasis using a syntactic construct called *mise en relief*: "C'est demain que vous voulez partir?".

  The acoustic realizations of sentence accents also vary among different languages. While Germanic languages, like German or English, highlight words within phrases mainly by means of pitch inflection and lengthening of syllabic nucleus and coda, this highlighting is achieved in French mainly by the so-called *accent d'insistance* (force accent), cf. [Koh06], which relies on initial consonant lengthening and an increase of signal intensity. In French, pitch inflection mainly signals prosodic phrase boundaries.

- *Indication of sentence modality* to communicate the intent of an utterance to the listener. In most languages, sentence modalities, like statement, exclamation, total (yes/no) question, partial (wh-) question, or parenthesis, are distinguished using different prosodic patterns. These patterns are mainly characterized by different intonation patterns, but also by different signal intensity patterns.

4.3 Approaches to Prosody Modeling

In the last decades, various approaches to model the acoustic parameters of speech prosody by computational means have been proposed. The advance in computer processing power and storage capacity resulted in a constantly increasing quality of these models. Starting with rule-based models in the seventies and eighties, prosodic modeling advanced over pattern concatenation models and statistical models in the nineties, to so-called "*using as-is*" prosody models in unit selection TTS, that are currently state-of-the-art. [vSMK08] provide an overview of the history of prosody models, including a brief introduction of the most influential models.

A further increase in quality of "using as-is" prosody models is restricted by the finding, "that even small text samples of a few sentences were virtually certain to contain (linguistic factor) vectors that occurred only once in the entire analysis (of 22 million phonetic segments)", cf. [vS93]. These vector frequency distributions belong to the *Large Number of Rare Events* (LNRE) class of distributions, introduced in [Khm87], for which the law of large numbers does not hold and that have the property of extremely uneven frequency distributions. See [vS97, Möb01] for a more detailed discussion of LNRE distributions in TTS synthesis.

Due to the LNRE property, it is almost impossible to record a speech database by the same speaker within a reasonable period of time, that contains all possible linguistic factor combinations. This is one of the constraints that restrict high-quality, unit-selection TTS to close-domain applications, cf. [vS97, Bla02]. Therefore, the focus of speech prosody research is currently moving back to pattern concatenation models and to statistical models.

The different approaches to speech prosody modeling can be characterized in terms of the *acoustic parameter(s)* they model, in terms of their *input representation of prosodic events*, like prosodic phrase boundaries or sentence accents, in terms of their *output representation of the acoustic parameter(s)*, and in terms of their *method to generate the output representation from the input representation*, as shown in Figure 4.1.

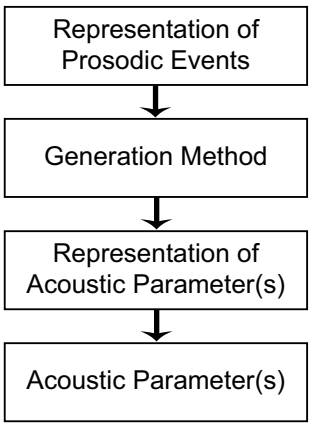

Figure 4.1: *General structure of speech prosody models.*

4.3.1 Acoustic Parameters

Most prosody models are designed to model one specific acoustic parameter only. Prosody research has concentrated thereby on modeling fundamental frequency and segmental duration, as these two acoustic parameters are generally considered to be much more important for the naturalness of synthetic speech than signal intensity, as confirmed by several studies, e.g., in [TBNA98]. However, given nearly perfect fundamental frequency and segment duration control, a very good intensity control would also become necessary, in order to model all sentence modalities and to further improve naturalness of synthetic speech.

4.3.2 Representations of Prosodic Events

The input representations of prosodic events mainly differ in how detailed prosodic events are described. They can be grouped into the following classes:

- Abstract phonological descriptions of prosodic events, e.g., the *phonological representation* used in polySVOX, that is described in Section 3.2.1.
- Qualitative phonetic descriptions of alignment and type of prosodic events. These descriptions can be rather coarse, as the distinction between high (H) and low (L) tones in the model of

Pierrehumbert [Pie80] and in the subsequent *tones and break indices* (ToBI) definition [SBP+92], or the qualitative description can be more detailed, like the *international transcription system for intonation* (INTSINT) [HDCE00] with eight discrete tone symbols.

- Quantitative phonetic descriptions of time alignment and size of prosodic events, as found, e.g., in the *Tilt* model [TB94, Tay00], in the *linear alignment model* [vSM00], or in the *phrase and accent command* models [Öhm67, Fuj81, Fuj83].

4.3.3 Representations of Acoustic Parameters

The various output representations of acoustic parameters can be separated into two major classes:

- *Direct representations* that directly derive the value of the acoustic parameter from the output of the model. The majority of prosody models, especially segment duration and signal intensity models, use this type of representation.

- *Superpositional representations* that form the value of the acoustic parameter as a superposition of several, hierarchical model outputs. This hierarchy often reflects linguistic structures, as sentence, phrase, syllable, and segment levels. A well-known example of a superpositional segment duration model is the syllable-based model of Campbell [Cam89, Cam92], that predicts syllable durations first and then fits segment durations to the syllables. Superpositional fundamental frequency models construct F_0 values as a combination of different F_0 components, like sentence, phrase, accent, or perturbation curves. Some influential models are the *phrase and accent command* models [Öhm67, Fuj81, MPH93], the *superposition of functional contours* (SFC) model [Aub93, HB00, BH05], the *linear alignment model* [vSM00], or the *multi-level unit sequence* model [vSKKM05].

Output representations of fundamental frequency additionally differ in how detailed the course of fundamental frequency values over time, the so-called *fundamental frequency contour*, is approximated:

- Early models predicted one value per prosodic event, e.g., per stressed syllable or per phrase boundary, like the model of Pierrehumbert [Pie81], or the *phrase and accent command* models [Öhm67, Fuj81], and roughly interpolated the course of F_0 between these events.

- Later approaches modeled the rises and falls of the F_0 contour and interpolated between them. Examples are the *IPO* approach [Col91] or the *rise/fall/connection* (RFC) model based on the *Tilt* model [TB94, Tay94, Tay00].

- Other models approximated the course of F_0 much more accurately by predicting several values for each syllable, e.g., [Tra95] used 5 values per syllable in the F_0 model of the *SVOX* system.

- Some approaches, finally, apply the original F_0 contours of recorded speech signals, like unit selection approaches, cf. [Cam94, CB95, BC95, HB96], that implicitly use the original F_0 contour of the selected speech unit, or approaches that combine stored parts of original F_0 contours, like the F_0 pattern concatenation approach reported in [Tra90, Tra92] or the *multi-level unit sequence* model [vSKKM05].

4.3.4 Generation Methods

The method to generate the output representation from the input representation is probably the most characterizing attribute of a prosody model. The numerous existing methods can be divided into the three following classes:

Knowledge-based or Rule-based Methods

Knowledge-based or rule-based methods are mainly found in early TTS systems. Explicit knowledge of linguistic experts is often modeled in form of a linear model.

Probably the most influential rule-based duration model is the Klatt duration model [Kla73, Kla79] that is part of the *MITalk* system [AHK87]. Other rule-based duration models are described, e.g., for English in [CUB73], for German in [Koh88], and for French in [O'S84].

Well-known, rule-based fundamental frequency models include *Pierrehumbert*'s model [Pie81], the *IPO approach* [Col91], the *Kiel Intonation Model (KIM)*, a rule-based F_0 model for German, established by Kohler [Koh90, Koh91], and the *phrase and accent command* models presented by Öhman and Fujisaki [Öhm67, Fuj81], that generate F_0 contours by means of impulse and step responses of second-order linear systems, which are driven by "accent" and "phrase" commands.

Data-based, Statistical Methods

Data-based, statistical methods apply machine learning techniques to automatically estimate the unknown parameters of the model from recorded speech data. Depending on the type of parameters being adjusted, two categories of statistical methods can be distinguished:

- Early statistical models applied *parametric estimation methods* in which a specific functional form for the model is assumed and the parameters of the model are then optimized by fitting the model to the data set.

- Most current statistical models apply *non-parametric estimation methods* in which no particular functional form is assumed and which allow the form for the model to be determined entirely by the data. Non-parametric methods that are commonly used for prosody models include decision trees in form of binary regression trees, so-called *classification and regression trees* (CART) that are described in [BFOS84], *artificial neural networks* (ANN), see [Lip87, Bis95] for a description, *multivariate adaptive regression splines* (MARS) introduced in [Fri91], and *hidden Markov models* (HMM) for which [Rab89] is a good introduction.

Statistical duration models based on parametric estimation methods include the French duration model by [KZ96] based on a *linear model*, the German duration model by [Hub90, Hub91], who applied a *generalized linear model* (GLM)or the *sums-of-products* approach presented in [vS92, vS93, vS94] for English duration prediction. The sums-of-products approach was later applied to German duration modeling [MvS96], and integrated as the first *multilingual* duration model into the Bell Labs multilingual TTS synthesis system [vSSM$^+$97].

CART-based statistical duration models were first presented by [Ril89, Ril92] for English segment duration prediction. This model was later applied for various languages, e.g., for Italian [MQ95], for German [Tra96], or for French [MDM98].

Duration models using ANNs were first reported by [Cam89, Cam92] for English syllable-based duration prediction. ANN-based segment duration models are presented by [Rie95] for German and by [MQ95] for Italian. [FVMVC99] present a comparison of CART- and ANN-based duration models for six languages, that shows for the ANN-based models smaller prediction errors in all six languages.

A MARS-based duration model for German segment duration prediction is presented in [Rie97]. A comparison of CART-, ANN- and MARS-based duration models for German in [Rie98] shows that MARS-based models have slightly smaller prediction errors than ANN-based models. And both types of models have significantly smaller prediction errors than CART-based duration models. However, MARS prediction errors turned out to be much more unevenly distributed than ANN prediction errors. Friedman pointed out in [Fri01], that *MARS predictions tend to be either very close to, or far from, the target values.*

Statistical fundamental frequency models based on parametric estimation methods are not so frequent. [BH96] present a statistical fundamental frequency model that applies a *linear model for regression* to generate F_0 contours from ToBI labels.

Non-parametric statistical fundamental frequency models include the Tilt-specification based RFC intonation model, developed by Taylor and Black [TB94, Tay94, Tay00], who applied a CART learning method to train a decision tree for every Tilt parameter of the model. This model was later also used for other languages, like, e.g., Italian [MQ95].

An early attempt to model F_0 using HMMs can be found in [LF86]. [YTM+99] describe the prosody model of the HMM-based speech synthesis system HTS, that was introduced in [MTKI96]. In the HTS system, the spectral shape, F_0, and segment duration are simultaneously modeled in a unified framework using so-called multi-space probability distribution HMMs, which are described in [TMMK02].

ANN-based F_0 models were already reported in [SG89, Sag90]. A very well-known, ANN-based F_0 model is the *recurrent neural network* (RNN)-based model of Traber [Tra90, Tra92, Tra95] that reached a

new level of naturalness in speech prosody. This model predicts 5 F_0 values per syllable directly from the phonological representation. It is applied in the SVOX system for German. [MJ01] present an *integrated model* (IGM) for German, that is based on Fujisaki's phrase and accent command model. They apply an ANN to estimate the parameters of the model. The *superposition of functional contours* (SFC) model for French, initiated in [Aub92, Aub93] and finally presented in [HB00, BH05], applies ANNs as generators of hierarchical F_0 contours that are superposed to form the F_0 contour of an utterance.

Data-based, Concatenative Methods

Data-based, concatenative methods apply the prosody of stored, natural speech units to synthesize the prosodic contours of new utterances.

An approach for German concatenative F_0 modeling has been described by Traber in [Tra90, Tra92], in which averaged natural F_0 patterns were concatenated in such a way that they formed the F_0 contour of a new utterance according to their phonological representation. This method is somewhat similar to the superpositional approach taken by Aubergé in [Aub90, Aub92, Aub93] for French. In this approach, hierarchical, mean contours taken from natural F_0 contours are superposed and concatenated to form the F_0 contour of an utterance. The *multi-level unit sequence* approach, introduced by van Santen in [vSKKM05], also applies a superpositional, concatenative method to generate the F_0 contour of a new utterance. Instead of averaged F_0 patterns, this approach superposes natural phrase, accent and residual F_0 contour components taken from a much larger prosody database.

Unit-selection TTS synthesis systems, introduced in [Cam94, CB95, BC95, HB96], also apply a concatenative method for prosody generation. This method implicitly concatenates natural prosodic patterns containing all three acoustic parameters of speech prosody (F_0, segment duration, and signal intensity) by concatenation of natural speech units without any or only slight prosodic modification. The quality of the generated prosodic contours is generally very high and obviously fits perfectly the speech signal. However, due to the LNRE property such prosody generation is restricted to close-domain applications only, cf. [vS97, Bla02].

4.4 Multilingual Prosody Modeling

All existing approaches for modeling prosody of multiple languages for speech synthesis have been concentrated so far on making the prosody models "language-independent", as it was formulated by van Santen in [vSSM+97]: to be *language-independent*, a prosody model

- must use the same executable for all languages, and
- it must be possible to model a new language by largely deriving language specific model parameters automatically from a speech database. The construction of a prosody model for a new language should not require modifications of other parts of the TTS system.

This language-independence is the main characteristic of multilingual prosody models. A *multilingual prosody model* is able to generate the prosodic contour for multiple languages, but in general not by the same voice. Switching between languages is only possible at sentence boundaries and is usually accompanied by voice switching. Seamless language switching and correct prosody modeling of foreign word or word group inclusions is therefore not possible.

The first multilingual duration model, a multilingual sums-of-products model, was presented in [vSSM+97] as part of the Bell Labs multilingual text-to-speech synthesis system [Spr97]. In the following years, the prosody models of other, mostly commercial TTS synthesis systems were extended to support multiple languages, like the Lernout&Hauspie TTS synthesis system in [FVMVC99, FVG+02], who compared CART-based with ANN-based prosody models for six different languages, or the unit selection based, multilingual IBM TTS synthesis system in [OFWR05], who tested English-German and English-Spanish unit selection TTS synthesis.

4.5 Polyglot Prosody Modeling

The limitations of multilingual prosody models restrict the usability of TTS synthesis systems to monolingual texts. The generation of an adequate prosody for mixed-lingual texts, sentences, or even words, requires a *polyglot prosody model* that is able to seamlessly switch between

languages and that applies the same voice for all languages. Listening experiments verified this finding. E.g., [OBK06] demonstrated the need of English prosody for the English inclusions in German sentences.

The requirements of a *polyglot prosody model* for polyglot TTS synthesis can be summarized as follows:

- First, for a prosody model to be *polyglot*,
  - the generation of prosodic contours must be done with *prosody models of the same speaker for all languages*, and
  - *seamless switching between languages* must be possible such that no rhythmic or melodic discontinuity is audible.
- And second, the model must be *language-independent* as defined in Section 4.4. E.g., it must be possible to extend a polyglot prosody model to cover an additional language without modifications of the model parameters for already supported languages.

Additionally, like any other prosody model, a polyglot prosody model should generate prosodic contours, that are

- as *natural sounding* as possible. Even for mixed-lingual text, the generated prosody should ideally be indistinguishable from natural prosody,
- and *robust* to linguistic factor vectors that are not covered by the speech database. As pointed out in Section 4.3, it is generally not feasible to record speech for all possible linguistic factor combinations. This is especially true for mixed-lingual texts.

Chapter 5

Natural Speech Data

This chapter describes the setup and the automatic segmentation and labeling of the natural speech corpora recorded from professional speakers: one bilingual natural speech corpus in German and in French that was used for the creation of the polyglot prosody control. And one monolingual German corpus that was already used in earlier works. This corpus was used for performance comparisons.

5.1 Introduction

The use of statistical models for monolingual speech prosody prediction requires the recording and careful annotation of natural speech data of this language. The annotation process includes transcription of the text to obtain the phonological representation of the utterances, and accurate phone segmentation, fundamental frequency (F_0) extraction, and signal intensity estimation of the speech data to obtain the physical prosodic parameters of the utterances. This set of phonological representations of the utterances as the input of speech prosody prediction and associated physical prosodic parameters as the desired output is termed "*prosody corpus*".

Polyglot speech prosody prediction must additionally be able to predict the physical prosodic parameters for all possible combinations of language mixtures. Following the traditional approach, a statistical

model for polyglot speech prosody prediction would require a polyglot prosody corpus whose size grows exponentially with the number of languages. Thus, the creation of a polyglot prosody corpus following this standard approach would become infeasible already for a decent number of languages.

The approach to polyglot speech prosody prediction presented in this thesis requires only monolingual prosody corpora. The size of the complete prosody corpus needed grows therefore only linearly with the number of languages.

5.2 Text Material and Recordings

The central intention behind the production of this polyglot prosody corpus was the creation of a polyglot speech prosody prediction for non-affective speech for the main languages spoken in Switzerland.

Therefore, two monolingual natural speech data sets, a German one and a French one, were recorded under studio conditions. The same professional, quadrilingual female speaker, who was chosen for the recording of the quadrilingual, single-speaker diphone inventory [THN+99], read about 1470 monolingual German sentences and about 1400 monolingual French sentences in a "neutral" style. In about 16 percent of the sentences the speaker was advised to emphasize a predefined word, but no further distinction of the type of emphasis, as, e.g., in [Koh06], was made. For test purposes, 77 mixed-lingual German sentences with English, French, or Italian inclusions, as well as 45 mixed-lingual French sentences with English, German, or Italian inclusions were additionally recorded.

The sentences were selected from different texts in such a way that they cover a large variety of sentence modalities (statements, wh-questions, yes/no-questions, declarative questions, alternative questions, and exclamations), several syntactic constructions with specific prosodic patterns (e.g., sentences including a parenthesis and sentences including a list of items), and different sentence lengths (single-word sentences, single-phrase sentences, and multi-phrase sentences). Thus, these sentences should include all necessary prosodic phenomena to construct a speech prosody prediction that is adequate for producing non-affective German and French prosody, cp. [vSMK08].

From each of the two monolingual natural speech data sets, the author finally selected 400 sentences to construct a German and a French monolingual speech prosody corpus. The selection criterion was a good coverage of the different sentence types. Table 5.1 shows a detailed description of the sentence type distributions for the German and for the French speech prosody corpus.

From the 400 sentences of the German speech prosody corpus, a set of 44 sentences was separated for testing speech prosody prediction. This set covers all sentence types of the German corpus. Out of the 400 sentences of the French speech prosody corpus, a test set of 48 sentences was selected, covering all sentence types of the French corpus. Table 5.2 lists the sentence types of the German and French test sets.

Additionally, a polyglot test set was constructed by selecting all mixed-lingual German sentences with French inclusions (20 sentences) and all mixed-lingual French sentences with German inclusions (8 sentences) from the recorded mixed-lingual sentences. All 120 utterances of these test sets were manually segmented and annotated by the author. The other 708 utterances of the German and the French corpora were automatically segmented and annotated, see Section 5.5. The combination of the monolingual German and monolingual French corpora and

	German		French	
	short	long	short	long
statement	33	90	30	108
statement with parenthesis	0	39	0	29
statement with list of items	0	21	0	30
wh-question	5	37	3	39
yes/no-question	13	34	0	24
question without inversion	15	32	8	22
alternative question	7	10	10	22
exclamation	23	41	12	63

Table 5.1: *Sentence type and sentence length distribution of the German and of the French prosody corpora. Single-word sentences are denoted as 'short', single-phrase and multi-phrase sentences are grouped as 'long'.*

the polyglot test set will be referred to as *polyglot prosody corpus*.

The reasons for restricting the polyglot prosody corpus to only 400 sentences of each language were twofold. First, the length of natural speech data recordings in each language is then approximately equivalent to the length of the recordings done for a former monolingual German speech prosody corpus used in [Tra95, Rie98]. Thus, it is possible to compare the quality of speech prosody prediction for the different corpora. Second, since the author had to manually correct the phonological representations of the complete corpus and to manually segment and annotate the test sets, a larger corpus would simply have been impossible to construct in a reasonable amount of time.

The former monolingual German speech prosody corpus was recorded, manually transcribed, segmented, and annotated within the SVOX project, cp. [Tra95, Rie98]. The corpus contains 186 sentences spoken by a professional male speaker in a "neutral" news-reader style. The sentences were taken from different texts from newspapers. All sentences are statements. From these 186 sentences, a test set of 55 sentences was separated. Table 5.3 displays the sentence length distribution of the complete corpus and of the test set.

	German		French	
	short	long	short	long
statement	4	5	2	12
statement with parenthesis	0	2	0	2
statement with list of items	0	2	0	7
wh-question	4	2	3	3
yes/no-question	4	3	0	2
question without inversion	4	2	2	3
alternative question	4	2	3	2
exclamation	4	2	3	4

Table 5.2: *Sentence type and sentence length distribution of the German and of the French test sets for speech prosody prediction. Single-word sentences are denoted as 'short', single-phrase and multi-phrase sentences are grouped as 'long'.*

	Complete corpus		Test set	
	short	long	short	long
statement	19	167	3	52

Table 5.3: *Sentence length distribution of the monolingual, German male corpus for speech prosody prediction, and of the test set taken from this corpus. Single-word sentences are denoted as 'short', single-phrase and multi-phrase sentences are grouped as 'long'.*

5.3 Fundamental Frequency Extraction

F_0 values of the natural speech data of the prosody corpora were computed every 5 ms using a pitch detection program developed by the ETH speech processing group. This pitch detection is based on combined information taken from the cepstrogram, from the spectrogram, and from the autocorrelation function of the speech signal. Signal sections judged as unvoiced by the algorithm are assigned virtual F_0 values by linear interpolation between neighboring voiced sections.

5.4 Transcription

The input to speech prosody prediction is the phonological representation (cp. Section 3.2.1) of the sentence to be uttered. Initial phonological representations of the sentences in the prosody corpora were obtained applying the transcription component of the polySVOX system to the text data of the corpora. These initial phonological representations contain the *standard phonetic transcription*, also called *canonical phonetic transcription*, of the sentences. The phonological information, i.e., phrase type, phrase boundary, and sentence accentuation, of these automatically generated representations was then manually corrected by the author to correspond to the recorded speech signals. This manual correction was done solely by the author in order to guarantee consistency over all corpora.

To speed up manual correction, the author implemented together with master students neural network based algorithms for automatic phrase type, phrase boundary, and syllable accent identification. These

algorithms were used to make a first correction of the phonological representations given the speech signals and the automatically generated phonological representations as inputs. Detailed information on this automatic identification procedure can be found, e.g., in [BS07].

5.4.1 Language Information

Four types of *language switch* were used: a language switch to English was denoted as \E\ in the phonological transcription, a switch to French as \F\, a switch to German as \G\, and a switch to Italian as \I\.

5.4.2 Sentence Accentuation Information

Three types of *accent* were distinguished:

Emphatic accents (accents associated with a very high pitch movement, lengthened syllable duration, and higher acoustic energy) were generally denoted as [E]. No further distinction between different types of emphasis, e.g., into emphasis for focus or emphasis for intensity as in [Koh06], was done.

A *pitch accent* (accents associated primarily with a major pitch movement and an optional lengthening of syllable duration) in the main accent position of a phrase was denoted as [1]. Other pitch accents within the phrase were marked as [2].

Non-pitch accents (accents associated mainly with a lengthened syllabic nucleus and coda, so-called *duration accents*, or with a lengthened syllabic onset and higher acoustic energy, so-called *force accents* or *accents d'insistance* in French, cf. [Koh03, Koh06]) on the main stress position of words were denoted as [3]. While the main stress position of words in English, German, and Italian is lexically specified, the main stress position of words in French depends on rhythmic criteria, cf. [Del66, Mer93, DC98, Mer99]. Non-pitch accents on other than the main stress position of words, and secondary and tertiary word accents were denoted as [4], and all unaccented syllables as [0].

5.4.3 Phrasing Information

Three types of *phrase boundary* were transcribed: one for sentence-final pauses (denoted as #{*:0} in the phrase boundary tag of the

phonological representation), that was placed before and after each sentence. The wildcard * stands for phrase type information. One for sentence-internal pauses (denoted as #{*:1}). And one for all other perceived sentence-internal phrase boundaries (denoted as #{*:2}). No distinction was made between breath-pauses (pauses with the speaker breathing) and non-breath-pauses.

Six different *phrase types* were transcribed. The distinction of phrase types was done with respect to the phrase final intonation contour, cf. [Dud84, Pie80, Del66, vE56], to the overall tone range of the intonation contour, and to overall acoustic energy: Phrases with a *complete fall* of the phrase final intonation contour and a low overall tone range were transcribed as #{T:*}. The wildcard * indicates the phrase boundary value. Phrases with a *non-complete to complete fall* of the phrase final intonation contour and medium overall tone range were transcribed as #{S:*}, or in case of a high overall tone range and high overall acoustic energy as #{E:*}. Phrases with a *progredient* phrase final intonation contour were transcribed as #{P:*}. Phrases with a *rising* phrase final intonation contour were transcribed as #{Y:*}, or in case of a short final fall after the rise as #{YC:*}.

Figure 5.4 and Figure 5.5 show on the left side the median, the 25, and the 75 percentiles of the time-normalized, linearized F_0 contours of all T, S, and E phrases and of all P, Y, and YC phrases, resp., of the German prosody corpus. On the right side, one characteristic F_0 contour is shown for each of these phrase types. Figure 5.6 and Figure 5.7 display the same information for the French prosody corpus.

The median intonation contour of a phrase type can be interpreted as a basis pattern for this phrase type, comparable with the so-called *phrase curves* of superpositional F_0 models [Fuj81, vSKKM05]. Figures 5.4, 5.5, 5.6, and 5.7 show that the six phrase types can be very well discriminated using the shape of the median intonation contour and overall tone range of the intonation contour indicated by the 25 and 75 percentiles. They also show that the intonation contours of German Y and YC phrases have in general a higher final rise than their French counterparts. French P phrases show a somewhat more pronounced

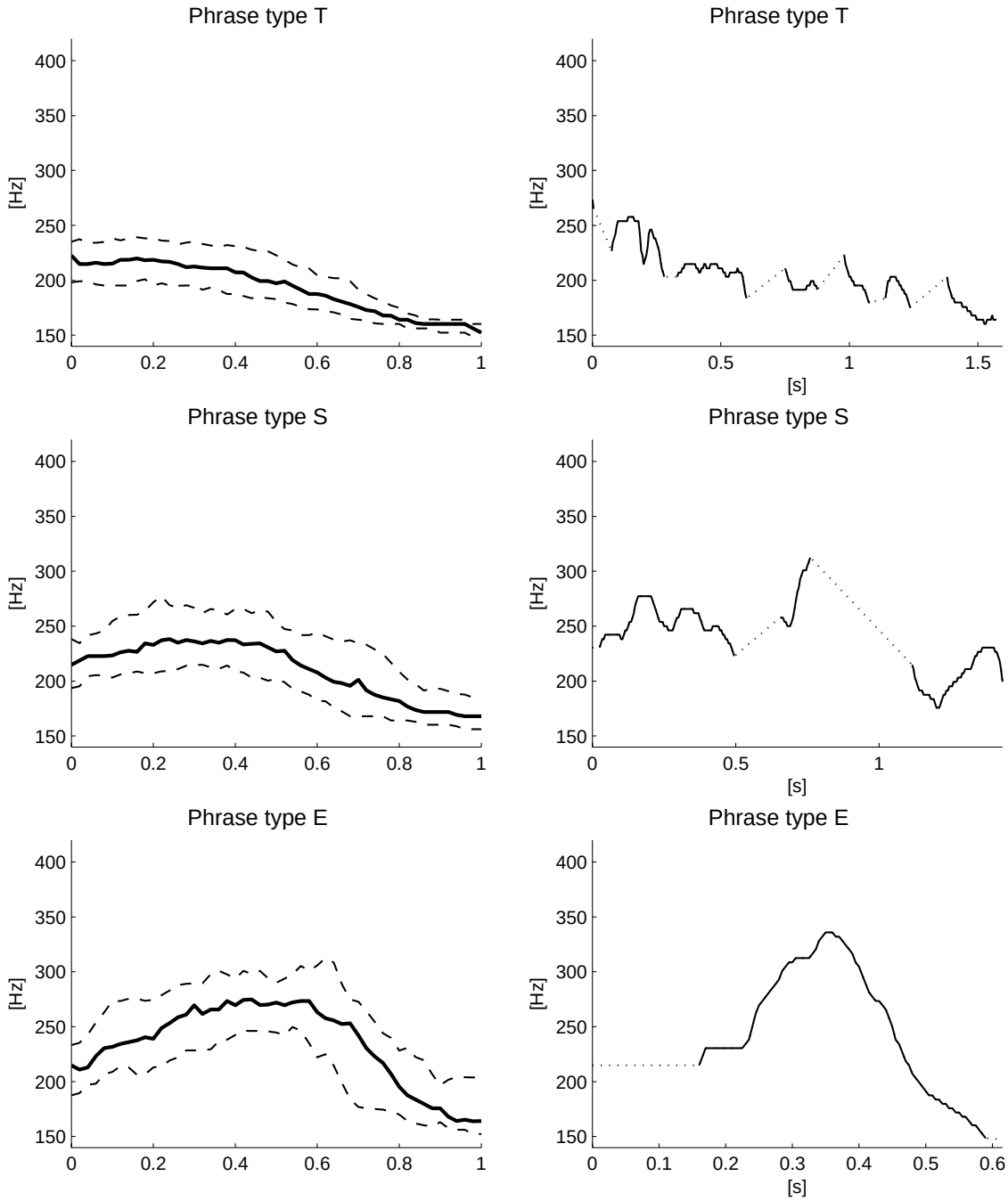

Table 5.4: *German phrase types T, S, and E: Median, the 25, and the 75 percentiles of the time-normalized, linearized F_0 contours of all T, S, and E phrases of the German prosody corpus are shown on the left side. On the right side, one characteristic F_0 contour is shown for each of these phrase types.*

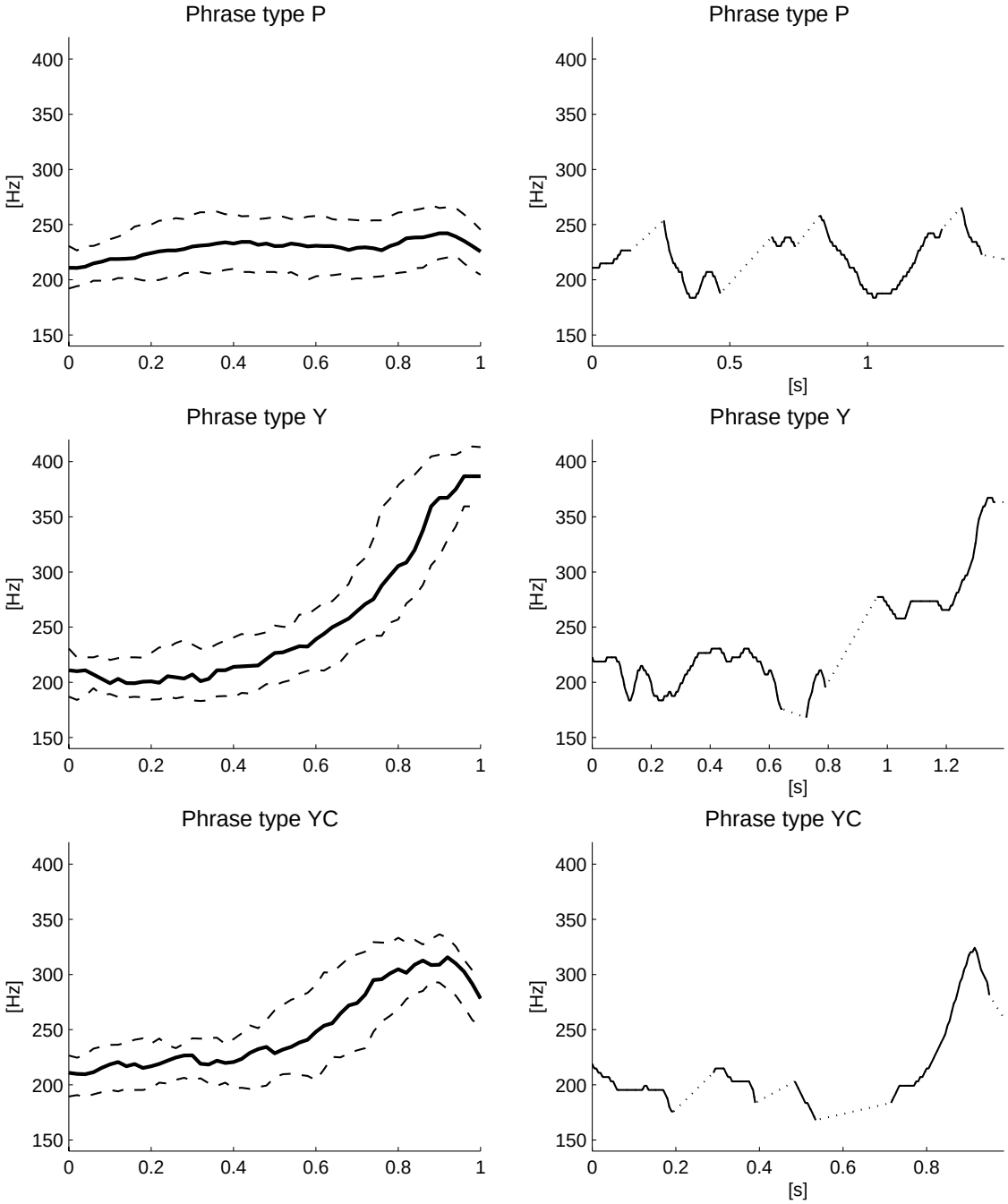

Table 5.5: *German phrase types P, Y, and YC: Median, the 25, and the 75 percentiles of the time-normalized, linearized F_0 contours of all P, Y, and YC phrases of the German prosody corpus are shown on the left side. On the right side, one characteristic F_0 contour is shown for each of these phrase types.*

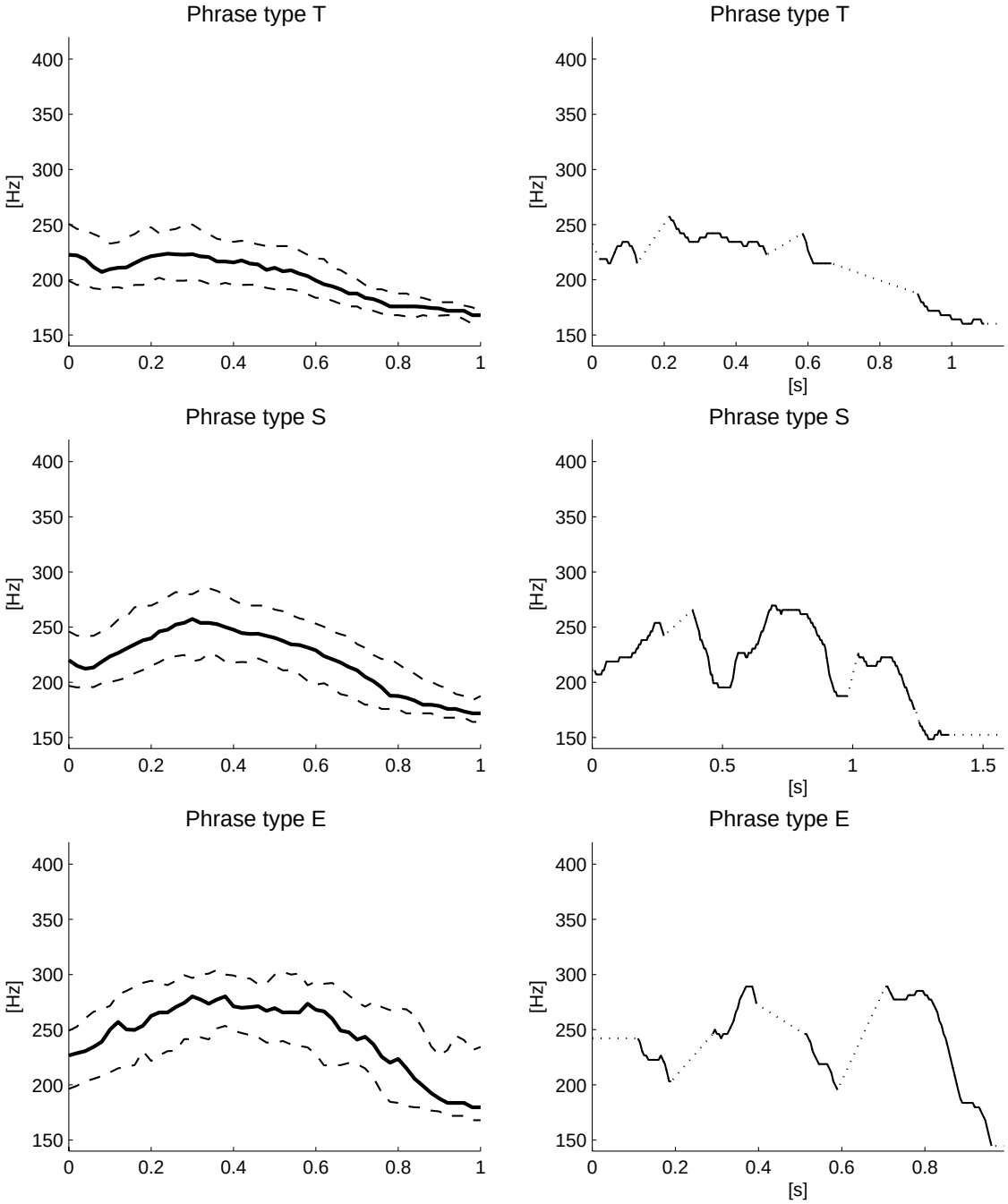

Table 5.6: *French phrase types T, S, and E: Median, the 25, and the 75 percentiles of the time-normalized, linearized F_0 contours of all T, S, and E phrases of the French prosody corpus are shown on the left side. On the right side, one characteristic F_0 contour is shown for each of these phrase types.*

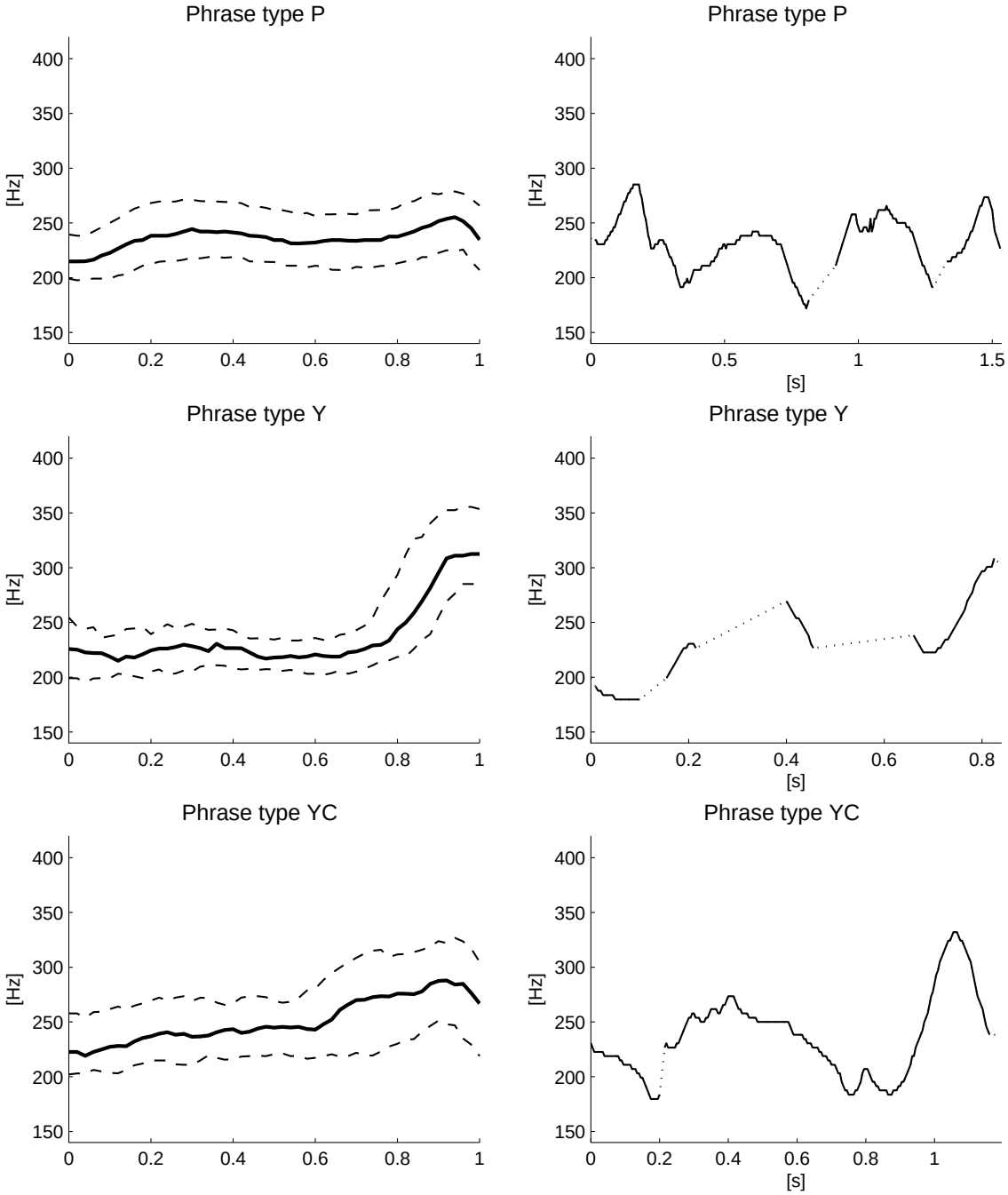

Table 5.7: *French phrase types P, Y, and YC: Median, the 25, and the 75 percentiles of the time-normalized, linearized F_0 contours of all P, Y, and YC phrases of the French prosody corpus are shown on the left side. On the right side, one characteristic F_0 contour is shown for each of these phrase types.*

initial and final pitch raising than German P phrases. A little more pronounced initial pitch raising is also observed in French T and S phrases when compared to German T and S phrases. The intonation contours of exclamation phrases seem to be in general quite similar.

5.5 Segmentation and Labeling

An accurate extraction of phone and speech pause durations requires an exact segmentation of the natural speech data of the prosody corpus into adjacent, non-overlapping speech or pause segments, and a correct assignment of labels to these segments indicating the segment type. This assignment is commonly termed "labeling".

Since the phonological representation contains the standard phonetic transcription of an utterance (see Section 5.4), it is convenient to use this standard transcription for automatically segmentation and labeling. However, several authors reported, e.g., [Rie98, vSSM+97], that statistical models for segment duration prediction, that were derived from automatically generated segmentations using the standard phonetic transcription, performed poorly. This was also the author's own experience in first experiments.

A *close phonetic transcription*, also referred to as *matched phonetic transcription*, that indicates pronunciation variants made by the speaker, results in a much better segmentation and labeling. Thus, a common practice to improve the quality of segment duration prediction models is to use manually segmented and labeled natural speech data, as it was done, e.g., in [MvS96] or in [Rie98].

5.5.1 Segment Types

Segment types correspond to the phone types determined in the transcription with two additions: plosives were segmented into their hold and burst parts that were labeled separately. While the burst part of a plosive is denoted by the same symbol used for the plosive phone type, a "`>`" denotes the preplosive pause. Speech pauses that correspond to phrase boundaries marked as `#{*:0}` or `#{*:1}` in the phonological representation of the utterances were labeled with the symbol "`/`". For a plosive following a speech pause, no preplosive pause was segmented.

	German	French
vowels	i iː i̯ ɪ y yː ʏ u uː u̯ ʊ e eː ø øː o oː ə ɛ ɛː œ ɔ a aː ɐ ɐ̯	i iː y yː u uː e ø øː o oː ə ɛ ɛː ɛ̃ ɛ̃ː œ œː œ̃ œ̃ː ɔ ɔː ɔ̃ ɔ̃ː a aː ɑ ɑː ɑ̃ ɑ̃ː
diphthongs	a͡i a͡u ɔ͡y	
consonants	p pʰ b t tʰ d k kʰ g ʔ m m̩ n n̩ ŋ r f v s z ʃ ç x h j l l̩	p b t d k g ʔ m n ɲ ŋ ʀ f v s z ʃ ʒ h j ɥ w l
affricates	p͡f t͡s t͡ʃ	
pauses	> /	> /

Table 5.8: *Phone and speech pause segment types used for transcription of the natural speech data of the German and of the French monolingual prosody corpora.*

Table 5.8 lists all segment types used for transcription of the natural speech data of the German and of the French prosody corpora.

5.5.2 Automatic Segmentation Procedure

Manual transcription and segmentation of the polyglot prosody corpus would have taken too much time. However, most existing automatic segmentation procedures are too inaccurate or they require already a close phonetic transcription as input. Therefore, the author developed a new automatic segmentation procedure for polyglot natural speech data that simultaneously delivers an highly accurate phonetic segmentation and a close phonetic transcription.

This segmentation procedure relies on iterative Viterbi search for best-matching pronunciation variants and on iterative retraining of phone hidden Markov models (HMMs). In contrast to existing, high-accuracy automatic segmentation procedures, e.g., [vSS99, Hos00] that

are two of the most accurate state-of-the-art segmentation systems, this procedure does not require very elaborate features, but only "standard" mel-frequency cepstral coefficients (MFCCs) and voicing information.

The segmentation procedure consists of two stages: First, context-independent three-state left-to-right phone HMMs with 8 Gaussian mixtures per state are trained on the natural speech data of the polyglot prosody corpus using the standard phonetic transcription of the utterances by applying a so-called "flat start" initialization, cf. [YEH+02].

For the second stage, a small set of language-dependent and speaker-dependent pronunciation variation rules is applied to the canonical transcriptions and a recognition network is generated for each utterance. Such a network includes all pronunciations allowed by the rules.

Then a Viterbi search determines the most likely path through the networks and thus delivers an adapted phonetic transcription of each utterance. These new transcriptions are used to retrain the HMMs that are in turn used in the next iteration for the Viterbi search. The procedure stops when the number of insertions, deletions, and replacements of phones between the current and the previous adapted transcriptions falls below some predefined threshold. Details on this segmentation procedure can be found in [RP05].

Since the length of the analysis window restricts the accuracy of boundary detection of certain segments, e.g., preplosive pauses, a post-processing was added to the second stage, that corrects segment boundary placement of specific segment classes based on the speech signal amplitude and voicing information.

The accuracy of automatic segmentation and labeling of the polyglot prosody corpus was evaluated on the 92 manually segmented sentences of the monolingual prosody test sets. These sentences contain altogether about 3680 segments. The deviation of the automatically generated segment boundary positions from the manual positions was less than 2 ms for 49%, less than 5 ms for 73%, less than 10 ms for 84%, and less than 20 ms for 93% of all segment boundaries. The length of 47% of the segments deviated less than 5% of their original length, 64% of the segments less than 10%, and 79% of the segments less than 20%. About 10% of the uttered phones did not correspond to the standard phonetic transcription of the utterances. The segmentation procedure automatically corrected approximately 40% of these pronunciation variations.

Chapter 6

Weighted ANN Ensembles for Prosody Modeling

6.1 Introduction

The requirements to a polyglot prosody model listed in Section 4.5 make data-driven, statistical generation methods the first choice for modeling polyglot prosody. Given prosody corpora of limited size, however, the statistical models that are applied in current TTS synthesis systems, like CART-, MARS-, HMM-, or ANN-based generation models, cf. Section 4.3, have one or more of the following disadvantages: their *prediction accuracy* is too low to achieve natural sounding prosody. Their *generalization capability* for input factor combinations that are not covered by the prosody corpora is not good enough. Or they are not *robust* enough against outliers of acoustic parameters due to errors in the segmentation, in the labeling, or in F_0 extraction of the prosody corpora.

A comparison of these models in [Rie98] showed that CART-based models are too inaccurate. MARS-based models achieve good accuracy, but have, according to [Fri01], a lower generalization capability and are sensitive to outliers. ANN-based models show good generalization capabilities and also good accuracy, but need in general more training data than the other methods to reach equal accuracy and are, in case

of limited training data, also sensitive to outliers that can result in unsteady output contours. E.g., [Tra95] reports that the RNN-based Model for F_0 generation needs a moving average filter to smooth the somewhat noisy network output F_0 contours.

Recent advances in machine learning show that weighted ensembles of ANNs for regression can significantly improve prediction accuracy, generalization capability, and robustness against outliers when compared to single networks, cf. [GVC05].

Factor relevance determination procedures can be used to remove (or "prune") less relevant input factors. Thus, the problem of small prosody corpora is less critical, and certain statistical models, like ANNs, also gain interpretability, as irrelevant input factors are removed.

6.2 Prediction Error Measures

The comparison of different prosody models, that are optimized on individual prosody corpora, requires some sort of error measure that appraises both the performance of the model and the relative complexity of the different regression tasks.

An error measure, that fulfills these requirements (cp. [GVC05]), and that is used here for all experiments, is the *normalized mean squared error (NMSE)*

$$NMSE_T(\mathbf{y}) = \frac{MSE_T(\mathbf{y})}{\sigma_D^2} \ . \tag{6.1}$$

The NMSE is defined as the mean squared error of the predictor $\mathbf{y}$ on the test set T divided by the variance σ_D^2 of the complete data set D. The *mean squared error (MSE)* is defined as

$$MSE_T(\mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{t}_i - \mathbf{y}(\mathbf{x}_i)\right)^2 \ , \tag{6.2}$$

with a test set T consisting of N input/target value pairs $\{\mathbf{x}_i, \mathbf{t}_i\}$.

The NMSE calculates the relative performance of the predictor $\mathbf{y}$ with respect to the complexity of the regression task, expressed by σ_D^2. The NMSE is always non-negative. A $NMSE = 1$ corresponds to

constant prediction of the data average. Predictors that perform better than constant prediction of the data average have a NMSE less than 1.

6.3 Weighted Neural Network Ensembles

A weighted ensemble is a set of independently trained statistical models, that are then often called *base learners*. The prediction output of the ensemble is a linear combination of the prediction outputs of the individual models

$$Y_M(\mathbf{x}) = \sum_{m=1}^{M} w_m \mathbf{y}_m(\mathbf{x}) \tag{6.3}$$

where M is the number of individual models, $\mathbf{y}_m$ is the prediction output of the m-th member, and w_m is a decreasing function of the prediction error of the m-th member over the whole training set. Thus, each ensemble member is weighted according to its individual performance.

6.3.1 Base Learners

As *base learners*, the author applied feed-forward ANNs for segment duration prediction, and recurrent ANNs (RNNs) for predicting F_0 contours. A feed-forward neural network, also known as multilayer perceptron (MLP), can be described as a series of functional transformations that are represented in case of two layers of weights by the network function

$$y_k(\mathbf{x}, \mathbf{w}) = g\left(\sum_{j=0}^{M} w_{kj}^{(2)} h\left(\sum_{i=0}^{D} w_{ji}^{(1)} x_i\right)\right) \tag{6.4}$$

and in case of three layers of weights by the network function

$$y_k(\mathbf{x}, \mathbf{w}) = g\left(\sum_{r=0}^{N} w_{kr}^{(3)} h\left(\sum_{j=0}^{M} w_{rj}^{(2)} h\left(\sum_{i=0}^{D} w_{ji}^{(1)} x_i\right)\right)\right) \quad , \tag{6.5}$$

where the set of all weight and bias parameters have been grouped together into a vector $\mathbf{w}$. $w_{j0}^{(1)}$, $w_{k0}^{(2)}$, $w_{r0}^{(2)}$, and $w_{k0}^{(3)}$ represent the bias parameters of the individual weight layers. $h(\cdot)$ is a differentiable, nonlinear activation function of the hidden units. $g(\cdot)$ is a differentiable activation function of the output units. The network function can be represented in form of a network diagram as shown in Figure 6.1.

In all experiments, $h = tanh$ was used as hidden unit activation function for all networks. The output unit activation function $g(\cdot)$ was the identity function. The author also tested *tanh* as output unit activation function, as it was done in [Tra95] and [Rie98]. *tanh* would prevent over-shooting of the prediction outputs. However, the identity function provides better approximation of extrema and a lower prediction error. Using ensemble models, no over-shooting was observed anyway.

Input and output values are rescaled by applying a linear normalization using mean and variance calculated with respect to the training

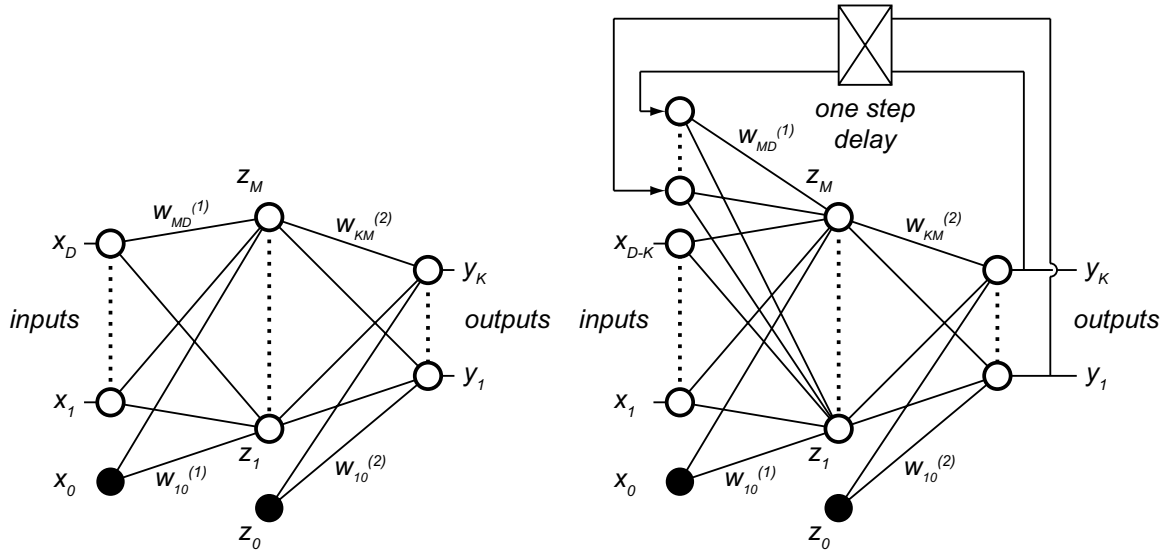

Figure 6.1: *Network diagrams for the feed-forward neural network with two layers of weights corresponding to (6.4) on the left and for a two-layer recurrent neural network with all output nodes fed back to the input layer of nodes on the right. The input, hidden, and output variables are represented by nodes, and the weight parameters are represented by links between the nodes. Bias weights are denoted by links coming from additional variables represented by filled nodes.*

set. Thus, the transformed variables have zero mean and unit standard deviation. The linear rescaling makes it possible to initialize network weights by random selection from a zero mean, unit variance isotropic Gaussian where the variance is scaled by the fan-in of the units as appropriate. Network weights of feed-forward ANNs are trained using the well-known *error back-propagation* procedure of [RHW86, Lip87]. RNNs are trained using *error back-propagation for sequences*, as described, e.g., in [Wer90]. Error back-propagation for sequences treats RNNs as feed-forward ANNs by "unfolding" them in time. Thus, the recurrent network architecture is equivalent to a feed-forward network with many sets of weights constrained to be the same. In both procedures, the *scaled conjugate gradient* algorithm introduced in [Møl93] is applied for optimization using a sum-of-squares error function of network outputs. These settings allow a very efficient training of the neural networks. [Bis95] provides a good introduction to feed-forward neural networks and a detailed description of the algorithms applied here.

6.3.2 Weighting Functions

As *weighting functions*, the author tested an exponential (6.6) and a potential weighting function (6.7), that were suggested in [GVC05], and for comparison also the arithmetic mean (6.8)

$$w_i = \frac{exp(-\alpha e_i)}{\sum_{j=1}^{M} exp(-\alpha e_j)} \ , \tag{6.6}$$

$$w_i = \frac{e_i^{-\alpha}}{\sum_{j=1}^{M} e_j^{-\alpha}} \ , \tag{6.7}$$

$$w_i = \frac{1}{M} \ , \tag{6.8}$$

where e_i is the prediction error of the i-th model. M is the number of individual models in the ensemble. α is a weighting coefficient. Setting $\alpha = 0$ results in the arithmetic mean for both weighting functions.

As an example for the influence of the weighting function on prediction error, Figure 6.2 shows the normalized mean squared error as a function of the weighting coefficient for German and for French segment duration prediction. Applying exponential or potential weighting result

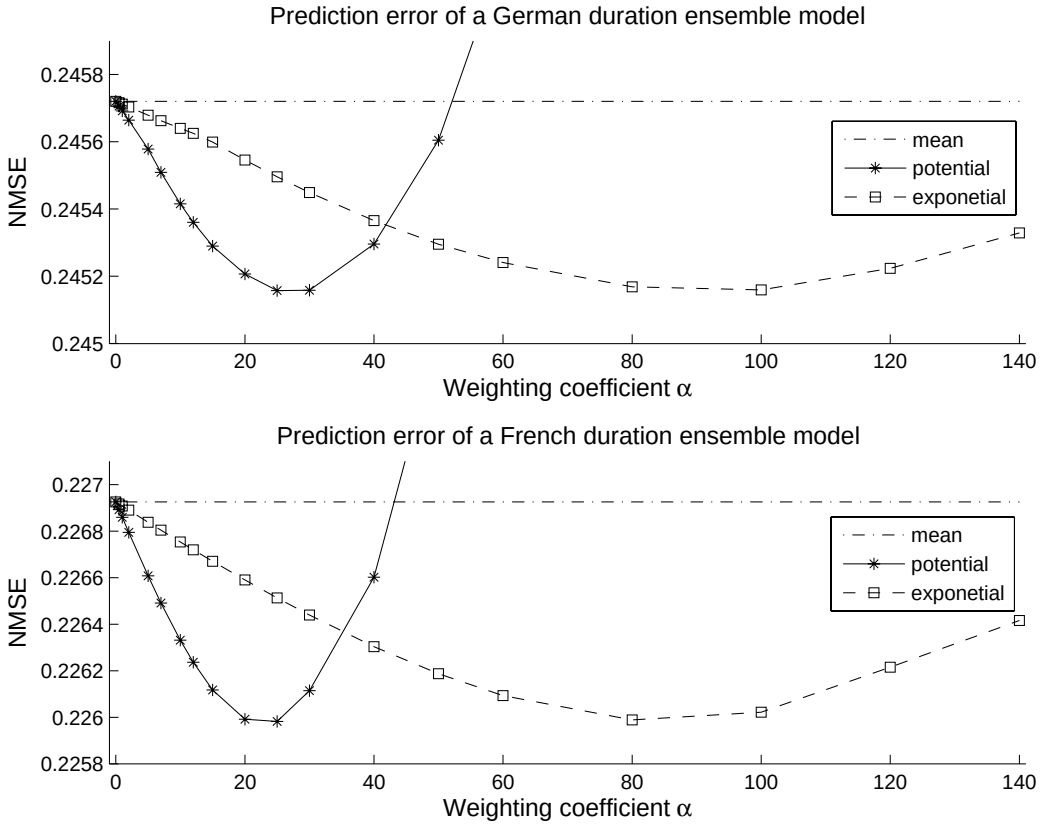

Figure 6.2: *Normalized mean squared error as a function of the weighting coefficient α using arithmetic mean, potential weighting, and exponential weighting function for German (above) and for French (below) segment duration prediction. The German model is an ensemble of 9 ANNs having 114 input factors each. The French model is an ensemble of 11 ANNs, each with 68 input factors.*

in nearly identical ensemble prediction errors and produce for small to medium values of α slightly better results than using the arithmetic mean. For large values of α, overfitting is observed, since only a few particular networks contribute to the ensemble output. Exponential weighting results in more robust error curves than potential weighting. Therefore, exponential weighting with a weighting coefficient $\alpha = 80$ was used for German and for French ensemble construction in all experiments.

6.3.3 Network Aggregation

[GVC05] present an evaluation of aggregation methods for ANN ensembles on several standard regression problems. The tested aggregation methods base either on *Boosting*, that was introduced for classification problems in [Sch90], and later extended for regression problems, e.g., in [Dru97] and in [Fri01], or they base on *Bagging* (short for "bootstrap aggregation"), which was introduced in [Bre96]. A *Weighted Bagging* algorithm (W-Bagging) and the so-called *Weighted Stepwise Ensemble Construction Algorithm* (W-SECA), another bagging-like algorithm, both introduced in [GVNC02], were among the top performers in nearly all test cases. These weighted bagging-based algorithms outperformed the boosting methods and also other regression methods that based, e.g., on Support Vector Machines (SVMs), which were introduced in [Vap95], or on Boosting using Radial Basis Functions (RBF) networks, which are described, e.g., in [RDB02]. For aggregation to be effective, however, the individual networks of the ensemble must be *both accurate and diverse*.

Diverse individual networks can be obtained, e.g., by varying the internal network structures or by using different adequately-chosen subsets of the training set to optimize the parameters of the individual networks. A vital element of success when using different subsets is the instability of the learning algorithm, cf. [Bre96]. ANNs are therefore well suited, as this instability comes naturally from the inherent randomness of the training algorithms of ANNs.

Network accuracy can be increased, e.g., by fitting the complexity of the network and the size of the training set optimally to each other. However, increasing the size of a prosody corpus is very elaborate and often not possible. Reducing the network complexity either requires to reduce (or "prune") the number of nodes or weights in the hidden layers, which means to reduce also the expressive power of the network, or to reduce the number of input nodes. This network pruning is described in details in Section 6.4.

For the ensemble construction of the prosody models, the author tested the W-SECA and the W-Bagging aggregation method. For W-Bagging, 6-fold cross validation was used instead of bootstrapping to obtain six different subsets of the training data. On each of these subsets, ANNs with eight different internal network structures, having one

Network Nr.	1	2	3	4	5	6	7	8	9	10	11
Layer 1	20	15	15	20	20	20	15	20	14	14	15
Layer 2	10	15	15	5		10	15	5	20	20	15

Table 6.1: *Network structure of each ANN member of the best ensemble shown in Fig 6.3. For each ensemble member, the number of nodes of the first and of the second hidden layer is given. Each ANN has 68 input nodes.*

or two layers of hidden nodes and different number of nodes, were trained; thus, in total, 48 ANNs were trained, that differ in the training data and in the internal network structure.

Figure 6.3 shows as a typical example the construction of the French

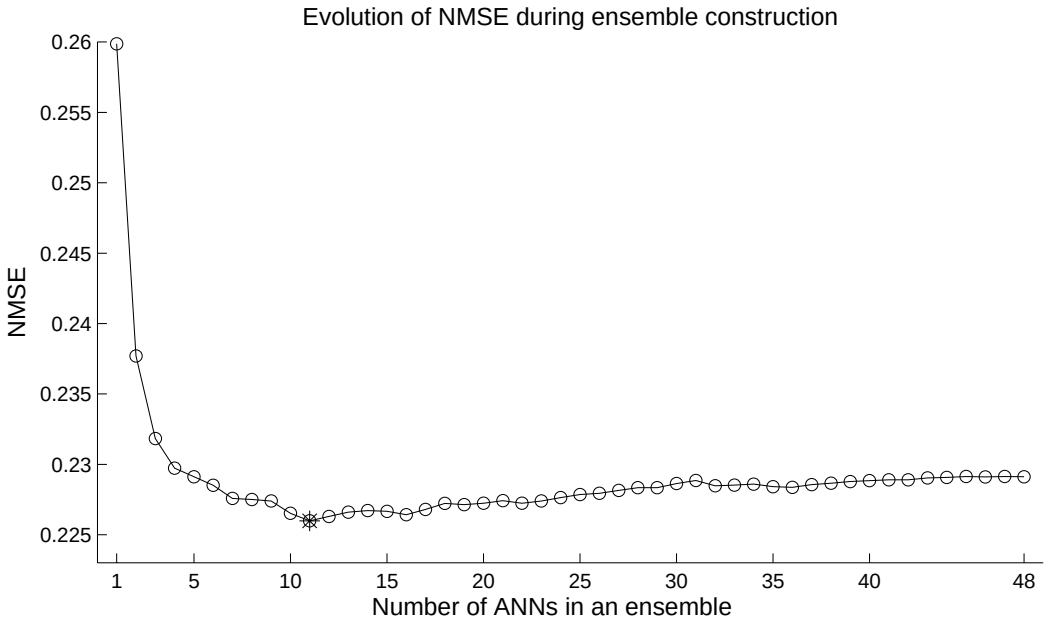

Figure 6.3: *Normalized mean squared error as a function of the number of ensemble members during ensemble construction. As a typical example, the construction of the French duration model is shown. Each ANN of the ensemble has 68 input factors. The best ensemble consists of 11 ANNs and its NMSE is denoted by a star. The ensemble with only 1 ANN is identical with the best single ANN-based model.*

duration model with 68 input factors using W-Bagging aggregation. The evolution of the NMSE of the ensemble on the test set is shown as a function of the number of ensemble members. The best ensemble consists of 11 ANNs. Table 6.1 shows the network structure of the individual ANNs. The prediction error of this ensemble-based duration model is 13% smaller than the prediction error of the best single ANN-based model, identical to the ensemble with only one ANN member in Figure 6.3. For German duration prediction, the improvement is about than 15%. W-SECA performed for the tested prosody corpora very similar to W-Bagging. As W-Bagging is less computational expensive than W-SECA, this method was finally applied.

6.4 Factor Relevance Determination

The optimal network structure for ANN-based prosody models is constrained by the following considerations:

- In ANN-based prosody models, the relevance of an individual input factor and the complexity of the regression problem is in general unknown. Therefore, most prosody models comprise a large number of input factors that might be somehow relevant. However, the inclusion of many input factors has a lot of drawbacks: e.g., the interpretation of the model is more difficult, irrelevant input factors may act as input noise, the generalization capability of the model is worsened, and the demand of training samples grows exponentially with the dimensionality of the factor space. This limitation is called the "curse of dimensionality", cf. [Bel61].

- Neural networks having a finite number of hidden units will approximate a given function with a residual error. [Jon92] has shown that this error decreases as the number of hidden units is increased. However, given a training set of finite size, the total number of network weights should be roughly $\frac{1}{10}$ of the total number of training points, cf. [DHS01]. A traditional method for optimizing the network structure is to initially train the network with a large number of hidden nodes and later prune irrelevant weights. Because of the limited size of the prosody corpora and the large number of input factors, the number of hidden nodes of

ANN-based prosodic models is in general very small. Therefore, traditional network pruning is not useful for ANN-based prosody models.

The key idea for optimizing the network structure of ANN-based prosody models is to find a trade-off between these two constraints: therefore, an initial model with a very large number of all possible input factors is trained. From this initial model, the input nodes of the least relevant factors, as determined by a factor relevance algorithm, are iteratively removed and simultaneously the number of hidden nodes is increased until the expressive power of the network optimally fits the evaluation data.

For factor relevance determination, the author applied an extension of the so-called "optimal brain surgeon" (OBS) algorithm introduced in [HSW93]. OBS is specially designed for pruning of ANNs and uses information from the Hessian to perform network pruning. The extension of OBS, the *Unit-OBS* algorithm, cf. [SR96], considers the outgoing weights of one node (unit) as a group of candidate weights: when all the weights of an unit can be deleted, the unit itself can be pruned. After pruning of an unit, the weights of the other units are corrected. To determine the relevance of the input factors, all input units are iteratively removed.

Factor relevance determination of RNNs is not defined by the OBS or by the Unit-OBS algorithms. If "unfolded" in time, a RNN can be trained similarly to an ANN. However, calculation of the inverse Hessian of such an unfolded RNN with about 50 000 input nodes was computationally not feasible. Therefore, an approximation was made by training the RNN to local minimum error first, and then, after removing the feed-back connections, applying Unit-OBS to the left over, feed-forward ANN structure of the RNN. The input nodes of the feed-back connections are not considered in factor relevance determination.

6.5 Prosody Ensemble Construction

The individual networks of an ensemble must be as accurate and as diverse as possible. Therefore, the ensembles for prosody control are finally constructed from individual networks, where each has a specific

set of *input factors*, a specific *network structure*, a specific *network size*, and is optimized on a specific *training set*.

This construction procedure of ANN- or RNN-based ensembles for prosody control can be summarized by the following steps:

1. Train a network with all input factors to local minimum error.
2. Determine the relevance of all input factors by applying the Unit-OBS algorithm on this network.
3. Train a set of networks for different numbers of the most relevant input factors. For each set of input factors, construct networks with a varying number of hidden layers and a varying number of hidden nodes, and train these networks by n-fold cross-validation on different training sets.
4. Aggregate individual networks to ensembles by testing all possible combinations of individual networks.
5. Select ensemble with lowest prediction error.

This construction procedure is completely automatic and was applied for the construction of F_0 and of duration ensemble models used for German and French prosody control.

Chapter 7

Polyglot Prosody Control

7.1 Model Architecture

The polyglot prosody model consists of independent F_0 control and segment duration control modules that generate from the phonological representation of an utterance the corresponding F_0 and segment duration values. Figure 7.1 displays a schematic overview of this model.

A *factor encoding* component converts the phonological representation into a language-independent input representation. An *output decoding* component converts the language-independent output representation into the actual acoustic parameter. These language-independent representations enable language switching between monolingual models and make it possible to add new models for others languages without requiring to modify the existing models. Language switching itself is triggered by the language tags of the phonological representation.

The polyglot F_0 control is described in Section 7.2 and the polyglot segment duration control in Section 7.3. Section 7.4 finally presents a perceptual evaluation experiment using the polyglot prosody model and a discussion of the results.

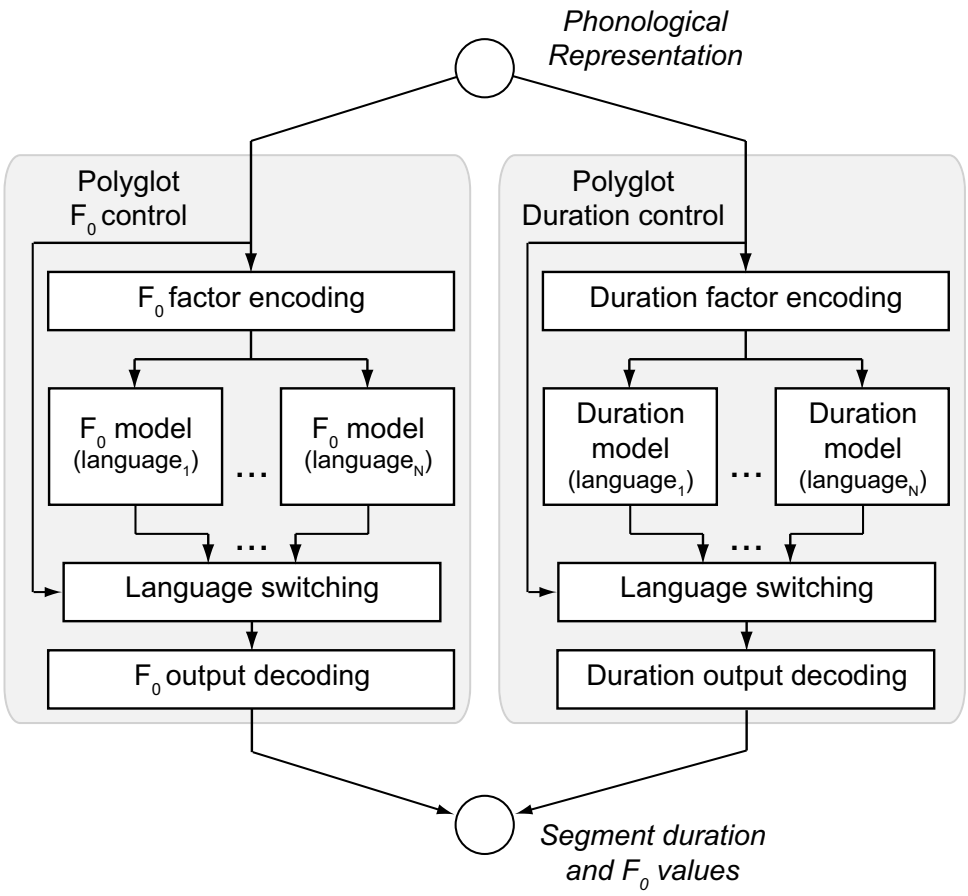

Figure 7.1: *Schematic representation of the polyglot prosody model: independent F_0 control and segment duration control modules generate from the phonological representation of an utterance the corresponding F_0 and segment duration values. Language-independent input and output representations, as output of factor encoding and as input to output decoding, resp., enable seamless language switching and make it possible to add new models for F_0 and segment duration generation for other languages without modifying the existing models.*

7.2 Fundamental Frequency Control

The polyglot F_0 control processes the phonological representation of a polyglot utterance as a sequence of syllable and boundary symbols. For each symbol, it generates a F_0 contour by applying the monolingual F_0 model that corresponds to the symbol's language.

7.2.1 Model Architecture

The polyglot F_0 control consists of a language-independent input factor representation, that is described in Section 7.2.3, a language and time independent F_0 output representation, which is presented in Section 7.2.4, and an independent, monolingual F_0 model for each individual language. In order to provide language switching between the individual monolingual models, the F_0 outputs of the preceding syllable are fed back to the inputs of the monolingual models. Figure 7.2

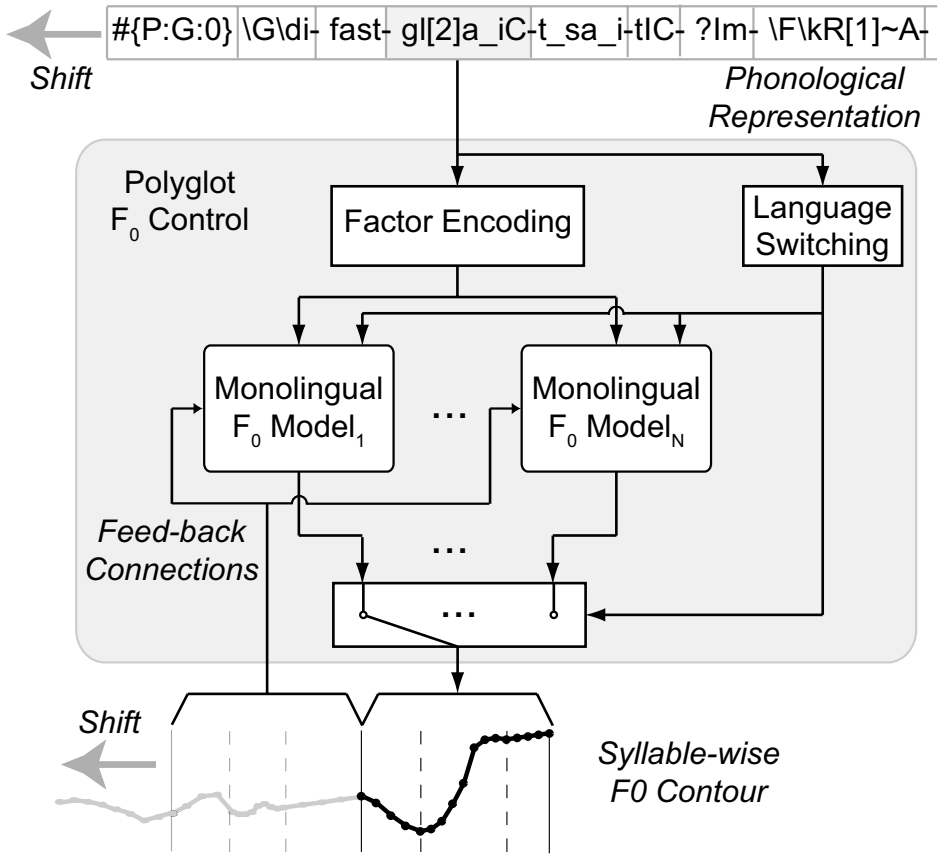

Figure 7.2: *Schematic representation of the polyglot F_0 model: for each syllable and for each boundary symbol of the phonological representation, a set of language independent input factors is extracted and encoded. The language information of the phonological representation selects the corresponding monolingual F_0 model to generate the F_0 output values. For language switching, the F_0 output values of the last syllable are fed back to the input of the monolingual models.*

gives a schematic overview of the polyglot F_0 model.

Each monolingual F_0 model is a weighted RNN ensemble that is constructed using the procedure presented in Section 6.5. Figure 7.3 displays a schematic representation of such a monolingual F_0 model. Each RNN has its own input factor selection that chooses the optimal

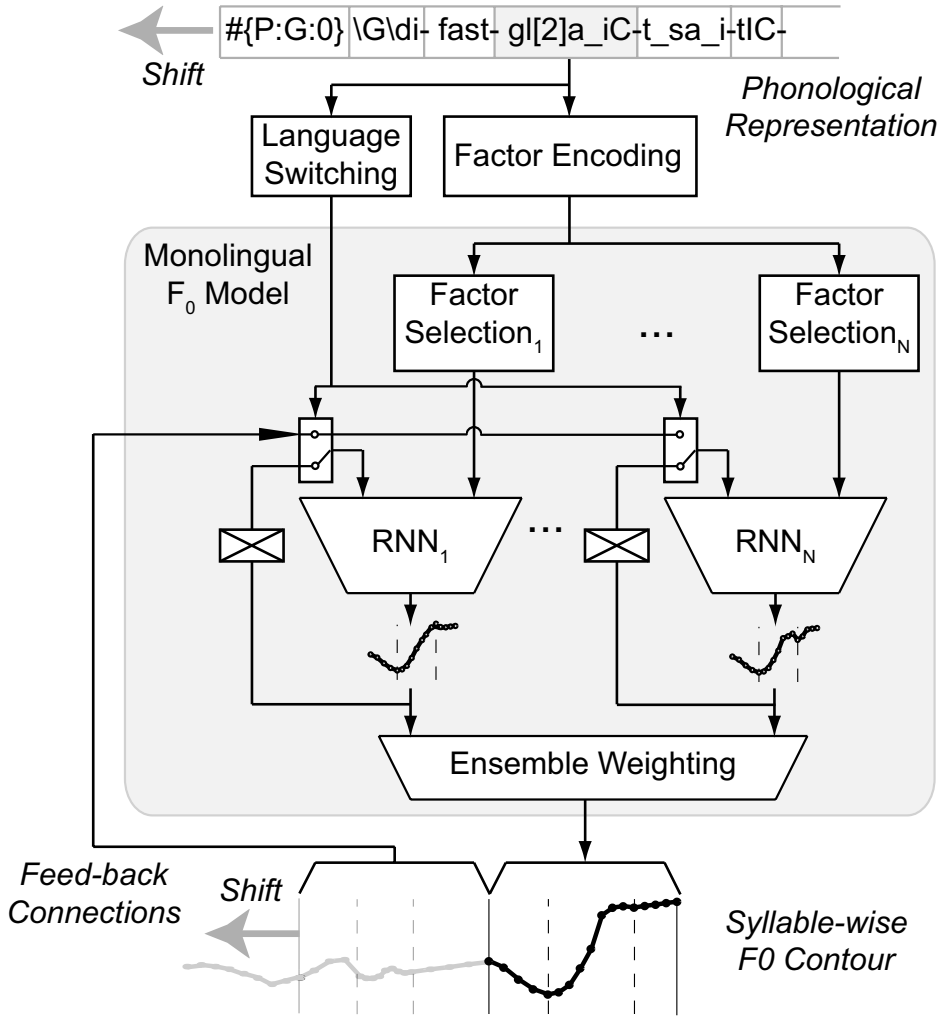

Figure 7.3: *Schematic representation of a monolingual F_0 model as a weighted RNN ensemble: each RNN member of the ensemble has its own input factor selection that chooses the optimal set of input factors for this network. Each RNN uses its own output as feed-back. These feed-back connections can be set to some external F_0 values in order to initialize the RNN at the start of an utterance or to set the general F_0 level, if it is used in the polyglot F_0 model.*

set of input factors for this network. The basic RNN structure is similar to the RNN-based F_0 model presented in [Tra95]. The network setup of the RNN ensemble members of the German and of the French F_0 models is given in Table 7.3.

7.2.2 Language Switching

The feed-back connections from the last hidden layer in the RNN-based F_0 model of Traber mainly serve to control the general level of F_0 whereas the more local phenomena are controlled by the direct input to the network, cf. [Tra95]. In order to enable language switching without audible melodic discontinuities, the feed-back connections of the RNN model of the preceding language can be used to initialize the recurrent input of the RNN model of the new language. Thus, the model for the new language continues at the same general level of F_0 as defined by the model of the preceding language.

The feed-back from the last hidden layer of the RNN, however, would make it impossible to switch between individual networks as the hidden layer outputs of individual networks are in general very different. Therefore, the author tested two different network configurations: the first configuration completely avoids feed-back connections and uses relative syllable position information within phrases as input instead. The second configuration feeds the final F_0 outputs of the network back to the input layer. As long as all networks are optimized on the same training set, both configurations make it possible to switch between networks without discontinuities in the intonation contours.

An experiment done by the author has shown that the feed-back of F_0 output values results in only slightly higher prediction errors than using the outputs of the last hidden layer. The use of relative position information instead of feed-back connections, however, results in seriously larger prediction errors.

In another experiment, the relevance of the relative position information factors was determined for two identical networks trained with and without feed-back connections: while these factors were in the top ten relevant factors for the feed-forward network, the relevance of these factors for the recurrent network was only mediocre. This result indicates that by using feed-back connections, additional information about the syllable position within an utterance is superfluous.

The finally used ensemble models for German and for French F_0 modeling therefore apply feed-back connections from the F_0 outputs and omit any position information factors. Each RNN member uses its own F_0 outputs as feed-back, as it is shown in Figure 7.3. These feed-back connections can be set to some external F_0 values: either to zero, in order to initialize the RNN at the start of an utterance, or to the F_0 values of the preceding syllable, in order to set the general F_0 level for language switching.

7.2.3 Input Representation

The phonological representation of an utterance is processed as a sequence of syllable and boundary symbols. Each input symbol is represented by a vector of 910 elements. All elements of this vector are set to zero by default. The values of ordinal factors are directly set in the vector. For categorical factors, a 1-out-of-n encoding is applied such, that each categorical factor is represented by n binary factors.

It is generally acknowledged, that the F_0 contour of a syllable depends on a relatively wide phonological context as far as accentuation and phrasing information is concerned, whereas the influence of segmental properties on the F_0 contour of a syllable is much more local. However, the correct size of these contexts for the different factors is unknown and depends on the prosodic phenomena to be modeled. The author therefore applied a context of *3 preceding and 6 subsequent symbols* for accentuation and phrasing information (equal to the context used in [Tra95]), and a context of *2 preceding and 2 subsequent symbols* for segmental properties. Starting with this large, initial factor set, the ensemble construction procedure of Section 6.5 was applied to automatically select the most relevant input factors and thereby the optimal context size of each factor.

For polyglot F_0 control, this input representation must be language independent. This means that no language specific segment types or phrase types can be used, but the language-independent description of manner and place of articulation of phones of the IPA and a basic, language-independent set of phrase types, as listed in Section 3.2.1. Also, information about syllable language or language switching position may not be part of the factor set. Language information is only used to switch between the monolingual F_0 models.

Factors describing accentuation and phrasing, syllable structure and segmental information, and sentence length and syllable position have been selected. For each of these categories, in the following all factors and their values are described.

Accentuation and Phrasing Factors

The factors for phrasing describe a phrase by boundary strength, by type, by length, and by its position within a sentence. The factors describing accentuation are the syllable stress levels. As very short phrases show a typical F_0 contour, in addition to "phrase length" a binary factor "short phrase" for phrases shorter than 4 syllables was used.

Accentuation and phrasing factor values	
syllable stress	`[E]`, `[1]`, `[2]`, `[3]`, `[4]`, unstressed
phrase boundary	`0`, `1`, `2`
phrase type	`P`, `T`, `S`, `Y`, `E`, `YC`, `F`
phrase length	number of syllables
short phrase	binary
first phrase	binary

For *boundary symbols*, only phrase boundary and phrase type factors are used. All other factor values are set to zero. For *syllable symbols*, all factors except phrase boundary factors are set. For accentuation and phrasing factors, a context of 3 preceding and 6 subsequent symbols is applied. Thus, in total, 190 input factors $((3 + 1 + 6) * 19)$ are used.

Syllable Structure and Segmental Factors

The factors for syllable structure describe for each syllable of how many segments syllable onset, nucleus, and coda are built.

Syllable structure factor values	
nucleus has 1 phone	binary
nucleus has 2 phones	binary
nucleus has more than 2 phones	binary
onset size	number of phones
coda size	number of phones

	Front(closing)	Central	Back(closing)
Close	i iː i̯ a͡i̯ y yː ɪ ʏ		ʊ u uː u̯ a͡u̯
Close-mid	e eː ø øː ɔ͡y	ə	o oː
Open-mid	ɛ ɛː ɛ̃ ɛ̃ː œ œː œ̃ œ̃ː		ɔ ɔː ɔ̃ ɔ̃ː
Open	a aː	ɐ ɐ̯	ɑ ɑː ɑ̃

Table 7.1: *IPA representation of vowels and diphthongs that are used in the polyglot prosody model. Each segment is described by a tuple of vowel height and vowel backness: e.g., [ø] is referred to as {Close-mid, Front}.*

Segmental factors differ between factors for consonants in syllable onset or coda, and factors for the segments in syllable nucleus. The nucleus is here defined as the sequence of vowels, semi-vowels, and syllabic consonants within one syllable. Diphthongs or triphthongs are split into the corresponding monophthongs, e.g., [a͡i̯] is split into [ai]. Aspirated plosives are split into the corresponding plosive followed by [h].

For the first and the second phone of the nucleus, the following segmental factors are used:

Nucleus factor values	
long vowel	binary
nasal vowel	binary
first formant position	low, middle, high
vowel characterization	tuple from Table 7.1
semi-vowel characterization	tuple from Table 7.2

The first and the last phone of onset and coda are described by the following segmental factors:

Onset/coda factor values	
segment type	consonant, glottal closure affricate, preplosive pause
voiced segment	binary
strong consonant	binary
consonant characterization	tuple from Table 7.2

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d				k g			ʔ
Nasal	m m̥			n n̥			ɲ	ŋ			
Trill				r					ʀ		
Tap											
Fricative		f v	θ ð	s z	ʃ ʒ		ç	x			h
Lateral fricative											
Approximant							ɥ j	w			
Lateral approximant				l l̥							
Affricate		p͡f		t͡s	t͡ʃ						

Table 7.2: *IPA representation of consonants, syllabic consonants, semi-vowels, and affricates that are used in the polyglot prosody model. Each segment is described by a tuple of manner of articulation and place of articulation: e.g., [ŋ] is referred to as {Velar, Nasal}.*

Segmental factors are only set for *syllable symbols*. For these factors, a context of 2 preceding and 2 subsequent symbols is used. This results, in total, in 715 segmental input factors $(5 * (5 + 19 * 2 + 25 * 4))$.

Sentence Length and Syllable Position Factors

In addition to phrase length, also factors describing sentence length are included. For testing feed-forward ANNs for F_0 modeling, three additional factors describing relative syllable position within phrase boundaries in the range [0..1] were included.

Sentence length & syllable position factor values	
sentence length	number of syllables
short sentence (less than 5 syllables)	binary
syllable position within 0 boundaries	position $\in$ [0..1]
syllable position within 1 boundaries	position $\in$ [0..1]
syllable position within 2 boundaries	position $\in$ [0..1]

For sentence length and syllable position information, no context is needed. Thus, this information is described by 5 input factors.

Using the Unit-OBS procedure described in Section 6.4, the relevance of these 910 input factors $(190 + 715 + 5)$ was determined for German and for French F_0 modeling separately. Appendix C lists the complete ranking of all input factors for German and for French F_0 modeling.

7.2.4 Output Representation

In order to make F_0 control independent from duration control, a time-independent representation of the F_0 contour is necessary. This can be achieved by applying a linear approximation of the original, linearized F_0 contour using a constant number of equidistant F_0 samples for each syllable.

The syllable-wise, ANN-based F_0 generation predicts such a constant number of F_0 samples per syllable that corresponds to the number of output nodes of the ANNs. The author tried various representations of the F_0 contour: a first experiment compared linear approximations of the F_0 contour using 5 (similar to the representation applied in [Tra95]),

11, and 17 equidistant F_0 samples for each syllable. This experiment showed that the use of 17 F_0 samples per syllable results in predicted F_0 contours that approximate the original F_0 contours more accurate than using representations with a fewer number of F_0 samples. The use of 17 F_0 samples also allows the generation of more natural sounding synthetic speech.

However, empirical findings concerning the timing of F_0 peaks within syllables due to segmental constraints [vSH94, vS02] or semantic constraints [Koh03] show that certain anchor points for positioning F_0 peaks within a syllable are necessary. A manual inspection of the F_0 contours of the syllables of the prosody corpora revealed for identical vowels roughly similar patterns in the nucleus part of the F_0 contours. Figure 7.4 shows the F_0 contours of the syllables [glaiç] and [bai] of the German prosody corpus. While the overall F_0 contours of these two syllables look rather different, the nucleus parts of both syllable have more similar F_0 patterns. To incorporate these findings into the F_0 model, the author introduced a "sub-syllabic" representation of the F_0 contour.

This *sub-syllabic representation* bases on a segmentation of each syllable into onset, nucleus, and coda. Onset and coda parts of the F_0 contour are each linearly approximated using 5 equidistant F_0 samples, the nucleus part of the F_0 contour is modeled by 9 equidistant F_0 samples. F_0 samples at onset-nucleus and nucleus-coda boundaries are identical. Thus, this representation also uses 17 F_0 samples in total. In case of an absent onset or coda, the respective 5 F_0 samples have the same value and lie upon each other. Figure 7.4 displays the application of this F_0 contour modeling on two accented syllables of the German prosody corpus.

Another experiment compared the sub-syllabic representation with the equidistant, linear approximation of the syllable F_0 contour, both using 17 F_0 samples per syllable. This comparison showed that the use of the sub-syllabic representation improves the prediction accuracy of the F_0 model considerably. This sub-syllabic representation also conforms very well to the requirements concerning the timing of F_0 peaks within syllables due to segmental constraints [vSH94, vS02] and semantic constraints [Koh03].

For neural network processing, the absolute F_0 values were *linearly normalized* to zero mean and a standard deviation of 0.33. This is

not really necessary, as the output nodes of the neural networks have a linear activation function. However, having target values with zero mean and having the majority of target values within the range of -1 to 1 speeds up convergence of neural network training.

For training the model, the F_0 values of syllable symbols correspond to the F_0 samples extracted for each syllable from the F_0 contour. Boundary input symbols have dummy F_0 values from a linear connection between the last value of the preceding syllable and the first value of the subsequent syllable.

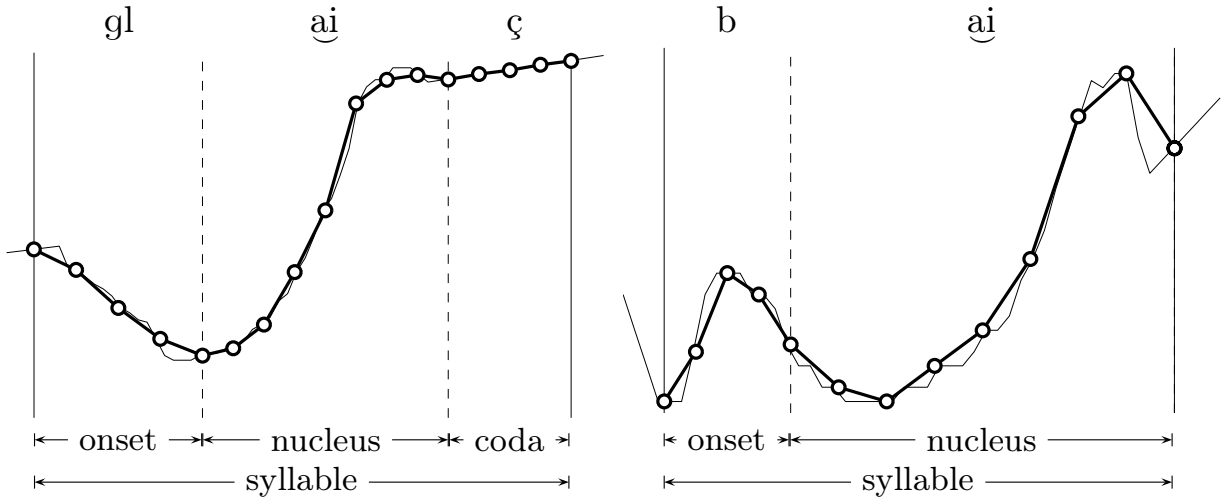

Figure 7.4: *Modeling of the original, linearized F_0 contour (thin line) of each of the syllables [gla͡iç] (left contour) and [ba͡i] (right contour) by 17 F_0 values (indicated as circles): Each syllable is segmented into onset, nucleus, and coda. Onset and coda parts of the F_0 contour are each linearly approximated (thick line) using 5 equidistant F_0 samples, the nucleus part of the F_0 contour is modeled by 9 equidistant F_0 samples. F_0 samples at onset-nucleus and nucleus-coda boundaries are identical. In case of an absent onset or coda, the respective 5 F_0 samples have the same value and lie upon each other, as indicated for the missing coda of the right syllable.*

7.2.5 F_0 Ensemble Construction

Factor Relevance Determination

The first steps of F_0 ensemble construction are the determination of factor relevance and the training of individual networks having different network configurations and different number of input factors. Figure 7.5 shows the NMSE as a function of the number of input factors for German and for French F_0 ensemble models. For this comparison, all members of an ensemble use the same number of input factors. The sequence of input factors starts with the most relevant to the left, and is *different for German and French*. The best German F_0 models have between 180 to 380 input factors, as shown in Figure 7.5. The best French F_0 models have about 440 to 550 input factors.

In both diagrams, it is clearly visible that the use of too few or of too many input factors generally results in higher prediction errors. This demonstrates the influences of the two considerations given above: a prosody model with too few input factors can only poorly describe the complexity of the problem. For French F_0 prediction, even a kind of "cut-off" point of about 220 input factors is observed, below that prediction error drastically increases. Too many input factors, however, result in too many weight parameters for the given size of the training set.

In order to look up, which input factors are used by the individual models, Appendix C provides a list of all input factors used for F_0 control together with their rank of relevance.

Network Aggregation

In the final step of F_0 ensemble construction, the best ensemble of all individual networks is determined for increasing ensemble size. The evolution of NMSE during ensemble construction for German and for French F_0 modeling is presented in Figure 7.6. The best German F_0 ensemble with RNNs with individual input factor sets has 8 network members. The best French F_0 ensemble consists of 9 RNNs. The network structures and the input factor sizes of these networks are given in Table 7.3. The best German F_0 ensemble achieves a NMSE of 0.2613, which is an improvement of about 7% compared to the best German F_0 ensemble with a fixed number of input factors, as shown in Figure 7.5.

The best French F_0 ensemble has a NMSE of 0.3148 and about 5% improvement compared to the best French F_0 ensemble with a fixed number of input factors.

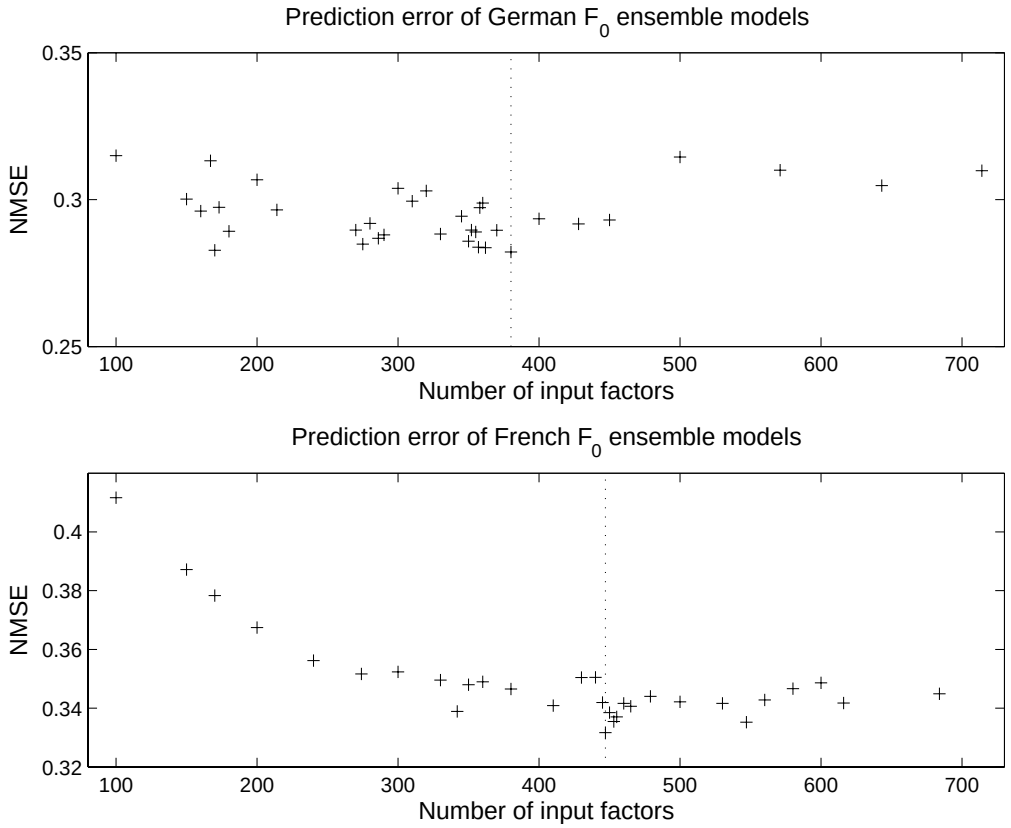

Figure 7.5: *NMSE of F_0 prediction models as a function of the number of input factors. The upper diagram displays prediction errors of German F_0 models, the lower one of French F_0 models. The sequence of input factors starts with the most relevant to the left, and is different for German and French. For each number of input factors, the prediction error of the best ensemble model is indicated by a cross. The best German F_0 prediction ensemble has 380 input factors. The best ensemble for French F_0 prediction uses 447 input factors. Both are indicated by a dotted line. For both languages, the number of ensemble member varies between 5 and 28.*

German F_0 ensemble									
Network Nr.	1	2	3	4	5	6	7	8	
Factors	170	380	350	352	160	270	290	370	
Layer 1	28	15	14	17	24	22	21	15	
Layer 2	27	22			22			27	
French F_0 ensemble									
Network Nr.	1	2	3	4	5	6	7	8	9
Factors	455	342	616	560	447	547	453	547	447
Layer 1	11	18	10	11	14	12	14	12	14
Layer 2									

Table 7.3: *Network structure of each RNN member of the best ensemble for German and for French F_0 control shown in Fig 7.6. For each ensemble member, the number of input factors and the number of nodes of the first and of an optional second hidden layer is given.*

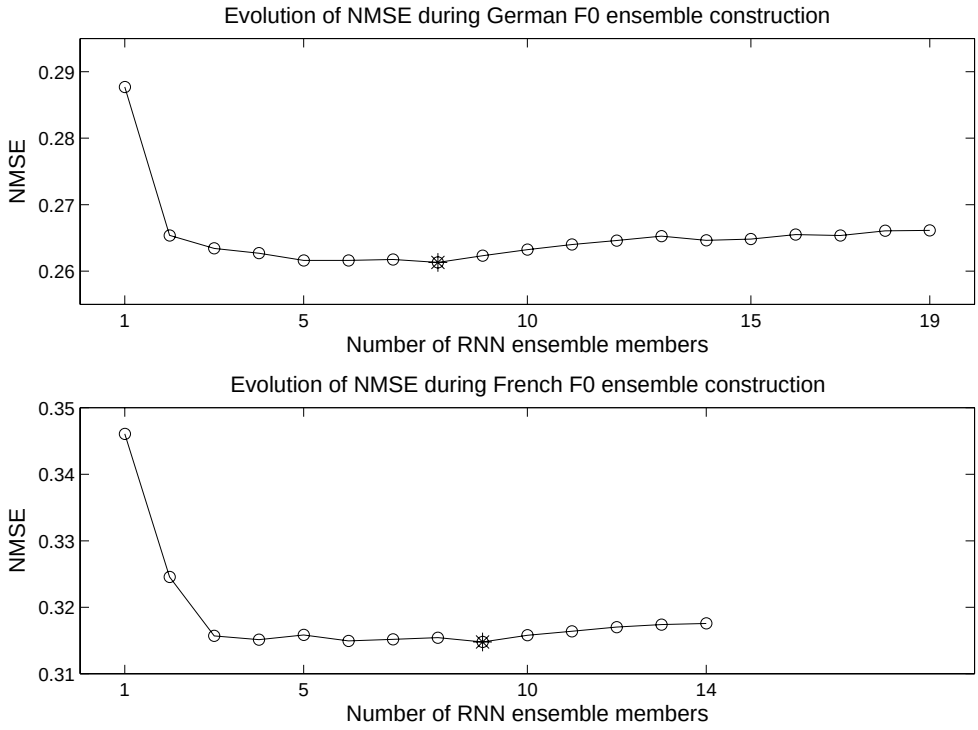

Figure 7.6: *Normalized mean squared error as a function of the number of ensemble members during F_0 ensemble construction. The ensemble model with lowest prediction error is indicated by a star. The best ensemble for German F_0 prediction consists of 8 RNNs. The best French F_0 ensemble model has 9 RNN members.*

7.3 Segment Duration Control

The polyglot segment duration control of the polySVOX system generates for each phone and for each pause of the phonological representation of a polyglot utterance the corresponding duration value. For each phone, it applies the appropriate monolingual duration model that corresponds to the language of the phone.

7.3.1 Model Architecture

Figure 7.7 shows an schematic overview of the polyglot segment duration control: it consists of a *factor encoding* module, that generates for each phone or pause of the phonological representation of a polyglot

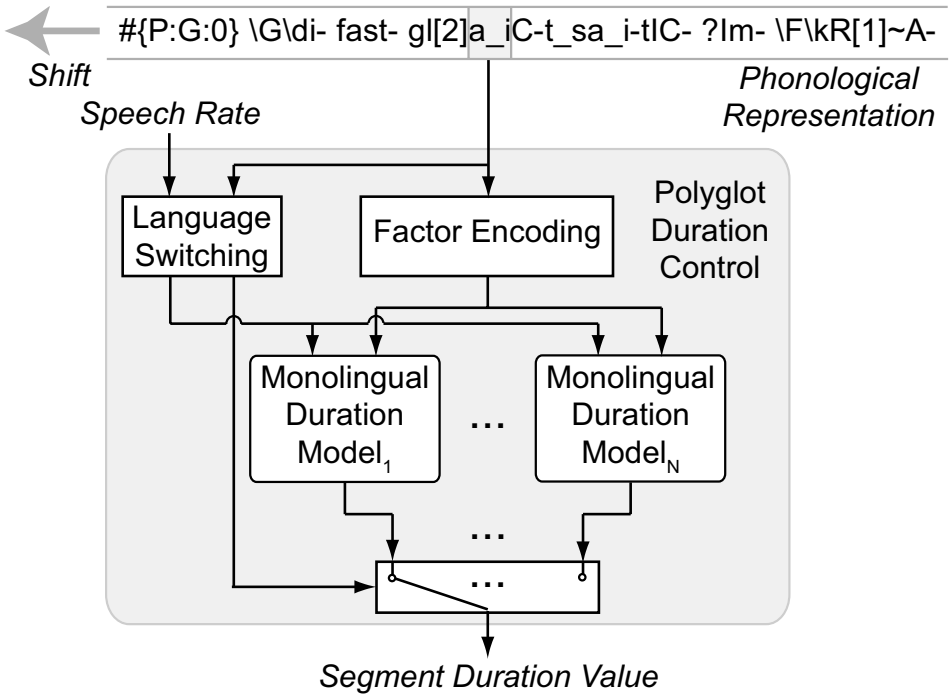

Figure 7.7: *Schematic representation of the polyglot duration model: for each phone or pause of the phonological representation, the corresponding input factors are extracted and encoded. The language information of the phonological representation selects the corresponding monolingual duration model to generate the segment duration output value. An additional input specifies the speech rate of the utterance.*

utterance a language-independent input factor representation, that is described in Section 7.3.4. A *language switching* component selects the appropriate model from a set of independent, monolingual segment duration models and sets the appropriate speech rate. The selected monolingual duration model finally generates the segment duration values. The normalization and encoding of the segment duration values is presented in Section 7.3.5.

Each monolingual duration model is a weighted ANN ensemble that is constructed using the procedure presented in Section 6.5. Figure 7.8 displays a schematic representation of such a monolingual duration model. Each ANN has its own input factor selection that chooses the optimal set of input factors for this network. The network setup of the ANN ensemble members of the German and of the French duration

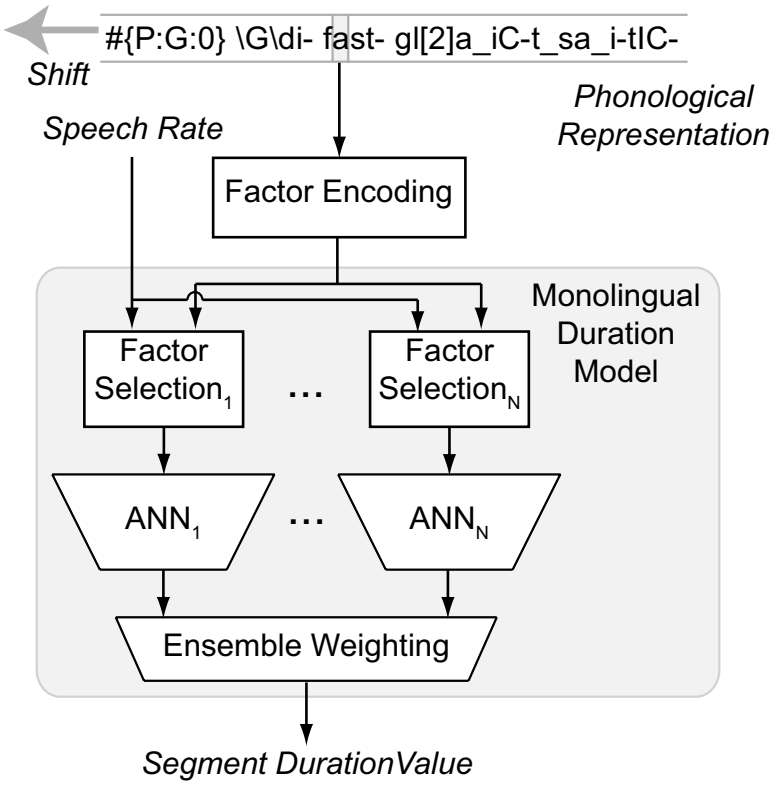

Figure 7.8: *Schematic representation of a monolingual duration model as a weighted ANN ensemble: each ANN member of the ensemble has its own input factor selection that chooses the optimal set of input factors for this network.*

models is given in Table 7.5.

7.3.2 Language Switching

Language switching within polyglot utterances must not result in audible rhythmic discontinuities. This requires that the general speech rates of both language specific duration models are similar. This could be achieved by recording prosody corpora of each language with similar, relatively constant speech rates having small variances. The speech rate of the German male prosody corpus displayed in Figure 7.9, e.g., exhibits such a "constant" speech rate with small variance (at least for longer utterances). However, as also visible in Figure 7.9, the variances of speech rate of the German and of the French female prosody corpora are considerable and much larger than of the male corpus. In first experiments, switching between a German and a French duration model trained on these two corpora resulted therefore most of the time in an audible change of speech rate.

In order to cope with the large variances in speech rate, the speech rate and the number of syllables of a sentence were provided as additional input factors to the ANNs. Doing so, also improved duration prediction performance of the individual networks considerably. And, the additional speech rate input made it possible to smoothly switch between the individual, monolingual duration models, simply by setting the same speech rate value as input for both duration models.

The speech rate input also makes it possible to speed up or slow down the tempo of synthesized speech by naturally lengthening or shortening the predicted duration values (at least for speech rates within the range of speech rate values of the prosody corpus). This is a nice improvement compared to state-of-the-art duration models with an implicit speech rate, where changes in tempo are done by a multiplication of the predicted duration values with a given factor.

7.3.3 Speech Pauses

Pauses within the speech signal can be classified into *intra-segmental pauses* that occur in connection with a phone, like preplosive pauses or glottal closures, and *inter-lexical pauses* or simply *speech pauses* that appear between words, cf. [Zel94]. Inter-lexical pauses can further be

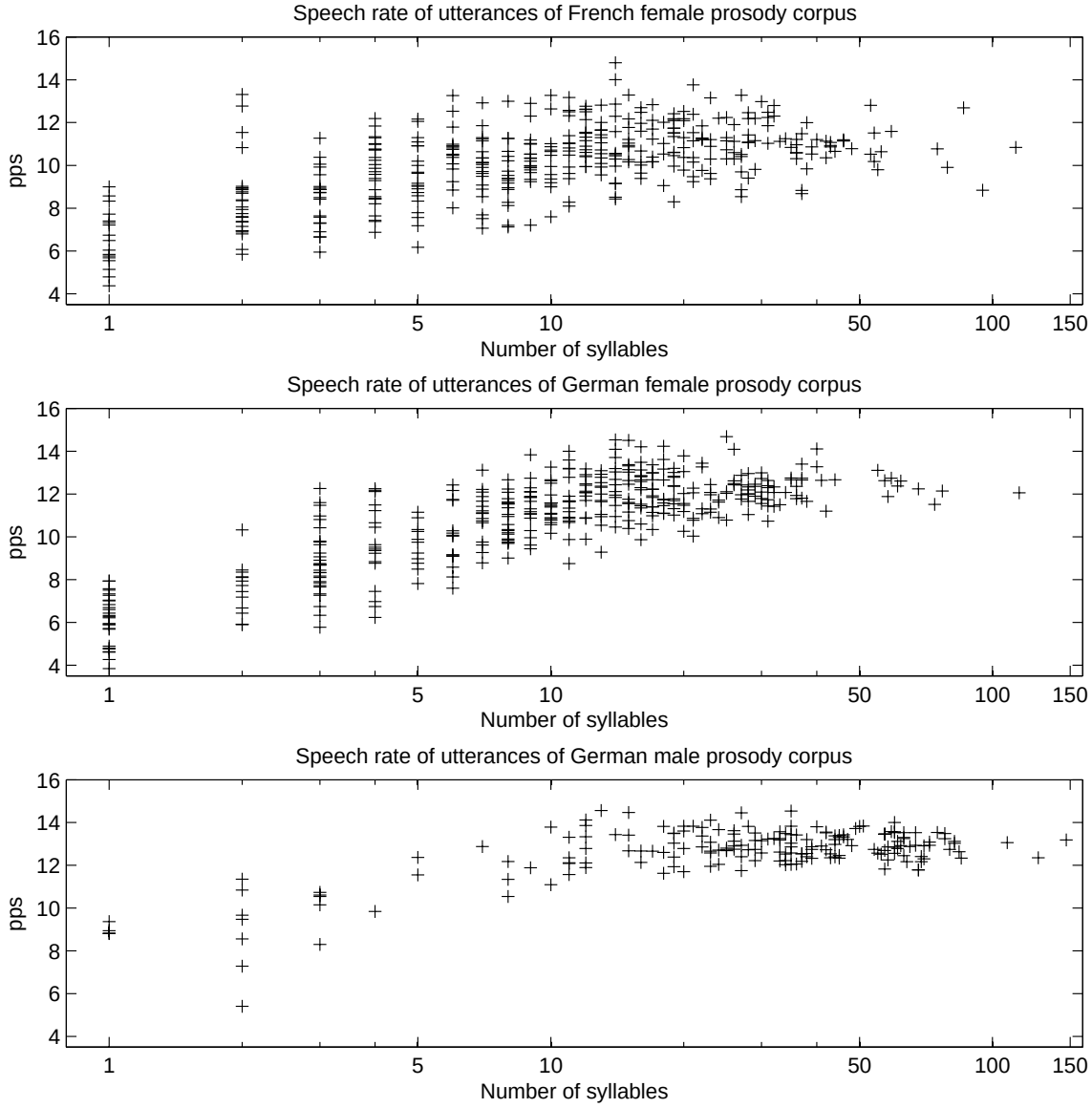

Figure 7.9: *Speech rate of the utterances of the French (top) and the German (center) female prosody corpora, and of the German male (bottom) prosody corpus as a function of the number of syllables of an utterance. The speech rate is calculated as average number of phones per second (pps). The number of syllables is given in logarithmic scale.*

subclassified into *end-pauses*, that mark the end of an utterance, and *non-end-pauses*, which occur inside utterances, cf. [Hub91].

The duration of an intra-segmental pause is influenced by the same

factors as the durations of phones. Duration control therefore treats preplosive pauses and glottal closures like standard phone segments.

The duration of an inter-lexical pause, however, depends on other factors. Beside of pragmatic, semantic, and rhythmic factors also the breathing requirements of the speaker must be considered. As the prosody corpora used here contain only utterances of single sentences and do not contain any longer paragraphs, no data for end-pauses was available. Also, similar to the investigations described in [Hub91] or in [Rie98], no relevant factors influencing the duration of non-end-pauses were found, beside of having the speaker breathing or not. Therefore, as it was also done in [Rie98], non-end-pauses were further subclassified into pauses with and without the speaker breathing. These are referred to as *breath-pauses* and *non-breath-pauses*, resp.

As there are no relevant factors, end-pauses and non-end-pauses are normally assigned a constant value, as it was done, e.g., in the duration models of Klatt [Kla79], of Huber [Hub91], or of Riedi [Rie98]. The polyglot duration control uses the same speech pause duration values for all languages. End-pauses and breath-pauses are assigned a constant duration value of 320 ms, which is roughly the mean value of all breath-pauses found in the German and in the French prosody corpora. Non-breath pauses are set to a constant duration value of 90 ms, again roughly the mean of all non-breath-pauses found in both corpora. The assignment of pause types to the phrase boundaries of the phonological representation is done similar to the procedure described in [Rie98].

7.3.4 Input Representation

From the phonological representation of an utterance, a sequence of phone and pause segments is extracted. The hold (preplosive pause) and the burst part of plosives are hereby treated as two separate segments. For plosives after a speech pause, no preplosive pause is extracted. Diphthongs, triphthongs, and affricates are each treated as one segment. Each of these segments is represented by a vector of 349 elements. Similar to F_0 control, all elements of this vector are set to zero by default. The values of ordinal factors are directly set in the vector. For categorical factors, a 1-out-of-n encoding is applied such, that each categorical factor is represented by n binary input factors.

Segment duration depends on a relatively local segmental context as

far as segment type information is concerned, as shown, e.g., in [Rie98]. The influence of accentuation and phrasing information, however, is wider. Similar to F_0 modeling, the correct size of the contexts for the different factors is unknown and depends on the prosodic phenomena to be modeled. Therefore a context of *2 preceding and 2 subsequent segments* is applied for segmental information. For accentuation and phrasing information, a context of *2 preceding and 2 subsequent syllables* is used. Starting with this large, initial factor set, the ensemble construction procedure of Section 6.5 is applied to automatically select the most relevant input factors and thereby the optimal context size of each factor.

For polyglot duration control, this input representation must be language-independent. Thus, no language specific segment types or phrase types are used, but the language-independent description of manner and place of articulation of phones of the IPA and a basic, language-independent set of phrase types, as given in Section 3.2.1. Also, information about syllable language or about language switching position may not be included in the factor set. Language information is only used to switch between the individual monolingual duration models.

Beside of segmental factors and factors describing accentuation and phrasing, additional factors of the syllable, foot, phrase, and sentence level have been selected. For each of these categories, in the following all factors and their values are described.

Segmental Factors

These factors describe the characteristics of a segment. They consist of a gross specification of the segment type and a detailed characterization of the articulation of vowels and of consonants according to the IPA specification given in Table 7.1 and in Table 7.2, resp. Additional information concerns lengthening, voicing, and the first formant position of the segment, and whether the segment is part of the syllable nucleus.

Segmental factor values	
segment type	vowel, gliding vowel, triphthong, consonant, affricate, glottal closure preplosive pause, speech pause
vowel characterization	tuple from Table 7.1
consonant characterization	tuple from Table 7.2
long segment	binary
voiced segment	binary
syllabic segment	binary
first formant position	low, middle, high

For the segmental factors, a context of 2 preceding and 2 subsequent segments is applied. This results in 200 input factors $((2 + 1 + 2) * 40)$ in total.

Accentuation, Phrasing, and Syllable Length Factors

These factors describe accentuation and phrasing information as well as the length of the syllable the segment is part of. The factors include the syllable stress level, the type of phrase boundary after the syllable, and the phrase type of the current phrase. Two additional factors describe the length of short and of long syllables separately, as in the German prosody corpus, a dependency of segment durations on syllable length is only visible in syllables with more than four phones.

Accentuation, phrasing, and syllable length factor values	
syllable stress	`[E]`, `[1]`, `[2]`, `[3]`, `[4]`, unstressed
phrase boundary	`0`, `1`, `2`, no boundary
phrase type	`P`, `T`, `S`, `Y`, `E`, `YC`, `F`
short syllable length	number of phones (if number < 5)
long syllable length	number of phones (if number > 4)

For accentuation, phrasing, and syllable length factors, a context of 2 preceding and 2 subsequent syllables is applied. For the current syllable and the four context syllables, 95 input factors $((2 + 1 + 2) * 19)$ are used in total.

Syllable Level Factors

These factors describe the position of the segment within the syllable and whether the segment is in the onset, the nucleus, or the coda of the syllable. No context is applied. Therefore, 5 input factors are used in total.

Syllable level factor values	
first phone in syllable	binary
phone position in syllable	phone number
syllable structure	onset, nucleus, coda

Foot Level Factors

A foot basically consists of one salient syllable and all non-salient syllables to the right (*left-headed foot*) or to the left (*right-headed foot*) until the next salient syllable or until a sentence or phrase boundary. A salient syllable is an accented syllable that carries at least word main stress, i.e., one of the accentuation levels `[E]`, `[1]`, `[2]`, or `[3]` in the phonological representation. Left-headed foots starting and right-headed foots ending at sentence or phrase boundaries may also have no salient syllable.

German speech rhythm is said to depend on left-headed foots, while French speech rhythm should base on right-headed foots. Thus, input factors for both foot types on the foot, the phrase, and the sentence level are included. As no context is applied, 26 input factors are used in total.

Foot level factor values	
syllable position in L-headed foot	salient syllable, 1., 2., 3., 4., 5., 6., 7., 8., >8. non-salient syllable
syllable position in R-headed foot	salient syllable, 1., 2., 3., 4., 5., 6., 7., 8., >8. non-salient syllable
length of L-headed foot	one syllable, short, long
length of R-headed foot	one syllable, short, long

Phrase Level Factors

These factors describe the phrase length in number of syllables and in number of foots, and the foot position within the phrase. No context is applied. In total, 7 input factors are used.

Phrase level factor values	
first L-headed foot	binary
L-headed foot position	foot number
first R-headed foot	binary
R-headed foot position	foot number
phrase length	number of syllables
phrase length	number of L-headed foots
phrase length	number of R-headed foots

Sentence Level Factors

On the sentence level, the input factors describe sentence length in number of syllables and in number of foots, and the foot position within a sentence. Additionally, the speech rate specified as phones per second averaged over the complete utterance is selected. In total, as no context is applied, 16 input factors are used.

Sentence level factor values	
L-headed foot position	sentence initial, sentence final, phrase initial, phrase central, phrase final, phrase with one foot
R-headed foot position	sentence initial, sentence final, phrase initial, phrase central, phrase final, phrase with one foot
sentence length	number of syllables
sentence length	number of L-headed foots
sentence length	number of R-headed foots
sentence speech rate	phones per second [pps]

The relevance of all 349 input factors $(200 + 95 + 5 + 26 + 7 + 16)$ was determined for German and for French segment duration modeling separately by applying the Unit-OBS procedure described in Section 6.4. Appendix C provides the complete relevance ranking of all input factors for German and for French duration modeling.

7.3.5 Output Representation

Neural networks, that are trained using the sum-of-squares error measure, achieve lowest prediction error if the target data is normally distributed, cf. [Bis95]. Therefore, the typical log-normal-like distribution of segment durations, as shown on the left side of Figure 7.10, is first

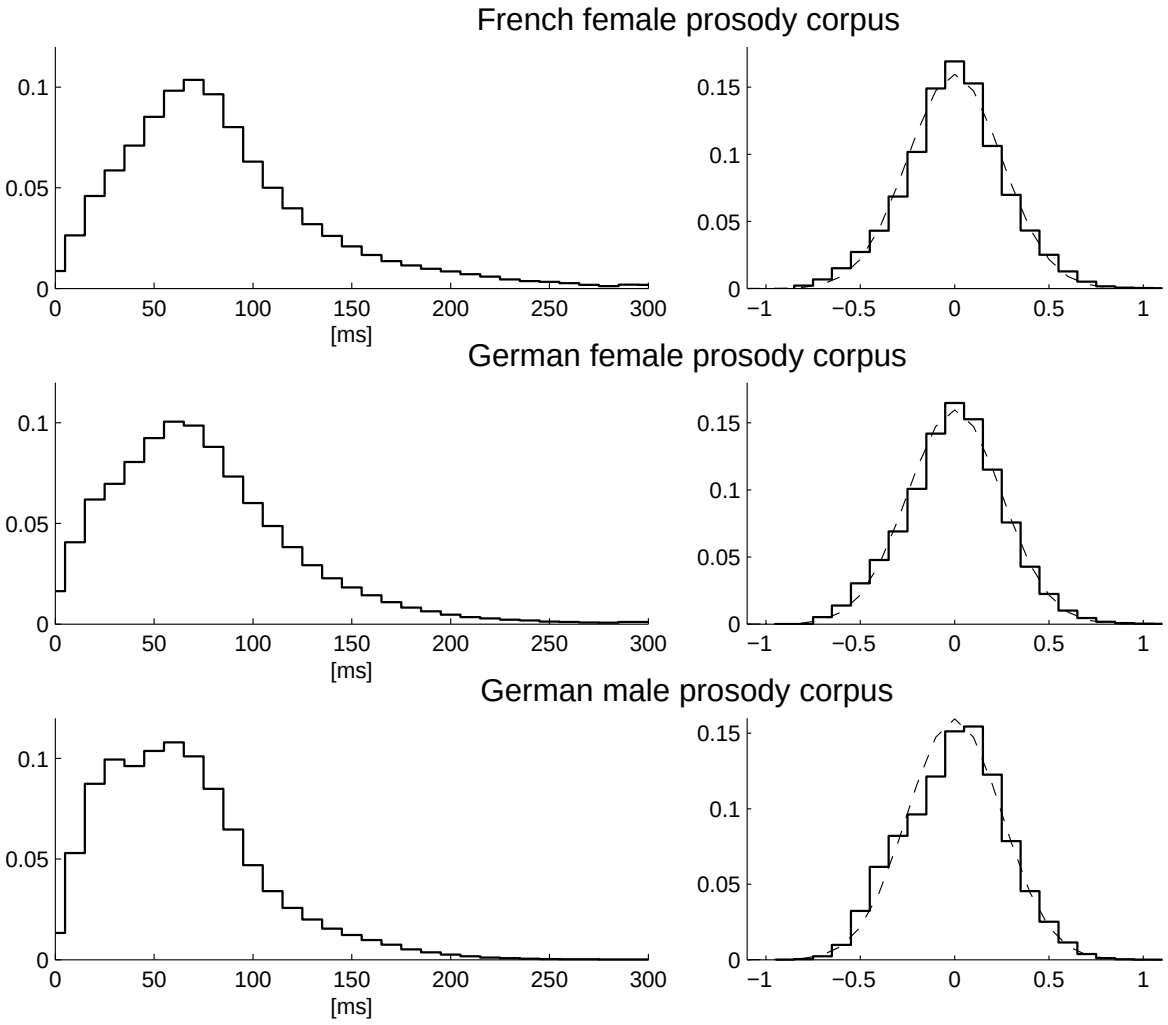

Figure 7.10: *Normalized histograms of segment durations of the German and the French female, and the German male prosody corpora. On the left side, the original segment durations are shown. On the right side, the transformed and normalized segment durations are displayed. The dashed line additionally indicates the corresponding normal distribution.*

	λ	μ	s
French female	0.3208	9.326	10.029
German female	0.3881	10.605	13.356
German male	0.3030	7.922	8.712

Table 7.4: *Parameters of segment duration transformation and normalization for the German and the French female, and the German male prosody corpus.*

transformed into a normal distribution using the Box-Cox transformation, cf. [BC64], and then *linearly normalized* to zero mean and a standard deviation of 0.25 using

$$\hat{y} = \frac{(\frac{y^{\lambda}-1}{\lambda} - \mu)}{s} \ , \tag{7.1}$$

with y as original and $\hat{y}$ as transformed and normalized segment duration value. Appropriate values for the parameters λ, μ, and s are given in Table 7.4 for all three prosody corpora.

After applying the transformation (7.1), all target values have zero mean and are within the range of -1 to 1, as it is displayed on the right side of Figure 7.10. Doing so, speeds up neural network training and optimizes prediction error, as experiments with different distribution transformations have verified. Also, it is possible to compare *linear* and *tanh* output activation functions using the same target data. Similar to F_0 modeling, a linear output activation function resulted in a lower prediction error.

7.3.6 Duration Ensemble Construction

Factor Relevance Determination

The first steps of duration ensemble construction are the determination of factor relevance and the training of individual networks having different network configurations and different number of input factors. Figure 7.11 shows the NMSE on the test set as a function of the number of input factors for German and for French duration ensemble models. For this comparison, all members of an ensemble use the same number

of input factors. The sequence of input factors starts with the most relevant to the left, and is *different between German and French*. The best German duration models were found with 110 to 118 input factors, as shown in Figure 7.11. The best French duration models have about 55 to 70 input factors and provide in general a lower prediction error than the German models.

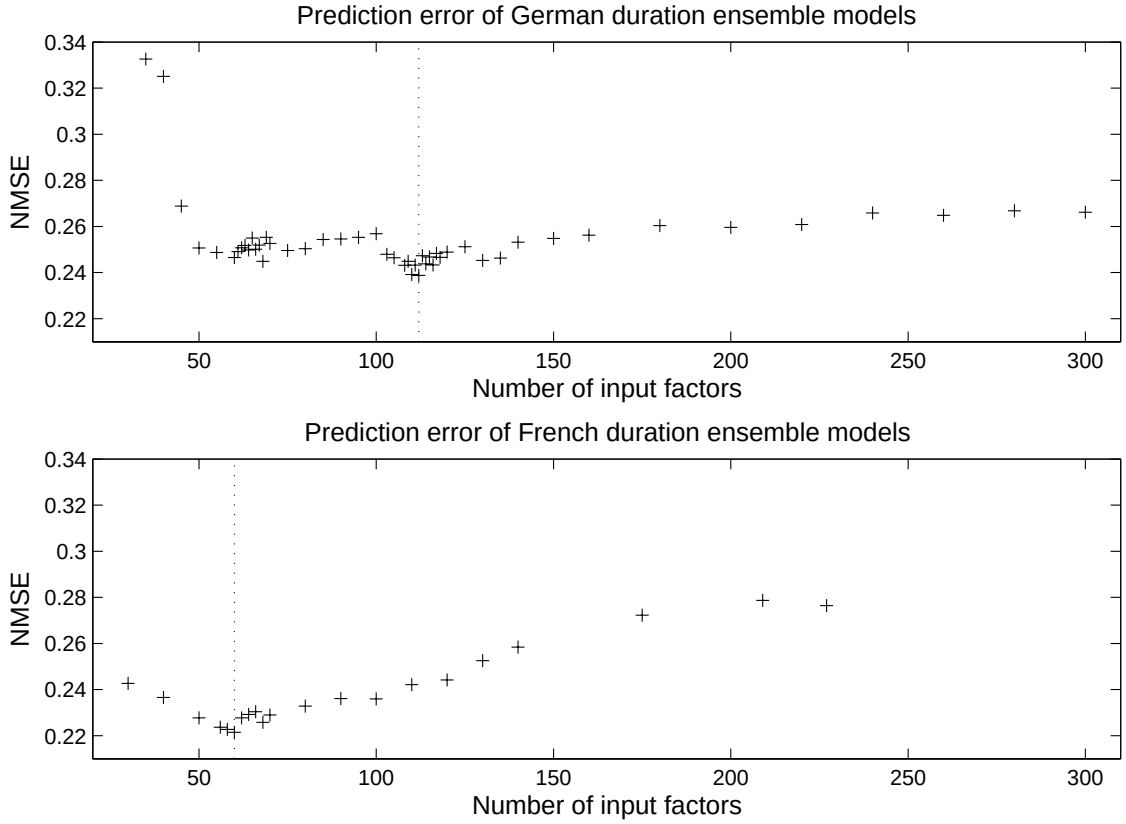

Figure 7.11: *NMSE of duration prediction models as a function of the number of input factors. The upper diagram displays prediction errors of German duration models, the lower one of French duration models. The sequence of input factors starts with the most relevant to the left, and is* different between German and French*. For each number of input factors, the prediction error of the best ensemble model is indicated by a cross. The best German duration prediction ensemble has 112 input factors. The best ensemble for French duration prediction uses 60 input factors. Both are indicated by a dotted line. For both languages, the number of ensemble member varies between 5 and 28.*

Similar to F_0 prediction, cp. Section 7.2.5, it is clearly visible in both diagrams, that the use of too few or of too many input factors generally results in higher prediction errors. In order to look up, which input factors are used by the individual models, Appendix C provides a list of all input factors used for duration control together with their rank of relevance.

Network Aggregation

In the final step of duration ensemble construction, the best ensemble of all individual networks is determined for increasing ensemble size. The evolution of NMSE during ensemble construction for German and for French duration modeling is presented in Figure 7.12.

The best German duration ensemble with individual input factor sets consists of 55 ANNs. The best French duration ensemble has 8 members. The network structures and the input factor sizes of the first ten German networks and of all French networks are given in Table 7.5. The best German duration ensemble has a NMSE of 0.2310 and slightly improves the best German duration ensemble with a fixed number of input factors by about 3%. The best French duration ensemble achieves

German duration ensemble										
Network Nr.	1	2	3	4	5	6	7	8	9	10
Factors	55	60	60	61	55	61	67	66	108	114
Layer 1	30	25	30	25	30	23	50	50	50	50
Layer 2						5				

French duration ensemble										
Network Nr.	1	2	3	4	5	6	7	8		
Factors	80	70	60	60	58	56	68	60		
Layer 1	20	15	16	20	16	20	14	17		
Layer 2		15	20	10	20	10	20	15		

Table 7.5: *Network structure of each ANN member of the best ensemble for German and for French duration control shown in Fig 7.12. For each ensemble member, the number of input factors and the number of nodes of the first and of an optional second hidden layer is given.*

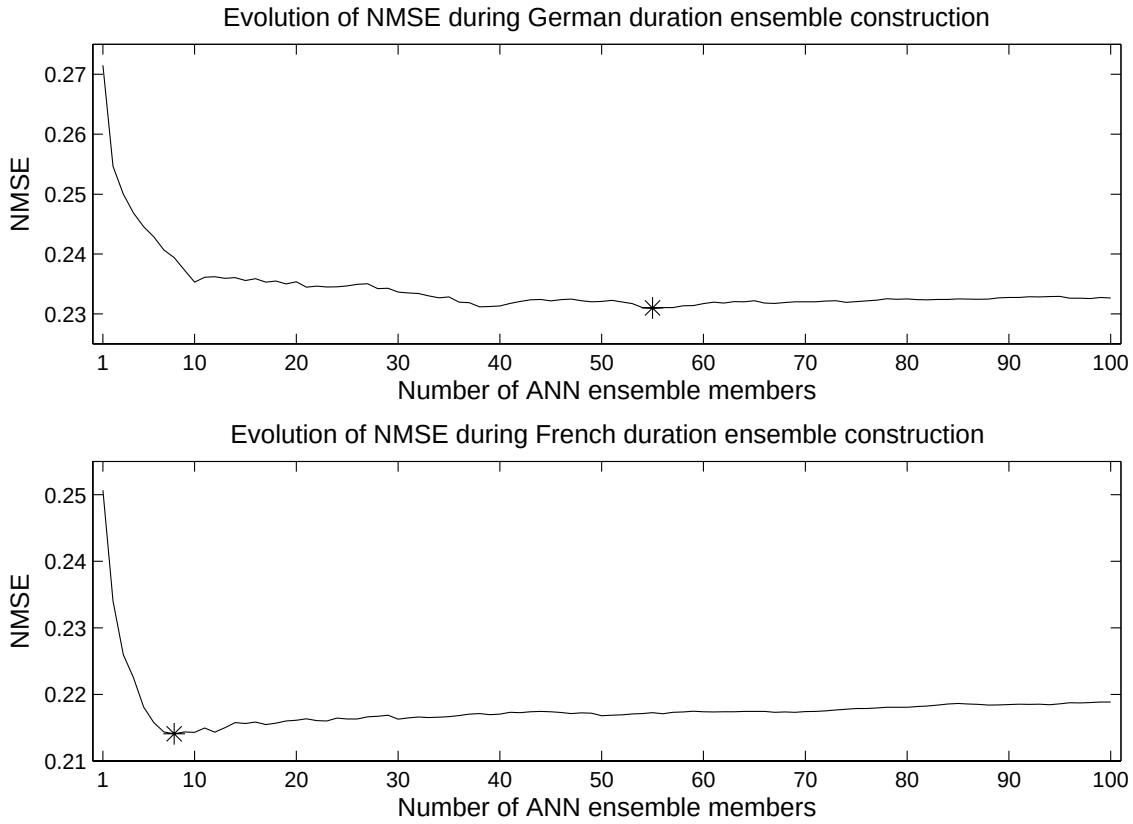

Figure 7.12: *Normalized mean squared error as a function of the number of ensemble members during duration ensemble construction. The ensemble model with lowest prediction error is indicated by a star. The best ensemble for German duration prediction consists of 55 ANNs. The best French duration ensemble model has 8 ANN members.*

a NMSE of 0.2141, which is also a slight improvement of about 3%.

7.4 Experiments and Discussion

7.4.1 Comparison with RNN- and MARS-based Prosody Models

For a comparison of the performance of ensemble prosody models with the performance of the RNN-based F_0 model (cf. [Tra92, Tra95]) and of the MARS-based duration model (see [Rie98]) that are currently applied in the German SVOX system, F_0 and duration ensemble models were trained on the same German male prosody corpus, cf. Section 5.2, as the reference models. For training, the identical setup of the training set and of the test set was used as reported for the reference models. Figure 7.13 shows the prediction error of F_0 and of duration ensemble models as a function of the number of input factors.

The best ANN-based ensemble model for duration prediction with 11 ANN members, each of them having 160 input factors, has a NMSE of 0.1866. This is about 12% improvement compared to the NMSE of 0.2108 of the best MARS-based duration model and about 16% improvement compared to the best ANN-based duration model, both presented in [Rie98]. The best RNN-based ensemble model for F_0 prediction consists of 4 RNN members, each using 356 input factors. This model has a NMSE of 0.3457 that corresponds to the root of the mean squared error (RMS) of 7.05 Hz. This is about 24% improvement compared to the RMS of 9.2 Hz of the best RNN-based F_0 model reported in [Tra95]. The RMS of the ensemble is now also below an "upper bound" of very good acoustic quality reported in [Tra92], where he states "that for our corpus sentences, a RMS prediction error of less than 8 Hz for a single sentence corresponds to a very good acoustic quality."

7.4.2 Perceptual Evaluation

The NMSE measure used to optimize F_0 and duration control of the polyglot TTS system not necessarily correlates with the subjective quality judgment of a human. Also, the F_0 and the duration models were optimized independently of each other, and if applied together in the TTS system, their deficiencies still may cumulate and thus impair the quality of the synthetic speech. In order to evaluate the quality of the complete polyglot prosody control, a small listening experiment was

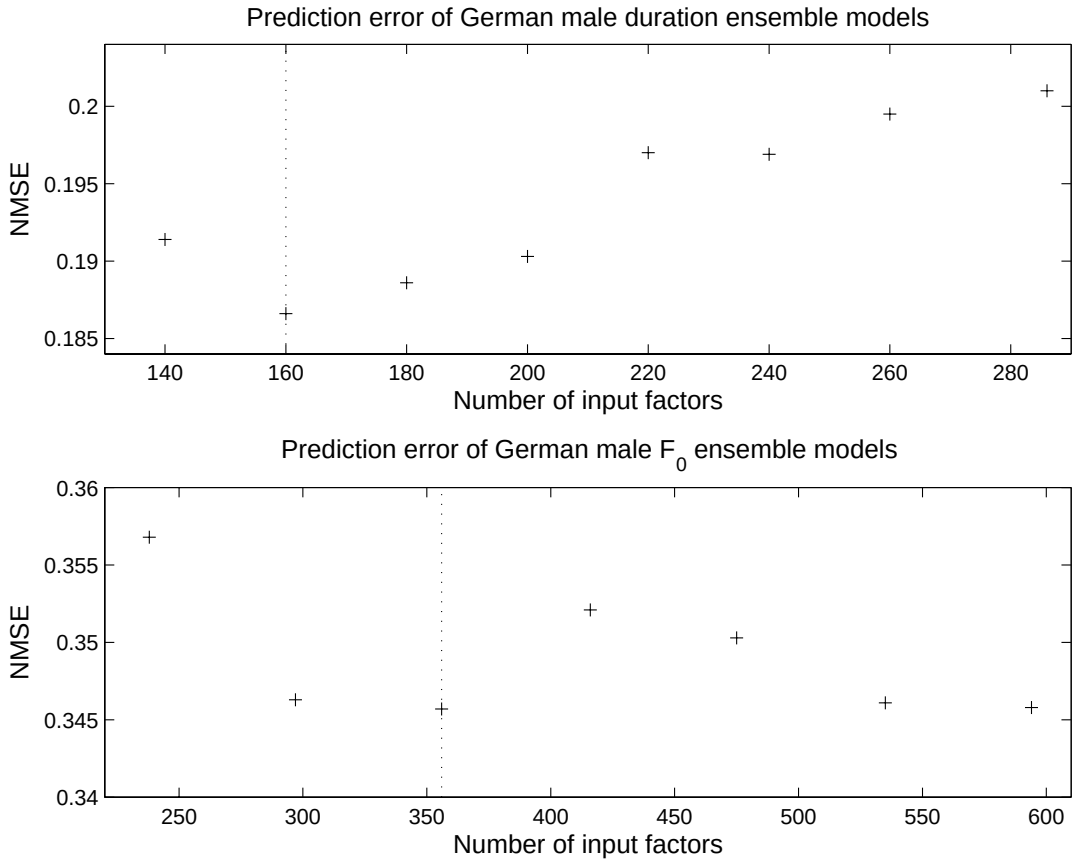

Figure 7.13: *NMSE of duration and of F_0 prediction models for the monolingual German male prosody corpus as a function of the number of input factors. The upper diagram displays the prediction errors of duration models, the lower one of F_0 models. The sequence of input factors starts with the most relevant to the left. The best ensemble for duration prediction uses 160 input factors and has 15 ANN members. The best ensemble for F_0 prediction has 356 input factors and consists of 4 RNN members.*

conducted with naive subjects and with members of the speech processing group.

Experiment Setup

The subjects were presented a total of 160 sentences. These sentences consisted of 40 German and 40 French sentences, each one with its

natural prosody and with the synthetic prosody predicted by the polyglot prosody control. The sentences were presented in random order. It was made sure that the natural and the synthetic versions of a sentence did not directly follow each other. The subjects had to indicate, whether they believe to have heard a sentence with synthetic or with natural prosody. All subjects are German native speakers with a good knowledge of French. Two of the subjects speak fluently French and German.

20 of the German sentences and 8 of the French sentences were taken from the polyglot test set. These mixed-lingual sentences contained either English, French or German foreign inclusions. The other sentences were taken from the German and from the French monolingual test sets (see Appendix D).

Prosody prediction started from the manual phonological transcriptions of the natural sentences, cf. Section 5.4. Doing so, the experiment tested prosody prediction alone, without possible errors in the phonological transcription that may be produced by text analysis. The predicted F_0 contour and the segment duration values of a sentence were transferred onto the natural speech signal using a TD-PSOLA-based analysis-synthesis procedure implemented by the author. Further details on this procedure can be found, e.g., in [MC90].

The subjects were instructed to especially pay attention to sentence melody and speech rhythm. Before starting the experiment, they listened to some sentences of the same type as the ones included in the experiments, in order to become familiar with the speaker's voice.

The experiment was carried out in one session. A simple computer program enabled the subjects to start the play back of a stimulus by pressing a button and to decide afterwards whether the heard sentence has natural prosody or not. Then, the program proceeded to the next sentence. No repetition of a sentence was possible. The stimuli were played back over a loud-speaker in a middle-sized room.

Results and Discussion

Table 7.6 shows the confusion matrices and the "recognition rates", i.e., the percentage of correct answers, of the individual subjects for all German and for all French sentences separately. Table 7.7 shows the results if only the mixed-lingual sentences of each language are consid-

ered. Similar to the listening experiments for the evaluation of German F_0 prediction described in [Tra95] and for the evaluation of German duration prediction presented in [Rie98], the experiment was designed such that completely random answers would result in a recognition rate of 50%.

In contrast to both former experiments, where the prediction of only one acoustic parameter was tested, the current experiment tested the prediction of F_0 *contour and segment duration values simultaneously*, which is a much more difficult task for prosody prediction, as already an error in only one of the predicted acoustic parameters will reveal the prosody contour as synthetic.

Following the interpretation in [Tra95], the best recognition rate of all German sentences of 56.25% suggests that about $2 * (100\% -$

Subject	Naive		German		French		Recogn. Rate [%]	
			nat.	synt.	nat.	synt.	German	French
1	no	nat.	28	12	26	14	46.25	43.75
		synt.	31	9	31	9		
2	no	nat.	22	18	23	17	42.5	50.0
		synt.	28	12	23	17		
3	no	nat.	30	10	28	12	51.25	48.75
		synt.	29	11	29	11		
4	no	nat.	27	13	21	19	50.0	42.5
		synt.	27	13	27	13		
5	no	nat.	26	14	20	20	56.25	48.75
		synt.	21	19	21	19		
6	yes	nat.	25	15	25	15	42.5	51.25
		synt.	31	9	24	16		
7	yes	nat.	27	13	26	14	48.75	53.75
		synt.	28	12	23	17		
total		nat.	185	95	169	111	48.21	48.39
		synt.	195	85	178	102		

Table 7.6: *Confusion matrices and recognition rates of subjects in the experiment for the evaluation of the quality of combined F_0 and duration prediction for all German and all French sentences (mono- and mixed-lingual). Random answers correspond to a recognition rate of 50%.*

56.25%) = 87.5% of all (monolingual and polyglot) synthetic German prosody contours sound natural if presented in isolation. The best recognition rate of all French sentences of 53.75% indicates that about 92.5% of all synthetic French prosody contours sound natural if presented in isolation. Considering only the mixed-lingual sentences, this is the case for about 80% of the mixed-lingual German sentences and for about 87.5% of the mixed-lingual French sentences. These results compare well with about 30% natural sounding, synthetic German F_0 contours predicted by the best ANN-based F_0 model reported in [Tra95] and with about 70% natural sounding, synthetic German duration sequences predicted by the best MARS-based model presented in [Rie98]. In those experiments, only one prosodic parameter (F_0 or segment duration) was predicted.

An interesting detail in Table 7.6 is that in total 69.6% of all synthetic and only 66.1% of all natural German prosody contours were

Subject	Naive		German		French		Recogn. Rate [%]	
			nat.	synt.	nat.	synt.	German	French
1	no	nat.	16	4	7	1	60.0	56.25
		synt.	12	8	6	2		
2	no	nat.	8	12	3	5	40.0	37.5
		synt.	12	8	5	3		
3	no	nat.	12	8	6	2	45.0	43.75
		synt.	14	6	7	1		
4	no	nat.	14	6	4	4	52.5	50.0
		synt.	13	7	4	4		
5	no	nat.	9	11	2	6	55.0	56.25
		synt.	7	13	1	7		
6	yes	nat.	12	8	3	5	42.5	43.75
		synt.	15	5	4	4		
7	yes	nat.	14	6	3	5	50.0	43.75
		synt.	13	7	4	4		
total		nat.	84	56	28	28	49.29	47.32
		synt.	86	54	31	25		

Table 7.7: *Confusion matrices and recognition rates in the experiment for the mixed-lingual German and the mixed-lingual French sentences alone.*

judged to sound natural. Similar, 63.6% of all synthetic and only 60.4% of all natural French prosody contours were judged as sounding natural. These numbers would indicate that the synthetic prosody contours predicted by the polyglot prosody control would sound in general more pleasant to humans than the natural prosody contours. However, the differences between the rates of synthetic and of natural prosody contours are statistically not significant.

These results show that it is possible to simultaneously produce natural sounding F_0 contours and segment duration values from abstract phonological information by means of the weighted RNN- and ANN-based ensembles. The synthetic prosody is nearly indistinguishable from natural prosody.

The results for the mixed-lingual sentences alone indicate that the polyglot prosody control presented in this chapter is able to produce natural sounding, polyglot prosody contours for sentences with foreign inclusions. They also indicate that it is possible to switch between monolingual prosody models at language boundaries without audible rhythmic or melodic discontinuities.

Chapter 8

Conclusions

8.1 Discussion

The first aim of this thesis was the creation of a modular TTS system architecture, that can be configured with an arbitrary number of independent, monolingual resources in order to form a polyglot TTS synthesis system. This aim has been achieved by building a linguistically motivated system architecture and applying general approaches to the different, language specific problems occurring in TTS synthesis. This architecture thereby strictly separates language-independent algorithms from language-dependent linguistic and acoustic data. Furthermore, following the linguistic view adopted as a basis for the ETH TTS project, a voice-independent part is separated from a voice-dependent part. The resulting polyglot TTS synthesis system, polySVOX, presented in this thesis consists of relatively simple, language-independent building blocks, that can each be configured with an arbitrary set of monolingual resources. The polySVOX system is, to the author's knowledge, the worldwide first "true" polyglot TTS synthesis system that comprises a complete mixed-lingual transcription stage and a complete polyglot phono-acoustical model.

The architecture of the individual building blocks bases on the architecture of the monolingual German TTS system SVOX presented in [Tra95]. Thus, also the second goal of this thesis, the reuse of the for-

malisms and algorithms of the SVOX system, was reached. However, morphological and syntactic analysis needed a complete redesign in order to support mixed-lingual input text. The formalisms of morpho-syntactic analysis were thereby taken from the SVOX system or extended, when necessary. Phonological processing and prosody control are completely new. Here, the formalisms of sentence accentuation were reused. Speech signal generation needed only minor modifications to support polyglot acoustic resources. The individual building blocks will be briefly discussed in the following sections.

Mixed-lingual Text Analysis

Mixed-lingual text analysis, presented in Chapter 2, provides a very detailed morphological and syntactic structure determination of mixed-lingual words and sentences. This includes the disambiguation of inter-lingual homographs, the grapheme-to-phoneme conversion of unknown words, of numbers, and of abbreviations in mixed-lingual sentences, and the identification of word and sentence boundaries, without extending the formalism and without adding new parsing methods. Whereas other systems apply an explicit preprocessing stage with, e.g., a separate language identification procedure, a separate word and sentence boundary identification procedure, or separate conversion procedures for grapheme-to-phoneme mapping problems, the polySVOX system integrates the *entire* text analysis within morpho-syntactic analysis.

The polySVOX text analysis was the worldwide first mixed-lingual text analysis applied in text-to-speech synthesis, cf. [BL04]. With a language identification rate of words of about 97.7%, this rule-based approach outperforms statistical approaches that are specially designed for the language identification task. E.g., [THRJ02] report for their neural network based algorithm a language identification rate of words of 86.6%. Very important for mixed-lingual text analysis is, however, the correct identification of foreign inclusions within mixed-lingual words, which is impossible for current statistical language identification approaches.

This mixed-lingual text analysis was tested with lexica and grammars that contain currently about 12 000 German, 6 500 English, 5 000 French, and 3 500 Italian lexicon entries, and about 1 900 German, 1 200 English, 1 300 French, and 800 Italian grammar rules. These

monolingual linguistic resources are completely independent. For each language, only about 20 grammar rules, so-called "inclusion grammar rules", are necessary to specify foreign inclusions of another language. Currently, about 240 inclusion grammar rules are used in total. Thus, the integration of linguistic resources of a new language, the third aim of this thesis, is very simple.

Mixed-lingual Phonological Processing

Mixed-lingual phonological processing, as described in Chapter 3, is strongly based on works in generative phonology. Mixed-lingual prosodic phrasing uses phonological words as minimal phrasing entities and builds larger phrases from smaller temporary phrases according to language-dependent constraints. These constraints comprise the syntactic structure of a sentence, the minimum length of an independent phrase, based on accent and syllable counts, and the balanced lengths of all independent phrases of a sentence. Mixed-lingual sentence accentuation is build on the following language-dependent principles: the nuclear stress rule, the rhythmic stress shift rule, and the accentual bipolarisation principle together with a dominance principle. The application of phonological transformations is based on language-dependent multi-context rules that specify phonological transformations within a syntactically constrained part of a sentence. The formalisms of the rules and patterns applied in phonological processing use the language information provided by morpho-syntactic analysis. This allows an easy integration of additional rules and patterns for a new language.

Polyglot Prosody Control

Polyglot prosody control described in Chapters 6 and 7 generates from the phonological representation of an utterance the corresponding fundamental frequency contour and segment duration sequence. For F_0 generation, a polyglot F_0 control is applied that comprises independent monolingual F_0 models. And for segment duration generation, a polyglot duration control is used that also contains independent monolingual duration models. All monolingual models use weighted ensembles of neural networks with an optimized set of input factors. This new approach to prosody control achieves impressive improvements even when

compared to the monolingual SVOX system that was regarded as one of the best TTS systems for German, cf. [Tra95]. For duration control, an improvement of the prediction error of about 12% compared to the best MARS-based duration model of [Rie98] was achieved, and for F_0 control, a prediction error improvement of about 24% compared to the best RNN-based F_0 model of [Tra95] was reached. These improvements made it possible, that in a perceptual evaluation about 90% of 80 different monolingual and mixed-lingual test sentences having synthetic prosody were judged indistinguishable from the corresponding original recordings with human prosody. These results also show that it is possible to switch between monolingual prosody models at language boundaries without audible rhythmic or melodic discontinuities.

Currently, prosody models for French and German are available. The integration of a prosody model for a new language is very simple and does not affect the existing prosody models. However, the prosody models must be trained on speech data recorded by the same speaker.

8.2 Outlook

The results achieved so far with the polySVOX TTS system are very convincing. But, of course, all parts of the system could be further improved. This includes the extension of the existing linguistic knowledge bases and the integration of additional linguistic resources for new languages.

The RNN- and ANN-based weighted ensemble models for prosody control provide impressive prediction results, even when optimized on a rather small prosody corpus. Optimizing them on larger prosody corpora could further improve the quality of prosody prediction.

The requirement of recording all monolingual prosody corpora by the same speaker restricts polyglot prosody control to a few languages that can be applied simultaneously. Therefore, a method to combine prosody models trained on speech data recorded from different speakers might be necessary in future.

Finally, a future application of the polySVOX system to a tone language could reveal the generality and the limitations of the language-independent architecture of the polySVOX system.

Appendix A

ASCII-Representation of IPA Symbols

The polySVOX system uses an ASCII representation - the ETH computer phonetic alphabet (ETHPA) - of the IPA (International Phonetic Association) phonetic symbols in order to process phonetic transcriptions. For readability reasons, the ETHPA symbols are defined to be as similar as possible to the IPA symbols. IPA symbols as well as ETHPA symbols can be put in strings. Such strings can unambiguously be split into phones again. These ETHPA forms occur in the main part of this thesis, and are also used in the phonetic sequence of lexical entries.

The following sections list the IPA symbols and the corresponding ETHPA symbols for English, French, German, and Italian. Each phone or diphthong is illustrated with some examples in graphemic and phonetic form, as given in the respective reference phonetic dictionary. As reference phonetic dictionaries, the author used for British and American English [JRHS03], for French [War96], for German [Dud05], and for Italian [Pon95].

A.1 English IPA Symbols

IPA	ETHPA	Example	
ə	`@`	another	[əˈnʌðə] [1]
ɚ	`@_r`	another	[əˈnʌðɚ] [2]
əʊ	`@_U`	nose	[ˈnəʊz] [1]
æ	`q`	hat	[ˈhæt]
ɑː	`A:`	cars	[ˈkʰɑːz] [1], [ˈkʰɑːrz] [2]
		pot	[ˈpʰɑːt] [2]
ɒ	`Q`	pot	[ˈpʰɒt] [1]
ʌ	`V`	cut, much	[ˈkʰʌt], [ˈmʌtʃ]
aɪ	`a_I`	buy	[ˈbaɪ]
aʊ	`a_U`	house	[ˈhaʊs]
b	`b`	bin	[ˈbɪn]
d	`d`	din	[ˈdɪn]
dʒ	`d_Z`	gin	[ˈdʒɪn]
ð	`D`	this, breathe	[ˈðɪs], [ˈbriːð]
e	`e`	pet	[ˈpʰet]
ɜː	`3:`	bird, furs	[ˈbɜːd], [ˈfɜːz] [1]
ɜˠː	`3_r:`	bird, furs	[ˈbɜˠːd], [ˈfɜˠːz] [2]
eə	`e_@`	there	[ˈtʰeə] [1]
eɪ	`e_I`	bay	[ˈbeɪ]
f	`f`	fat	[ˈfæt]
g	`g`	give, bag	[ˈgɪv], [ˈbæg]
h	`h`	hit	[ˈhɪt]
i	`i`	happy	[ˈhæpi]
iː	`i:`	ease	[ˈiːz]
ɪ	`I`	pit	[ˈpʰɪt]
ɪə	`I_@`	here	[ˈhɪə] [1]
j	`j`	youth, yes	[ˈjuːθ], [ˈjes]
k	`k`	skat	[ˈskɑːt]
kʰ	`k_h`	key	[ˈkʰiː]
l	`l`	life, whale	[ˈlaɪf], [ˈweɪl]
l̩	`=l`	bottle	[ˈbɒtl̩] [1], [ˈbɑːtl̩] [2]
m	`m`	map	[ˈmæp]
n	`n`	nap	[ˈnæp]
n̩	`=n`	vision	[ˈvɪʒn̩]
ŋ	`N`	thing	[ˈθɪŋ]

IPA	ETHPA	Example	
ɔː	`O:`	core	[kʰɔː] [1], [kʰɔːr] [2]
o͡ʊ	`o_U`	nose	[ˈno͡ʊz] [2]
ɔ͡ɪ	`O_I`	boy	[ˈbɔ͡ɪ]
p	`p`	speed	[ˈspiːd]
pʰ	`p_h`	pin	[ˈpʰɪn]
r	`r`	ring, stress	[ˈrɪŋ], [ˈstres]
s	`s`	sip, mouse	[ˈsɪp], [ˈma͡ʊs]
ʃ	`S`	ship, brush	[ˈʃɪp], [ˈbrʌʃ]
t	`t`	street	[ˈstriːt]
tʰ	`t_h`	time	[ˈtʰa͡ɪm]
t͡ʃ	`t_S`	chin	[ˈt͡ʃɪn]
θ͡	`T`	thin, breath	[ˈθ͡ɪn], [ˈbreθ]
u	`u`	influential	[ˌɪnfluˈent͡ʃl̩]
uː	`u:`	lose	[ˈluːz]
ʊ	`U`	put, book	[ˈpʰʊt], [ˈbʊk]
ʊ͡ə	`U_@`	durable	[ˈdjʊ͡ərəbl̩] [1]
v	`v`	vat	[ˈvæt]
w	`w`	well	[ˈwel]
x	`x`	loch	[ˈlɒx] [3]
z	`z`	zip, fees	[ˈzɪp], [ˈfiːz]
ʒ	`Z`	vision	[ˈvɪʒn̩]

[1] British English
[2] American English
[3] Scottish

A.2 French IPA Symbols

IPA	ETHPA	Example	
ə	`@`	fortement	[fɔʀtəmɑ̃]
(ə)	`(@)`	petit	[p(ə)ti] [1]
a	`a`	tabac, patte	[taba], [pat(ə)]
aː	`a:`	bagage	[bagaːʒ(ə)]
ɑ	`A`	bât, pâte	[bɑ], [pɑt(ə)]
ɑː	`A:`	bagare	[bagɑːʀ(ə)]
ɑ̃	`~A`	temps	[tɑ̃]
ɑ̃ː	`~A:`	ange	[ɑ̃ːʒ]
b	`b`	bon, robe	[bɔ̃], [ʀɔb(ə)]
d	`d`	dans, chaude	[dɑ̃], [ʃoːd(ə)]
e	`e`	ému, ôté	[emy], [ote]
ɛ	`E`	perdu	[pɛʀdy]
ɛː	`E:`	treize	[tʀɛːz(ə)]
ɛ̃	`~E`	matin	[matɛ̃]
ɛ̃ː	`~E:`	linge	[lɛ̃ːʒ]
f	`f`	feu, chef	[fø], [ʃɛf]
g	`g`	gare, bague	[gɑːʀ(ə)], [bag(ə)]
h	`h`	halte, hop	[halt(ə)], [hɔp] [2]
i	`i`	lit, ami	[li], [ami]
iː	`i:`	lige	[liːʒ(ə)]
j	`j`	yeux, paille	[jø], [pɑːj(ə)]
ɲ	`J`	agneau, vigne	[aɲo], [viɲ(ə)]
k	`k`	carte, barque	[kaʀt(ə)], [baʀk(ə)]
l	`l`	long, bal	[lɔ̃], [bal]
m	`m`	madame, femme	[madam(ə)], [fam(ə)]
n	`n`	nous, bonne	[nu], [bɔn(ə)]
ŋ	`N`	camping	[kɑ̃piŋ]
o	`o`	beau, galop	[bo], [galo]
oː	`o:`	chaude	[ʃoːd(ə)]
ɔ	`O`	obstacle	[ɔpstakl(ə)]
ɔː	`O:`	corps	[kɔːʀ]
ɔ̃	`~O`	bon	[bɔ̃]
ɔ̃ː	`~O:`	ronde	[ʀɔ̃ːd(ə)]
ø	`2`	deux	[dø]
øː	`2:`	creuse	[kʀøːz(ə)]

IPA	ETHPA	Example	
œ	`9`	neuf	[nœf]
œː	`9:`	peur	[pœːʀ]
œ̃	`~9`	lundi, parfum	[lœ̃di], [paʀfœ̃]
p	`p`	patte, cap	[pat(ə)], [kap]
ʀ	`R`	rue, venir	[ʀy], [v(ə)niːʀ]
s	`s`	sœur, passe	[sœːʀ], [pɑːs(ə)]
ʃ	`S`	chat, poche	[ʃa], [pɔʃ(ə)]
t	`t`	tête, net	[tɛt(ə)], [nɛt]
u	`u`	roue	[ʀu]
uː	`u:`	ajour	[aʒuːʀ]
v	`v`	vent, rêve	[vɑ̃], [ʀɛːv(ə)]
w	`w`	oui, nouer	[ʔwi], [nwe]
y	`y`	élu, punir	[ely], [pyniːʀ]
yː	`y:`	pur	[pyːʀ]
ɥ	`H`	huile, nuire	[ɥil(ə)], [nɥiːʀ(ə)]
z	`z`	zéro, rose	[zeʀo], [ʀoːz(ə)]
ʒ	`Z`	jardin, piège	[ʒaʀdɛ̃],[pjɛːʒ(ə)]
ʔ	`?`	les haricots	[le ʔaʀiko]

[1] optional schwa

[2] within interjections

A.3 German IPA Symbols

IPA	ETHPA	Example	
a	a	hat	[ˈhat]
aː	a:	Bahn	[ˈbaːn]
ɐ	6	Ober	[ˈʔoːbɐ]
ɐ̯	^6	Uhr	[ˈʔuːɐ̯]
a͡i	a_i	weit	[ˈva͡it]
a͡u	a_u	Haut	[ˈha͡ut]
b	b	Ball	[ˈbal]
ç	C	ich	[ˈʔɪç]
d	d	dann	[ˈdan]
d͡ʒ	d_Z	Gin	[ˈd͡ʒɪn]
e	e	Methan	[meˈtaːn]
eː	e:	Beet	[ˈbeːt]
e͡ɪ	e_I	Frey	[ˈfre͡ɪ] [1]
ɛ	E	hätte	[ˈhɛtə]
ɛː	E:	wähle	[ˈvɛːlə]
ə	@	halte	[ˈhaltə]
f	f	Fass	[ˈfas]
g	g	Gast	[ˈgast]
g͡g	g_g	Rüegger	[ˈryə͡ggər] [2]
h	h	hat	[ˈhat]
i	i	vital	[viˈtaːl]
iː	i:	viel	[ˈfiːl]
i̯	^i	Studie	[ˈʃtuːdi̯e]
i̯ə	i_@	Dietikon	[ˈdi̯əti̯ˌkoːn] [1]
ɪ	I	bist	[ˈbɪst]
j	j	ja	[ˈjaː]
k	k	Skandal	[skanˈdaːl]
kʰ	k_h	kalt	[ˈkʰalt]
l	l	Last	[ˈlast]
l̩	=l	Nabel	[ˈnaːbl̩]
m	m	Mast	[ˈmast]
m̩	=m	grossem	[ˈgroːsm̩]
n	n	Naht	[ˈnaːt]
n̩	=n	baden	[ˈbaːdn̩]
ŋ	N	lang	[ˈlaŋ]

IPA	ETHPA	Example	
o	o	Moral	[moˈraːl]
oː	o:	Boot	[ˈboːt]
o͡	^o	loyal	[lo͡aˈjaːl]
ɔ	O	Post	[ˈpɔst]
ɔ͡y	O_y	Heu	[ˈhɔ͡y]
ø	2	Ökonom	[ˌʔøkoˈnoːm]
øː	2:	Öl	[ˈʔøːl]
œ	9	göttlich	[ˈgœtlɪç]
p	p	Spatz	[ˈʃpat͡s]
p͡f	p_f	Pfahl	[ˈp͡faːl]
pʰ	p_h	Pakt	[ˈpʰakt]
r	r	Rast	[ˈrast]
r͡r	r_r	Karren	[ˈkar͡rən]
s	s	Hast	[ˈhast]
ʃ	S	Schal	[ʃaːl]
t	t	Stier	[ʃtiːr]
tʰ	t_h	Tal	[tʰaːl]
t͡s	t_s	Zahl	[t͡saːl]
t͡ʃ	t_S	Matsch	[ˈmat͡ʃ]
u	u	kulant	[kuˈlant]
uː	u:	Hut	[ˈhuːt]
u͡	^u	aktuell	[akˈtu͡ɛl]
u͡i	u_i	pfui	[ˈp͡fu͡i]
ʊ	U	Pult	[ˈpʰʊlt]
ʊ͡ə	U_@	Ruedi	[ˈrʊ͡ədi] [1]
v	v	was	[ˈvas]
x	x	Bach	[ˈbax]
y	y	Mykene	[myˈkeːnə]
yː	y:	Rübe	[ˈryːbə]
y̆	^y	Etui	[ʔeˈty̆iː]
y͡ə	y_@	Blüemlisalp	[ˌbly͡əmlisˈalp] [1]
ʏ	Y	füllt	[ˈfʏlt]
z	z	Hase	[ˈhaːzə]
ʒ	Z	Genie	[ʒeˈniː]
ʔ	?	beamtet	[bəˈʔamtət]

[1] Swiss German diphthong
[2] strong Swiss German [g]

A.4 Italian IPA Symbols

IPA	ETHPA	Example	
a	a	parete	[paˈreːte]
aː	a:	pane	[ˈpaːne]
b	b	bambina	[bamˈbiːna]
b͡b	b_b	repubblica	[reˈpub͡blika]
d	d	ladina	[laˈdiːna]
d͡d	d_d	freddezza	[freˈd͡dett͡sa]
d͡dz	d_d_z	mezzi	[ˈmɛd͡dzi]
d͡dʒ	d_d_Z	oggi	[ˈɔd͡dʒi]
d͡z	d_z	zona	[ˈd͡zɔːna]
d͡ʒ	d_Z	Genova	[ˈd͡ʒɛːnova]
e	e	terreno	[teˈr͡reːno]
eː	e:	nero	[ˈneːro]
ɛ	E	mezzo	[ˈmɛd͡dzo]
ɛː	E:	bene	[ˈbɛːne]
f	f	fumo	[ˈfuːmo]
f͡f	f_f	caffè	[kaˈf͡fɛ]
g	g	gondola	[ˈgondola]
g͡g	g_g	aggressivo	[ag͡greˈs͡siːvo]
i	i	bilancio	[biˈlant͡ʃo]
iː	i:	lira	[ˈliːra]
i͡	^i	inizio	[iˈnit͡ti͡o]
ɲ	J	gnocco	[ˈɲɔk͡ko]
ɲ͡ɲ	J_J	prognosi	[ˈprɔɲ͡ɲozi]
k	k	vacanza	[vaˈkant͡sa]
k͡k	k_k	bocconi	[boˈk͡koːni]
l	l	lama	[ˈlaːma]
l͡l	l_l	midollo	[miˈdol͡lo]
ʎ	L	figlio	[ˈfiːʎo]
ʎ͡ʎ	L_L	bottiglia	[boˈt͡tiʎ͡ʎa]
m	m	menù	[meˈnu]
m͡m	m_m	mamma	[ˈmam͡ma]
n	n	Napoli	[ˈnaːpoli]
n͡n	n_n	nonno	[ˈnɔn͡no]
ŋ	N	banca	[ˈbaŋka]
o	o	posata	[poˈzaːta]

IPA	ETHPA	Example	
oː	`o:`	volo	[ˈvoːlo]
ɔ	`O`	ricordo	[riˈkɔrdo]
ɔː	`O:`	cosa	[ˈkɔːza]
p	`p`	presto	[ˈprɛsto]
p͡p	`p_p`	scialuppa	[ʃaˈlup͡pa]
r	`r`	Rimini	[ˈriːmini]
r͡r	`r_r`	carro	[ˈkar͡ro]
s	`s`	salsa	[ˈsalsa]
s͡s	`s_s`	deflusso	[deˈflus͡so]
ʃ	`S`	scena	[ˈʃɛːna]
ʃ͡ʃ	`S_S`	riuscita	[riuˈʃ͡ʃiːta]
t	`t`	cantata	[kanˈtaːta]
t͡s	`t_s`	zitto	[ˈt͡sit͡to]
t͡ʃ	`t_S`	cena	[ˈt͡ʃeːna]
t͡t	`t_t`	viadotto	[viaˈdot͡to]
t͡t͡s	`t_t_s`	merluzzo	[merˈlut͡t͡so]
t͡t͡ʃ	`t_t_S`	nocciola	[noˈt͡t͡ʃɔːla]
u	`u`	lumaca	[luˈmaːka]
uː	`u:`	luna	[ˈluːna]
u̯	`^u`	acqua	[ˈak̂u̯a]
v	`v`	vivace	[viˈvaːt͡ʃe]
v͡v	`v_v`	provvidenza	[prov͡viˈdɛnt͡sa]
z	`z`	sbarra	[ˈzbar͡ra]

A.5 Suprasegmental Symbols

In addition to the phonetic symbols, the following ETHPA symbols of suprasegmentals occur in the lexicas and the phonetic word transcriptions. These are common for all languages.

IPA	ETHPA		
ˈ	'	(apostrophe)	Primary stress
ˌ	,	(comma)	Secondary stress
.	-		Syllable boundary
	( )		Optional phone markers

Appendix B

Grammars and Lexica

A grammar is a collection of grammar rules. Each grammar rule consists of a head, which denotes a constituent, the production symbol `'==>'`, and a body, which denotes a list of subconstituents.

```
headcons ==> [ { subcons } ] * [ penalty ] [ { keywords } ]
```

An empty subconstituent list denotes the empty production. The body is terminated by the `'*'` symbol. A grammar rule can optionally be followed by an integer penalty value. If this penalty value is missing, a default value of 1 is assumed. These penalty values are added during rule application in the parser and are used to select the optimal solution out of a number of ambiguous solutions.

The keywords `':INV'`, `':WORD_END'`, `':SENT_END'`, and `':PARA_END'` may optionally be set after the penalty value or the `'*'` sign, if the penalty value is missing. `':INV'` makes the corresponding node of a rule invisible in the resulting syntax tree. `':WORD_END'`, `':SENT_END'`, and `':PARA_END'` are used for the identification of syntactic word, sentence, and paragraph boundaries, respectively, as explained in Section 2.3.

Each constituent is composed of a constituent identifier and a list of feature terms associated with the constituent. Language specific constituent identifiers have got the suffix `'_E'`, `'_F'`, `'_G'`, or `'_I'` that

stands for English, French, German, or Italian, resp. Feature terms can be atoms or variables. Variables start with a `'?'` followed by an identifier. The variable `'?'` itself is the "anonymous variable", which is usually applied as a "don't care" marker. Atoms are constants, whose identifiers must not start with a `'?'`. Term unification operates on all variables with identical identifiers within one grammar rule.

A lexicon is a collection of so-called preterminal constituents together with their associated terminal elements (i.e., the individual words or morphemes). Each lexicon entry consists of a constituent name followed by the graphemic and phonemic representation of the terminal element, and optionally followed by a penalty value and keywords:

```
cons "graphem_repr" "phonem_repr" [ penalty ] [ { keywords } ]
```

The following subsections list all English, French, German and Italian grammar rules and lexicon entries that are necessary to understand the examples given in this article. The grammars as well as the lexica presented here are by far not complete. The complete, quadrilingual set of all grammars and lexica of the polySVOX system comprises currently about 3 700 grammar rules and about 27 000 lexicon entries.

B.1 English lexicon and grammars

English lexicon

```
[L1]   PRGTRM ()     "<PB>"     ""          0    :WORD_END
[L2]   PCT_E (f,s)   "."        ""              :WORD_END
[L3]   PCT_E (f,s)   ". "       ""              :WORD_END
[L4]   TRM_E (?)     " "        ""          0    :WORD_END
[L5]   TRM_E (?)     ""         ""          100
[L6]   TRM_E (abbr)  ""         ""
[L7]   HYP_E ()      "-"        ""
[L8]   HYP_E ()      "- "       ""              :WORD_END

[L9]   PERSS_E (sg,p3,n,s)                  "it"       "'It"
[L10]  PERSS_E (pl,p1,n,o)                  "'s"       "z"
[L11]  PREPS_E ()                           "in"       "'In"
[L12]  PREF_E ()                            "in+"      "'In-+"
[L13]  PREF_E ()                            "up+"      "'Vp-+"

[L14]  NTS_E (ntcl1)                        "saint"    "s'@nt+"
[L15]  NTS_E (abbr)                         "st"       "s'@nt+"
```

```
[L16]  NTE_E (ntcl1)                        ""         ""
[L17]  NTE_E (abbr)                         "."        ""
[L18]  NTE_E (abbr)                         ""         ""

[L19]  NS_E (ncl7,sgen1,n)                  "dat+"     "d'e_It+"
[L20]  NS_E (ncl1,sgen1,n)                  "hat+"     "h'qt+"
[L21]  NS_E (ncl1,sgen1,n)                  "input+"   "'InpUt+"
[L22]  NS_E (ncl1,sgen1,n)                  "nation+"  "n'e_IS(@)n+"
[L23]  NS_E (ncl1,sgen1,n)                  "street+"  "str'i:t+"
[L24]  NS_E (abbr,nosgen,n)                 "st"       "str'i:t+"
[L25]  NPRS_E (ncl1,sgen1,n)                "asia+"    "'e_IS@+"
[L26]  NPRS_E (ncl1,sgen1,f)                "mary+"    "m'e_@ri+"
[L27]  NE_E (ncl1,sg)                       ""         ""
[L28]  NE_E (ncl1,pl)                       "s"        "s"
[L29]  NE_E (ncl7,sg)                       "e"        ""
[L30]  NE_E (ncl7,pl)                       "es"       "s"
[L31]  NE_E (abbr,sg)                       ""         ""
[L32]  NE_E (abbr,sg)                       "."        ""
[L33]  NGE_E (sgen1,sg)                     "'s"       "z"

[L34]  VS_E (emute1,pres)                   "dat+"     "d'e_It+"
[L35]  VS_E (emute1,pres)                   "di+"      "d'a_I+"
[L36]  VS_E (s,pres)                        "input+"   "'InpUt+"
[L37]  VS_E (s,pres)                        "put+"     "p'Ut+"
[L38]  VS_E (emute1,pres)                   "welcom+"  "w'elk@m+"
[L39]  AUXBS_E (sg,p3,ind,pres,yes)         "'s"       "z"
[L40]  AUXHS_E (sg,p3,ind,pres,yes)         "'s"       "z"
[L41]  VE_E (emute1,pres,ind,pres,sg,p1)    "e"        ""
[L42]  VE_E (emute1,pres,ind,pres,sg,p2)    "e"        ""
[L43]  VE_E (emute1,pres,ind,pres,sg,p3)    "es"       "z"

[L44]  AS_E (as1)                           "great"    "gr'e_It"
[L45]  ASE_E (as1,pos)                      ""         ""
[L46]  ASE_E (as1,comp)                     "er"       "@(r)"
[L47]  ASE_E (as1,sup)                      "est"      "@st"
```

English submorphemic lexicon

```
[L48]  OCONS_E (s,?)     "b"  "b"
[L49]  OCONS_E (s,?)     "c"  "k"
[L50]  OCONS_E (s,?)     "l"  "l"
[L51]  OCONS_E (s,?)     "m"  "m"
[L52]  CCONS_E (s,?)     "b"  "b"
[L53]  CCONS_E (s,?)     "m"  "m"
[L54]  CCONS_E (m,nf)    "mb" "mb"
[L55]  CCONS_E (s,f)     "mb" "m"
[L56]  SVOW_E (ln,?)     "a"  "'e_I"
[L57]  SVOW_E (sh,?)     "a"  "'q"
[L58]  SVOW_E (ln,?)     "i"  "'a_I"
[L59]  SVOW_E (sh,?)     "i"  "'I"
[L60]  SVOW_E (ln,?)     "o"  "'@_U"
```

```
[L61]  SVOW_E (sh,?)      "o"  "'Q"
[L62]  SVOW_E (ln,?)      "u"  "j'u:"
[L63]  SVOW_E (sh,?)      "u"  "'V"
[L64]  UVOW_E ()          "a"  "@"
[L65]  UVOW_E ()          "i"  "I"
[L66]  UVOW_E ()          "o"  "@"
[L67]  UVOW_E ()          "u"  "j@"
[L68]  UNSUFF_E (n,p,nf) "ia" "I@"
```

English word grammar

```
[R1]   PRGTRM ()                  ==> PRGTRM () * :SENT_END
[R2]   PCT_E (?F,?T)              ==> PCT_E (?F,?T) * :SENT_END
[R3]   N_E (?N,?G,?SG)            ==> NOUN_E (?N,?G,?SG,?NCL)
                                      NGE_OPT_E (?SG,?N)
                                      TRM_E (?NCL) *
[R4]   NOUN_E (?N,?G,?SG,?NCL)    ==> NS_E (?NCL,?SG,?G)
                                      NE_E (?NCL,?N) * :INV
[R5]   NPR_E (?N,?G,?SG)          ==> NPRS_E (?NCL,?SG,?G)
                                      NE_E (?NCL,?N)
                                      NGE_OPT_E (?SG,?N)
                                      TRM_E (?NCL) *
[R6]   NGE_OPT_E (?SG,?N)         ==> * 0 :INV
[R7]   NGE_OPT_E (?SG,?N)         ==> NGE_E (?SG,?N) * 0 :INV
[R8]   NT_E ()                    ==> NTS_E (?NTCL)
                                      NTE_E (?NTCL)
                                      TRM_E (?NTCL) *
[R9]   AUXB_E (?N,?P,?M,?T,?POS) ==> AUXBS_E (?N,?P,?M,?T,?POS)
                                      TRM_E (std) *
[R10]  AUXH_E (?N,?P,?M,?T,?POS) ==> AUXHS_E (?N,?P,?M,?T,?POS)
                                      TRM_E (std) *
[R11]  PERS_E (?NR,?P,?G,?C)      ==> PERSS_E (?NR,?P,?G,?C)
                                      TRM_E (std) * 0


[R12]  NS_E (ncl1,?,?)               ==> ESTEM_E (n) *
[R13]  ESTEM_E (n)                   ==> CSYL_E (?,m,sh,r,i,nf)
                                         UNSUFF_E (n,p,nf) *
[R14]  CSYL_E (?O,?C,?PR,r,?IP,?FP) ==> SYLL_E (?O,n,ln,u,?IP,nf)
                                         SYLL_E (s,?C,?PR,s,ni,?FP) *
[R15]  SYLL_E (?O,?C,?PR,s,?IP,?FP) ==> ONSCL_E (?O,?IP)
                                         SVOW_E (?PR,?FP)
                                         CODCL_E (?C,?FP) *
[R16]  SYLL_E (?O,?C,?PR,u,?IP,?FP) ==> ONSCL_E (?O,?IP)
                                         UVOW_E ()
                                         CODCL_E (?C,?FP) *
[R17]  ONSCL_E (n,?)                 ==> * 1 :INV
[R18]  ONSCL_E (s,?P)                ==> OCONS_E (s,?P) * 1 :INV
[R19]  ONSCL_E (d,?P)                ==> OCONS_E (d,?P) * 1 :INV
[R20]  ONSCL_E (m,?P)                ==> OCONS_E (m,?P) * 2 :INV
[R21]  CODCL_E (n,?)                 ==> * 0 :INV
```

```
[R22]  CODCL_E (s,?P)                ==> CCONS_E (s,?P) * 2 :INV
[R23]  CODCL_E (d,?P)                ==> CCONS_E (d,?P) * 2 :INV
[R24]  CODCL_E (m,?P)                ==> CCONS_E (m,?P) * 2 :INV
```

English sentence grammar

```
[R25]  PRGTRM ()                ==> PRGTRM () * :PARA_END
[R26]  S_E (?T)                 ==> PERS_E (?N,?P,?,s)
                                    VP_E (ind,?T,?N,?P,?,fin)
                                    PP_E ()
                                    PCT_E (f,s) *
[R27]  S_E (?T)                 ==> NP_E (?N,?G)
                                    VP_E (ind,?T,?N,?P,?,fin)
                                    NP_E (?,?)
                                    PCT_E (f,s) *
[R28]  VP_E (inf,?T,?N,?P,?,?) ==> AUXB_E (?N,?P,inf,?T,pos) *
[R29]  PP_E ()                  ==> PREP_E (?)
                                    NP_E (?,?) *
[R30]  NP_E (?N,?G)             ==> NPRP_E (?,?)
                                    N_REP_E (?N,?G) *
[R31]  N_REP_E (?N,?G)          ==> N_E (?N,?G,?) * :INV
[R32]  N_REP_E (?N,?G)          ==> N_E (?,?,?)
                                    N_REP_E (?N,?G) * :INV
[R33]  NPRP_E (?N,?G)           ==> NT_E (?)
                                    NPR_REP_E (?N,?G) * :INV
[R34]  NPR_REP_E (?N,?G)        ==> NPR_E (?N,?G,?) * :INV
[R35]  NPR_REP_E (?N,?G)        ==> NPR_E (?,?,?)
                                    NPR_REP_E (?N,?G) * :INV
```

English paragraph grammar

```
[R36]  P_E ()     ==> S_REP_E () PRGTRM () *
[R37]  S_REP_E () ==> S_E (?) * :INV
[R38]  S_REP_E () ==> S_E (?)
                      S_REP_E () * 5 :INV
```

English inclusion grammars

```
[R39]  NOUN_E (?NR,?,?,?) ==> NOUN_F (?NR,?) * 150
[R40]  N_E (?NR,?,?)      ==> N_F (?NR,?) * 100
[R41]  NPR_E (?,?,?)      ==> PRN_F () * 100
[R42]  ADJ_E (?)          ==> ADJ_F (?,?,?) * 100
[R43]  NP_E (?NR,?G)      ==> NP_F (?NR,?,?G) * 80
[R44]  NP_E (?NR,?)       ==> NP_F (?NR,?,?) * 90
[R45]  N_E (?,?,?)        ==> NP_F (?,?,?) * 90
[R46]  NOUN_E (?NR,?,?,?) ==> NOUN_G (?,?NR,?) * 120
```

```
[R47]  N_E (?NR,?,?)       ==> N_G (?NR,?,?) * 100
[R48]  NPR_E (?,?,?)       ==> PRN_G (?,?,?) * 100
[R49]  NPR_E (?,?,?)       ==> NP_G (?,?,?,?,?) * 110
[R50]  ADJ_E (?)           ==> ADJ_G (?,?,?,?,?) * 110
[R51]  NP_E (?,?)          ==> NP_G (?,?,?,?,?) * 90
[R52]  N_E (?,?,?)         ==> NP_G (?,?,?,?,?) * 90
```

B.2 French lexicon and grammars

French lexicon

```
[L69]  PRGTRM ()     "<PB>"     ""          0   :WORD_END
[L70]  PCT_F (f,s)   "."        ""              :WORD_END
[L71]  PCT_F (f,s)   ". "       ""              :WORD_END
[L72]  TRM_F (?)     " "        ""          0   :WORD_END
[L73]  TRM_F (?)     ""         ""          100
[L74]  TRM_F (abbr)  ""         ""          1

[L75]  NS_F (sgm1,sgf1,plm1,plf1)         "ami+"     "ami+"
[L76]  NS_F (non,sgf2,non,plf1)           "date+"    "dat(@)+"
[L77]  NS_F (non,sgf2,non,plf1)           "femme+"   "fam+"
[L78]  NS_F (sgm1,non,plm1,non)           "film+"    "film+"
[L79]  NS_F (non,sgf2,non,plf1)           "nation+"  "nasj~o+"
[L80]  PRNS_F ()                          "chirac+"  "SiRak+"
[L81]  NESG_F (m,sgm1)                    ""         ""
[L82]  NESG_F (f,sgf1)                    "e"        ""
[L83]  NESG_F (f,sgf2)                    ""         ""
[L84]  NEPL_F (m,plm1)                    "s"        ""
[L85]  NEPL_F (f,plf1)                    "s"        ""

[L86]  VS_F (g1,scl1a,nonrefl,?,non)      "dat+"     "dat+"
[L87]  VE_F (g1,scl1a,ind,pres,sg,pers1)  "e"        "(@)"
[L88]  VE_F (g1,scl1a,ind,pres,sg,pers2)  "es"       "(@)"
[L89]  VE_F (g1,scl1a,ind,pres,sg,pers3)  "e"        "(@)"
[L90]  VE_F (g1,scl1a,ind,pres,pl,pers1)  "ons"      "~o"
[L91]  VE_F (g1,scl1a,ind,pres,pl,pers2)  "ez"       "e"
[L92]  VE_F (g1,scl1a,ind,pres,pl,pers3)  "ent"      "(@)"

[L93]  AS_F (v,sgm1,sgf7,plm1,plf1,non)   "bon+"     "b~o+"
[L94]  AS_F (n,sgm1,sgf1,plm1,plf1,adv1)  "fatal+"   "fatal+"
[L95]  AS_F (a,sgm1,sgf3,plm1,plf1,non)   "grand+"   "gR~a+"
[L96]  AS_F (n,sgm1,sgf1,plm1,plf1,non)   "noir+"    "nwaR+"
[L97]  AE_F (sg,m,sgm1)                   ""         ""
[L98]  AE_F (sg,f,sgf1)                   "e"        "(@)"
[L99]  AE_F (sg,f,sgf3)                   "e"        "d(@)"
[L100] AE_F (sg,f,sgf7)                   "ne"       "n(@)"
[L101] AE_F (pl,m,plm1)                   "s"        ""
[L102] AE_F (pl,f,plf1)                   "s"        ""
```

French word grammar

```
[R53]  PRGTRM ()           ==> PRGTRM () * :SENT_END
[R54]  PCT_F (?F,?T)       ==> PCT_F (?F,?T) * :SENT_END
[R55]  N_F (?NR,?G)        ==> NOUN_F (?NR,?G)
                               TRM_F (?) *
[R56]  NOUN_F (sg,?G)      ==> NS_F (?SGM,?,?,?)
                               NESG_F (?G,?SGM) * :INV
[R57]  NOUN_F (sg,?G)      ==> NS_F (?,?SGF,?,?)
                               NESG_F (?G,?SGF) * :INV
[R58]  NOUN_F (pl,?G)      ==> NS_F (?SGM,?,?PLM,?)
                               NESG_F (?G,?SGM)
                               NEPL_F (?G,?PLM) * :INV
[R59]  NOUN_F (pl,?G)      ==> NS_F (?,?SGF,?,?PLF)
                               NESG_F (?G,?SGF)
                               NEPL_F (?G,?PLF) * :INV
[R60]  ADJ_F (?POS,?N,?G) ==> AS_F (?POS,?SGM,?,?,?,?)
                               AE_F (?N,?G,?SGM)
                               TRM_F (?) *
[R61]  ADJ_F (?POS,?N,?G) ==> AS_F (?POS,?,?SGF,?,?,?)
                               AE_F (?N,?G,?SGF)
                               TRM_F (?) *
[R62]  ADJ_F (?POS,?N,?G) ==> AS_F (?POS,?SGM,?,?PLM,?,?)
                               AE_F (?,?G,?SGM)
                               AE_F (?N,?G,?PLM)
                               TRM_F (?) *
[R63]  ADJ_F (?POS,?N,?G) ==> AS_F (?POS,?,?SGF,?,?PLF,?)
                               AE_F (?,?G,?SGF)
                               AE_F (?N,?G,?PLF)
                               TRM_F (?) *
```

French sentence grammar

```
[R64]  NP_F (?N,?P,?G) ==> N_F (?N,?P,?G)
                           ADJ_F (n,?N,?G) *
```

French inclusion grammars

```
[R65]  ADJ_F (?,?,?)             ==> ADJ_E (?) * 100
[R66]  PRN_F ()                  ==> NPR_E (?,?,?) * 100
[R67]  V_F (?,?,?NR,?P,?,?,non) ==> V_E (?,?,?NR,?P) * 180
[R68]  N_F (?NR,?)               ==> N_G (?NR,?,?) * 100
[R69]  ADJ_F (?,?N,?)            ==> ADJ_G (?,?N,?,?,?) * 100
```

B.3 German lexicon and grammars

German lexicon

```
[L103] PRGTRM ()     "<PB>"     ""          0 :WORD_END
[L104] PCT_G (f,s)   "."        ""          :WORD_END
[L105] PCT_G (f,s)   ". "       ""          :WORD_END
[L106] TRM_G (?)     " "        ""          0 :WORD_END
[L107] TRM_G (?)     ""         ""          100
[L108] TRM_G (abbr)  ""         ""          1

[L109] NS_G (sk3,pk7,m)                 "film+"     "'fIlm+"
[L110] NS_G (sk1,pk4,f)                 "nation+"   "na't_sˆio:n+"
[L111] NS_G (sk2,non,m)                 "ton+"      "'to:n+"
[L112] NES_G (sk1,?)                    ""          ""
[L113] NES_G (sk2,n)                    ""          ""
[L114] NES_G (sk2,g)                    "s"         "s"
[L115] NES_G (sk2,d)                    ""          ""
[L116] NES_G (sk2,a)                    ""          ""
[L117] NES_G (sk3,n)                    ""          ""
[L118] NES_G (sk3,g)                    "es"        "@s"
[L119] NES_G (sk3,d)                    ""          ""
[L120] NES_G (sk3,d)                    "e"         "@"
[L121] NES_G (sk3,a)                    ""          ""
[L122] NEP_G (pk4,?)                    "en"        "@n"
[L123] NEP_G (pk7,n)                    "e"         "@"
[L124] NEP_G (pk7,g)                    "e"         "@"
[L125] NEP_G (pk7,d)                    "en"        "@n"
[L126] NEP_G (pk7,a)                    "e"         "@"

[L127] VS_G (v1,a,v,non)                "datier+"   "da'ti:r+"
[L128] VS_G (v7,a,v,non)                "geb+"      "'ge:b+"
[L129] VS_G (v6,a,auxh,non)             "hab+"      "'ha:b+"
[L130] VS_G (v6,b,auxh,non)             "ha+"       "'ha+"
[L131] VS_G (v1,a,v,a)                  "lös+"      "'l2:z+"
[L132] VE_G (v1,a,ind,pres,sg,pers1)    "e"         "@"
[L133] VE_G (v1,a,ind,pres,sg,pers2)    "st"        "st"
[L134] VE_G (v1,a,ind,pres,sg,pers3)    "t"         "t"
[L135] VE_G (v6,a,ind,pres,sg,pers1)    "e"         "@"
[L136] VE_G (v6,b,ind,pres,sg,pers2)    "st"        "st"
[L137] VE_G (v6,b,ind,pres,sg,pers3)    "t"         "t"
[L138] VE_G (v12,a,ind,pres,sg,pers1)   "iere"      "'i:r@"
[L139] VE_G (v12,a,ind,pres,sg,pers2)   "ierst"     "'i:rst"
[L140] VE_G (v12,a,ind,pres,sg,pers3)   "iert"      "'i:rt"
[L141] P1SUFF_G ()                      "end+"      "@nd+"
[L142] P2PREF_G ()                      "ge+"       "g@+"
[L143] P2E_G (v1)                       "et"        "@t"
[L144] P2E_G (v1)                       "t"         "t"
[L145] P2E_G (v12)                      "iert"      "'i:rt"

[L146] AS_G (pos,non,non)               "schwarz+"  "'Svart_s+"
[L147] AE_G (typ2,n,sg,f)               "e"         "@"
```

```
[L148] AE_G (typ2,g,sg,f)              "en"       "@n"
[L149] AE_G (typ2,d,sg,f)              "en"       "@n"
[L150] AE_G (typ2,a,sg,f)              "e"        "@"

[L151] ARTDEFS_G (n,sg,m)              "der"      "'de:r"
[L152] ARTDEFS_G (n,sg,f)              "die"      "'di:"
[L153] PREF_G (v,p3,sep)               "ab+"      "'?ap+"
[L154] PREF_G (v,p3,sep)               "an+"      "'?an+"
[L155] CONJS_G (sub,front,c)           "als"      "'?als"
```

German word grammar

```
[R70]  PRGTRM ()              ==> PRGTRM () * :SENT_END
[R71]  PCT_G (?F,?T)          ==> PCT_G (?F,?T) * :SENT_END
[R72]  N_G (?C,?NR,?G)        ==> NOUN_G (?C,?NR,?G)
                                  TRM_G (?) *
[R73]  NOUN_G (?C,sg,?G)      ==> NS_G (?SGCL,?,?G)
                                  NES_G (?SGCL,?C) * :INV
[R74]  NOUN_G (?C,pl,?G)      ==> NS_G (?,?PLCL,?G)
                                  NEP_G (?PLCL,?C) * :INV
[R75]  P2_G (?,?)             ==> PREF_OPT_G (v,?,?)
                                  P2PREF_G ()
                                  VS_G (?VCL,?,v,?)
                                  P2E_G (?VCL)
                                  TRM_G (?) *
[R76]  P2_G (?,?)             ==> PREF_OPT_G (v,?,?)
                                  VS_G (?VCL,?,v,?)
                                  P2E_G (?VCL)
                                  TRM_G (?) *
[R77]  PREF_OPT_G (?U,?T,?S) ==> * 0 :INV
[R78]  PREF_OPT_G (?U,?T,?S) ==> PREF_G (?U,?T,?S) * 0 :INV
```

German sentence grammar

```
[R79]  NP_G (?C,?NR,?P,?G,?NT) ==> DET_G (?C,?NR,?G,?F,?TYP)
                                   NPNUC_G (?C,?NR,?P,?G,?TYP,?NT) *
[R80]  NP_G (?C,?NR,?P,?G,?NT) ==> DET_G (?C,?NR,?G,?F,?TYP)
                                   ADJ_G (?C,?NR,?G,?GR,?TYP)
                                   NPNUC_G (?C,?NR,?P,?G,?TYP,?NT) *
```

German inclusion grammars

```
[R81]  AS_G (p)               ==> AS_E (?,pos) * 150
[R82]  NS_G (sk2,pk1,?)       ==> NS_E (?,?,?) * 150
[R83]  NS_G (sk2,pk2,?)       ==> NS_E (?,?,?) * 150
[R84]  VS_G (v1,a,v,?)        ==> VS_E (?,pres) * 150
```

```
[R85]  PREF_G (?,p3,sep)           ==> PREF_E () * 100

[R86]  VS_G (v12,a,v,?REF)         ==> VS_F (?,?,?REF,?,non) * 160

[R87]  N_G (?NR,?,?)               ==> N_E (?NR,?,?) * 100
[R88]  PRN_G (?,?,?)               ==> NPR_E (?,?,?) * 100
[R89]  PRN_G (?,?NR,?)             ==> NP_E (?NR,?) * 110
[R90]  V_G (?,?,?,?,?)             ==> V_E (?,?,?,?) * 100
[R91]  ADJ_G (?,?,?,?,?)           ==> ADJ_E (?) * 110
[R92]  NP_G (?,?NR,?,?,?)          ==> NP_E (?NR,?) * 80
[R93]  NPNUC_G (?,?NR,pers3,?,?,?) ==> NP_E (?NR,?) * 90
[R94]  PP_G (?,?,?,?)              ==> PP_E () * 100

[R95]  N_G (?NR,?,?)               ==> N_F (?NR,?) * 110
[R96]  PRN_G (?,?,?)               ==> PRN_F () * 110
[R97]  PRN_G (?,?NR,?)             ==> NP_F (?NR,?,?) * 110
[R98]  V_G (?,?,?,?,?)             ==> V_F (?,?,?NR,?,?,?,non) * 120
[R99]  ADJ_G (?,?,?,?,?)           ==> ADJ_F (?,?,?) * 110
[R100] NP_G (?,?NR,?,?,?)          ==> NP_F (?NR,?,?) * 90
[R101] NPNUC_G (?,?NR,pers3,?,?,?) ==> NP_F (?NR,?,?) * 90
[R102] PP_G (?,?NR,?,?)            ==> PP_F (?NR,?) * 110

[R103] N_G (?,?,?)                 ==> N_I (?,?) * 110
[R104] PRN_G (?C,?NR,?G)           ==> NPR_I (?) * 110
[R105] PRN_G (?,?NR,?)             ==> NP_I (?NR,?,?) * 110
[R106] ADJ_G (?,?,?,?,?)           ==> ADJ_I (?,?) * 110
```

B.4 Italian lexicon and grammars

Italian lexicon

```
[L156] PRGTRM ()     "<PB>"     ""          0    :WORD_END
[L157] PCT_I (f,s)   "."        ""              :WORD_END
[L158] PCT_I (f,s)   ". "       ""              :WORD_END
[L159] TRM_I (?)     " "        ""          0    :WORD_END
[L160] TRM_I (?)     ""         ""          100
[L161] TRM_I (abbr)  ""         ""          1

[L162] NS_I (null,m)     "caff‘e+" "kaf_f'E+"
[L163] NS_I (e,m)        "latt+"   "l'at_t+"
[L164] NE_I (e,sg,m)     "e"       "e"
[L165] NE_I (e,pl,m)     "i"       "i"
[L166] NE_I (null,sg,m)  ""        ""
[L167] NE_I (null,pl,m)  ""        ""

[L168] AS_I (o)          "dat+"    "d'a:t+"
[L169] AE_I (o,sg,m)     "o"       "o"
[L170] AE_I (o,pl,m)     "i"       "i"
[L171] AE_I (o,sg,f)     "a"       "a"
```

```
[L172] AE_I (o,pl,f)    "e"       "e"
```

Italian word grammar

```
[R107] PRGTRM ()     ==> PRGTRM () * :SENT_END
[R108] PCT_I (?F,?T) ==> PCT_I (?F,?T) * :SENT_END
[R109] N_I (?N,?G)   ==> NOUN_I (?N,?G)
                         TRM_I (?) *
[R110] NOUN_I (?N,?G)==> NS_I (?CL,?G)
                         NE_I (?CL,?N,?G) * :INV
[R111] ADJ_I (?N,?G) ==> AS_I (?CL)
                         AE_I (?CL,?N,?G)
                         TRM_I (?) *
```

Appendix C

Input Factors for Prosody Control

C.1 Input Factors for Duration Control

The following tables list all 349 input factors extracted for segment duration modeling. The individual factors are described in Section 7.3.4. The tables contain the rank of each input factor, as estimated by factor relevance determination. The table entry of a factor that did not exist in the prosody corpus is left empty. The rank of a factor used by all duration models of a specific language has a dark grey background, e.g., **10**. The rank of a factor used at least by one duration model is indicated with a light grey background, e.g., 66.

German					Factor	French				
phone						phone				
-2	-1	0	1	2		-2	-1	0	1	2
139	107	53	167	87	vowel	171	**8**	**1**	165	197
218	52	**44**	**28**	148	gliding vowel	131	118	272	152	271
					triphthong					
106	215	**49**	189	76	consonant	202	215	**51**	**50**	166
151	**39**	66	104	296	affricate	148	275	300	299	295
48	255	**14**	267	277	glottal closure	273	279	282	140	287
220	172	179	**30**	**36**	preplosive pause	106	216	269	164	170
176	102		**41**	82	speech pause	203	**20**		**6**	124
298	**10**	**20**	304	**19**	voiced	65	**31**	**23**	297	301
245	97	**2**	226	269	long	278	**39**	**7**	60	276
300	**29**	**45**	**11**	121	syllabic	168	204	213	142	172
262	**32**	**1**	147	136	plosive	**41**	286	144	308	289
156	111	78	72	265	nasal	**40**	126	**4**	**47**	291
193	77	**6**	94	69	trill	224	296	105	283	306
					tap					
241	99	74	**24**	187	fricative	63	**22**	**2**	90	290
					lateralfricative					
197	229	169	133	276	approximant	293	**26**	**9**	97	159
181	86	**5**	141	261	lateralapproximant	292	127	307	285	103

Table C.1: *Segmental input factors for segment duration control.*

German					Factor	French				
phone						phone				
-2	-1	0	1	2		-2	-1	0	1	2
223	152	65	170	248	bilabial	201	**17**	104	192	122
126	109	**15**	84	154	labiodental	66	**16**	**12**	**49**	120
					dental					
71	127	**12**	108	230	alveolar	64	**13**	**44**	**48**	123
213	110	242	163	249	postalveolar	246	**21**	**43**	91	121
					retroflex					
153	114	**16**	59	250	palatal	96	**25**	158	**52**	212
309	57	228	51	103	velar	116	**24**	267	**53**	210
219	201	129	**37**	180	uvular	247	**10**	268	**46**	**38**
					pharyngeal					
157	135	**9**	171	207	glottal	135	**18**	**19**	**27**	211
146	**27**	**7**	164	211	front / frontclosing	113	93	**15**	209	100
143	**25**	**8**	61	79	central / centring	167	229	266	114	205
125	256	202	160	144	back / backclosing	112	67	**14**	208	99
93	274	58	**47**	210	close	132	259	257	227	101
174	**21**	**17**	216	306	close-mid	176	128	217	225	62
209	252	**13**	100	206	open-mid	175	151	258	228	98
134	**22**	183	234	200	open	177	150	56	226	102
132	62	**18**	**26**	227	low F1	133	149	**33**	180	139
221	**43**	**38**	119	188	middle F1	173	219	**35**	119	129
64	67	68	**42**	185	high F1	174	218	**34**	181	270

Table C.2: *Segmental input factors for segment duration control.*

German					Factor	French				
syllable						syllable				
-2	-1	0	1	2		-2	-1	0	1	2
145	80	**4**	85	131	unstressed	262	187	**11**	88	190
140	246	**23**	50	101	stress [1]	264	84	195	141	189
116	253	60	123	63	stress [2]	265	86	193	238	182
175	120	124	128	73	stress [3]	83	130	196	89	183
158	222	**33**	81	118	stress [4]	263	188	194	239	191
199	196	91	98	88	stress [E]	115	87	**36**	241	61
		54	96	191	phrase boundary 0	157	153	214	240	109
137	232	55	113	192	phrase boundary 1	221	186	**30**	237	68
208	89	75	70	112	phrase boundary 2	220	161	95	236	108
105	95	**31**	**34**	**35**	no phrase boundary	81	185	**3**	**28**	69
194	168	190	235	117	phrase type P	77	253	73	76	117
161	90	239	238	263	phrase type S	79	254	178	243	260
243	231	173	251	268	phrase type T	78	162	74	75	125
177	182	186	236	275	phrase type E	223	256	137	136	94
259	204	212	244	214	phrase type Y	80	143	163	242	261
278	254	237	198	162	phrase type YC	222	255	179	72	111
307	83	165	283	302	syllen: short	138	298	57	**54**	110
299	155	**46**	224	297	syllen: long	288	284	**29**	**55**	134

Table C.3: *Accentuation, phrasing, and syllable length input factors for segment duration control.*

German	Factor	French
291	phone pos	302
92	first phone	**32**
40	onset	245
130	nucleus	169
260	coda	244

Table C.4: *Syllable level input factors for segment duration control.*

German	Factor	French
233	L-headed syl pos: salient syl of a foot	234
159	L-headed syl pos: 1. non-salient syl of foot	85
166	L-headed syl pos: 2. non-salient syl of foot	233
225	L-headed syl pos: 3. non-salient syl of foot	154
282	L-headed syl pos: 4. non-salient syl of foot	156
280	L-headed syl pos: 5. non-salient syl of foot	198
279	L-headed syl pos: 6. non-salient syl of foot	232
281	L-headed syl pos: 7. non-salient syl of foot	235
272	L-headed syl pos: 8. non-salient syl of foot	
	L-headed syl pos: $>$ 8. non-salient syl of foot	
195	L-headed foot length: only one salient syl in foot	280
301	L-headed foot length: short	147
290	L-headed foot length: long	231
203	R-headed syl pos: salient syl of a foot	207
178	R-headed syl pos: 1. non-salient syl of foot	**37**
217	R-headed syl pos: 2. non-salient syl of foot	**42**
305	R-headed syl pos: 3. non-salient syl of foot	155
257	R-headed syl pos: 4. non-salient syl of foot	200
240	R-headed syl pos: 5. non-salient syl of foot	184
142	R-headed syl pos: 6. non-salient syl of foot	206
270	R-headed syl pos: 7. non-salient syl of foot	199
308	R-headed syl pos: 8. non-salient syl of foot	
	R-headed syl pos: $>$ 8. non-salient syl of foot	
122	R-headed foot length: only one salient syl in foot	146
292	R-headed foot length: short	145
284	R-headed foot length: long	294

Table C.5: *Foot level input factors for segment duration control.*

German	Factor	French
56	L-headed foot nr: first foot of phrase	251
288	L-headed foot nr: foot nr in phrase	305
247	L-headed foot pos: sentence initial foot	277
149	L-headed foot pos: sentence final foot	107
115	L-headed foot pos: phrase initial foot	230
205	L-headed foot pos: phrase final foot	250
184	L-headed foot pos: phrase central foot	252
150	L-headed foot pos: phrase with one foot	304
286	R-headed foot nr: first foot of phrase	249
258	R-headed foot nr: foot nr in phrase	281
287	R-headed foot pos: sentence initial foot	160
285	R-headed foot pos: sentence final foot	92
289	R-headed foot pos: phrase initial foot	82
294	R-headed foot pos: phrase final foot	303
295	R-headed foot pos: phrase central foot	248
293	R-headed foot pos: phrase with one foot	274
271	phrase length [syl]	59
266	phrase length [L-foot]	71
303	phrase length [R-foot]	58
138	sentence length [syl]	**45**
264	sentence length [L-foot]	70
273	sentence length [R-foot]	309
3	speech rate [pps]	**5**

Table C.6: *Phrase and sentence level input factors for segment duration control.*

C.2 Input Factors for F_0 Control

The following tables list all 910 input factors extracted for F_0 modeling. The individual factors are described in Section 7.2.3. The tables contain the rank of each input factor, as estimated by factor relevance determination. The table entry of a factor that did not exist in the prosody corpus is left empty. The rank of a factor used by all duration models of a specific language has a dark grey background, e.g., **10**. The rank of a factor used at least by one duration model is indicated with a light grey background, e.g., 66.

syllable										Factor
-3	-2	-1	0	1	2	3	4	5	6	
242	568	248	411	527	333	492	472	646	652	phrase boundary 0
513	213	477	352	**102**	**143**	511	363	318	440	phrase boundary 1
570	359	172	293	234	188	398	223	315	**100**	phrase boundary 2
76	**20**	**49**	422	**136**	**99**	279	**129**	316	504	phrase type P
15	260	**81**	**111**	**19**	481	**134**	238	167	565	phrase type T
36	**32**	**11**	**13**	**12**	**78**	**18**	292	210	384	phrase type S
10	**26**	**5**	**3**	**1**	**9**	**27**	**16**	**47**	348	phrase type Y
253	**139**	355	**51**	**55**	219	538	551	651	299	phrase type E
96	**84**	**73**	**28**	**25**	**64**	**59**	**89**	**66**	335	phrase type YC
			201	523	325	496	463	643	654	phrase type F
31	286	173	**33**	**79**	**17**	186	337	**82**	237	first phrase
351	**40**	217	169	**147**	**62**	**52**	**39**	322	**120**	phrase length [syl]
88	601	553	**69**	229	214	**58**	189	369	447	short phrase
60	**54**	**14**	**4**	368	524	529	525	639	473	stress [E]
575	271	**57**	**8**	**43**	420	311	350	382	598	stress [1]
526	240	**85**	**35**	**122**	409	613	367	517	413	stress [2]
433	434	302	**157**	609	399	468	174	332	695	stress [3]
277	262	203	**23**	465	190	403	679	376	421	stress [4]
600	170	**92**	**7**	184	393	336	341	540	478	unstressed

Table C.7: *Input factors for German F_0 control.*

syllable										Factor
-3	-2	-1	0	1	2	3	4	5	6	
130	**192**	**32**	**67**	**162**	463	504	559	649	350	phrase boundary 0
403	**242**	**44**	**12**	**14**	481	**90**	**154**	614	**336**	phrase boundary 1
104	**108**	466	**86**	**172**	**180**	**65**	**91**	455	365	phrase boundary 2
57	**166**	405	**194**	**22**	**185**	589	650	437	**126**	phrase type P
158	**105**	**39**	**56**	**148**	**63**	**112**	**306**	**132**	**41**	phrase type T
114	**74**	**134**	**26**	**23**	**280**	**264**	**115**	460	**260**	phrase type S
43	**10**	**1**	**6**	**8**	**36**	**174**	**16**	**143**	**252**	phrase type Y
116	**62**	**111**	**15**	**11**	**48**	**64**	**46**	439	390	phrase type E
82	**20**	**88**	**19**	**30**	**37**	**300**	**68**	**29**	**178**	phrase type YC
			110	**99**	433	519	574	602	**339**	phrase type F
377	**81**	**24**	**2**	**159**	**33**	343	387	488	532	first phrase
190	**137**	358	**61**	**28**	**42**	**156**	**113**	**181**	**97**	phrase length [syl]
319	474	**72**	**106**	**142**	**21**	**234**	**59**	**276**	**98**	short phrase
381	**25**	**147**	**40**	**71**	472	393	**216**	**291**	**183**	stress [E]
256	**198**	**7**	**5**	**54**	583	**179**	**182**	**203**	616	stress [1]
327	**151**	**92**	**27**	**51**	**294**	432	421	478	**207**	stress [2]
573	349	**76**	**38**	**186**	512	**233**	352	**217**	**314**	stress [3]
375	489	**103**	454	435	**124**	431	580	499	531	stress [4]
258	**312**	**18**	**9**	**94**	**282**	**263**	544	**341**	457	unstressed

Table C.8: *Input factors for French F_0 control.*

German					Factor	French				
syllable						syllable				
-2	-1	0	1	2		-2	-1	0	1	2
561	510	437	346	258	nucleus 1 phone	646	534	**131**	528	571
628	563	**53**	547	**135**	nucleus 2 phones	553	380	442	520	523
					nucleus >2 phones					
295	**115**	206	**128**	**41**	onset size	**80**	464	**340**	395	**315**
235	**130**	**90**	**83**	**77**	coda size	**302**	**53**	**77**	**96**	**152**

Table C.9: *Input factors for German and for French F_0 control.*

German					Factor	French				
syllable						syllable				
-2	-1	0	1	2		-2	-1	0	1	2
429	297	207	168	485	Nuc[1] long	**269**	**73**	**160**	386	**309**
566	689	587	701	653	Nuc[1] nasal	**146**	**58**	424	**100**	**219**
					Nuc[1] trill					
500	655	623	691	669	Nuc[1] approximant	587	599	576	548	620
597	438	602	686	615	Nuc[1] lateralapproximant					
607	678	644	692	554	Nuc[1] bilabial					
584	664	519	626	690	Nuc[1] alveolar					
484	650	618	675	668	Nuc[1] palatal	550	493	630	586	506
706	704	705	680	707	Nuc[1] velar	461	537	444	382	**313**
537	498	257	212	358	Nuc[1] front(closing)	482	**303**	**144**	**320**	533
665	501	631	670	619	Nuc[1] central/centring	615	565	**222**	**335**	485
548	343	165	166	199	Nuc[1] back(closing)	**299**	**167**	**274**	**155**	561
555	280	404	415	245	Nuc[1] close	490	357	**149**	645	**278**
546	585	567	684	632	Nuc[1] close-mid	401	**170**	**244**	**266**	427
424	405	274	160	252	Nuc[1] open-mid	**250**	**295**	436	**176**	**332**
457	195	183	423	442	Nuc[1] open	353	**245**	**125**	**187**	447
595	256	505	608	560	Nuc[1] low F1	402	**285**	**66**	633	**289**
194	408	208	272	361	Nuc[1] middle F1	**215**	471	**268**	**206**	**161**
379	381	356	416	458	Nuc[1] high F1	**95**	347	**275**	**129**	**326**
443	310	314	392	386	Nuc[2] long	**257**	**118**	**60**	522	**229**
573	685	583	700	648	Nuc[2] nasal	430	**45**	**17**	**70**	**52**
					Nuc[2] trill					
					Nuc[2] approximant	675	564	657	562	658
596	445	610	681	617	Nuc[2] lateralapproximant					
611	677	634	694	564	Nuc[2] bilabial					
578	662	518	630	696	Nuc[2] alveolar					
					Nuc[2] palatal	665	578	652	568	660
					Nuc[2] velar					
606	395	687	470	414	Nuc[2] front(closing)	345	366	**85**	**169**	473
649	489	621	666	579	Nuc[2] central/centring	592	535	**195**	**308**	396
449	275	495	187	222	Nuc[2] back(closing)	479	378	**136**	**249**	**230**

Table C.10: *Input factors for German and for French F_0 control.*

German					Factor	French				
syllable						syllable				
-2	-1	0	1	2		-2	-1	0	1	2
572	645	198	480	227	Nuc[2] close	541	487	359	452	356
676	200	338	674	569	Nuc[2] close-mid	486	492	**101**	**141**	**293**
276	**107**	284	467	**91**	Nuc[2] open-mid	582	448	**317**	364	605
431	254	**127**	588	574	Nuc[2] open	**138**	**119**	413	469	**189**
702	391	385	400	558	Nuc[2] low F1	558	**238**	**292**	368	497
232	**131**	461	417	418	Nuc[2] middle F1	**235**	529	**164**	451	**329**
599	171	471	699	535	Nuc[2] high F1	397	399	383	**267**	637
660	557	**116**	539	594	Onset[1] consonant	549	539	**165**	**232**	634
233	661	226	592	562	Onset[1] affricate	683	678	687	679	670
331	406	624	545	300	Onset[1] glottal closure	659	668	**262**	627	636
119	**144**	530	**95**	247	Onset[1] preplosive pause	385	394	**213**	**87**	498
106	**46**	**2**	**6**	211	Onset[1] voiced	**202**	**47**	**4**	**3**	**121**
714	713	715	716	711	Onset[1] strong					
441	372	342	**74**	**97**	Onset[1] plosive	**196**	**328**	**316**	**243**	598
380	534	436	**142**	324	Onset[1] nasal	468	426	456	**225**	606
298	**121**	371	353	340	Onset[1] trill	411	354	342	**304**	556
					Onset[1] tap					
612	290	330	251	215	Onset[1] fricative	**123**	406	**310**	369	551
					Onset[1] lateralfricative					
					Onset[1] approximant					
75	**138**	**103**	**72**	191	Onset[1] lateralapproximant	**184**	370	**157**	**279**	484
37	202	479	482	506	Onset[1] bilabial	**283**	**208**	360	362	445
321	**105**	193	**29**	486	Onset[1] labiodental	**231**	**128**	**117**	**205**	588
					Onset[1] dental					
265	**113**	**151**	464	487	Onset[1] alveolar	517	480	**197**	428	572
282	533	**67**	178	571	Onset[1] postalveolar	494	458	603	527	577
					Onset[1] retroflex					
663	521	688	366	580	Onset[1] palatal	673	626	638	656	621
590	373	164	402	531	Onset[1] velar	**50**	410	**78**	**241**	434

Table C.11: *Input factors for German and for French F_0 control.*

German					Factor	French				
syllable						syllable				
-2	-1	0	1	2		-2	-1	0	1	2
659	460	673	698	637	Onset[1] uvular	557	495	418	**322**	552
					Onset[1] pharyngeal					
466	364	476	419	544	Onset[1] glottal	674	667	**239**	651	639
124	216	494	320	528	Onset[end] consonant	570	419	**188**	459	503
577	**158**	270	**140**	347	Onset[end] affricate	682	681	686	680	666
301	287	**145**	425	490	Onset[end] glottal closure	661	663	**240**	618	632
					Onset[end] preplosive pause					
153	454	**42**	**133**	231	Onset[end] voiced	**305**	**83**	**13**	**153**	**150**
712	709	710	717	708	Onset[end] strong	676	669	684	677	685
589	428	192	509	246	Onset[end] plosive	**324**	**255**	423	**247**	542
98	288	182	410	536	Onset[end] nasal	422	415	443	351	566
86	205	255	175	**155**	Onset[end] trill	524	361	**273**	510	**288**
					Onset[end] tap					
532	328	181	**126**	278	Onset[end] fricative	**163**	475	**139**	**270**	470
					Onset[end] lateralfricative					
					Onset[end] approximant					
502	266	304	**48**	604	Onset[end] lateralapproximant	400	**177**	**84**	**330**	404
123	**112**	586	549	**132**	Onset[end] bilabial	**223**	507	**173**	**210**	**333**
452	**149**	394	**30**	614	Onset[end] labiodental	**193**	**311**	**290**	**281**	**107**
					Onset[end] dental					
196	249	264	522	177	Onset[end] alveolar	**284**	514	**211**	**168**	441
273	**137**	239	**22**	281	Onset[end] postalveolar	467	540	591	429	516
					Onset[end] retroflex					
640	389	642	455	581	Onset[end] palatal	671	619	644	655	622
349	427	603	236	493	Onset[end] velar	584	**227**	**75**	**120**	**214**
667	450	672	697	635	Onset[end] uvular	554	**237**	412	543	**286**
					Onset[end] pharyngeal					
426	469	334	309	520	Onset[end] glottal	672	664	**248**	654	643
543	497	499	412	627	Coda[1] consonant	641	538	593	648	505
					Coda[1] affricate					
					Coda[1] glottal closure					
439	228	**141**	294	629	Coda[1] preplosive pause	501	**79**	**199**	376	**337**
110	**94**	323	163	456	Coda[1] voiced	**109**	**102**	**140**	**69**	**175**

Table C.12: *Input factors for German and for French F_0 control.*

German					Factor	French				
syllable						syllable				
-2	-1	0	1	2		-2	-1	0	1	2
					Coda[1] strong					
244	541	446	430	483	Coda[1] plosive	384	**191**	**272**	511	**204**
383	387	161	362	345	Coda[1] nasal	**253**	597	**297**	**133**	**246**
582	459	306	**104**	475	Coda[1] trill	536	594	**224**	560	555
					Coda[1] tap					
658	197	283	**50**	474	Coda[1] fricative	530	617	355	440	515
					Coda[1] lateralfricative					
					Coda[1] approximant					
515	296	313	230	357	Coda[1] lateralapproximant	628	**135**	**228**	**218**	**226**
241	268	390	**117**	605	Coda[1] bilabial	**251**	**49**	388	579	**277**
550	365	451	**71**	388	Coda[1] labiodental	425	414	389	379	595
					Coda[1] dental					
305	360	401	370	657	Coda[1] alveolar	416	**301**	502	496	374
552	622	682	625	693	Coda[1] postalveolar	408	**265**	398	372	**221**
					Coda[1] retroflex					
444	**114**	576	204	**93**	Coda[1] palatal	625	653	612	607	631
593	**148**	**38**	647	267	Coda[1] velar	635	590	611	**296**	604
285	**152**	**108**	**146**	**109**	Coda[1] uvular	613	547	420	477	508
					Coda[1] pharyngeal					
					Coda[1] glottal					
432	329	462	263	491	Coda[end] consonant	662	629	600	601	483
225	176	516	542	620	Coda[end] affricate					
					Coda[end] glottal closure					
683	638	636	503	508	Coda[end] preplosive pause	575	624	640	585	563
291	**34**	**65**	209	**87**	Coda[end] voiced	373	**89**	**35**	**93**	**31**
					Coda[end] strong					
308	374	339	**154**	250	Coda[end] plosive	500	**325**	**338**	**259**	407
375	**101**	377	435	556	Coda[end] nasal	**209**	567	521	**331**	449
591	**159**	259	**21**	453	Coda[end] trill	518	**323**	438	391	**298**
					Coda[end] tap					
488	317	303	162	378	Coda[end] fricative	**287**	392	**236**	409	569
					Coda[end] lateralfricative					
					Coda[end] approximant					
185	243	220	507	**118**	Coda[end] lateralapproximant	526	476	**321**	509	465

Table C.13: *Input factors for German and for French F_0 control.*

German					Factor	French				
syllable						syllable				
-2	-1	0	1	2		-2	-1	0	1	2
326	**56**	327	261	559	Coda[end] bilabial	**261**	**171**	**271**	**212**	**307**
150	224	641	514	396	Coda[end] labiodental	462	346	446	**334**	581
					Coda[end] dental					
407	269	354	307	512	Coda[end] alveolar	453	**254**	**201**	609	513
656	671	633	616	703	Coda[end] postalveolar	546	**200**	**220**	596	363
					Coda[end] retroflex					
289	**61**	344	448	319	Coda[end] palatal	642	647	610	608	623
221	179	**125**	**80**	397	Coda[end] velar	367	371	348	417	491
312	**44**	**156**	**70**	**63**	Coda[end] uvular	545	450	344	525	**318**
					Coda[end] pharyngeal					
					Coda[end] glottal					

Table C.14: *Input factors for German and for French F_0 control.*

German	Factor	French
180	sentence length [syl]	**122**
24	short sentence	**55**
68	pos within 0 boundaries	127
45	pos within 1 boundaries	34
218	pos within 2 boundaries	145

Table C.15: *Input factors for German and for French F_0 control.*

Appendix D

Perceptual Evaluation Test Sentences

In the following, the 40 German and the 40 French sentences used for the perceptual evaluation described in Section 7.4.2 are listed. In this list, foreign inclusions are bounded by brackets and are indexed according to their language either as ($_{\text{E}}$English), ($_{\text{F}}$French), or ($_{\text{G}}$German).

D.1 German Test Sentences

1. Wir wollen ($_{\text{E}}$Manager) an der Spitze, die Ideen haben, Mut und Freude.
2. Ich hätte ($_{\text{F}}$Laurence Côte) beneiden können.
3. Die Reise führte über ($_{\text{F}}$Lyon) nach ($_{\text{F}}$Vézelay).
4. "($_{\text{F}}$Le monde à l'envers)" war ein sehr guter Film.
5. Ist dir der Name "($_{\text{F}}$Antoine Duplan)" bekannt?
6. Wann kommt ($_{\text{F}}$Isabelle Candelier) denn endlich?
7. Hast du dich über das Konzert von ($_{\text{F}}$Céline Dion) gefreut?
8. Sie kennt die Werke von ($_{\text{F}}$Marcel Pagnol) sehr gut.

9. Im Kino läuft heute Abend "(F Chat noir, chat blanc)".
10. Ich verstehe die Idee von (F Jean-Claude Longet) nicht.
11. Ich habe "(F L'éternité et un jour)" bestellt.
12. Nachdem im "(F Grand Hôtel de l'Opéra)" keine Plätze mehr frei waren, wurden wir vorläufig in der Pension "(F L'Auberge des Montagnes)" untergebracht.
13. Schon zum zweiten mal organisierte er eine "(F Soirée Américaine)".
14. Erst letzte Woche sei dies bekannt geworden, meldete darauf das Magazin "(E Facts)"; und selbst die angesehene Tageszeitung "(F Le Monde)" (F parl)ierte vom "(F première fois)".
15. "(F À la recherche du temps perdu)" von (F Marcel Proust) ist stellenweise etwas langatmig, aber trotzdem sollte man es gelesen haben.
16. Ich hatte mich verlaufen und war statt im (F Boulevard du Théâtre) in der (F Avenue de St-Paul) gelandet.
17. Wenn ihr mit uns zusammen "(F Le jardin de Célibidache)" anschauen würdet, würde uns das sehr freuen.
18. Der Weg führte wie letztes Jahr durch "(F Saint Jean pied de port)".
19. Schon der zweite Teil von "(F La cage aux folles)" war viel schlechter als der Originalfilm, aber der dritte war wirklich miserabel.
20. Nach einer längeren Rundreise in der (F Bretagne) fuhren wir von (F Nantes) aus den (F Loire)-Schlössern entlang Richtung Süden und dann ganz nach Osten.
21. Kommt jemand, um den Ehemann abzuholen?
22. Peking.
23. Verwandlungen?

24. Aber nein, man schafft dieses Interesse mit den Medien.
25. Gehen oder kommen?
26. Vermittlung?
27. Nett ist hier niemand.
28. Bewunderung oder Vergessenheit?
29. Bedauernswert.
30. Weiss.
31. Trinkst du nach dem Essen am liebsten eine Tasse Tee, einen Branntwein oder einen starken Kaffee?
32. Können wir noch feststellen, wo sich das Grab Alexanders befand?
33. Lassen Sie Ihrer Fantasie freien Lauf und begeistern Sie sich für Ihre Ideen!
34. Dümmlich?
35. Beladen!
36. Die Türkei auf dem Weg zur neuen Regionalmacht?
37. Anderseits, muss Sport letztlich pädagogisch motiviert sein, um sinnvoll zu sein?
38. Vertiefungen?
39. Drittens.
40. Letztere wurden dann mit Fett, speziell Talg, zu Seife verkocht.

D.2 French Test Sentences

1. À Zurich, le (GKunsthaus) recèle bien des trésors du Moyen Âge et présente les différents courants artistiques d'Europe et des États-Unis du vingtième siècle.

2. À la mi-mars, le (E Tokyo Game Show) sera l'occasion de nouvelles annonces pour ces "(E world game companies)".

3. Avec trois mois de retard, les troupes viennent de recevoir une lettre signée Martine (G Brunschwig Graf), conseillère d'État, responsable du DIP.

4. Le cinéma demeure, au pays de (G Fritz Lang) et de (G Fassbinder), une variante un peu archaïque de l'audiovisuel, sauf s'il s'agit de productions hollywoodiennes.

5. Parmi ces musiciens travaillent Jacques (G Wildberger), (G Klaus Huber), (G Jürg Wyttenbach), (G Rudolf Kelterborn) et (G Ernst Pfiffner).

6. À (G Buch am Irchel), dans le canton de Zurich, on ressent encore les liens avec la terre, mais aussi la liberté des grands espaces.

7. Au cas où cela se produirait, tout le monde déménagerait de bon coeur à la (G Potsdamer Platz), l'an prochain, pour célébrer le cinquantième Festival de Berlin.

8. C'est dans ce climat qu'ont émergé (G Max Frisch) et (G Friedrich Dürrenmatt) qui, par la force de leur interrogations, ont marqué pour longtemps le théâtre de leur temps.

9. Toutes les voies convergent vers le Vieux Port, point de ralliement obligé des visiteurs où se pressent cafés et restaurants proposant bouillabaisse et autres spécialités de poisson.

10. Mange!

11. L'abonnement mensuel est particulièrement avantageux.

12. Maman a préparé une galette pour jeudi.

13. Qui peut répondre à cette demande urgente?

14. Son maître saisit un bâton.

15. Cinq, quatre, trois, deux, un, zéro.

16. La jeune fille se peigne devant sa glace.

17. Signe.
18. Oui?
19. Je n'en veux aucun!
20. Il est désormais accablé par son travail.
21. C'est pas trop tôt!
22. Une dernière nouvelle?
23. Bleu ou blanc?
24. Est-ce pour toi une question de principe ou d'habitude?
25. Les soldats reçoivent une instruction militaire, ils font l'exercice, ils font aussi des manoeuvres, ils montent la garde.
26. Elle ou lui?
27. Connaissez-vous ce rapport et ne pensez-vous pas qu'il serait intéressant de le consulter?
28. Le soir et le week-end, le centre est déserté.
29. J'aimerais bien vous y voir!
30. La modestie n'est pas la moindre de mes qualités.
31. A New York, vous n'avez pas encore de la neige en ce moment?
32. Vous pourriez me dire ce qu'il y a à la télévision ce soir?
33. Quand?
34. "Proposez-vous des spécialités cruelles telles que cuisses de grenouille, foie gras, potage aux ailerons de requin, crustacés et truites gardées en bassin?" leur a-t-il demandé.
35. En épidémiologie, par exemple, on commence à s'apercevoir que, faute d'avoir expérimenté certains médicaments sur les femmes, on est parfois arrivé à des résultats aberrants.

36. A l'issue d'un processus de quelques mois, le pouvoir nigérian revient aux civils.

37. Philosophie.

38. Après ce que tu as vécu récemment, considères-tu le mariage comme encore possible ou as-tu abandonné tout espoir?

39. Pour la première fois à Genève depuis 1990, aucune troupe de théâtre, danse ou musique n'a été choisie pour recevoir une aide du canton.

40. Puisqu'il nous dérange, offrons-lui un trône!

Bibliography

[AHK87] J. Allen, M. S. Hunnicutt, and D. Klatt. *From Text to Speech: The MITalk System*. Cambridge University Press, Cambridge, England, 1987.

[Att05] M. Atterer. *Experiments on the Prediction of Prosodic Phrasing for German Text-to-Speech Synthesis*. PhD thesis, Univ. Stuttgart, 2005.

[Aub90] V. Aubergé. Semi-automatic constitution of a prosodic contour lexicon for the text-to-speech synthesis. In *Proceedings of the ESCA Workshop on Speech Synthesis*, pages 215–218, Autrans, France, September 1990.

[Aub92] V. Aubergé. Developing a structured lexicon for synthesis of prosody. In G. Bailly and C. Benoît, editors, *Talking Machines: Theories, Models and Applications*, pages 307–322. Elsevier North-Holland, 1992.

[Aub93] V. Aubergé. Prosody modeling with a dynamic lexicon of intonative forms: Application for text-to-speech synthesis. In *Proceedings of ESCA Workshop on Prosody*, pages 62–65, Lund, Sweden, September 1993.

[BC64] G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society*, Series B(26):211–246, 1964.

[BC95] A. W. Black and N. Campbell. Optimising selection of units from speech databases for concatenative synthesis.

In *Proceedings of Eurospeech'95*, pages 581–584, Madrid, Spain, September 1995.

[BdMF01] P. Boula de Mareüil and F. Floricic. On the pronunciation of acronyms in French and Italian. In *Proceedings of Eurospeech 2001*, pages 1923–1926, Aalborg, Denmark, September 2001.

[Bel61] R. Bellman. *Adaptive Control Processes.* Princeton University Press, 1961.

[BF90] J. Bachenko and E. Fitzpatrick. A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, 16(3):155–170, September 1990.

[BFOS84] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth, 1984.

[BH96] A. W. Black and A. Hunt. Generating F_0 contours from the ToBI labels using linear regression. In *Proceedings of ICSLP'96*, pages 1385–1388, 1996.

[BH05] G. Bailly and B. Holm. SFC: A trainable prosodic model. *Speech Communication*, 46:348–364, 2005. description of superposition of functional contours (SFC) model.

[Bie66] M. Bierwisch. Regeln für die Intonation deutscher Sätze. *Studia Grammatica*, VII:99–201, 1966.

[Bis95] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, 1995.

[BL04] A. W. Black and K. A. Lenzo. Multilingual text-to-speech synthesis. In *Proceedings of the ICASSP 2004*, Montreal, Canada, 2004.

[Bla02] A. W. Black. Perfect synthesis for all of the people all of the time. In *Proceedings of IEEE TTS Workshop*, Santa Monica, CA, 2002.

[Bol89] D. Bolinger. *Intonation and its Uses. Melody in Grammar and Discourse*. Edward Arnold, London, 1989.

[Bre96] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, August 1996.

[BS07] J. Baumberger and J. Sonnenmoser. Erkennung von Phrasengrenzen und Phrasentyp in Sprachsignalen. Semesterarbeit am TIK, ETH Zürich, Sommersemester 2007.

[Cam89] W. N. Campbell. Syllable-level duration determination. In *Proceedings of Eurospeech'89*, pages 698–701, Paris, France, September 1989.

[Cam92] W. N. Campbell. Syllable-based segmental duration. In G. Bailly and C. Benoît, editors, *Talking Machines: Theories, Models and Applications*, pages 211–224. Elsevier North-Holland, 1992.

[Cam94] W. N. Campbell. Prosody and the selection of units for concatenation synthesis. In *Proceedings of second ESCA Workshop on Speech Synthesis*, pages 61–64, Mohonk Mountain House, NY, USA, September 1994.

[CB95] W. N. Campbell and A. Black. Prosody and the selection of source units for concatenative synthesis. In J. P. H. van Santen, R. Sproat, J. Olive, and J. Hirschberg, editors, *Progress in Speech Synthesis*. Springer Verlag, 1995.

[CCL90] C. H. Coker, K. W. Church, and M. Y. Liberman. Morphology and rhyming: Two powerful alternatives to letter-to-sound rules for speech synthesis. In *Proceedings of ESCA Workshop on Speech Synthesis*, pages 83–86, Autrans, France, September 1990.

[Col91] R. Collier. Multi-lingual intonation synthesis. *Journal of Phonetics*, 19(1), January 1991.

[Cry69] D. Crystal. *Prosodic Systems and Intonation in English*, volume 1 of *Cambridge Studies in Linguistics*. Cambridge University Press, Cambridge, UK, 1969.

[CT94] W. B. Cavnar and J. M. Trenkle. N-gram based text categorization. In *3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–169, April 1994.

[CUB73] C. H. Coker, N. Umeda, and P. Browman. Automatic synthesis from ordinary English text. *IEEE Transactions on Audio and Electroacoustics*, AU-21(3):293–298, June 1973.

[DC98] A. Di Cristo. Intonation in French. In D. Hirst and A. Di Cristo, editors, *Intonation Systems: A Survey of Twenty Languages*, pages 195–218. Cambridge University Press, Cambridge, UK, 1998.

[DCDCV97] A. Di Cristo, P. Di Cristo, and J. Véronis. A metrical model of rhythm and intonation for French text-to-speech synthesis. In *Proceedings of ESCA Workshop on Intonation*, pages 83–86, 1997.

[Del66] P. Delattre. Les dix intonations de base du français. *French review*, 40:1–14, 1966.

[DHS01] R. O. Duda, P. E. Hart, and D. G. Storck. *Pattern Classification.* John Wiley & Sons, 2nd edition, 2001.

[Dru97] H. Drucker. Improving regressors using boosting techniques. In *Proceedings of 14th International Conference on Machine Learning*, pages 107–115, 1997.

[Dud84] *Duden "Grammatik der deutschen Gegenwartssprache",* 4. Auflage. Bibliographisches Institut. Mannheim, Wien, Zürich, 1984.

[Dud05] *Duden "Aussprachewörterbuch",* 6. Auflage. Bibliographisches Institut. Mannheim, Leipzig, Wien, Zürich, 2005.

[Dut93] T. Dutoit. *High Quality Text-To-Speech Synthesis of the French Language.* PhD thesis, Faculté Polytechnique de Mons, October 1993.

[Fri91] J. H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19(1):1–141, 1991.

[Fri01] J. H. Friedman. Greedy function approximation: A gradient boosting machine. IMS 1999 Reitz Lecture, April 2001.

[Fuj81] H. Fujisaki. Dynamic characteristics of voice fundamental frequency in speech and singing. Acoustical analysis and physiological interpretations. *STL-QPSR*, 22(1):1–20, 1981.

[Fuj83] H. Fujisaki. Dynamic characteristics of voice fundamental frequency in speech and singing. In Peter F. MacNeilage, editor, *The production of speech*, pages 39–55. Springer, New York, 1983.

[FVG+02] J. Fackrell, H. Vereecken, C. Grover, J.-P. Martens, and B. Van Coile. Corpus-based development of prosodic models across six languages. In E. and Keller, editor, *Improvements in Speech Synthesis*, chapter 16, pages 176–185. John Wiley & Sons, 2002.

[FVMVC99] J. Fackrell, H. Vereecken, J.-P. Martens, and B. Van Coile. Multilingual prosody modelling using cascades of regression trees and neural networks. In *Proceedings of Eurospeech'99*, pages 1835–1838, Budapest, Hungary, September 1999.

[Gig95] E. Giguet. Categorization according to language: A step toward combining linguistic knowledge and statistic learning. In *4th International Workshop of Parsing Technologies*, Prague, Czech Republic, September 1995.

[Gre95] G. Grefenstette. Comparing two language identification schemes. In *Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data*, pages 1–6, Rome, Italy, December 1995.

[GVC05] P. M. Granitto, P. F. Verdes, and H. A. Ceccatto. Neural network ensembles: Evaluation of aggregation algorithms. *Artificial Intelligence*, 163(2):139–162, 2005.

[GVNC02] P. M. Granitto, P. F. Verdes, H. D. Navone, and H. A. Ceccatto. Aggregation algorithms for neural network ensemble construction. In *Proceedings of SBRN 2002*, pages 178–183, 2002.

[HB96] A. Hunt and A. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of ICASSP'96*, pages 373–376, Atlanta, Georgia, USA, 1996.

[HB00] B. Holm and G. Bailly. Generating prosody by superposing multi-parametric overlapping contours. In *Proceedings of ICSLP*, pages 203–206, Beijing, China, 2000.

[HDC84] D. Hirst and A. Di Cristo. French intonation: A parametric approach. *Die Neueren Sprachen*, 83(5):554–569, 1984.

[HDCE00] D. J. Hirst, A. Di Cristo, and R. Espesser. Levels of representation and levels of analysis for the description of intonation systems. In M. Horne, editor, *Prosody: Theory and Experiment Studies*. Kluwer Academic Press, Dordrecht, 2000.

[Hos00] J.-P. Hosom. *Automatic Time Alignment of Phonemes using Acoustic-Phonetic Information*. PhD thesis, Oregon Graduate Institute of Science and Technology, May 2000.

[HPT98] K. Huber, B. Pfister, and C. Traber. POSSY: Ein Projekt zur Realisierung einer polyglotten Sprachsynthese. In *DAGA-Tagungsband, S. 392–393*, 1998.

[HSW93] B. Hassibi, D. G. Stork, and G. J. Wolff. Optimal brain surgeon and general network pruning. In *IEEE International Conference on Neural Networks*, volume 1, pages 293–299, San Francisco, CA, USA, March 1993.

[HT01] J. Häkkinen and J. Tian. N-gram and decision tree based language identification for written words. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 335–338, Italy, 2001.

[Hub90] K. Huber. A statistical model of duration control for speech synthesis. In *Proceedings of the EUSIPCO*, pages 1127–1130, Barcelona, Spain, September 1990.

[Hub91] K. Huber. *Messung und Modellierung der Segmentdauer für die Synthese deutscher Lautsprache*. Diss. Nr. 9535, Institut für Elektronik, ETH Zürich, Juli 1991.

[Jon92] L. K. Jones. A simple lemma on greedy approximation in hilbert space and convergence rates for projection pursuit regression and neural network training. *Annals of Statistics*, 20(1):608–613, 1992.

[JRHS03] D. Jones, P. Roach, J. Hartmann, and J. Setter. *Cambridge English Pronouncing Dictionary*. Cambridge University Press, Cambridge, United Kingdom, 16th edition, 2003.

[Kah76] D. Kahn. *Syllable-based Generalizations in English Phonology*. PhD thesis, MIT, 1976.

[Khm87] E. V. Khmaladze. The statistical analysis of large number of rare events. Technical Report MS-R8804, Department of Mathematical Statistics, CWI, Amsterdam, Netherlands, 1987.

[Kip66] P. Kiparsky. Über den deutschen Akzent. *Studia Grammatica*, VII:69–98, 1966.

[Kla73] D. H. Klatt. Interaction between two factors that influence vowel duration. *Journal of the Acoustical Society of America*, 54(4):1102–1004, October 1973.

[Kla79] D. H. Klatt. Synthesis by rule of segmental durations in English sentences. In Lindblom & Öhman, editor, *Frontiers of Speech Communication Research*, pages 287–299. Academic Press, 1979.

[Koh88] K. J. Kohler. Zeitstrukturierung in der Sprachsynthese. In *ITG Fachbericht, Digitale Sprachverarbeitung - Prinzipien und Anwendungen*, pages 165–170. VDE Verlag, Berlin, 1988.

[Koh90] K. J. Kohler. Improving the prosody in German text-to-speech output. In *Proceedings of the ESCA Workshop on Speech Synthesis*, pages 189–192, Autrans, France, September 1990.

[Koh91] K. J. Kohler. Prosody in speech synthesis: the interplay between basic research and TTS application. *Journal of Phonetics*, 19(1):121–138, 1991.

[Koh03] K. J. Kohler. Neglected categories in the modelling of prosody – pitch timing and non-pitch accents. In *Proceedings of 15th ICPhS*, pages 2925–2928, Barcelona, Spain, 2003.

[Koh06] K. J. Kohler. What is emphasis and how is it coded? In *Proceedings of Speech Prosody 2006*, pages 748–751, Dresden, Germany, 2006.

[Kos83] K. Koskenniemi. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production.* PhD thesis, Department of General Linguistics, University of Helsinki, Finland, 1983.

[KZ96] E. Keller and B. Zellner. A timing model for fast French. *York Papers in Linguistics*, 17:53–75, 1996.

[Lav94] J. Laver. *Principles of Phonetics*, volume 1 of *Cambridge Textbooks in Linguistics.* Cambridge University Press, Cambridge, UK, 1994.

[LC91] M. Liberman and K. Church. Text analysis and word pronunciation in text-to-speech synthesis. In S. Furui and M. Sondhi, editors, *Advances in Speech Signal Processing*, chapter 24, pages 791–831. Marcel Dekker, Inc., 1991.

[LF86] A. Ljolje and F. Fallside. Synthesis of natural sounding pitch contours in isolated utterances using hidden Markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing (ASSP)*, 34:1074–1080, 1986.

[Lip87] R. P. Lippmann. An introduction to computing with neural nets. *IEEE Transactions on Acoustics, Speech and Signal Processing (ASSP)*, 4(2):4–2, April 1987.

[MC90] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communications*, 9(5–6):453–467, December 1990.

[McA89] M. McAllister. The problems of punctuation ambiguity in fully automatic text-to-speech conversion. In *Proceedings of Eurospeech 89*, volume 1, pages 538–541, Paris, France, September 1989.

[MDM98] F. Malfère, T. Dutoit, and P. Mertens. Automatic prosody generation using suprasegmental unit selection. In *Proceedings of 3rd ESCA Workshop on Speech Synthesis*, pages 323–328, Jenolan Caves House, Australia, 1998.

[Mer93] P. Mertens. Accentuation, intonation et morphosyntaxe. *Travaux de Linguistique*, 26:21–69, 1993.

[Mer99] P. Mertens. Un algorithme pour la génération de l'intonation dans la parole de synthèse. In *Proceedings of TALN 1999*, Cargèse, July 1999.

[MJ01] H. Mixdorff and O. Jokisch. Building an integrated prosodic model of German. In *Proceedings of Eurospeech 2001*, pages 947–950, Aalborg, Denmark, September 2001.

[Möb01] B. Möbius. Rare events and closed domains: Two delicate concepts in speech synthesis. In *4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Perthshire, Scotland, 2001.

[Møl93] M. F. Møller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6(4):525–533, 1993.

[MPH93] B. Möbius, M. Pätzold, and W. Hess. Analysis and synthesis of German F_0 contours by means of Fujisaki's model. *Speech Communication*, 13:53–61, 1993.

[MQ95] F. Mana and S. Quazza. Text-to-speech oriented automatic learning of Italian prosody. In *Proceedings of Eurospeech'95*, pages 589–592, September 1995.

[MTKI96] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai. Speech synthesis using HMMs with dynamic features. In *Proceedings of ICASSP 1996*, pages 389–392, Atlanta, USA, May 1996.

[MvS96] B. Möbius and J. P. H. van Santen. Modeling segmental duration in German text-to-speech synthesis. In *Proceedings of ICSLP*, pages 2395–2398, Philadelphia, 1996.

[OBK06] P. Olaszi, T. Burrows, and K. Knill. Investigating prosodic modifications for polyglot text-to-speech synthesis. In *ISCA Tutorial and Research Workshop on Multilingual Speech and Language Processing (MultiLing 2006)*, Stellenbosch, South Africa, April 2006.

[OFWR05] J. B. Ordinas, V. Fischer, and C. Waast-Richard. Multilingual models in the IBM bilingual text-to-speech systems. In *Proceedings of Interspeech 2005*, pages 1485–1488, Lisbon, Portugal, September 2005.

[Öhm67] S. Öhman. Word and sentence intonation: A quantitative model. *STL-QPSR*, 8(2-3):20–54, 1967.

[O'S84] D. O'Shaughnessy. A multispeaker analysis of durations in read french paragraphs. *Journal of the Acoustical Society of America*, 76(6):1664–1672, 1984.

[Pie80] J. B. Pierrehumbert. *The Phonology and Phonetics of English Intonation.* PhD thesis, Massachusetts Institute of Technology, September 1980.

[Pie81] J. B. Pierrehumbert. Synthesizing intonation. *Journal of the Acoustical Society of America*, 70(4):985–995, 1981.

[Pon95] *Pons-Kompaktwörterbuch Italienisch-Deutsch,* 2. Auflage. Ernst Klett Verlag, 1995.

[PR03] B. Pfister and H. Romsdorfer. Mixed-lingual text analysis for polyglot TTS synthesis. In *Proceedings of Eurospeech'03*, pages 2037–2040, Geneva, Switzerland, September 2003.

[PW80] F. C. N. Pereira and D. H. D. Warren. Definite clause grammars for language analysis - a survey of the formalism and a comparison with augmented transition networks. *Artificial Intelligence*, 13:231–278, 1980.

[Rab89] L. Rabiner. Tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, pages 257–286, February 1989.

[RDB02] G. Rätsch, A. Demiriz, and K. P. Bennett. Sparse regression ensembles in infinite and finite hypothesis spaces. *Machine Learning*, 48:189–218, February 2002.

[RHW86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. *Parallel Distributed Processing*, 1:318–362, 1986.

[Rie95] M. Riedi. A neural network-based model of segmental duration for speech synthesis. In *Proceedings of Eurospeech'95*, pages 599–602, September 1995.

[Rie97] M. Riedi. Modeling segmental duration with multivariate adaptive regression splines. In *Proceedings of Eurospeech'97*, pages 2627–2630, September 1997.

[Rie98] M. Riedi. *Controlling Segmental Duration in Speech Synthesis Systems*. PhD thesis, No. 12487, Computer Engineering and Networks Laboratory, ETH Zurich (TIK-Schriftenreihe Nr. 26, ISBN 3-906469-05-0), February 1998.

[Ril89] M. Riley. Some applications of tree-based modelling to speech and language. In *Proc. DARPA Speech and Natural Language Workshop*, pages 339–352, Cape Cod, Mass., 1989.

[Ril92] M. Riley. Tree-based modelling of segmental durations. In G. Bailly and C. Benoît, editors, *Talking Machines: Theories, Models and Applications*, pages 265–273. Elsevier North-Holland, 1992.

[Roa91] P. Roach. *English Phonetics and Phonology.* Cambrigde University Press, Cambrigde, 1991.

[RP04] H. Romsdorfer and B. Pfister. Multi-context rules for phonological processing in polyglot TTS synthesis. In *Proceedings of Interspeech 2004 – ICSLP*, pages 737–740, Jeju Island, Korea, October 2004.

[RP05] H. Romsdorfer and B. Pfister. Phonetic labeling and segmentation of mixed-lingual prosody databases. In *Proceedings of Interspeech 2005*, pages 3281–3284, Lisbon, Portugal, 2005.

[RP06] H. Romsdorfer and B. Pfister. Character stream parsing of mixed-lingual text. In *ISCA Tutorial and Research Workshop on Multilingual Speech and Language Processing (MultiLing 2006)*, Stellenbosch, South Africa, April 2006.

[RP07] H. Romsdorfer and B. Pfister. Text analysis and language identification for polyglot text-to-speech synthesis. *Speech Communication*, 49(9):697–724, September 2007.

[RPB05] H. Romsdorfer, B. Pfister, and R. Beutler. A mixed-lingual phonological component which drives the statistical prosody control of a polyglot TTS synthesis system. In S. Bengio and H. Bourlard, editors, *Machine Learning for Multimodal Interaction*, pages 263–276. Springer-Verlag Heidelberg, January 2005.

[Rus90] T. Russi. *A Framework for Syntactic and Morphological Analysis and its Application in a Text-to-Speech System.* PhD thesis, No. 9328, Electronics Laboratory, ETH Zurich, December 1990.

[Rus92] T. Russi. A framework for morphological and syntactic analysis and its application in a text-to-speech system for German. In G. Bailly and C. Benoît, editors, *Talking Machines: Theories, Models and Applications*, pages 163–182. Elsevier North-Holland, 1992.

[Sag90] Y. Sagisaka. On the prediction of global F_0 shape for japanese text-to-speech. In *Proceedings of ICASSP'90*, pages 325–328, 1990.

[SBP+92] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wigthman, P. Price, J. Pierrehumbert, and J. Hirschberg. ToBI: A standard for labeling English prosody. In *Proceedings of ICSLP*, volume 2, pages 867–870, 1992.

[SCGC96] R. Sproat, S. Chilin, W. Gale, and N. Chang. A stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics*, 3(22):377–404, 1996.

[Sch90] R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

[Sch91] J. C. Schmitt. Trigram-based method of language identification. U.S. Patent number: 5062143, October 1991.

[Sch01] M. Schröder. Emotional speech synthesis - a review. In *Proceedings of Eurospeech 2001*, pages 561–564, Aalborg, Denmark, September 2001.

[Sch08] M. Schröder. Expressive speech synthesis: Past, present, and possible futures. In J. Tao and T. Tan, editors, *Affective Information Processing*. Springer, 2008.

[Sel81] E. O. Selkirk. On prosodic structure and its relation to syntactic structure. In T. Fretheim, editor, *Nordic Prosody II: Papers from a Symposium*, pages 111–140, Trondheim, 1981.

[SG89] M. S. Scordilis and J. N. Gowdy. Neural network based generation of fundamental frequency contours. In *Proceedings of ICASSP'89*, pages 219–222, 1989.

[SK06] T. Schultz and K. Kirchhoff. *Multilingual Speech Processing*. Elsevier, 2006.

[Spr97] R. Sproat, editor. *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*. Kluwer Academic, Dordrecht, 1997.

[SR96] A. Stahlberger and M. Riedmiller. Fast network pruning and feature extraction using the unit-OBS algorithm. In *Neural Information Processing Systems*, pages 655–661, Denver, Colorado, USA, December 1996.

[Tay94] P. Taylor. The rise/fall/connection model of intonation. *Speech Communication*, 15(1-2):169–186, October 1994.

[Tay00] P. Taylor. Analysis and synthesis of intonation using the Tilt model. *Journal of the Acoustical Society of America*, 107(3):1697–1714, 2000.

[TB94] P. Taylor and A. Black. Synthesizing conversational intonation from a linguistically rich input. In *Proceedings of ESCA Workshop on Speech Synthesis*, pages 175–178, Mohonk, New York, USA, 1994.

[TBNA98] J. Trouvain, W. J. Barry, C. Nielsen, and O. Andersen. Implications of energy declination for speech synthesis. In *Proceedings of SSW3*, pages 47–52, Jenolan Caves House, Australia, November 1998.

[THN+99] C. Traber, K. Huber, K. Nedir, B. Pfister, E. Keller, and B. Zellner. From multilingual to polyglot speech synthesis. In *Proceedings of Eurospeech'99*, pages 835–838, September 1999.

[THRJ02] J. Tian, J. Häkkinen, S. Riis, and K. J. Jensen. On text-based language identification for multilingual speech recognition systems. In *Proceedings of ICSLP 2002*, Denver, Colorado, USA, September 2002.

[TMMK02] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi. Multi-space probability distribution HMM. *IEICE Transactions on Information and Systems*, E85-D(3):455–464, March 2002.

[Tra90] C. Traber. F_0 generation with a database of natural F_0 patterns and with a neural network. In *Proceedings of the ESCA Workshop on Speech Synthesis, Autrans (France)*, September 1990.

[Tra92] C. Traber. F_0 generation with a database of natural F_0 patterns and with a neural network. In G. Bailly and C. Benoît, editors, *Talking Machines: Theories, Models and Applications*, pages 287–304. Elsevier North-Holland, 1992.

[Tra95] C. Traber. *SVOX: The Implementation of a Text-to-Speech System for German.* PhD thesis, No. 11064, Computer Engineering and Networks Laboratory, ETH Zurich (TIK-Schriftenreihe Nr. 7, ISBN 3 7281 2239 4), March 1995.

[Tra96] C. Traber. Datengesteuerte Prosodiegenerierung mittels automatischer Lernverfahren. In *Fortschritte der Akustik - DAGA'96*, pages 86–89. VDE Verlag, Berlin, 1996.

[Tra97] C. Traber. Improvements of the morpho-syntactic analysis of the SVOX text-to-speech system. Projektbericht, Institut für Technische Informatik und Kommunikationsnetze, ETH Zürich, Mai 1997.

[Tra05] H. Traunmüller. Paralinguale Phänomene (paralinguistic phenomena). In U. Ammon, N. Dittmar, K. Mattheier, and P. Trudgill, editors, *SOCIOLINGUISTICS: An International Handbook of the Science of Language and Society*, chapter 76. Walter de Gruyter, Berlin/New York, 2005.

[TS04] J. Tian and J. Suontaustaa. Scalable neural network based language identification from written text. In *Proceedings of the ICASSP 2004*, Montreal, Canada, May 2004.

[Vap95] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, 1995.

[vE56] O. von Essen. *Grundzüge der hochdeutschen Satzintonation.* A. Henn Verlag, Ratingen, Germany, 1956.

[vS92] J. P. H. van Santen. Deriving text-to-speech durations from natural speech. In G. Bailly and C. Benoît, editors, *Talking Machines: Theories, Models and Applications*, pages 275–285. Elsevier North-Holland, 1992.

[vS93] J. P. H. van Santen. Timing in text-to-speech systems. In *Proceedings of Eurospeech'93*, pages 1397–1404, Berlin, Germany, September 1993.

[vS94] J. P. H. van Santen. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, 8:95–128, 1994.

[vS97] J. P. H. van Santen. Combinatorial issues in text-to-speech synthesis. In *Proceedings of Eurospeech'97*, pages 2511–2514, Rhodes, Greece, September 1997.

[vS02] J. P. H. van Santen. Quantitative modeling of pitch accent alignment. In *Proceedings of Speech Prosody 2002*, pages 107–112, Aix-en-Provence, France, April 2002.

[vSH94] J. P. H. van Santen and J. Hirschberg. Segmental effects on timing and height of pitch contours. In *Proceedings of ICSLP'94*, pages 719–722, Yokohama, Japan, September 1994.

[vSKKM05] J. P. H. van Santen, A. Kain, E. Klabbers, and T. Mishra. Synthesis of prosody using multi-level unit sequences. *Speech Communication*, 46:365–375, 2005.

[vSM00] J. P. H. van Santen and B. Möbius. A quantitative model of F_0 generation and alignment. In A. Botinis, editor, *Intonation: Analysis, Modelling and Technology*, pages 269–288. Kluwer Academic, Dordrecht, 2000.

[vSMK08] J. P. H. van Santen, T. Mishra, and E. Klabbers. Prosodic processing. In J. Benesty, M. Sondhi, and Y. Huang, editors, *Handbook of Speech Processing*, chapter 23, pages 471–487. Springer, Berlin Heidelberg, 2008.

[vSS99] J. P. H. van Santen and R. Sproat. High-accuracy automatic segmentation. In *Proceedings of Eurospeech'99*, pages 2809–2812, Budapest, Hungary, 1999.

[vSSM+97] J. P. H. van Santen, C. S. Shih, B. Möbius, E. Tzoukermann, and M. Tanenblatt. Multi-lingual duration modeling. In *Proceedings of Eurospeech'97*, pages 2651–2654, Rhodes, Greece, September 1997.

[War96] L. Warnant. *Orthographie et Prononciation en Français*. Duculot, 1996.

[Wer90] P. J. Werbos. Backpropagation through time: What it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

[YEH+02] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK Book*. Cambridge University Engineering Departement, Cambridge, 2002.

[YTM+99] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proceedings of Eurospeech'99*, pages 2347–2350, Budapest, Hungary, September 1999.

[Zel94] B. Zellner. Pauses and the temporal structure of speech. In E. Keller, editor, *Fundamentals of Speech Synthesis and Speech Recognition*, chapter 3, pages 41–62. John Wiley & Sons, 1994.

Curriculum Vitae

1973	Born on July 27 in Gmunden, Austria.
1984–1992	Gymnasium Gmunden, Austria.
1992–2001	Studies in electrical engineering at the Technical University of Vienna, Austria.
1993–2000	Studies in economics at the Vienna University of Economics, Austria.
spring 2001	Diploma in electrical engineering (Dipl. Ing.)
1998–2001	Engineer at Philips Speech Processing, Vienna, Austria.
2002–2009	Research assistant (since 2003 PhD student) with the Speech Processing Group at the Computer Engineering and Networks Laboratory, ETH Zurich.