Production and Perception of Prosodic Events — Evidence from Corpus-based Experiments

Von der Philosophisch-Historischen Fakultät der Universität Stuttgart
zur Erlangung der Würde eines Doktors der Philosophie (Dr. phil.)
genehmigte Abhandlung

Vorgelegt von

Antje Schweitzer

aus Heidenheim-Schnaitheim

Hauptberichter: Prof. Dr. Bernd Möbius
1. Mitberichter: Prof. Dr. Grzegorz Dogil
2. Mitberichter: Prof. Dr. Elmar Nöth

Tag der mündlichen Prüfung: 10. Dezember 2010

Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
2010

Erklärung

Hiermit erkläre ich, dass ich, unter Verwendung der im Literaturverzeichnis aufgeführten Quellen und unter fachlicher Betreuung, diese Dissertation selbständig verfasst habe.

(Antje Schweitzer)

Danksagung

Ich möchte mich an dieser Stelle bei den vielen Menschen bedanken, ohne deren Hilfe diese Arbeit niemals zustande gekommen wäre.

Allen voran bei meinem Doktorvater Bernd Möbius; für seine Geduld und sein Vertauen in mich; für fruchtbare Diskussionen und Kommentare und dafür, dass er mir den nötigen Freiraum gab; ganz besonders auch dafür, dass ich ihn jederzeit mit Fragen überfallen konnte, die auch noch prompt beantwortet wurden.

Bei Prof. Grzegorz Dogil, der mir immer den Rücken stärkte und mir in den Ohren lag, meine Ergebnisse zu veröffentlichen. Bei Elmar Nöth, der sich sehr kurzfristig als Gutachter zur Verfügung stellte (auch wenn ihm dafür nur ein Bier in Aussicht gestellt wurde).

Bei Michael Walsh, der diese Arbeit Korrektur gelesen hat und dabei mit vielen wertvollen inhaltlichen Kommentaren beigetragen hat (und dafür noch nicht mal ein Bier haben wollte). Bei Sabine Schulte im Walde, die mir Tipps und Skripts zur Auswertung meiner Ergebnisse gab, obwohl sie selbst genug um die Ohren hatte. Bei Bernd Schwald, der mir half, eine vernünftige mathematische Notation für die Evaluierungsmethode zu finden. Bei meiner Bürogenossin Katrin Schneider für ihre immerwährende Hilfsbereitschaft, und dass sie es mit Gelassenheit ertrug, wenn ich beim Arbeiten mit mir selbst oder mit meinem Computer redete. Außerdem bei allen anderen Kollegen der Experimentellen Phonetik, nicht zuletzt für die freundschaftliche Atmosphäre.

Besonderer Dank gilt natürlich meiner Familie; insbesondere meinem Mann, der es mir nicht übel nahm, dass ich die Abende mit dieser Arbeit statt mit ihm verbrachte.

Teile bzw. Aspekte dieser Arbeit wurden durch das BMBF im Rahmen der Projekte SmartKom und SmartWeb sowie durch die DFG im Rahmen des Projekts “Prosodieproduktion” gefördert; ohne diese Projekte wäre diese Arbeit so nicht möglich gewesen.

Contents

List of Abbreviations

ERB	Equivalent Rectangular Bandwidth
F0	Fundamental Frequency
F1	First Formant
F2	Second Formant
GToBI	German Tones and Break Indices
GToBI(S)	German Tones and Break Indices (Stuttgart version)
IMS	Institute of Natural Language Processing
IPA	International Phonetic Alphabet
LNRE	Large Number of Rare Events
MLM	Multilevel Exemplar Model
PaIntE	Parametrized Intonation Events
POS	Part-of-Speech Tag
STTS	Stuttgart-Tubingen Tagset
ToBI	Tones and Break Indices
TTS	Text-to-speech Synthesis
VOT	Voice Onset Time

List of Figures

List of Tables

Abstract

This thesis explores perception and production of prosody by way of corpus experiments. Following Dogil and Möbius (2001) I suggest to apply Guenther and Perkell's speech production model for the segmental domain (Guenther 1995; Guenther et al. 1998; Perkell et al. 2001) to the prosodic domain. Guenther and Perkell argue that the targets in speech production take the form of multidimensional regions in auditory space (Guenther 1995; Guenther et al. 1998) or auditory-temporal space (Perkell et al. 2000). Speakers establish these target regions in speech acquisition, as well as internal models for mapping from articulator reference frame to a perceptual planning reference frame. I suggest that Guenther and Perkell's model is compatible with exemplar theory, and that the target regions can be derived in an exemplar-theoretic fashion.

The key idea in exemplar theory as applied to speech (e.g. Lacerda 1995; Goldinger 1996, 1997, 1998; Johnson 1997; Pierrehumbert 2001, 2003) is that speakers have access to memory traces ("exemplars") of previously perceived instances of speech in which almost full phonetic detail is retained. Linguistic knowledge on various linguistic levels then arises from abstracting over the stored exemplars (Pierrehumbert 2001, 2003). I suggest that target regions in the sense of Guenther and Perkell are established in the same way: They are implicitly derived from the range of values that is observed for the stored exemplars in the relevant dimensions.

Applying Guenther and Perkell's model to the prosodic domain, I assume that the prosodic categories are the categories posited by GToBI(S) (Mayer 1995) in adaptation of the Tone Sequence Model (Pierrehumbert 1980) to German. As for the dimensions of the target regions pertaining to these categories, I suggest a measure of local speech rate, viz. duration z-scores, as the temporal dimension, and tonal parameters describing the shape of F0 contours related to prosodic categories, the so-called PaIntE parameters, as tonal dimensions. The duration z-scores are obtained by standardizing phone durations using phoneme-specific means and standard deviations. The tonal PaIntE parameters are derived by approximating the F0 contour in a three-syllable window around the syllable of interest using the PaIntE model (Möhler and Conkie 1998). According to Guenther and Perkell, the relevant dimensions are perceptual dimensions. To motivate the perceptual relevance of duration z-scores and

PaIntE parameters, realizations of prosodic categories in a large database are investigated by examining their distributions for each parameter. It is shown that the parameters capture well-known aspects of the realization of prosodic events, such as phrase-final lengthening related to prosodic phrases, the differences in the alignment of peaks and those between rise and fall amplitudes for the different categories as predicted by GToBI(S), the optimal alignment of peaks with syllable structure (House 1996), but also more recent findings such as the influence of vowel height on the alignment of peaks in German H*L accents (Jilka and Möbius 2007). Confidence tests confirm that for the prosodic categories, the parameter distributions observed in the corpus differ significantly. This is taken as evidence that the parameters play a role in perception.

To further motivate this claim, I show that the parameters are useful in detecting prosodic categories automatically. Exemplar theory would suggest that if all relevant perceptual dimensions are known, it should be straightforward to detect clouds corresponding to phonetic categories using clustering techniques. In this vein, Pierrehumbert (2003) reviews clustering results obtained by Kornai (1998) where clusters of F1/F2 data corresponded well to vowel categories. She posits that stable categories are characterized by "well-defined clusters or peaks in phonetic space" (Pierrehumbert 2003, p. 210). To detect clusters corresponding to prosodic categories, I conducted clustering experiments using a prosodically annotated corpus of a male speaker. For each syllable in the corpus, 29 attributes involving duration z-scores and PaIntE parameters as well as derived parameters and some additional higher-linguistic attributes were extracted. The resulting data were clustered using various clustering algorithms and various numbers of clusters.

To begin with, the experimental results show that it is in general possible to identify clusters which correspond well to prosodic categories. Furthermore, if the clusters correspond to categories, they should generalize to new data. For evaluating the generalizability of the clusterings I suggest a new procedure which evaluates clusterings on independent test data using a classification accuracy measure which models exemplar-theoretic categorization: According to exemplar theory, categorizing new instances in speech perception is based on the stored exemplars and their categories (Lacerda 1995; Johnson 1997; Pierrehumbert 2001, 2003), i.e., categorization should be possible based on the detected clusters and their categories. Two clustering algorithms, namely SimpleKMeans and FarthestFirst, perform similarly well with respect to this measure, reaching classification accuracies of slightly more than 85% on independent test data. This is clearly above the baseline of around 78%.

As for the number of clusters, these scores are reached for approximately 1600 clusters in case of SimpleKMeans, and for approximately 2000 clusters in case of FarthestFirst, suggesting that these are appropriate numbers of clusters. Such high numbers of clusters may be unexpected at first. However, a one-

to-one correspondence of clusters to phonetic categories cannot be expected (Pierrehumbert 2003, p. 211). Also, there were altogether 29 attributes used for clustering. Thus, the clusters are detected in a 29-dimensional space. Relative to the dimensionality of the clustering space, 1600 to 2000 clusters seems more appropriate than on first glance.

Finally, prosodic categorization is simulated using supervised machine learning methods to classify new exemplars based on the same parameters as in the clustering experiments, again to corroborate their perceptual relevance. Several classification algorithms yield results of approx. 78% accuracy on the word level for pitch accents, and approx. 88% accuracy on the word level for phrase boundaries, which compare very well to results reported in other recent studies, particularly to results on German. The word level accuracies for pitch accents correspond to approximately 87.5% on the syllable level, which is slightly but not dramatically better than the accuracies of around 85% obtained above for the clusterings. The classifiers generalize well to similar data of a female speaker in that they perform equally well as classifiers trained directly on the female data. In contrast to most other studies, the classifiers predict the full set of GToBI(S) labels rather than just two classes. These classifiers have been integrated into a prototype of a tool for automatic prosodic labeling. Some examples of automatic prosodic annotations produced by this tool are given to illustrate its usefulness in automatic prosodic labeling.

In summary, the main contributions of this thesis are, (i), the application of an exemplar-theoretic interpretation of Guenther and Perkell's speech production model to the prosodic domain, (ii), a set of perceptually relevant parameters which capture tonal and temporal aspects of the implementation of prosodic events, (iii), an extensive investigation of the GToBI(S) prosodic categories in terms of these parameters, (iv), a measure to evaluate the generalizability of cluster results to new data, (v), a prototype of a tool for automatic prosodic labeling.

Deutsche Zusammenfassung

Diese Arbeit beschäftigt sich mit Korpusexperimenten zur Perzeption und Produktion von Prosodie. Ich folge Dogil und Möbius (2001) und schlage vor, Guenther and Perkells Sprachproduktionsmodell für die segmentale Ebene (Guenther 1995; Guenther et al. 1998; Perkell et al. 2001) auf die Ebene der Prosodie zu übertragen. Guenther und Perkell vertreten die Meinung, dass Produktionsziele in der Sprachproduktion durch multidimensionale Zielregionen im auditorischen (Guenther 1995; Guenther et al. 1998) oder auditorisch-temporalen Raum (Perkell et al. 2000) repräsentiert werden. Sprecher erlernen diese Zielregionen beim Spracherwerb, ebenso wie interne Modelle, die es dem Sprecher erlauben, Produktionsgesten von einem artikulatorischen Referenzrahmen auf einen perzeptuellen Referenzrahmen abzubilden. Ich schlage vor, dass Guenther und Perkells Modell mit der Exemplartheorie kompatibel ist, und dass die Zielregionen mithilfe exemplartheoretischer Prozesse erlernt werden können.

Die zentrale Idee bei der Anwendung der Exemplartheorie auf Sprache (z.B. Lacerda 1995; Goldinger 1996, 1997, 1998; Johnson 1997; Pierrehumbert 2001, 2003) ist, dass Sprecher Zugriff auf Spuren sprachlicher Einheiten im Gedächtnis haben, auf sogenannte Exemplare. Es wird angenommen, dass diese Exemplare phonetische Details fast in vollem Umfang beinhalten. Linguistisches Wissen auf unterschiedlichen Ebenen entsteht dann durch Abstraktion über die gespeicherten Exemplare (Pierrehumbert 2001, 2003). Ich schlage vor, dass die Zielregionen in Guenther und Perkells Modell auf dieselbe Weise etabliert werden können: sie werden implizit durch die Bandbreiten der Werte bestimmt, die die gespeicherten Exemplare in den relevanten Dimensionen aufweisen.

Bei der Anwendung von Guenther und Perkells Modell auf die Prosodie nehme ich an, dass die prosodischen Kategorien die Kategorien sind, die GToBI(S) (Mayer 1995) in Adaption des Tonsequenzmodells (Pierrehumbert 1980) auf das Deutsche vorschlägt. Als temporale Dimension der Zielregionen für diese Kategorien schlage ich ein Maß für lokale Sprechgeschwindigkeit vor, nämlich z-transformierte Lautdauern, und als tonale Dimensionen die sogenannten PaIntE Parameter, die die Form der F0-Kontur für prosodische Kategorien beschreiben. Die z-transformierten Lautdauern ergeben sich durch Stan-

dardisierung der Lautdauern mit phonemspezifischen Mittelwerten und Standardabweichungen. Die tonalen PaIntE Parameter werden durch Approximation der F0-Kurve durch das PaIntE Modell (Möhler und Conkie 1998) in einem Drei-Silben-Fenster um die betreffende Silbe herum ermittelt.

Nach Guenther und Perkell sind die relevanten Dimensionen perzeptuelle Dimensionen. Um die perzeptuelle Relevanz der z-transformierten Lautdauern und der PaIntE Parameter zu motivieren, werden prosodische Kategorien in einer großen Datenbank hinsichtlich ihrer Distributionen für diese Parameter untersucht. Es wird gezeigt, dass die Parameter bekannte Aspekte der Realisierung prosodischer Ereignisse erfassen, wie z.B. phrasenfinale Längung im Zusammenfang mit prosodischen Phrasengrenzen, von GToBI(S) vorhergesagte Unterschiede zwischen den prosodischen Kategorien hinsichtlich der Alignierung des F0-Gipfels und der Amplituden von F0-Anstieg und F0-Fall, die optimale Alignierung der F0-Gipfel mit der Silbenstruktur gemäß House (1996), aber auch neuere Erkenntnisse wie den Einfluss der Vokalhöhe auf die Alignierung des Gipfels bei deutschen H*L Akzenten (Jilka and Möbius 2007). Konfidenztests bestätigen, dass die Parameterdistributionen für die unterschiedlichen Kategorien signifikant unterschiedlich sind. Dies wird als Hinweis darauf interpretiert, dass die Parameter in der Prosodieperzeption eine Rolle spielen.

Um diese These weiter zu erhärten, zeige ich, dass die Parameter bei der automatischen Entdeckung prosodischer Kategorien nützlich sind. Die Exemplartheorie legt nahe, dass es relativ direkt möglich sein sollte, Exemplarwolken, die phonetischen Kategorien entsprechen, mithilfe von Clusteringtechniken zu entdecken, sofern alle relevanten perzeptuellen Dimensionen bekannt sind. So diskutiert Pierrehumbert (2003) Clusteringergebnisse von Kornai (1998), bei denen Cluster in F1/F2 Daten gut den Vokalkategorien entsprechen. Sie postuliert, dass stabile Kategorien durch wohl definierte Cluster oder Maxima im phonetischen Raum charakterisiert sind (Pierrehumbert 2003, p. 210). Um Cluster zu entdecken, die den prosodischen Kategorien entsprechen, wurden in dieser Arbeit Clusteringexperimente mit Daten eines prosodisch annotierten Korpus eines männlichen Sprechers durchgeführt. Für jede Silbe im Korpus wurden 29 Attribute extrahiert, darunter z-transformierte Lautdauern und PaIntE Parameter ebenso wie daraus abgeleitete Parameter und einige zusätzliche höherlinguistische Attribute. Diese Daten wurden mit unterschiedlichen Clusteringverfahren sowie unterschiedlichen Vorgaben für die Clusteranzahl geclustert.

Zunächst einmal zeigen die Ergebnisse dieser Experimente, dass es möglich ist, Cluster zu identifizieren, die prosodischen Kategorien entsprechen. Weiterhin sollten diese Cluster, wenn sie Kategorien entsprechen, auch auf andere Daten übertragbar sein. Um die Übertragbarkeit der Clusterings zu überprüfen schlage ich eine neue Prozedur vor, die die Clusterings auf unabhängigen Testdaten mithilfe eines Maßes für die Klassifikationsgenauigkeit evaluiert, wobei dieses Maß exemplartheoretische Kategorisierung modelliert: Nach Ansicht der Exemplartheorie beruht die Kategorisierung neuer Einheiten in der Sprach-

perzeption auf den gespeicherten Exemplaren und ihren Kategorien (Lacerda 1995; Johnson 1997; Pierrehumbert 2001, 2003), d.h., eine Kategorisierung sollte mithilfe der entdeckten Cluster und ihrer Kategorien möglich sein. Zwei Clusteringalgorithmen, SimpleKMeans und FarthestFirst, liefern ähnlich gute Ergebnisse hinsichtlich dieses Maßes. Es werden Klassifikationsgenauigkeiten von etwas mehr als 85% auf unabhängigen Testdaten erreicht. Das ist deutlich über der Baseline von etwa 78%.

Was die Anzahl der Cluster betrifft, so werden diese Genauigkeiten für etwa 1600 Cluster im Fall von SimpleKMeans und für etwa 2000 Cluster im Fall von FarthestFirst erreicht, was nahe legt, dass diese Clusteranzahlen die angemessensten sind. Eine solch hohe Anzahl von Clustern mag zunächst unerwartet sein. Allerdings kann eine eins-zu-eins-Entsprechung der Cluster zu phonetischen Kategorien nicht erwartet werden (Pierrehumbert 2003, p. 211). Zudem wurden für das Clustering insgesamt 29 Attribute verwendet, d.h., die Cluster werden in einem 29-dimensionalen Raum gesucht. Relativ zu der Dimensionalität des Clusteringraums können Clusteranzahlen von 1600 bis 2000 Clustern als angemessen betrachtet werden.

Desweiteren wird die prosodische Kategorisierung mithilfe von überwachten Machine Learning-Verfahren modelliert. Dabei werden neue Exemplare anhand derselben Daten und Parameter wie bei den Clusterexperimenten klassifiziert; auch hier, um ihre perzeptuelle Relevanz zu bestätigen. Mehrere Klassifikationsalgorithmen liefern Ergebnisse von etwa 78% Genauigkeit auf Wortebene für Pitchakzente, und etwa 88% Genauigkeit auf Wortebene für Phrasengrenzen. Diese Ergebnisse können sich mit Ergebnissen aus anderen neueren Studien durchaus messen, besonders mit Ergebnissen zum Deutschen. Die Genauigkeit auf Wortebene für Pitchakzente entspricht etwa 87.5% Genauigkeit auf Silbenebene und ist somit nur wenig besser als die Genauigkeit von etwa 85%, die sich bei der Klassifikation in den Clusterexperimenten ergab. Die Klassifikatoren lassen sich auf ähnliche Daten einer weiblichen Sprecherin gut übertragen: sie liefern ebenso gute Ergebnisse wie Klassifikatoren, die direkt auf den Daten der weiblichen Sprecherin trainiert wurden. Im Gegensatz zu den meisten anderen Studien zur Klassifikation von prosodischen Ereignissen versuchen die Klassifikatoren die volle Menge der GToBI(S) Ereignisse zu erkennen, anstatt nur zwei Klassen von Ereignissen. Die Klassifikatoren wurden in den Prototyp eines Werkzeugs für automatische prosodische Annotation integriert. Zur Illustration der Qualität der automatischen Annotation werden einige Beispielannotationen, die mit diesem Werkzeug generiert wurden, besprochen.

Zusammenfassend lässt sich sagen, dass folgende Aspekte dieser Arbeit zur aktuellen Forschung im Bereich der Produktion und Perzeption von Prosodie beitragen: (i) die Anwendung einer exemplartheoretischen Interpretation von Guether und Perkells Sprachproduktionsmodell auf die Prosodie; (ii) eine Menge perzeptuell relevanter Parameter, die tonale und temporale Aspekte

der Implementierung prosodischer Ereignisse des Deutschen erfasst; (iii) eine ausführliche Untersuchung der GToBI(S) Kategorien hinsichtlich dieser Parameter; (iv) ein Maß für die Übertragbarkeit von Clusteringergebnissen auf neue Daten; und (v) der Prototyp eines Werkzeugs für automatische prosodische Annotation.

Chapter 1

Introduction

Prosody research in the past decades has been motivated by at least two different ambitions. On the one hand, phonologically oriented models (e.g. Pierrehumbert 1980; Ladd 1983; 't Hart et al. 1990; Kohler 1991) have attempted to describe prosody by identifying a finite set of linguistically meaningful prosodic events, and often by examining the linguistic functions of these events. On the other hand, the advent of speech interfaces necessitated modeling prosody in speech synthesis in order to increase naturalness, giving rise to a variety of intonation models which can be used to predict concrete F0 contours in synthesizing utterances (Fujisaki and Hirose 1984; Taylor 1998; Möhler and Conkie 1998). These two avenues of research are not mutually exclusive. For instance, the Kiel Intonation model put forth by Kohler (1991) applies parametric rules to generate F0 contours for speech synthesis from a set of five intonation events. The IPO model introduced by 't Hart et al. (1990) was also originally developed for speech synthesis. Also, Möhler and Conkie (1998) use a parametrization technique to generate F0 contours for speech synthesis, however, the input to their model is in the form of intonation events as posited by a German adaption of Pierrehumbert's (1980) model (Mayer 1995).

Phonologically oriented models of intonation which postulate a finite set of prosodic events have to deal with the variation which can be observed in realizations of these events. This variation makes it hard to specify the exact properties of prosodic events without remaining too vague. Speech technology-oriented models on the other hand do not have to model this variation—it is sufficient if the model generates one possible, valid F0 contour. Variation is not a central issue in such models, although modeling variation may increase naturalness. Thus, speech synthesis-oriented models can afford to be much more exact in specifying the properties of prosodic events, at the expense of reduced generality.

Probably the most prominent and most wide-spread exponent of a phonologically oriented model is the Tone Sequence Model. It is based on Pierrehumbert's (1980) analysis of American English intonation and posits meaningful,

categorically distinct, autonomous intonation events. The phonetic implementation of these events is described in some detail in the ToBI labeling guidelines (Silverman et al. 1992; Beckman and Ayers 1994). The names of the events are composed of two symbols L and H because they are claimed to be composed of high (H) and low (L) pitch targets. For pitch accents, a diacritic * indicates which of these targets are aligned with the stressed syllable, and for phrase accents and phrase boundaries, the symbols - and % indicate alignment with the edge of the phrase. In this way, the names of the intonation events already code some aspects of their phonetic implementation, and some more details are explicated in the labeling guidelines. However, the exact realization remains vague.[1]

More concrete descriptions of the phonetic implementation of the American English ToBI accents are given by Jilka et al. (1999). They describe the tonal realization of these ToBI events much more systematically, establishing rules for converting pitch accents and phrasal tones to concrete F0 contours. These rules determine the alignment of each target in the voiced part of the associated syllable and in their position relative to the pitch range at the point where they occur. However, these rules were intended for speech synthesis, and in this respect they describe typical phonetic implementation without claiming to cover all valid realizations.

In this thesis, I will use a speech synthesis-oriented intonation model, viz. the PaIntE model introduced by Möhler and Conkie (1998), to examine realizations of intonation events of a phonologically oriented model, viz. Mayer's (1995) adaptation of the Tone Sequence Model to German. I will also investigate temporal properties of the intonation events posited by Mayer (1995), and in this respect it is more appropriate to refer to these events as prosodic events rather than intonation events. Thus, a substantial part of this thesis will focus on the production of prosodic events, examining both tonal and temporal aspects of their implementation in a large speech corpus. In particular, I will suggest to apply a speech production model for the segmental domain (Guenther et al. 1998; Perkell et al. 2001) to the prosodic domain.

In the past fifteen years, exemplar theory has gained considerable attention in phonetic research (e.g. Lacerda 1995; Goldinger 1996, 1997, 1998; Johnson 1997; Pierrehumbert 2001, 2003). The key idea in exemplar theory is that all perceived speech is stored in memory in the form of so-called "exemplars" and that linguistic knowledge arises from speakers' abstractions over these exemplars. For instance, abstract properties of linguistic categories can be seen as the aggregate properties of all exemplars of the category. An implication is that it should be possible to model linguistic knowledge by abstracting over data

[1]Note that Pierrehumbert (1981) also introduces an algorithm to determine concrete F0 contours for speech synthesis; however, input to this algorithm is in the form of F0 target points determined by some other component, rather than in the form of the prosodic events postulated by Pierrehumbert (1980).

from large speech corpora, given that the data are represented by the perceptual properties which listeners employ for storing exemplars. Therefore, in investigating the productions of prosodic events as described above, particular attention will be paid to perception and the question of whether the investigated properties may be useful in perception in that they are sufficiently distinct for each category. I will also suggest that the speech production model proposed by Guenther and colleagues (Guenther et al. 1998; Perkell et al. 2001) is compatible with exemplar-theoretic ideas in that the target regions posited by the authors could be implicitly defined by stored exemplars.

Exemplar theory assumes that the exemplars are used in perception to categorize new events. Thus, if the tonal and temporal properties discussed in this thesis are perceptually relevant, it should be possible to build classifiers which categorize new events on the basis of these properties. It should even be possible to detect the categories automatically, since they are represented by accumulations of similar exemplars ("exemplar clouds"). In the core chapter of this thesis, I will pursue these two ideas, by modeling categorization in human perception using a speech corpus. First, I will use clustering techniques to automatically identify accumulations of similar exemplars, which then should correspond to prosodic categories. Second, I will model human categorization of prosodic events based on these data, comparing various machine learning schemes.

I will assume here that the prosodic categories are the categories posited by Mayer's (1995) adaptation of the Tone Sequence Model to German. It should be noted that even though the intonation events of the Tone Sequence Model are referred to as "intonational categories" (Beckman and Ayers 1994), their categorical status is not yet well-established. However, there is some evidence for categorical perception of American English ToBI categories (Pierrehumbert and Steele 1989; Redi 2003). For German, categorical perception of GToBI(S) boundary tones has been shown (Schneider et al. 2009), but categorical perception of pitch accents of the GToBI(S) system has not been established yet.[2] Still, the intonation events of GToBI(S) are at least good candidates for intonation categories. Moreover, there exist large corpora which have been annotated according to this standard, providing experimental material for investigating the phonetic implementation of these "candidate" categories.

In using clustering techniques to automatically identify prosodic categories, the aims are slightly different from other clustering applications. This is because in detecting prosodic categories, one does not only want to detect categories that are inherent to the data set at hand, but generally valid, universal categories that generalize to other data sets. To assess generalizability, I will

[2]However, categorical perception for two of the three peaks posited by the Kiel Intonation Model has been shown (Kohler 1991). In terms of GToBI(S), these three peaks might correspond to HH*L, H*L, and L*H accents.

introduce an evaluation method which has not been used before, at least not to my knowledge. I will argue that it is better suited to the present problem than other, more established, evaluation measures in clustering.

The experiments on modeling human categorization of prosodic events will lead to a prototype of a tool for automatic prosodic labeling. Such an application is of great interest in speech technology because manual prosodic labeling is extremely time-consuming and, additionally, notorious for its subjectivity. A tool for automatic labeling would not only speed up the labeling process decisively, it would also ensure greater objectivity.

Summing up this introduction, the main contributions of this thesis are, (i), the application of an exemplar-theoretic interpretation of Guenther and Perkell's speech production model to the prosodic domain, (ii), a set of perceptually relevant parameters which capture tonal and temporal aspects of the implementation of prosodic events, (iii), an extensive investigation of the GToBI(S) prosodic categories in terms of these parameters, (iv), a measure to evaluate the generalizability of cluster results to new data, (v), a prototype of a tool for automatic prosodic labeling.

The thesis is organized as follows. Before actually turning to the production and perception of prosody, chapter 2 will shortly review some theoretical background on perception and production in the segmental domain which is relevant for the experiments in this thesis, viz. the speech production models of Keating (1990) and Guenther and Perkell (Guenther et al. 1998; Perkell et al. 2001) as well as exemplar-theoretic models including their treatment of speech production. At the end of this chapter, a review of prosody in exemplar-theoretic models will lead over to the prosodic domain and to chapter 3, which serves to introduce the two intonation models that are combined in analyzing prosodic events in this thesis, GToBI(S) and PaIntE. Even though the PaIntE model will only be relevant for later chapters, discussing both models in the same chapter makes it easier to illustrate how PaIntE is expected to capture relevant aspects of the tonal implementation of the GToBI(S) events. The following two chapters are intended to establish the parameters that I claim to be relevant in the perception of prosody. I will first turn to temporal aspects of prosody in chapter 4 and suggest z-scores of speech segment durations as a measure of local speech rate to capture the temporal properties of the GToBI(S) events.[3] In chapter 5, I will then use the PaIntE model to examine the tonal properties of the GToBI(S) events. Having established the relevant temporal and tonal parameters, I will turn to modeling categorization and present corpus experiments on clustering and classification of GToBI(S) events in chapter 6.[4] Finally, the results of this thesis will be discussed in chapter 7.

[3]Part of the experiments discussed in this chapter have been published in Schweitzer and Möbius (2003) and Schweitzer and Möbius (2004).

[4]The results of the experiments on classification have been published in Schweitzer and Möbius (2009).

Chapter 2

Perception and production in the segmental domain

This chapter deals with current models of speech production and perception in the segmental domain. I will briefly review two models of speech production, viz. Keating's (1990) window model of coarticulation in section 2.1, and Guenther and Perkell's speech production model (Guenther et al. 1998; Perkell et al. 2001) in section 2.2. Even though the latter model is a model of production, it is strongly linked to perception because it assumes that the targets in production are perceptual targets rather than articulatory targets. I will then turn to exemplar models in section 2.3, which treat both perception and production. In particular, I will propose in section 2.3.5 that the three types of models are compatible with each other assuming a slightly modified exemplar-theoretic view of how production targets are derived.

2.1 Keating's window model of coarticulation

Keating (1990) suggests that targets in speech production are ranges of possible values, which Keating calls windows (Keating 1990, p. 455). She claims that such target windows exist for each feature value by which a segment is characterized. Segments differ in the widths of these windows, with narrow windows corresponding to little contextual variation (little coarticulation), and wide windows corresponding to much contextual variation (strong coarticulation). The width of the window for each segment is determined by the range of values observed for that segment across different contexts. Thus, this window is fixed for each segment and does not vary further across contexts. In production, speakers interpolate between consecutive windows in a way that the resulting path is a continuous, smooth contour which traverses all windows

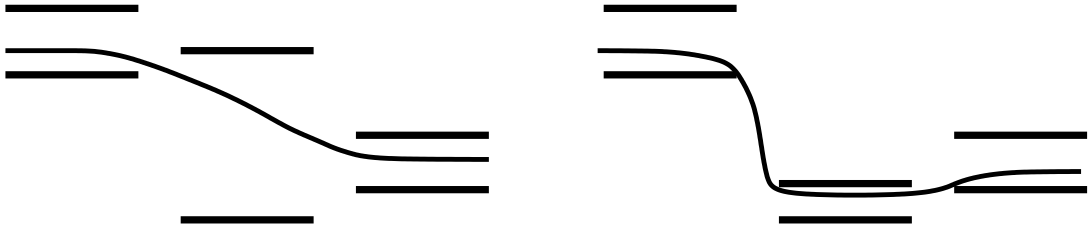

Figure 2.1: Articulation contours corresponding to two sequences of three segments each. The left and right segments in the two panels are identical. The middle segment in the left panel is subject to strong coarticulation and exhibits a wide window of possible values, while the middle segment in the right panel is less affected by coarticulation and is characterized by a much narrower window. The articulation contours are interpolated through consecutive windows to result in a continuous, smooth contour which requires minimal articulatory effort (adapted from Keating (1990, p. 457, fig. 26.1))).

but requires minimal articulatory effort.[1]

For instance, figure 2.1 displays the target windows for two sequences of three segments each. The two sequences differ only in the middle segment. In the left sequence, the middle segment is associated with a relatively wide window, while the middle segment in the right panel exhibits a much narrower window. The wider window in the left panel allows for a smooth interpolation between the target windows for the two surrounding segments which is almost unaffected by the window of the middle segment, while the narrow window in the right panel clearly contributes to the contour. Where exactly the contour traverses the window is determined by the context. In the left panel, the contour traverses the window of the middle segment from the upper limit to approximately the middle of the window because this allows for a smooth contour from the higher range of the left segment through the window of the middle segment to the lower window of the right segment. In the right panel, the contour traverses the middle segment only in the upper range of the window because the two adjacent segments' windows are higher than the one for the middle segment.

Keating (1990) exemplifies the window concept using windows in articulatory space; however, she points out that these windows may as well be determined in acoustic or perceptual dimensions, with contextual variation in one space not necessarily corresponding to variation in another space (p. 456, footnote 1). Also, she notes that for many features, there may be no 1-to-1 relation

[1] Similarly, Lindblom (1990) introduced the notion of hypoarticulation, which refers to speakers' tendency to produce speech with minimal articulatory effort. According to Lindblom (1990), hypoarticulation is counteracted by hyperarticulation, which aims at optimal perception to ensure successful communication.

between features and physical dimensions.

Byrd (1996) adapts the window model of coarticulation to account for variation in the timing of articulatory gestures as posited by articulatory phonology (Browman and Goldstein 1986, 1992). Contrary to traditional articulatory phonology, she assumes that there is not only one particular phase angle at which gestures must be coupled; instead she assumes that there exists a range of permissible values for this angle, i.e. a phase "window" in the sense of Keating (1990). Factors such as speech rate or prosodic structures are called "influencers" (Byrd 1996, p. 149) and are assumed to affect all phase windows in the same way. However, her concept of a phase window deviates from Keating's (1990) window concept in that the windows are represented by probability densities of phase angles, i.e. certain angles are more likely than other angles, and the influencers weight these densities further (Byrd 1996, pp. 150–151). Keating (1990), on the other hand, states that the windows represent an "undifferentiated range representing the contextual variability of a feature value" (Keating 1990, p. 455)—i.e. it is not intended that certain values in that range are more likely than others. In any case, Byrd (1996) demonstrates that the window concept may also be used to integrate temporal aspects with Keating's (1990) model. However, while the latter does not make any assumptions about the dimensions of the windows in speech production, as stated above, Byrd (1996) assumes gestural scores as formulated by articulatory phonology as underlying targets.

2.2 Guenther and Perkell's model

Guenther and Perkell however in a number of publications have argued that the targets in speech production take the form of multidimensional regions in auditory space (Guenther 1995; Guenther et al. 1998) or auditory-temporal space (Perkell et al. 2000). They build on Keating's (1990) window theory, positing a multidimensional region of acceptable values for each speech category and claiming that the exact trajectory through these regions is determined by economic constraints.

While Byrd's (1996) adaptation of Keating (1990) builds on articulatory phonology (Browman and Goldstein 1986, 1992) and the task dynamic model (Saltzman and Munhall 1989), Guenther and colleagues explicitly reject the idea advocated by these models that articulatory scores are the underlying targets in speech production. Their claim that the targets are expressed in an auditory reference frame, rather than a muscle length, articulator, tactile, or constriction reference frame is substantiated by several observations. Firstly, since individuals differ in muscle lengths and articulator shapes, the mapping from muscle lengths or articulator positions to vocal tract shape would have to be learned by each individual. Given the complexity of this mapping, Guenther

et al. (1998) argue that it is unclear how this should be accomplished if not with the help of auditory feedback, in which case the reference frame would be auditory. They claim that feedback on the state of the articulators, as well as tactile and proprioceptive feedback, is used in speech acquisition to establish internal models for mapping from articulator reference frame to planning reference frame, which can be used in speech production in place of perceptual feedback if this is not available (Guenther et al. 1998, p. 617–618).

Second, they show that a model which uses auditory targets for production rather than articulator targets still can exhibit stable, approximately invariant articulator configurations, as evidenced by their DIVA model (Guenther et al. 1998, p. 621). Third, perturbation experiments such as bite block and lip tube experiments show that speakers are able to use new articulator configurations to preserve phonemic identity. Experiments from a study on lip tube perturbation in the production of French /u/ indicate that these new articulator configurations do not aim at preserving vocal tract shape for /u/, which were clearly different from the normal shape for /u/ for most speakers (Guenther et al. 1998, p. 623, citing Savariaux et al. (1995)).

A further argument comes from studies on the production of American English /r/, which can be produced by different articulator configurations, not only between speakers, but also by the same speaker. These articulator configurations are often referred to as "bunched" /r/ and "retroflex" /r/. Figure 2.2, from Guenther et al. (1998, p. 627, fig. 12), illustrates that, if one assumes an articulator position target for /r/, one must assume disjoint target regions for /r/, while in acoustic-auditory space, there will be one convex target region.

Perkell et al. (2000) further substantiate how the internal models posited by Guenther et al. (1998) are used in speech production. They hypothesize that auditory feedback can not be "used for closed-loop error correction in the intrasegmental control of individual articulatory movements, because the feedback delay is too large" (Perkell et al. 2000, p. 238). In addition they observe that many people continue to speak intelligibly after total hearing loss. Thus, auditory feedback is not directly used for controlling speech production, instead, the internal model maps from vocal-tract shape, as determined by orosensory feedback and the outflow of articulator commands, to acoustic properties. To acquire (and maintain) this model, a teaching signal in the form of acoustic input is required, in addition to the feedback mentioned above, however, once the model is established, the acoustic signal is not permanently required (Perkell et al. 2000, p. 238–239).

According to Perkell et al. (2000, p. 250), the internal models determine the "phonemic settings" which serve to distinguish phonemes, while auditory feedback is necessary to make changes in the "postural settings" for suprasegmental properties such as speaking rate, mean F0, and F0 range. Data from speakers with profound hearing loss who received cochlear implants support this assumption: in the segmental domain, formant values, which are expected

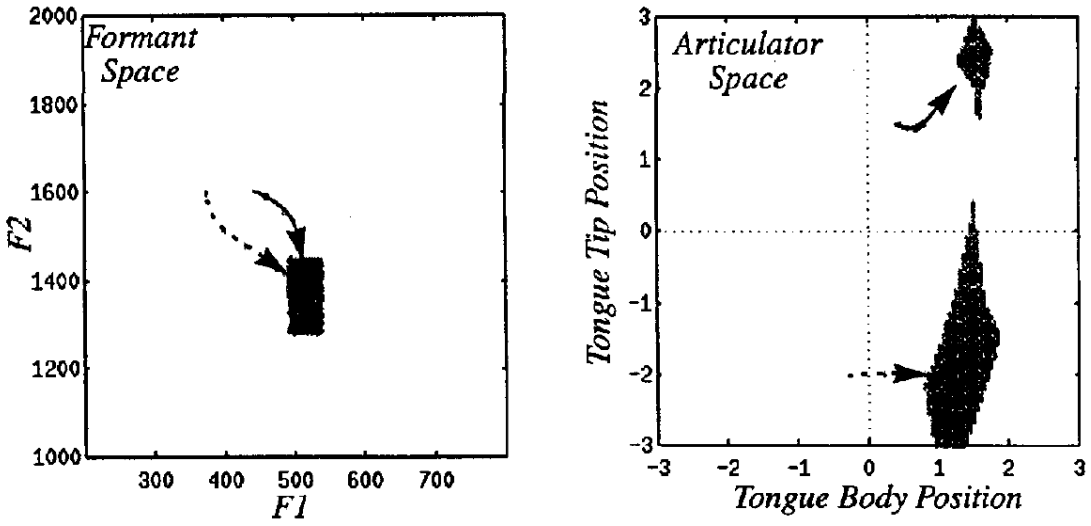

Figure 2.2: Target region for American English /r/ assuming an acoustic-auditory target region (left panel) or an articulatory target region (right panel) (reproduced from Guenther et al. (1998, fig. 12), with kind permission of the authors). The arrows indicate the direction from which the region is approached following /d/ (solid lines) or following /g/ (dashed lines).

to pertain to the phonemic settings implemented by the internal model, are relatively stable even after hearing loss and are only adapted in the few cases where deviations from the norm had occurred, while all subjects adapt the postural settings after activation of the cochlear implant Perkell et al. (2000, pp. 250–255).

Building on Guenther and Perkell's model (Guenther et al. 1998; Perkell et al. 2000), Dogil and Möbius (2001) propose that not all prosodic aspects belong to the postural settings. They hypothesize that phonologically distinctive functions of prosody as implemented by pitch accents or boundary tones are phonemic and thus are implemented in an internal model, while discourse functions belong to the postural settings. This is supported by the observation that after adult hearing loss, "text coherence (discourse and utterance intonation) is known to be lost early but intra-syllabic settings (tones, pitch accents) tend to be stable, even though the parameter F0 is involved in both domains" (Dogil and Möbius 2001, p. 667). This thesis will examine the phonemic settings of prosodic categories by establishing perceptual parameters and target regions in prosody production in chapters 4 and 5.

In applying Guenther and Perkell's speech production model to prosody, I will adopt an exemplar-theoretic interpretation of their model. In this vein, I will argue in section 2.3.5 that Guenther and Perkell's model is compatible with exemplar-theoretic models. Before doing so, I will discuss the main aspects

of exemplar theory, particularly those which are relevant for this claim in the following sections.

2.3 Exemplar theory

The key idea in exemplar theory as applied to speech (e.g. Lacerda 1995; Goldinger 1996, 1997, 1998; Johnson 1997; Pierrehumbert 2001, 2003) is that speakers have access to memory traces ("exemplars") of previously perceived instances of speech in which almost full phonetic detail is retained. Linguistic knowledge on various linguistic levels then arises from abstracting over the stored exemplars (Pierrehumbert 2001, 2003). Categorizing new instances in speech perception is based on the stored exemplars and their categories (Lacerda 1995; Johnson 1997; Pierrehumbert 2001, 2003); in speech production, production targets are derived from them (Pierrehumbert 2001, 2003).

In this section, I will describe some aspects of exemplar theory, as far as they are relevant for the experiments in this thesis. Section 2.3.1 will motivate the claim that speech units are stored in memory including much more detail than traditional models assume. Then I will explain in section 2.3.2 how categories in speech can be thought of as clouds of stored exemplars. In section 2.3.3, I will discuss exemplar-theoretic categorization, and explain the exemplar-theoretic view on speech production in section 2.3.4. I will suggest in section 2.3.5 how Guenther and Perkell's model (Guenther et al. 1998; Perkell et al. 2000), which was discussed in the preceding subsection, can be combined with an exemplar-theoretic approach. The last section of this chapter, section 2.3.6, will leave the segmental domain and address prosody in exemplar-theoretic models.

2.3.1 Storing exemplars in memory

According to exemplar theory, every instance of speech units that a listener has perceived is stored in memory, as a memory trace, or exemplar. In contrast to abstractionist views, exemplar theory assumes that much phonetic detail of these instances, including redundant detail, is retained in the exemplars. This view is supported by studies which show that subjects' performance in word recognition and word identification tasks is better for words which have been presented in the same voice before (e.g. Palmeri et al. 1993; Goldinger 1996, 1997). For these words, subjects' performance is better than for words which have been presented in another voice before. Such a voice-dependent training effect can only be explained if one assumes that not only the abstract word is stored in memory, but also details of the specific voice. These details now help to recognize or identify words that have been heard before with greater accuracy.

For instance, Goldinger (1997) let subjects identify words in noise, first in a study session, and later in a test session. The time interval between study and test session varied between 5 minutes, one day, and one week. The number of voices used was two, six, or ten. In most conditions, Goldinger (1997) observed better identification in the test session than in the study session, and the increase in identification accuracy was consistently much stronger for words which were presented in the same voice as in the study session. This shows that voice detail beyond just the abstract word form must have been stored in the subjects' memory, and this detail enhances perception.

The effect was still present when the test session occurred a week after the study session, showing that this detail must still be accessible even after a week. The effect was also present if the voice was not the same but perceptually similar to the voice in which the word had been presented in the study session: the advantage in identification was correlated with perceptual similarity as determined by multidimensional scaling.

There is no consensus yet which properties exactly are stored for each exemplar. Pierrehumbert (2003) for instance illustrates exemplar-theoretic categorization of vowels based on formant values. Similarly, Johnson (1997, p. 157) manipulates formant values in an experiment on exemplar-theoretic vowel identification. However, he himself notes that he is "being purposefully vague" about exactly which auditory properties are relevant in an exemplar model, and that he considers different properties at different points, including acoustic properties such as formant values, F0, and durations, but also critical-band activation levels, spectral templates, or auditory-based spectra (Johnson 1997, p. 149, footnote 2).

Wade et al. (2010) assume that exemplars are coded by amplitude envelopes of the speech signal across different frequency bands in their simulations modeling the selection of exemplars as targets for production. While amplitude envelopes are clearly relevant in speech perception, as evidenced by the successful application of cochlear implants for instance, I would maintain that even finer-detailed properties of speech are stored. However, Wade et al. (2010) do not claim that the amplitude envelopes are all that is stored; instead they note that these should at least contain "some of the information that is actually stored and considered by humans" (Wade et al. 2010, p. 232).

Assuming that each exemplar is stored in memory, including a remarkable amount of phonetic detail, immediately leads to the question of memory capacity. For instance, Pierrehumbert (2003, p. 180) assumes that a speaker has been exposed to about 18,000 hours of speech or 200 million words by the time he reaches adulthood. Given that each word consists of at least one, but usually several, phonemes, the number of phonemes will be in the order of one billion. Each phoneme exemplar is characterized by a number of phonetic properties that must be fine-grained enough to retain voice characteristics, formant values, F0, etc., and, as suggested by Johnson (1997, p. 151), possibly even their

variation over time in the immediate context.

Johnson (1997, p. 152) refers to this as the "head-filling-up problem". He points out that, assuming a connectionist exemplar model such as Kruschke's (1992) model, not all exemplars have to be stored as separate items. In such a model, a map represents the complete space spanned by the relevant perceptual dimensions. Due to the granularity of perception, this map consists of a finite number of locations in perceptual space. Each location is associated with each category. New instances thus will not be explicitly stored, but they will affect the association weight: the strength of the association between their location in perceptual space and their category will increase.

Very similarly, Pierrehumbert (2001) assumes that instances which can not be distinguished in perception will be stored as identical, and thus "an individual exemplar—which is a detailed perceptual memory—does not correspond to a single perceptual experience, but rather to an equivalence class of perceptual experiences" (Pierrehumbert 2001, p. 141). She assumes that perceiving instances that are not distinguishable from already stored exemplars will increase the strength of the exemplar. This has a similar effect as increasing the base activation level in Johnson's (1997) model.

Another factor that reduces the necessary storage capacity is that exemplars decay over time, i.e., the strength (Johnson 1997) or base activation level (Pierrehumbert 2001) decreases over time. Thus, more recent exemplars tend to be stronger or more activated than older exemplars.

All but one exemplar model cited in this subsection assume that exemplars are stored in some kind of multidimensional map in which the dimensions correspond to phonetic or auditory properties in speech perception, or to articulatory properties in speech production (Goldinger 1997; Johnson 1997; Pierrehumbert 2001, 2003). This entails that exemplars that are phonetically, auditorily, or articulatorily similar will be stored close to each other. Since no language exploits phonetic space evenly, the exemplars will not be evenly distributed across the map but rather form clouds or accumulations at various points.

The exception is work by Wade et al. (2010), who assume that exemplars are stored in the sequence in which they occurred, i.e., they are not stored close to phonetically similar exemplars, but close to the exemplars in their immediate context.

2.3.2 Phonetic categories in exemplar theory

Pierrehumbert (2003, p. 179) describes phonetic categories as regions in multidimensional phonetic space. She claims that in perception, this space is the perceptually encoded acoustic space, whereas in production, it can be seen as a gestural space for articulatory gestures.

Speech acquisition is then seen as acquiring for each category its probability distribution over the phonetic space (Pierrehumbert 2003, p. 184). These probability distributions are derived from the distributions of memory traces in perceptual space, i.e., from "clouds" of exemplars, which are associated with category labels. They are acquired gradually: when new exemplars are perceived, they are categorized based on the category labels of the surrounding exemplars, and then become memory traces themselves. Thus, new exemplars update the distribution of the category that they represent (Pierrehumbert 2003, pp. 185–186).

In discussing how the categories are initiated in speech acquisition, Pierrehumbert (2003) mentions results obtained by Kornai (1998), who showed that unsupervised clustering of F1/F2 data for vowels yields clusters that are "extremely close to the mean values for the 10 vowels of American English" (Pierrehumbert 2003, p. 187). She interprets theses results as supporting evidence that detection of phonetic categories in human speech acquisition may be guided by identifying regions in perceptual space which correspond to peaks in population density. She also cites experiments by Maye and Gerken (2000) and Maye et al. (2002) in which participants interpreted stimuli in a continuum as belonging to two distinct categories if the stimuli exhibited a bimodal distribution over the continuum (Pierrehumbert 2003, p. 187). In general, she assumes that "well-defined clusters or peaks in phonetic space support stable categories, and poor peaks do not." (Pierrehumbert 2003, p. 210).

However, she then concedes that, while the distributions for phoneme categories may be quite distinct for phonemes in the same contexts, there may be overlap between distributions for different phonemes in different contexts, for instance, there may be the same amount of breathiness for a vowel in one context as for /h/ in another context. She therefore suggests that "positional allophones appear to be a more viable level of abstraction for the phonetic encoding system than phonemes in the classic sense" (Pierrehumbert 2003, p. 211). This means that while the underlying distributions for phoneme categories are expected to overlap in phonetic space, the underlying distributions of positional allophones should be more clearly distinct.

2.3.3 Exemplar-theoretic categorization

According to Lacerda (1995, pp. 142-143) and Pierrehumbert (2003, pp. 205–208), new instances are categorized by comparing them to the stored exemplars. They are categorized as belonging to the category which is most frequent in the neighborhood.[2] This means that categorization is determined by the dis-

[2]In a more elaborate description of the categorization process provided in Pierrehumbert (2001), the contribution of exemplars in the neighborhood is weighted by their recency, which allows for explanation of changes in exemplar distributions as they occur in category acquisition

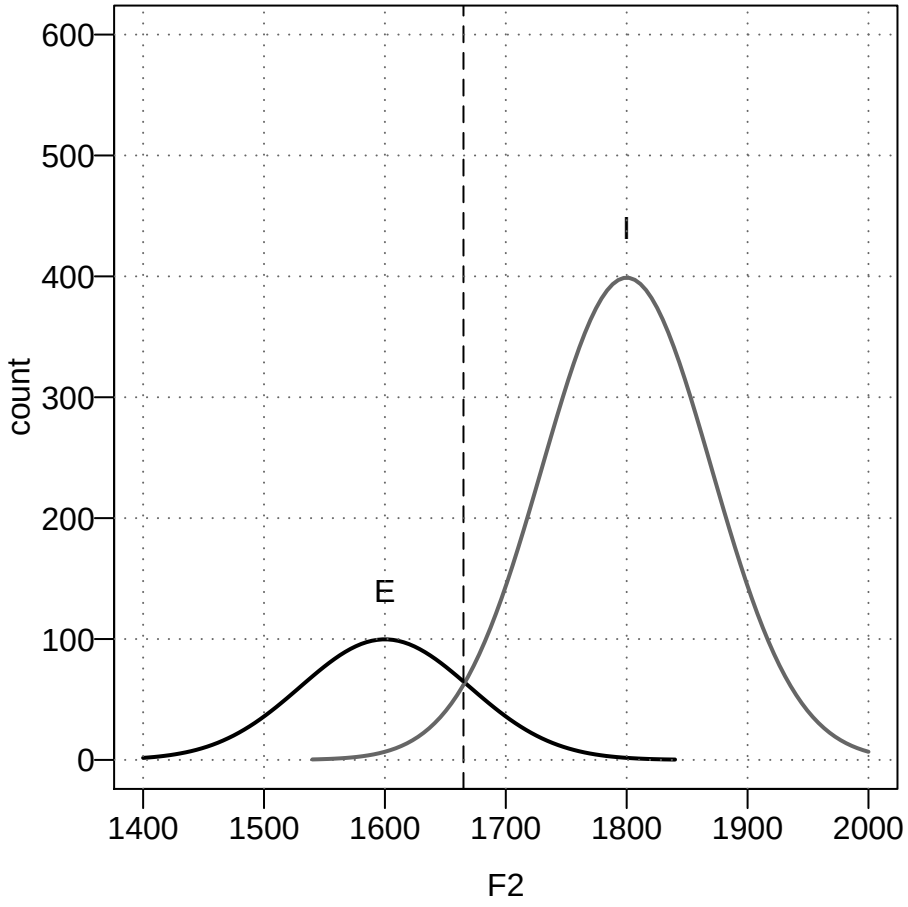

Figure 2.3: F2 distributions of stored /E/ and /I/ instances. The dashed line indicates the categorization threshold: new instances with F2 values below the threshold are categorized as instances of /E/, while new instances with F2 values above the threshold as instances of /I/ (after Pierrehumbert (2003, fig. 6.6)).

tribution of instances of competing categories, which is illustrated in figure 2.3, after Pierrehumbert (2003, fig. 6.6). For the sake of illustration, the problem is simplified to categorizing a new vowel instance based on its location in just one dimension (the F2 dimension) instead of its location in a multidimensional phonetic space. The new instance is categorized as belonging to the category to which the majority of surrounding memory traces belongs. This gives rise to a categorization threshold which is located at the point where the two distribution curves intersect. The threshold is indicated by the vertical line in figure

or in diachronic phonological processes. However, for the present thesis, assuming that only the frequencies of the surrounding exemplars are relevant for strength of activation will suffice.

2.3. New instances with F2 values below the threshold will be categorized as instances of /E/, while new instances with F2 values above the threshold will be categorized as instances of /I/.

Johnson (1997) following Nosofsky (1988) models exemplar-theoretic categorization slightly differently and assumes that new instances activate existing exemplars in the following way. Each stored exemplar has a base activation level, which is multiplied by the degree of auditory similarity between the new instance and the exemplar. Optionally, noise can be added to the resulting activation. Then, evidence for each category is the sum of activation of all exemplars of this category. In determining the auditory similarity, a sensitivity constant is employed which effectively reduces the impact of distant exemplars. Johnson (1997, p. 148) notes that this means that "the similarity function provides a sort of K nearest-neighbors classification". However, there are some differences, such as the fact that all exemplars have a base activation level. Also, in calculating the Euclidean distance, which contributes to calculating auditory similarity, attention weights are used. These weights determine to which extent each auditory property affects auditory similarity.

An advantage of the exemplar-theoretic account of categorization is that it easily models the well-known perceptual magnet effect (Kuhl 1991) without assuming that each category is represented by a so-called prototype. The magnet effect refers to the fact that subjects are better at discriminating speech stimuli that are acoustically close to category boundaries than they are at discriminating stimuli that are in the center of a category, i.e., close to the category prototype, even if the acoustic distance between the stimuli is the same in both cases. According to Kuhl (1991), this warping of the granularity of perception, or discrimination sensitivity, in the vicinity of category prototypes is caused by the (metaphoric) magnet effect of the prototype on surrounding stimuli.

As explicated by Lacerda (1995), the effect can easily be accounted for by an exemplar model. In his model, discrimination sensitivity "is based on the local variation in the number of exemplars coming from different categories" (Lacerda 1995, p. 143). It is obvious that the variation is very low in the middle of the category distribution, where all exemplars have the same category label, while it is high at boundaries between categories, thus low discrimination sensitivity is expected in the middle of the category distribution, and high sensitivity is expected at the boundaries.

Under such an account, the effect is caused by the accumulation of exemplars itself rather than by an abstract prototype. Assuming an abstract prototype necessitates a mechanism to rearrange the prototype in speech acquisition as well as in diachronic language changes. An exemplar account on the other hand does not need to assume such a mechanism: changes in the exemplar clouds provide a straightforward way to account for shifts in the location of the magnet. These changes arise because listeners are continuously exposed to new exemplars while older exemplars decay. Thus, if the newly perceived exemplars

are consistently shifted in one direction, the exemplar cloud will gradually be shifted in the same direction.

2.3.4 Production in exemplar theory

Once it has been established that exemplars are used in speech perception, it is not far-fetched to wonder whether they may also serve as targets in production. After all, as Johnson (1997, p. 153) notes, part of the stored exemplars are one's own exemplars. Since for these, the articulatory gestures could be stored along with the exemplars, they could directly be used as targets in speech production. Also, it has been discussed in section 2.2 that mappings from directions in perceptual space to directions in articulator space can be established during speech acquisition, which allows for direct use of perceptual targets rather than motor targets in speech production (Guenther et al. 1998).

In fact, there is much evidence that exemplars must play a role in speech production. Goldinger (1997, 1998, 2000) has shown in several shadowing and word naming experiments that subjects imitate phonetic detail of words that they have been exposed to before. For instance in a pilot study reported in Goldinger (1997, pp. 49–51), subjects heard words produced by ten speaker and "shadowed" them, i.e., they repeated them as quickly but clearly as possible. Before, in a baseline condition, they had read the words off a computer. The results showed that in the shadowing condition, subjects tended to adapt their pitch from their baseline pitch towards the pitch of the stimulus token, i.e., in the shadowing condition, they imitated stimulus tokens to some degree. In a post-hoc analysis, Goldinger (1997) found that imitation in terms of pitch tracking and duration matching was stronger for low-frequency words. Under an exemplar-theoretic account in which stored exemplars influence production, this effect is expected since for low-frequency words, the stimuli constitute a higher portion of the stored exemplars and should thus contribute more, while for high-frequency words, the specific details of the stimulus tokens are obscured by the many stored exemplars of that particular word.

In a more extensive and more carefully controlled follow-up study, Goldinger (1998) confirmed these results. In this later study, similarity of shadowing tokens to stimuli tokens was not assessed using phonetic measures such as pitch or duration; instead similarity was established by perception tests. The setup was as follows. Participants first read a list of words, and their renditions were recorded and saved as a baseline which productions from the actual experiment could be compared against later. In the experiment itself, listening and shadowing blocks alternated. In the listening blocks, subjects heard stimulus words between 0 and 12 times. In the shadowing blocks, they again heard the stimuli and had to repeat them as quickly as possible in half of the trials; in the other half, they had to repeat them after a delay of three to four seconds.

To assess similarity of stimulus token and shadowing token, these were presented to independent listeners in a perception test, together with the baseline tokens recorded earlier. Listeners heard AXB sequences of three tokens, with the stimulus token in the middle (X) position and the baseline and shadowing tokens balanced between A and B positions. They had to decide whether A or B was more similar to the X token, i.e, whether the test subjects' shadowing token or their baseline token was more similar to the stimulus token.

The results indicate that imitation occurred in the immediate shadowing condition. Here, listeners confirmed that the shadowing tokens were more similar to the stimulus tokens than were the baseline tokens. The proportion of shadowing tokens that was judged as more similar to the stimulus tokens increased with the number of repetitions that the shadowing subjects had been exposed to before shadowing, i.e., if subjects had heard the stimulus token more often before, the imitation effect was stronger. The proportion also increased with decreasing word frequency of the stimulus token, i.e., again, the effect was stronger for low-frequency words. This is exactly what an exemplar model would predict: the higher the proportion of exemplars of the stimulus in relation to all exemplars of that word, the stronger the effect. The proportion can either be high because the number of exemplars of that word is low in general, which is the case for low-frequency words, or because the stimulus has been presented more often.

In the delayed shadowing condition, the imitation effect was less strong. For high-frequency and medium high-frequency words, listeners did not detect shadowing tokens at above-chance level. However, for medium low-frequency and low-frequency words, the effect was present but weaker than in the immediate shadowing condition; listeners detected similarity at slightly above-chance level. The exemplar-theoretic explanation for this effect is that the stimulus token activates exemplars that are similar to it. In immediate shadowing, these will be the exemplars that influence production. However, during the delay time, the exemplars that have been activated in turn activate further exemplars that are similar to them, and this recursion continues until the token is actually produced. For the higher-frequency tokens, this will cause the produced token to be influenced by a considerable number of exemplars of that word, obscuring the contribution of the initial stimulus, while for lower-frequency tokens, where much less similar tokens exist, some of the stimulus' details will still be perceptible in the produced token.

To rule out that the imitation is an artifact of the shadowing task, i.e., to rule out that participants imitate the stimuli because they feel obliged to, or, as Goldinger (1998, p. 256) puts it, "frivolously imitate voices while shadowing", Goldinger (2000) replicated the results using a printed word naming task in which participants read words aloud one day before and seven days after a training session. In the training session, they heard stimulus words and had to identify them by clicking on a printed version of them on the screen. As in

the Goldinger (1998) experiment, stimuli differed in the number of repetitions in the training session, and in their word frequencies. The results were similar to the results in the delayed shadowing condition in the earlier experiment. They confirm an imitation effect for medium low-frequency and low-frequency words even in printed word naming seven days after exposure to the stimuli. For medium high-frequency words, the effect was quite weak and only present for stimuli that had been repeated several times. For high-frequency words, the effect was not present. Taken together, the results confirm the earlier results excluding possible artifacts of the shadowing design. They even extend the earlier results in that the effect is still present seven days after exposure to the stimuli.

Such imitation effects have also been found in two shadowing studies (Shockley et al. 2004; Nielsen 2008) in which voice onset time (VOT) of the stimuli had been manipulated. The aim of these studies was to assess whether the imitation of one specific phonetic feature can be triggered. Indeed, subjects adapted their VOTs from their baseline towards the stimuli; however, to a lesser extent than in the stimuli (Shockley et al. 2004) and only if this did not endanger phonetic contrast to other categories (Nielsen 2008).

Taken together, these studies provide convincing evidence that the exemplars play a role in speech production, in addition to their relevance for perception. Pierrehumbert (2001) provides an account of how the exemplars are used in production. In her view, the decision to produce some category activates the exemplars associated with the corresponding category label and makes them available as production targets. According to Pierrehumbert (2001, p. 145), factors such as social or stylistic context may result in selective activation of just parts of the exemplar cloud, but "the aggregate behavior of the system over all situations may be modeled as a repeated random sampling from the entire aggregate of exemplars" (Pierrehumbert 2001, p. 145). In this random selection of one particular exemplar as a production target, activation strength is considered, i.e., strongly activated exemplars are more likely to be chosen. Leaving aside other factors than frequency in determining the activation strength of each exemplar, this is equivalent to saying that frequent exemplars are more likely to be chosen.

Since no speaker has perfect motor control in production, Pierrehumbert (2001) suggests that the production process can be viewed as the production of the selected exemplar with added noise. The production result itself of course becomes part of the exemplar cloud. Adding noise in production complicates matters to some extent since this would cause the exemplar clouds to spread continuously over time. According to Pierrehumbert (2001), the spreading effect is countered by a process called entrenchment. Entrenchment is thought to model the reduction in variability which can be caused by practice. This is achieved by assuming that not a single exemplar is selected as a production target, but a region in phonetic space, and all exemplars within that region

contribute to production to the degree of their activation. Thus, the concrete production target is derived by activation-weighted averaging over a group of exemplars (Pierrehumbert 2001, p. 150). Averaging over a group of exemplars has the desired effect of reducing variability.

As for the exact specification of exemplars in production, Pierrehumbert (2001, p. 145) leaves open the question of whether the exemplars may have a dual acoustic-motor nature, i.e., whether the properties which are stored for each exemplar also include motor information, or whether the motor programs necessary to produce the acoustic targets are derived on-line. As mentioned above, Johnson (1997, p. 153) suggests that at least for those exemplars that come from one's own productions, it is conceivable that the articulatory gestures could be stored along with the exemplars. However, the results of the studies cited above clearly show that subjects' productions are influenced by productions of other speakers. Assuming that articulatory gestures are the targets in speech production, there must be a way to derive articulatory gestures from perceived instances that are not one's own productions.

A slightly different model of exemplar-theoretic production has been suggested by Wade et al. (2010), who assume that the context is taken into account in selecting exemplars for production. They assume that in producing an utterance, the corresponding sequence of segments is produced from left to right. For each segment, a segment exemplar is selected as a production target; in producing this target, noise is added, reflecting noise caused by the articulation process. The resulting production then becomes part of the exemplar cloud. In selecting the appropriate exemplar, the acoustic context to the left as well as the symbolic context (i.e., the segment categories) to the right of the segment to be currently produced are considered. Exemplars are weighted based on how well they match the current context (Wade et al. 2010, pp. 229–230).

Finally, the Multilevel exemplar model (MLM) proposed by Walsh et al. (2010) is worth noting here, in general because it presents a slightly different view on exemplar-based perception and production, and in particular because it accounts for syllable frequency effects on syllable and segment durations discussed in section 4.4.3. According to the MLM, basic constituents (e.g., segments) and more complex units (e.g., syllables) compete in perception and production. In perception and production, units are activated as well as the constituents from which they are composed. If the activation level of the unit exceeds a certain threshold, it can be used for production and perception; else, the constituents are used instead. In production, for frequent units, the high number of stored exemplars of that type will ensure that the unit level activation will usually be high enough to permit unit-level production, i.e., an exemplar of the cloud representing the unit will be randomly selected for production, while for infrequent units, the activation level will be lower, therefore a sequence of constituent-level exemplars will be assembled for production, each randomly selected from its own exemplar cloud. The concept of two competing

routes in production, upon which the MLM model is based, is discussed in more detail in section 4.4.

To summarize, there is a considerable number of studies showing that perceived exemplars are relevant for production, and a number of exemplar models which account for this. What is common to these models is that they assume that the exemplars including rich detail are stored in memory, and that they serve as targets in production, usually by random selection of one specific exemplar as a production target.

2.3.5 An exemplar version of Guenther and Perkell's model

The speech production models proposed by Keating (1990) and Guenther and Perkell (Guenther et al. 1998; Perkell et al. 2001) on the one hand and exemplar models of production as discussed in the previous section on the other hand assume slightly different views of the production process, but are by and large compatible with each other. Guenther and Perkell's model posits that the targets in prosody production are perceptual targets; Keating (1990) illustrates the window concept using windows in articulatory space but explicitly states that these windows might as well be determined in perceptual space, thus both sorts of models are compatible with the exemplar-theoretic view that production targets are defined in perceptual space.

The exemplar models assume that speakers do no abstract over the exemplars for production or perception, i.e., they do not derive prototypes for the perception of speech categories, or derive abstract multidimensional target regions for production. Instead, most exemplar models assume that single exemplars, which are assumed to be selected randomly from a subset of exemplars with high activation, serve as concrete targets in production. Abstraction occurs only implicitly, and on-line: the aggregation of activated exemplars represents a more abstract production target. Keating (1990) and Guenther and Perkell (Guenther et al. 1998; Perkell et al. 2001) however seem to posit that windows or target regions are established in speech acquisition and then are represented as abstract production targets.

However, if one assumes that in exemplar-theoretic production, the activated exemplars are not used as a pool of exemplars from which one is selected for production, but rather that they define for each dimension an acceptable range of values for the category to be produced, i.e., that they implicitly define a multidimensional target region, this is very much in accordance with Guenther and Perkell's model.

Both sorts of models, Guenther and Perkell's as well as exemplar-theoretic models, posit a strong link between perception and production. According to Guenther and Perkell's perspective, this link is implemented in the form of internal models which map from directional vectors in perceptual space to direc-

tions in articulatory space. Exemplar models of course have to assume a similar mapping. Even though Johnson (1997, p. 153) notes that for one's own exemplars, the articulatory gestures could be stored along with the exemplars, it cannot be the case that only these are used in production because productions of other speakers clearly influence speech production, as attested for instance in the results of the shadowing and word naming experiments discussed above.

Thus, exemplar models and the speech production models of Keating (1990) and Guenther and Perkell (Guenther et al. 1998; Perkell et al. 2001) are easily united if one assumes that in exemplar-theoretic production, the activated exemplars implicitly define perceptual target regions, instead of serving as a collection of single, concrete exemplars. I will assume this point of view on production throughout the present thesis. The next two chapters are intended to shed more light on the target regions associated with the GToBI(S) events under such an assumption, but before, I will address prosody in exemplar theory, proceeding from the segmental to the suprasegmental domain.

2.3.6 Prosody in exemplar models

To my knowledge, few studies have looked at prosody in an exemplar-theoretic framework. As discussed in section 2.3.4, Goldinger (1997) found that in shadowing experiments, subjects tended to adapt their pitch from their baseline pitch towards the pitch of the stimulus token, and to match the durations of their productions to the stimulus token. The effect was stronger for low-frequency words. This immediately indicates that pitch and duration are stored with the exemplars, and that these properties are retained in production. However, pitch was tracked averaging over words or stimuli, and durations were measured in terms of stimulus durations. Thus, it is not clear if the results are due to the speaker imitating prosodic categories (the "phonemic" settings proposed by Perkell et al. (2000) mentioned above) or more global aspects such as pitch range or overall duration ("postural" settings in Perkell et al.'s (2000) terminology). The former is certainly expected from an exemplar-theoretic perspective—if categories are represented as collections of stored exemplars, their defining properties must be retained when storing them, otherwise categorization would not be possible based on stored exemplars.

Further evidence for exemplar storage of prosodic categories comes from two recent studies which have found frequency effects in the realization of pitch accents in German (K. Schweitzer et al. 2009) and American English (K. Schweitzer et al., accepted). In an exemplar-theoretic framework, the difference is accounted for by the difference in the amount of stored exemplars of each pitch accent. However, an earlier study (Walsh et al. 2008) found no effect of syllable frequency on pitch accent variability.

In joint work by Calhoun and me (Calhoun and Schweitzer, accepted), we

propose that words and short phrases in American English are stored with their intonation contours, and that discourse meanings of highly frequent word-contour pairings can spread by analogy to less frequent pairings. To substantiate this, we used PaIntE (cf. section 3.2) to parametrize the contours, and calculated duration z-scores for the segments (cf. chapter 4). Representing the contours by attributes derived from these parameters, we identified 15 "typical" contours using clustering techniques, i.e., the investigated contours were more specific than just the five pitch accent types of the American English ToBI inventory (Beckman and Ayers 1994). Evidence for the storage of words together with their contours then comes from the fact that certain words and contours formed collocations, i.e., they appear together more often than would be expected based on their individual frequencies. In a perception experiment, we confirmed that the discourse meanings of the most frequent pairings spread to other word-contour pairings, which constitutes further evidence that the contours must have been lexicalized.

Taken together, these studies suggest that intonation is stored along with the exemplars; in particular, there is evidence that phonemic aspects of intonation, such as pitch accent realizations, are stored, rather than just postural aspects such as averaged pitch.

This thesis is not intended to directly address the issue of whether prosody is stored or not; rather, I assume that the exemplar key ideas are correct and, basing on this assumption, I suggest prosodic properties that are good candidates for actually being stored. These properties will be introduced in the following chapters: Chapter 3 will introduce the PaIntE parameters (Möhler and Conkie 1998), and chapter 4 will introduce duration z-scores as a measure of local speech rate. I will demonstrate in chapters 4 and 5 that these parameters are useful by systematically analyzing realizations of prosodic events in a large speech corpus and showing that they capture valid aspects of the tonal and temporal implementation of prosodic events. Further evidence for the relevance of the parameters is presented in chapter 6, where I will successfully use them to model exemplar-theoretic categorization of prosodic events.

Chapter 3

Modeling intonation

In this chapter, I will describe the two intonation models that are combined in analyzing prosodic events in this thesis. First, I will give a more detailed account of the phonologically oriented GToBI(S) in section 3.1 including detailed descriptions of the typical pitch contours of the pitch accents and boundary tones, before explicating the speech synthesis-oriented PaIntE model in section 3.2. Here, I will only address each model separately. How they can be combined to analyze the phonetic implementation of the GToBI(S) prosodic events in more detail will be addressed later in chapter 5, where I will illustrate how the prosodic events posited by GToBI(S) are realized in terms of PaIntE parameters.

3.1 GToBI(S)

The GToBI(S) (Mayer 1995) prosodic labeling system is intended to integrate Féry's (1993) analysis of German intonation and the ToBI labeling conventions (Silverman et al. 1992; Pitrelli et al. 1994). ToBI stands for "Tones and Break Indices". GToBI is abbreviated for German ToBI, and the (S) refers to the Stuttgart version of GToBI as described by Mayer (1995). What distinguishes GToBI(S) from the more wide-spread GToBI dialect originally developed in Saarbrücken (Grice and Baumann 2002; Grice et al. 2005) is that it was intended for a prosodic module for Discourse Representation Theory (Kamp and Reyle 1993) and thus its aim was to phonologically distinguish intonation events by their function in the domain of discourse interpretation (Mayer 1995, p.1). The most conspicuous consequence of this approach is probably that there is no distinction between an L+H* and an L*+H pitch accent in the GToBI(S) system.

GToBI(S) provides 5 basic types of pitch accents, L*H, H*L, L*HL, HH*L, and H*M, which are claimed to serve different functions in the domain of discourse interpretation. They can be described as rise, fall, rise-fall, early peak, and stylized contour, respectively. For each pitch accent, the tone associated

with the accented syllable is marked by the * diacritic and is called the starred tone. For high (H) starred tones, the pitch contour reaches a target high in the speaker's register, for low (L) starred tones, it reaches a target low in the speaker's register. Tones preceding the starred tones are called leading tones and are associated with the pre-accented syllable, whereas tones following the starred tones are called trail tones and are associated with the post-accented syllable (or syllables, in the case of L*HL). For high leading or trail tones, the pitch contour reaches a high target on the associated syllable, and for low leading or trail tones, it reaches a low target on the associated syllable.

There are allotonic variants of L*H and H*L in pre-nuclear contexts; in these, only the starred tone (L* in L*H, and H* in H*L) is realized on the pitch accented syllable, while the trail tone (H in L*H, and L in H*L) is realized on the syllable preceding the following pitch accent (*partial linking*), or even omitted completely (*complete linking*).

Schematic diagrams of the pitch contours of the GToBI(S) pitch accents are given in figure 3.1. Boxes represent syllables, accented syllables are highlighted in gray color. Dotted lines indicate one or several interceding syllables. The two panels in the first row display two realizations of an underlying L*H accent. On the left, the standard variant is depicted. It is characterized by a low target which is reached in the course of the accented syllable, followed by a rise to a high trail tone which is reached on the post-accented syllable. In the partial linking variant, which is depicted on the right, the contour also reaches a low target on the accented syllable, but the following high trail tone is split off and realized only immediately before the next pitch-accented syllable. The contour between low target and high trail tone is interpolated across interceding syllables (indicated by the dotted line). The trail tone is labeled by ..H on the syllable on which it is realized (this syllable is labelled as pre-accented syllable in figure 3.1). The complete linking variant is not explicitly given in the figure, but its schematic would correspond to just the left part of the complete linking diagram, with the pitch contour interpolated to the following pitch accent.

The schematics for the H*L variants given in the middle row of figure 3.1 are exactly symmetric to the L*H variants, exchanging lows for highs and highs for lows: The standard H*L depicted on the left is characterized by a high target which is reached in the course of the accented syllable, followed by a fall to a low trail tone which is reached on the post-accented syllable. In the partial linking variant, which is depicted on the right, the contour also reaches a high target on the accented syllable, but the following low trail tone is split off and realized only immediately before the next accented syllable. This trail tone is labeled by ..L. Again, the complete linking variant is not explicitly given in the figure, but its schematic would correspond to just the left part of the complete linking diagram.

The panels on the bottom row show two more pitch accents which occur less frequently. As illustrated in the bottom left schematic, for the HH*L accent,

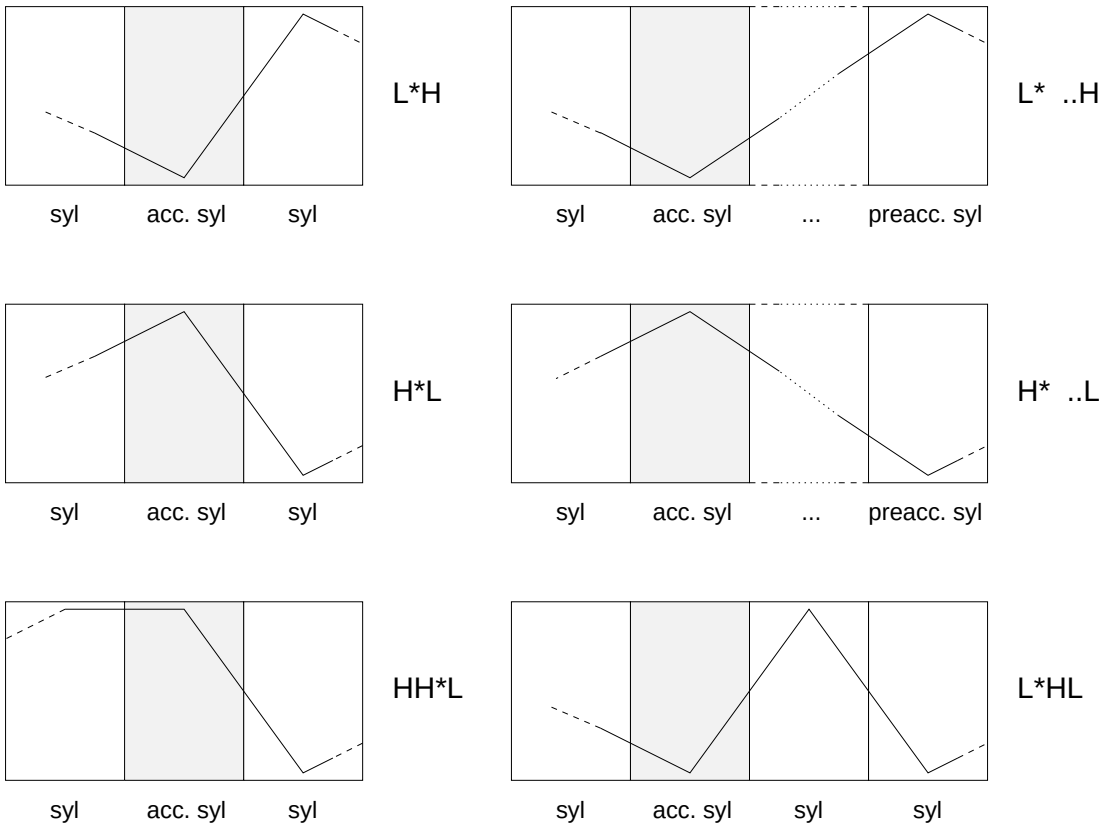

*Figure 3.1: Schematic diagrams of the pitch contours of the GToBI(S) pitch accents. Boxes represent syllables, accented syllables are highlighted in gray color. Dotted lines indicate one or several interceding syllables. Dashed lines indicate that the actual contour arises from interpolation to targets of preceding or following accents. The very rare H*M is not depicted here.*

the pitch contour reaches a high target on the syllable preceding the accented syllable. It should be noted that this syllable must be phonologically weak, i.e., unstressable. The contour continues in the upper range of the speaker's register into the accented syllable and then falls to a low target which is reached on the post-accented syllable. The bottom right panel shows a diagram of the L*HL pitch accent. Here, the pitch contour reaches a low target on the accented syllable, then rises to a high target on the post-accented syllable only to fall to a low target again on the following syllable.

If the bitonal pitch accents (L*H, H*L) occur in phrase-final position, the targets are realized within this one syllable, i.e., the complete fall or the complete rise occurs on the pitch-accented syllable. Analogously for the tritonal L*HL accent: if there is only one post-accented syllable, the high and low targets of the trail tones are realized on that syllable; if the pitch-accented syllable itself

is phrase-final, all three targets are realized on the accented syllable.

As does ToBI for American English (Silverman et al. 1992; Pitrelli et al. 1994), GToBI provides a diacritic "!" to indicate downsteps (i.e., an H* target which is realized significantly lower than a preceding H* target in the same phrase). Mayer (1995) notes, however, that although it is recommended to label downsteps, it is not clear whether the downstepped pitch accents differ in meaning from their non-downstepped counterparts (Mayer 1995, p.8).

As for prosodic phrasing, GToBI(S) distinguishes between intonation phrases and intermediate phrases. Each intonation phrase consists of one or more intermediate phrases. Intonation phrases are terminated by boundary tones, which are realized on the phrase-final syllable. There are three different boundary tones: %, H%, and L%. In the case of %, the pitch contour is determined by spreading of the preceding trail tone; i.e., it ends either low or high in the register, depending on the trail tone of the preceding pitch accent. For L%, which in German never occurs after a rising pitch accent (Mayer 1995, p. 10), the pitch contour is low before the boundary tone already and falls further below the register of the phrase. For H%, on the other hand, the pitch contour either exceeds the register following a high trail tone, or at least rises in spite of a preceding trailing L. Intermediate phrase boundaries which do not coincide with intonation phrase boundaries exhibit no obvious tonal movement and are labeled by "-".

Schematic pitch contours corresponding to the GToBI(S) boundary tones are indicated in figure 3.2. The upper panel shows contours for H% (upper line) and % (lower line) following an earlier L*H accent: in this case, the contour is high in the speaker's register even before the boundary, it rises even further for H% to end beyond the speaker's normal register, while it stays high inside the speaker's register for the tonally unspecified %. Recall that L% boundary tones do not occur after L*H in German. The lower panel shows contours for all three boundary tones following an earlier H*L accent. The same contours would be expected in all contexts where the preceding accent has a low trail tone. In the case of H%, the contour ends high in the speaker's register (upper line); for the unspecified %, it ends low in the speaker's register (middle line); the contour for L% ends even below the speaker's register (lower line).

Summarizing this section, GToBI(S) (Mayer 1995) identifies a set of prosodic events. As is evident from the fact that the names of these events are composed of the tonal targets associated with the events, GToBI(S) roughly describes the course of the pitch contour in the vicinity of these prosodic events. The PaIntE model, in contrast, allows for a more exact description of the pitch contour. I will turn to this model in the following section.

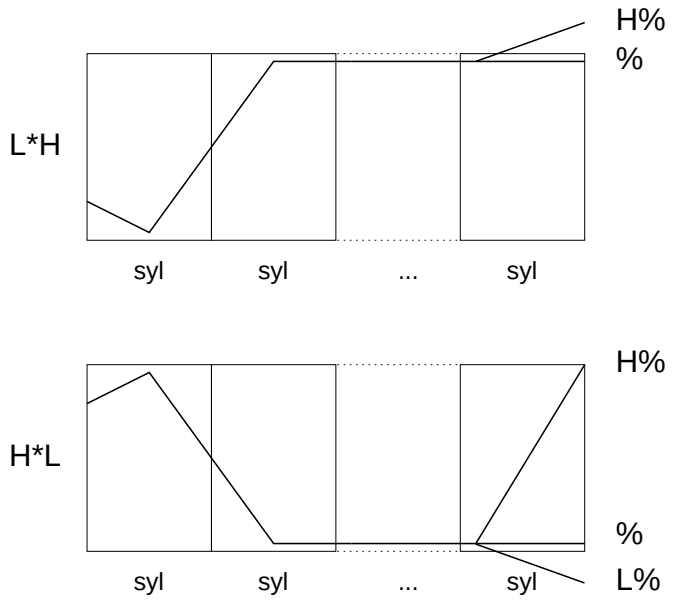

Figure 3.2: Schematic diagrams of the pitch contours of the GToBI(S) boundary tones in two different contexts: after an accent with a high trail tone (upper panel) and after an accent with a low trail tone (lower panel). Boxes represent syllables, and dotted lines indicate one or several interceding syllables.

3.2 PaIntE

I will claim in the subsequent chapters that the PaIntE parameters quantify perceptually relevant aspects of tonal contours and in this way can be regarded as the tonal dimension in prosody perception. In this vein, chapter 5 will illustrate how the prosodic events posited by GToBI(S) are realized in terms of PaIntE parameters. Here, I will shortly describe the model, and discuss some technical aspects which are not documented in the pertinent publications. Some of them are worth noting because I have changed them for the present experiments, and some might be interesting particularly with respect to future improvements of the approximation reliability.

The PaIntE model (Möhler and Conkie 1998; Möhler 2001) parametrizes intonation contours using six linguistically motivated parameters. PaIntE stands for “Parametrized Intonation Events”. It was invented by Möhler and Conkie (1998) and was originally intended for F0 modeling in speech synthesis. The basic idea was to approximate the F0 contour in a certain window around syllables that are known to carry a pitch accent or a boundary tone using a linguistically motivated approximation function. This function is composed of a rising and a falling sigmoid function. Mathematically, it is a function f of time,

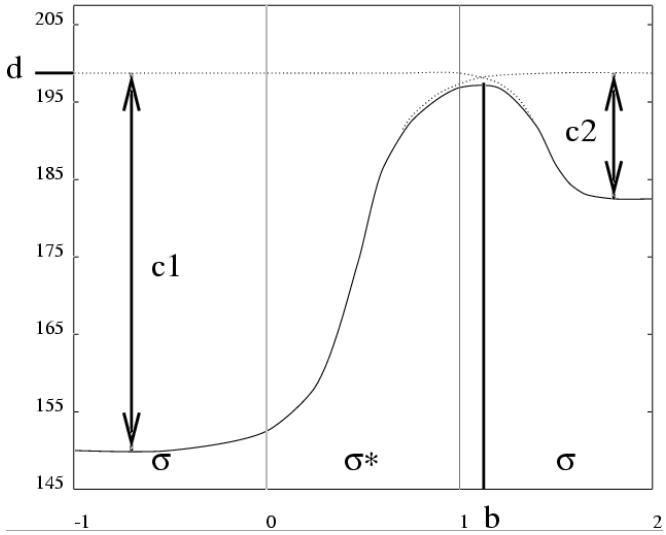

Figure 3.3: Schematic of the PaIntE approximation function reproduced from Möhler and Conkie (1998). The approximation window represents three syllables. The accented syllable is indicated by the asterisk (σ). The peak height is determined by parameter d, amplitudes of range and fall correspond to parameters* c1 *and* c2*, respectively, and the peak alignment depends on the* b *parameter. Parameters* a1 *and* a2 *represent the (amplitude-normalized) steepness of the rising and falling sigmoid.*

with f(x) approximating observed F0 values. It is defined as follows:

$$f(x) = d - \frac{c1}{1 + exp(-a1(b - x) + \gamma)} - \frac{c2}{1 + exp(-a2(x - b) + \gamma)} \tag{3.1}$$

The exact contour is determined by the six parameters *a1*, *a2*, *b*, *c1*, *c2*, and *d*. Parameters *a1* and *a2* represent the (amplitude-normalized) steepness of the rising and falling sigmoid, respectively, while *c1* and *c2* specify the amplitudes of the sigmoids. Parameter *d* can be interpreted as approximating the absolute peak height in Hertz, and *b* determines the temporal alignment of the peak. A more extensive discussion of the interpretation of the PaIntE parameters is provided in section 5.1, where I discuss each parameter separately using parameter values observed in a large prosodically annotated database for illustration.

In a more recent version of the PaIntE model (Möhler 2001), the x-axis can be normalized in different ways. Using *sylnorm* normalization, which was also used by Möhler and Conkie (1998), the time axis inside the approximation window is normalized in a way that syllable boundaries occur at integer values, with the accented syllable beginning at 0 and ending at 1. In this case,

b determines the temporal alignment of the peak in terms of relative position in the normalized duration of the syllables in the approximation window. The experiments presented in the following chapters were carried out using *sylnorm* normalization, and a schematic of the PaIntE function with *sylnorm* normalization in a three-syllable window is given in figure 3.3.

Alternatively, using *anchornorm* normalization, each syllable is split into three parts representing the (unvoiced) onset of the syllable, its sonorant nucleus, which is defined as containing the nucleus and possibly preceding voiced consonants in the onset, and, finally, the coda. Each syllable in the approximation window is then normalized to length one with the same values for syllable boundaries as in the *sylnorm* case. Syllable-internally, the onset is adjusted linearly to a length of 0.5 times the syllable duration, the nucleus to 0.3, and the coda to 0.2 times the syllable duration. However, I did not use *anchornorm* normalization in this thesis because a pre-test of the experiments presented in chapter 6 indicated better results for *sylnorm* normalization.

3.2.1 Prosodic context

In Möhler's (2001) implementation, the length of the approximation window varies between one and three syllables. It is influenced by prosodic context in three ways. First, the window is not extended beyond phrase boundaries. Second, depending on the type of pitch accent, the syllable preceding the accented syllable is only included in the approximation window for so-called "early" accents for which the peak is expected to occur relatively early. These early accents were defined to be L+H*, H+!H*, and HH*L. The former two are accents used in the Saarbrücken dialect of GToBI, while HH*L is used in the Stuttgart dialect. I have added H* and H*L as the Stuttgart variants of L+H* and H+!H* as well as their downstepped versions, and I have introduced an option to override these specifications in order to always use the preceding syllable irrespective of whether an early accent is involved or not. Finally, parametrization is carried out for syllables known to be accented only, i.e., knowledge about the location of pitch accents is a prerequisite.

However, in the application of PaIntE presented in this thesis, the idea is to examine which perceptual properties distinguish different prosodic events from each other, and which properties distinguish syllables that are not related to prosodic events from syllables that are related to prosodic events. This entails that I do not want to use any knowledge of prosodic structure in the data when extracting the PaIntE parameters.

To this end, the PaIntE source code has been modified to allow for parametrization without any assumptions about, or references to, prosodic properties derived from the prosodic labels. I have introduced an option to parametrize every syllable instead of pitch-accented or phrase-final syllables

only. In this case, PaIntE will always use a three-syllable window, irrespective of the prosodic context of the particular syllable, with the only exception that approximation windows are not extended across silences.

3.2.2 F0 smoothing

F0 contours are estimated from the speech signal using ESPS's `get_f0`. All F0 contours are smoothed before PaIntE approximation using `smooth_f0` provided with the Edinburgh Speech Tools (Edinburgh Speech Tools Library 1999). `smooth_f0` is a median smoother which interpolates across unvoiced regions, but not across silences.

3.2.3 Approximation methods

In preparing the approximation, PaIntE looks for an F0 peak (*max*) in the middle of the approximation window first, as illustrated by the upper panel of figure 3.4. It looks for this peak in a reduced start window between 0.2 and 1.4 in the *sylnorm* case. This window is indicated by the vertical dashed lines in the upper panel of figure 3.4. PaIntE then searches for minima to the left (*lmin*) and to the right (*rmin*) of *max* within this window. Then, if *max* and/or *lmin* and/or *rmin* turn out to have been found right at the boundaries of the window segment, as was the case for *lmin* and *rmin* in the example illustrated here, it looks for them beyond these boundaries up to the approximation window's original boundaries. In the example in figure 3.4 for instance, *lmin* and *rmin* were located at the boundaries of the reduced start window. After extending the search beyond the start window's boundaries (lower panel), *lmin* is found well inside the original approximation window at approx. 0, and *rmin* is found a the right boundary of the original window. The locations of maximum and minima are used to determine which of three approximation methods is appropriate in that particular context, as described below.

Mean F0 approximation. No PaIntE approximation takes place if there are less than two voiced frames for the current window or if *lmin* and *rmin* as determined above are less than 5 frames apart. In these cases, PaIntE reverts to a very simple approximation called *meanf0* by just determining the mean F0 value in that window as the *d* parameter; the five other PaIntE parameters are set to 0.

PaIntE approximation. In the standard case, in which a peak has been found, i.e. neither *lmin* nor *rmin* coincide with *max*, the approximation is carried out using the PaIntE function as defined in equation 3.1 above. This is called the *pfun* method. In this case, the approximation window is adjusted: it covers

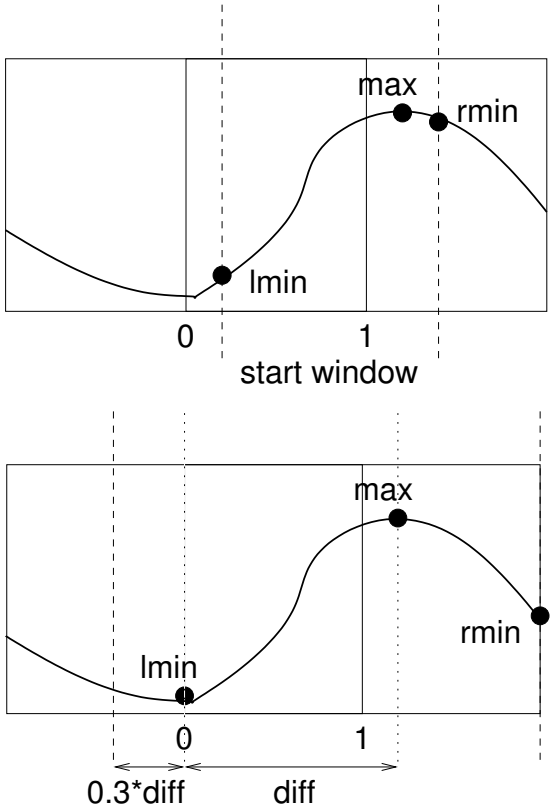

*Figure 3.4: Reducing the approximation window. Three-syllable original approximation window of a hypothetical F0 contour. PaIntE determines maximum (*max*) and left and right minima (*lmin, rmin*) in a reduced start window between 0.2 and 1.4 (dashed lines in the upper panel). If necessary, the search is then extended beyond the boundaries of the start window (lower panel). Here,* diff *indicates the time span from* lmin *to* max *(dotted lines). The final approximation window starts 0.3 times* diff *before* lmin *(left dashed line), but does not extend beyond* rmin*, which is already at the boundary of the original approximation window (right dashed line).*

at least the time span from *lmin* to *rmin*; if possible, it reaches 0.3 times the temporal difference between *lmin* and *max* beyond *lmin* to the left, and 0.3 times the difference between *max* and *rmin* beyond *rmin* to the right. This is illustrated in the lower panel of figure 3.4. Here, *lmin* lies well inside the original approximation window, and the window can be extended to the left. It is extended by 0.3 times the difference between *lmin* and *max*, i.e. 0.3 the duration of the rise. However, it cannot be extended beyond *rmin* to the right because *rmin* is already at the boundary of the original approximation window. The boundaries of the final approximation window are indicated by the vertical

dashed lines in figure 3.4.

Single sigmoid approximation. If *lmin* or *rmin* coincide with the max, i.e., a rise or fall have been detected, but not a clear peak, the PaIntE approximation is modified to leave out one of the two sigmoids (because there is no detectable peak). Depending on which sigmoid is left out, this is called *rise sigmoid* or *fall sigmoid* method. In this case, the *a* parameter of the missing sigmoid is set to -1, and the *c* parameter is set to 0. The remaining parameters are determined by the single sigmoid approximation. Before the actual approximation, the approximation window length is adjusted to be 0.6 times the temporal difference between the minimum which does not coincide with the *max* and the *max*, and the window is positioned in a way that the *max* is at a third of the window in the case of a fall, leaving two thirds of the window for the fall, or at two thirds of the window in the case of a rise, leaving the first two thirds of the window for the rise. However, it does not extend beyond the boundaries of the original approximation window.

3.2.4 Check for plausibility

After the approximation, the results are checked. If the parameters from the *pfun* approximation are not satisfactory, the approximation is repeated using one of the two simpler single-sigmoid methods described above. The results are not satisfactory if any of the following criteria applies.

(i) $a1 < 0.1$ or $a2 < 0.1$ in the *sylnorm* case

(ii) $c1 < 2$ or $c2 < 2$

(iii) b is more than half the window length of the approximation window outside the approximation window

(iv) 0 iterations took place

(v) the mean standard error between original F0 values and approximated values is too big (> 50)

(vi) the approximation window was suspiciously short (less than 10 frames)

Criteria (i) and (ii) obviously aim to identify cases where the peak is not very pronounced, which suggests that a single-sigmoid approximation is more appropriate. In these cases, the results are not dismissed because they are not plausible; rather, the other method seems better suited. Criterion (iii) is the only one that dismisses approximation results based on conspicuous parameter values. The last criterion, (vi), was introduced by me because I had tracked

down many outliers to cases where the approximation window contained less than 10 frames.

In the case of a single-sigmoid approximation, the results are checked using criteria (iii) to (vi). In the original implementation, only criteria (iv) and (v) were used in the single-sigmoid case; I have added criteria (iii), (vi) for better consistency with the *pfun* case. If the results are not satisfactory based on these criteria, the approximation window is adjusted based on the approximation results, and the single-sigmoid approximation is repeated. If the results are still not satisfactory then, the approximation is repeated using the *meanf0* method.

3.2.5 Future improvements

There are two aspects in which the PaIntE procedure should be modified in the future. Firstly, many apparent errors in the approximation can be traced back to unfortunate results of the F0 smoothing algorithm. For instance, in case of pitch halving errors in the raw F0 values, if there are too many subsequent errors, which is often the case for glottalized initial vowels in stressed German syllables, these erroneous values yield quite pronounced valleys in the smoothed F0 contour. These valleys are picked up in the approximation and are then often misinterpreted as rising accents.

Also, microprosodic effects caused by inherent pitch properties of certain phonemes are not filtered out by the current smoothing algorithm. For instance, the voiced fricative /v/ causes small, visible valleys in the F0 contour, which persist into the smoothed contour, which may cause amplitudes of *c1* or *c2* which are similar to cases of pitch accents with low amplitudes.

As described above, before the actual approximation, PaIntE tries to identify a peak in the middle of the approximation window between 0.2 and 1.4, i.e., from 20% of the preceding syllable to 40% of the current syllable. It only looks for earlier or later peaks if it finds no clear peak in this initial window. However, if a peak caused by microprosodic effects is present in the initial window, later or earlier peaks will not be detected and thus not approximated.

These undesired effects make a better F0 smoothing solution appealing, such as the one proposed by Reichel and Winkelmann (2010). They have shown that their F0 smoothing algorithm is superior to simple mean smoothers (and other pitch smoothing algorithms such as polynomial fitting as described by van Santen et al. (2004) or MOMEL (Hirst and Espesser 1993)) in that it is more effective in removing microprosodic effects.

A second aspect concerns the distinction of pitch and F0. F0 is perceived as pitch, and from a theoretical perspective, it is thus more interesting to approximate pitch rather than F0. Since it has been demonstrated that the perception of F0 can be adequately modeled using the ERB scale (Hermes and van Gestel 1991), using this scale for approximation would be closer to human perception

than the Hz scale that is currently used in PaIntE.

Still, I will illustrate how the PaIntE parameters despite these shortcomings capture perceptually relevant tonal aspects of the implementation of GToBI(S) events in chapter 5, and how they can successfully be used in modeling exemplar-theoretic classification in chapter 6. Before that, I will introduce duration z-scores as a temporal parameter in the perception of prosody in the following chapter.

Chapter 4

The temporal targets in prosody production

This chapter deals with temporal targets in prosody production. In applying Guenther and Perkell's model to the production of temporal aspects of prosody, as proposed in chapter 2, three questions have to be answered. These questions are: (i) what is the temporal dimension of perceptual space, (ii) which are the relevant prosodic events, and (iii) which are the target regions corresponding to these events.

Question (ii) has been discussed briefly in the introduction already. In analogy to the segmental domain, the relevant events should be prosodic categories that are perceptually distinct. I have stated in the introduction that the GToBI(S) events (Mayer 1995) are good candidates for prosodic categories, even though for pitch accents and intermediate phrase boundaries, their categorical status has not been established yet using the categorical perception paradigm. For the present purpose, I will still assume that the relevant events are phrase boundaries and pitch accents according to GToBI(S) (Mayer 1995). GToBI(S) was described in some detail in section 3.1. For the considerations and experiments on temporal aspects of prosody presented in this chapter however I will only distinguish between intermediate boundaries (ip) and intonation phrase boundaries (IP), but not between different types of pitch accents.

Regarding question (i), which are the relevant prosodic events, Perkell et al. (2000) have already integrated time as an additional dimension in their model, however, not with regard to prosody. Adding time as a dimension for segmental target regions allows for dynamic target regions which change in the course of a segment. For instance, they illustrate the idea using a vowel-stop-vowel sequence in which the auditory-acoustic targets for the stop change dynamically from pre-closure to closure to post-closure (Perkell et al. 2000, p. 236, fig. 1).

Similarly, Byrd (1996) suggests that phase windows describe the variation observed in the timing of articulatory gestures in segmental articulation. She

claims that prosodic context is an "influencer" which affects all phase windows in the same way (cf. section 2.1). Thus, the temporal properties of prosodic context can be thought of as orthogonal to segmental timing. In this vein, I suggest that in prosody production, the temporal dimension is a dimension in its own right, which does not serve to model dynamic changes of segmental parameters over time, or which is only relevant in the production of some segments or prosodic events; rather it constitutes its own, independent dimension, independent of segmental aspects. Including time as an additional dimension in the same way as Perkell et al. (2000) did for the segmental domain does not reflect the role of temporal aspects in prosody production. Instead, I will introduce z-scores of speech unit durations as a measure of local speech rate in section 4.2, arguing that this measure can be regarded as the temporal dimension, particularly from an exemplar-theoretic point of view.

To answer question (iii) I will claim in section 4.3 that the z-score measure can be regarded as the temporal dimension of perceptual space for prosodic events by explicating how well-known effects of prosodic context can be modeled using this measure. I will conclude that the z-score distributions for units related to prosodic events can be regarded as target regions in the production and perception of temporal aspects of prosodic events. Part of the experiments discussed in this section have been published in Schweitzer and Möbius (2003).

Section 4.4 will address an application of the proposed model. It has been claimed (Whiteside and Varley 1998a) that frequent and infrequent syllables may be produced using different strategies. I will argue that the model would indeed predict differences in the production and will quantify these differences using real speech data. Aspects of this section have been published in Schweitzer and Möbius (2004).

Since all experiments described in this chapter have been carried out on the same corpus of speech data, I will take the time to describe the specifics of the corpus in some detail in section 4.1 before turning to the actual experiments. I will put particular focus on syllable frequencies in discussing the corpus because these will be of interest for the experiments concerning frequent and infrequent syllables, and since the corpus has been designed to specifically cover infrequent speech units, frequency distributions will deviate to some extent from what would be expected from randomly selected text.

4.1 Description of the speech data

The speech corpus used in this chapter is the MS corpus originally recorded for unit selection speech synthesis. It was designed to cover at least all German diphone types and all phoneme types in different contexts. To this end, phoneme/context vectors for each sentence in a large collection of newspaper articles were predicted using the IMS German Festival TTS system. For each

segment, the phoneme/context vector describes (i) its phonemic identity, (ii) its position in the syllable (onset or rhyme), (iii) syllabic stress on the corresponding syllable, (iv) type, or absence, of pitch accent on the syllable, (v) type, or absence, of boundary tone or phrase accent on the syllable, (vi) position of the syllable in the phrase (initial, medial, or final) and (vii) word class of the related word (function word or content word). From these sentences a subset with the same coverage as the full set in terms of diphone types and a good coverage of different phoneme/context vectors was extracted. Sentences containing diphone types that were not found in the corpus, but theoretically allowed by German phonotactics, were manually added, as was some application specific text material. It was not attempted to manually construct sentences for the missing vector types.

The corpus was read by a professional male speaker. Each utterance was annotated on the segment, syllable and word level by forced alignment with manually corrected transcriptions, and manually prosodically labeled according to GToBI(S). A detailed description of GToBI(S) can be found in section 3.1. The automatic segmentation into words, syllables, and segments was manually corrected afterwards in three steps. First, outliers with respect to segmental durations were systematically identified and corrected, if necessary. Second, further segmentation errors were corrected in the process of manual prosodic annotation. Third, diagnostic evaluation of intelligibility problems in synthesis revealed further cases of incorrect segmentations that had not been discovered up to this point. The resulting speech data amounts to more than 150 minutes of speech and contains approximately 94,000 segments, 34,000 syllables, and 17,000 words.

The final version of the MS corpus used for the experiments in this chapter contains 2,682 different phoneme/context vectors. The corresponding histogram is shown in figure 4.1. The distribution of phoneme/context vectors is a typical instance of an LNRE distribution (Large Number of Rare Events, cf. Baayen 2001), with many vectors occurring only once. The most frequent vector occurs 1,914 times, as indicated by the leftmost data point in figure 4.1. The nine next frequent vectors exhibit token frequencies between 1,226 and 714. The frequencies of the 10 most frequent vectors are plotted beside the corresponding data points in figure 4.1.

Histograms of boundary tones and pitch accents in the MS corpus are given in figure 4.2. There are altogether 5,182 intonation phrase boundaries and intermediate phrase boundaries in the corpus, of which 74 were labeled with a "?" diacritic to indicate labeler uncertainty. Since they constitute a portion of only approximately 1.4%, the "?" diacritics were ignored, so that each of the uncertain phrase boundaries was mapped to its "certain" counterpart. Most of these uncertainties concern intermediate phrase boundaries; only 6 of the 3,970 intonation phrase boundaries were marked as uncertain. Turning to pitch accents, there were 7,930 pitch accents, 198 of which were marked as uncertain,

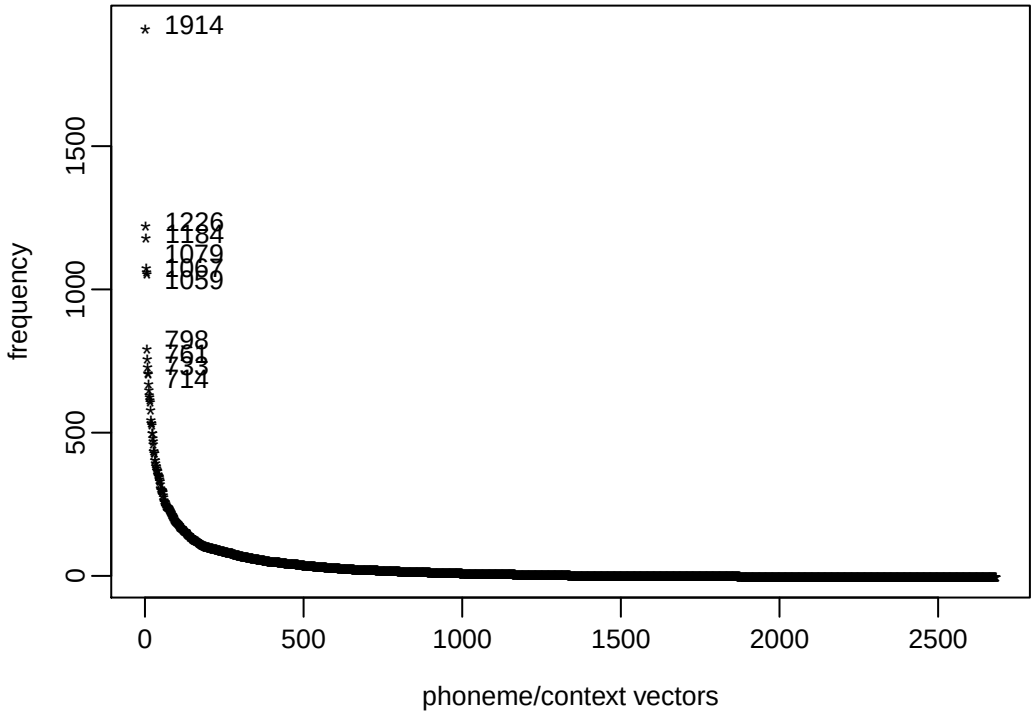

Figure 4.1: Histogram of phoneme/context vectors. The vectors are sorted from left to right according to their frequency of occurrence, with the most frequent vector on the left. Frequencies are indicated on the y-axis. For the ten most frequent vectors, their frequencies are additionally plotted beside the corresponding data points.

corresponding to a portion of approximately 2,5%. Reasoning as above, they were all mapped to the respective pitch accent. In an additional 27 cases, it was unclear whether the syllable was accented at all. The label provisioned for such cases is "*?". No mapping took place in these cases.

It is obvious from figure 4.2 that some prosodic events were rare even in a corpus of this size. Specifically, L% occurred only 13 times, HH*L 4 times, and H*M only once. L* and L*HL were slightly more frequent with 87 and 101 occurrences, respectively[1]. Partial linking occurred so seldom (29 times

[1]The low frequency of HH*L and H*M is presumably due to the corpus consisting of read speech only. In the case of L%, its rare occurrence might be attributed to the fact that the sequence L*H L% is not possible in German. Thus, L% will usually only occur following H*L, L*HL, HH*L, and in all these cases, the pitch contour is in the lower range of the register already, making it hard to decide objectively whether the boundary tone is a simple %, or a L%. Since the course of the imaginary baseline is not always perfectly clear, labelers may be tempted to resort to the "default" % in many such cases.

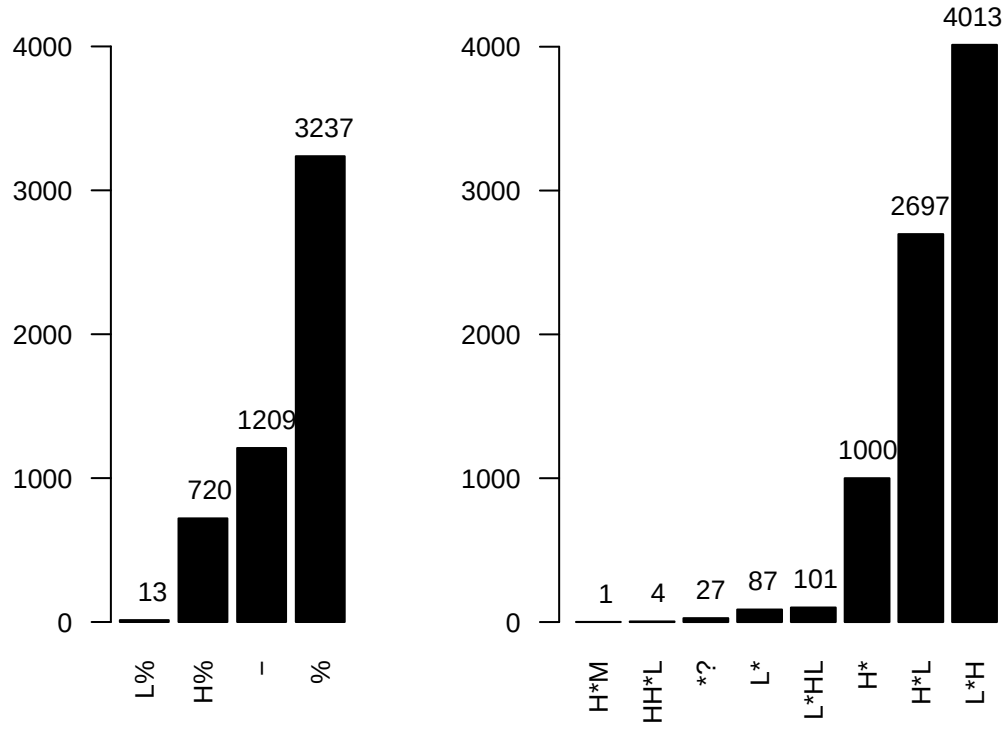

Figure 4.2: Histograms of boundary tones (left panel) and pitch accents (right panel).

including uncertain instances) that it seemed appropriate to not distinguish it from complete linking, which was more frequent (1,058 instances).

I will use the MS corpus to examine temporal aspects related to prosodic events in section 4.3, and also to address the hypothesis put forth by Whiteside and Varley (1998a) that frequent and infrequent syllables may be produced using different strategies (section 4.4). Before turning to these experiments, I will introduce z-scores of speech unit durations as a measure of local speech rate which can be used to quantify perceptually relevant temporal aspects of prosodic events in the following section.

4.2 A measure of local speech rate

When turning to the temporal aspects of prosody, two acoustic parameters are often considered, viz. duration of speech units, and frequency of speech units (i.e. their number in a certain time interval). The speech units are usually taken to be segments, syllables, or words. The two parameters are of course related: frequency can be inferred from duration, and, conversely, mean duration (but

not individual duration of each speech unit) can be deduced from frequency. Pfitzinger (2001) gives an extensive overview over the various measures that are cited in the literature.

Frequency of speech units is often used in automatic speech recognition as a measure of speech rate, motivated by the insight that extreme speech rates severely degrade recognition scores, which necessitates the adaptation of the speech models to different speech rates. In these cases, speech rate is usually determined over longer stretches of speech, such as complete utterances or even dialogs, and thus is called *global speech rate* by Pfitzinger (1996, 2001).

Pfitzinger distinguishes *global speech rate* from *local speech rate*. The latter is of higher interest for the analysis of dynamic prosodic phenomena and can be obtained by moving a window stepwise through the relevant utterance and calculating the frequency of speech units in each window. Pfitzinger (1998, p. 1088) notes that the window size should be at least as large as the longest speech unit in the data to avoid outliers. He uses a window of 625 ms for his data.

Individual durations of speech units on the other hand are often examined in approaches that deal with speech synthesis. In such approaches, the duration of each segment in an utterance is usually predicted according to its phoneme identity as well as its prosodic and segmental context. Campbell and Isard (1991) and Campbell (1992) predict segment durations indirectly from syllable durations.

Since individual segment durations are determined to a great extent by the inherent duration of the respective phoneme, the analysis of absolute segment durations is less attractive for prosodic research. Instances of different phonemes can not be compared to each other. On the other hand, examining individual durations offers a finer resolution than the window-based approach to local speech rate described above, where durations are averaged over a window which is longer than the segments: When focusing on temporal properties that are relevant for prosodic events, it is desirable to have a resolution that is as fine as possible, because one is interested in local changes in the temporal parameters in the vicinity of prosodic events, i.e., changes that take place even within syllables. For instance, it might be of interest whether the presence of a pitch accent influences all segments within a syllable to the same extent. Thus, the durational approach seems preferable, but is problematic because of the influence of inherent phoneme durations.

This influence is illustrated in figure 4.3. Phoneme specific constraints are visible in the distributions of durations of different phonemes: the distributions look similar but have different means and standard deviations. The left panel shows a histogram of the durations of 426 exemplars of [OY]. Most of them have been realized with durations around 135 ms, but some are shorter than 50 ms, and several instances are up to 250 ms long. The histogram of segment durations of 824 realizations of [aU] in the right panel of figure 4.3 looks sim-

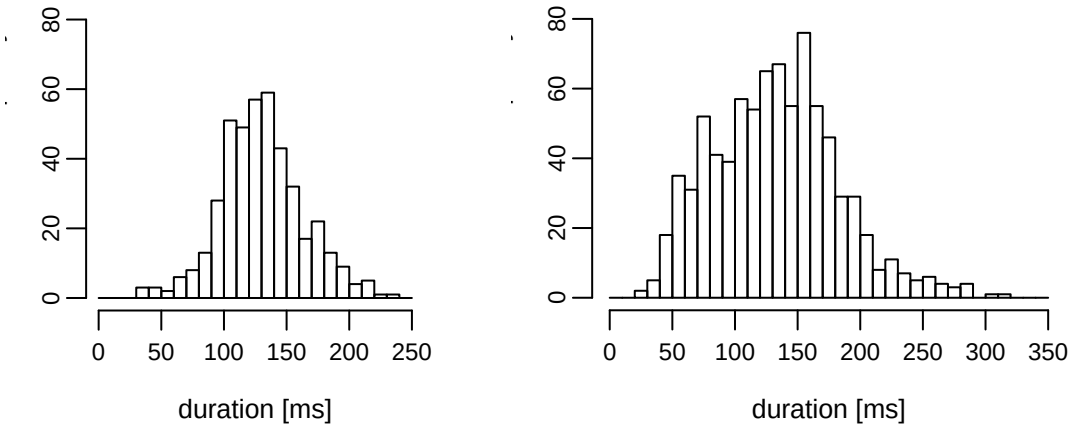

Figure 4.3: Histograms of durations for [OY] (left panel) and [aU] (right panel).

ilar: the mean duration of [aU] was approximately the same as for [OY], but there is more variation in the durations of [aU], they exhibit a higher standard deviation. Obviously, [aU] can vary more in its duration than [OY].

This property has been termed *elasticity* by Campbell and Isard (1991) and Campbell (1992). The idea is that different phonemes can be lengthened or shortened to different extents and that this phoneme specific ”elasticity” manifests itself in the distribution of observed segment durations and their standard deviation, respectively. Campbell and Isard (1991) claim that all segments within a syllable are lengthened or shortened by the same factor if their different “elasticities” are taken into account. They exploit this to predict segment durations from given syllable durations.

To demonstrate the elasticity effect, figure 4.4 shows a plot of the mean durations and standard deviations of the segments in the database. Mean durations are plotted along the x-axis, standard deviations along the y-axis. The phoneme symbols are indicated in SAMPA notation. Phonemes that lie on an imaginary vertical line thus have the same mean but different elasticities or standard deviations, e.g. [f] and [s], [m] and [n], and [aU] and [OY], for instance, while phonemes on an imaginary horizontal line have the same standard deviation but different means.

In analyzing prosodic categories, one would expect different categories to cause different degrees of shortening or lengthening of the involved speech units. The degree of shortening or lengthening in terms of absolute duration would of course also depend on the elasticity of the particular speech unit. A small amount of lengthening observed for a speech unit that is known to be less

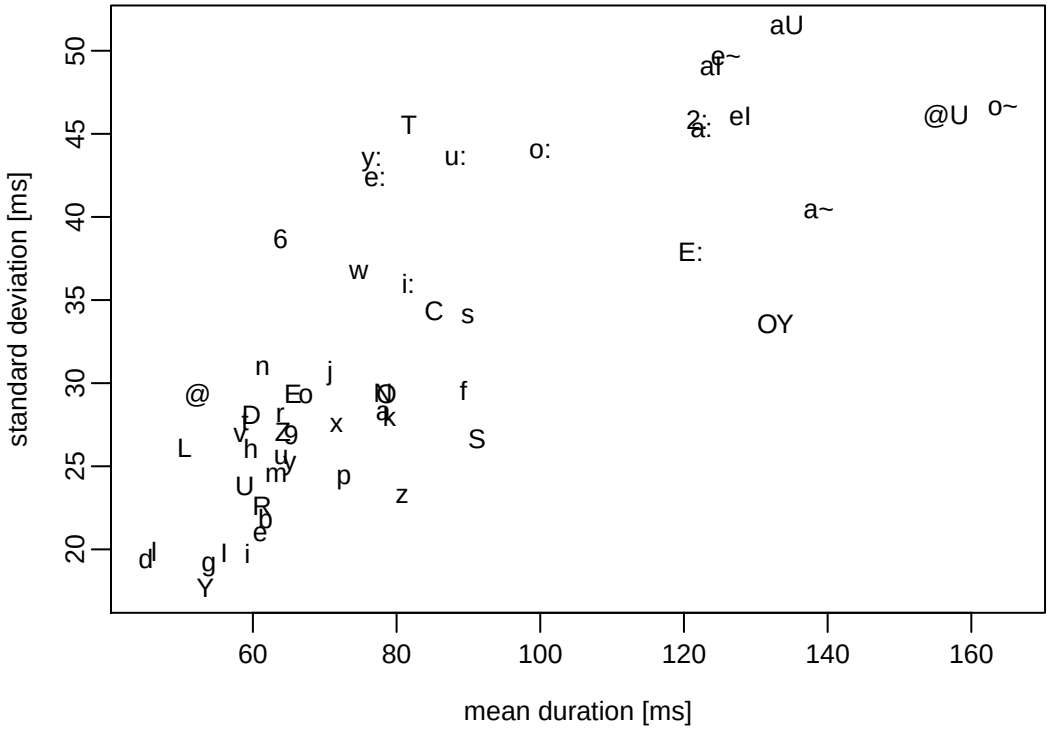

Figure 4.4: "Elasticity" of different phoneme classes in the MS corpus. Mean durations are plotted against standard deviations. Phoneme symbols are indicated in SAMPA.

elastic would be similar to stronger lengthening of a unit that is more elastic. To be able to compare such units, the degree of lengthening or shortening must be seen relative to the elasticity of the unit.

This can be captured by using speech unit specific *z-scores* of durations instead of absolute durations, as suggested by Campbell and Isard (1991), who have used z-scores to determine segment durations that add up to a given syllable duration in the framework of a text-to-speech system. Reversing the process, I propose to use z-scores in the analysis of speech unit durations and, more generally, local speech rate.

Converting absolute scores to z-scores is a common statistical transformation: absolute values are replaced by their deviation from the mean, and divided by the standard deviation. To get speech unit specific z-scores, for instance phoneme specific z-scores, mean and standard deviations are calculated for each phoneme class separately, and in transforming a particular exemplar to its z-score, mean and standard deviation of the respective phoneme class are used. If p_i is a realization of a phoneme of class p, let $dur(p_i)$ be its duration, $\mu(p)$ the mean duration of all phonemes of class p, and $\sigma(p)$ their standard

deviation. Then the z-score of p_i, $zscore(p_i)$, is defined as

$$(1) \qquad zscore(p_i) = \frac{dur(p_i) - \mu(p)}{\sigma(p_i)}$$

Using z-scores, it is possible to compare exemplars of different phoneme classes to each other, because phoneme specific mean and standard deviation are eliminated. Thus z-scores are another possible way to measure local speech rate, which offers an excellent resolution.[2] Some smoothing may be necessary to alleviate outliers. These may occur because the precision of the values depends on a consistently segmented database.

From an exemplar-theoretic point of view, the z-score can be interpreted as the position of a particular exemplar in the distribution of exemplars of the same category: It indicates the distance of the respective exemplar from the distribution's mean relative to its standard deviation. This distance can be interpreted as a measure for the extent of lengthening or shortening of the exemplar compared to other exemplars. Thus, instead of comparing absolute durations of two segments for instance, the positions of both segments in the distribution of durations are compared. The use of z-scores instead of absolute durations allows one to assess the amount of lengthening or shortening pertaining to a particular exemplar not only with respect to other realizations of the same type but with respect to all realizations of all types.

Thus the z-score approach to local speech rate is no arbitrary measure that happens to solve the problem of phoneme specificity (or more generally speech unit specificity). On the contrary, in the framework of exemplar theory, it is plausible to postulate that z-scores are relevant in the perception of temporal aspects of prosody: the z-score of an exemplar simply describes its position in the exemplar cloud. Listeners have access to the z-score of an exemplar by comparing it to the other exemplars in the same cloud. Note that assuming that listeners compare exemplars to other exemplars is not an ad hoc stipulation; a basic assumption of exemplar theory is that in perception, every single new exemplar is compared to stored exemplars (cf. section 2.3.3). Approaching an answer to question (i), what is the temporal dimension of perceptual space in prosody perception, I suggest that z-scores of speech unit durations can be regarded as relevant in the perception of prosodic events. I will leave open the question of the exact nature, or rather, of the exact length of the speech units here. In the following sections, I will look at segments and syllables in more detail: The next section will deal with the z-score distributions of syllables and segments related to prosodic events in order to illustrate how the target regions can be implicitly defined by the observed distributions. Then, I will demonstrate in section 4.4 how corpus data may be used to corroborate the hypothesis that

[2]The same approach has been proposed by Heid (1998, p. 296), also motivated by Campbell and Isard (1991).

both segments and syllables may be the underlying speech units in production. However, a definitive answer to the question is beyond the scope of this thesis.

4.3 Target regions for temporal z-scores

The perceptually relevant dimensions and their respective acoustic correlates are not uncontroversial. While it has been demonstrated that the perception of fundamental frequency can be adequately modeled using the ERB scale (Hermes and van Gestel 1991), the relation between segment or syllable durations and perceived local speech rate is less well-established. I have elaborated in the preceding section that z-scores are an adequate measure of local speech rate. The duration z-scores can be interpreted as coding the location of an exemplar in the exemplar cloud, as elaborated in the previous section. Comparing exemplars to other stored exemplars is an essential process in exemplar theory. In perception, new exemplars are necessarily compared to stored tokens when they are categorized. Thus, I suggest that this justifies regarding the duration z-scores as a potential perceptual dimension of prosodic target regions.

Note that even though z-scoring is a normalization process, and even though in this way, assessing the location of an exemplar in the exemplar cloud can be seen as a kind of normalization, too, this is different from a speaker normalization process which manipulates the auditory input. Johnson (1997) has argued against a speaker normalization process of this kind, because there is no evidence for a separate process, and because it is not necessary to assume such a process in exemplar theory: listening to a particular speaker will increase the attentional weights of stored exemplars of this speaker, which causes new input to be perceived with reference to speaker-specific stored exemplars. Similarly, I suggest that the temporal properties of a new exemplar will be perceived with reference to the temporal properties of stored exemplars of the same type, and that this can be expressed by the duration z-scores.

To further substantiate the idea of duration z-scores as a perceptual dimension, I will demonstrate that well-known effects of prosodic context on local speech rate, viz. phrase-final lengthening and lengthening of pitch-accented speech, are clearly visible in the z-score probability distributions, or "density plots",[3] of the associated speech units, both on the segmental and on the syllable level.

The effect of these two prosodic factors on segmental durations are clearly visible in the z-score distributions of segments in such contexts. Figure 4.5 shows the distributions for all segments across all contexts (dotted line), for pitch-accented nuclei (dashed line), and for phrase-final segments (solid line).

[3] All density plots were generated using the *density* function implemented in R (R Development Core Team 2009), which computes kernel density estimates for a sample of a population in order to estimate the probability distribution of the entire population.

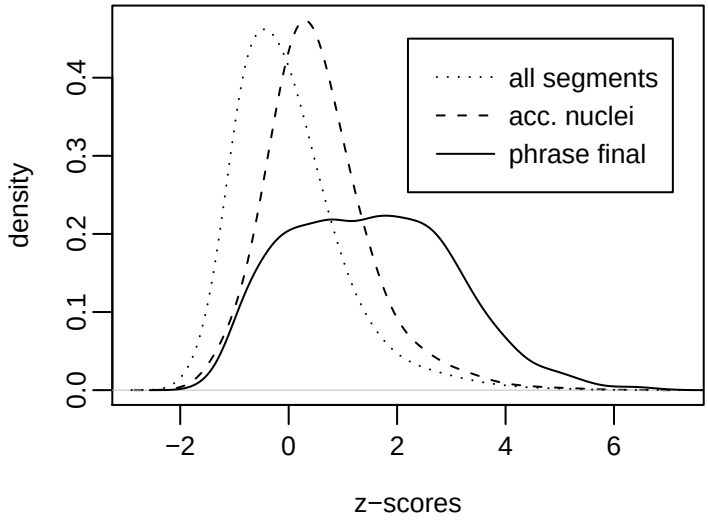

Figure 4.5: Z-score density functions for all phonemes (dotted line), nuclei of accented syllables (dashed), and phrase-final phonemes (solid). Z-scores of accented nuclei and phrase-final phonemes are significantly higher than the average.

It is obvious that the distribution for all segments differs from the distribution for pitch-accented nuclei: the latter is clearly shifted to the right indicating that the z-scores of accented nuclei are typically higher than those of average segments. The effect is even stronger for phrase-final segments. It should be noted that, apart from being shifted to the right, their distribution is less narrow indicating more variation. Both effects are significant[4], but the difference of mean z-scores in the two prosodically marked contexts, compared to the mean z-scores across all contexts, is higher for phrase-final segments than for accented nuclei: the means are 0.002, 0.547 and 1.502 for all segments, accented nuclei and phrase-final segments, respectively.[5] The difference in means is also significant for accented nuclei vs. phrase-final segments, at the same significance level as above.[6]

The distribution of z-scores for phrase-final segments in figure 4.5 seems to be bimodal, as can be seen from the two "bumps" in the distribution indicated

[4]All statistical tests reported in this chapter have been conducted using the R statistics package (R Development Core Team 2009). Distributions have been compared pairwise using the Wilcoxon rank sum test with continuity correction. The confidence level was adjusted from 0.95 to 0.999 to allow for up to roughly 50 repeated Wilcoxon tests.

[5]Calculation of the 99.9% confidence intervals for pairwise differences of the means reveals that the true differences in means lie between 0.539 and 0.606 for all segments compared to pitch-accented nuclei, and between 1.420 and 1.547 for all segments compared to phrase-final segments.

[6]The confidence interval for the difference in means is between 0.845 and 1.015.

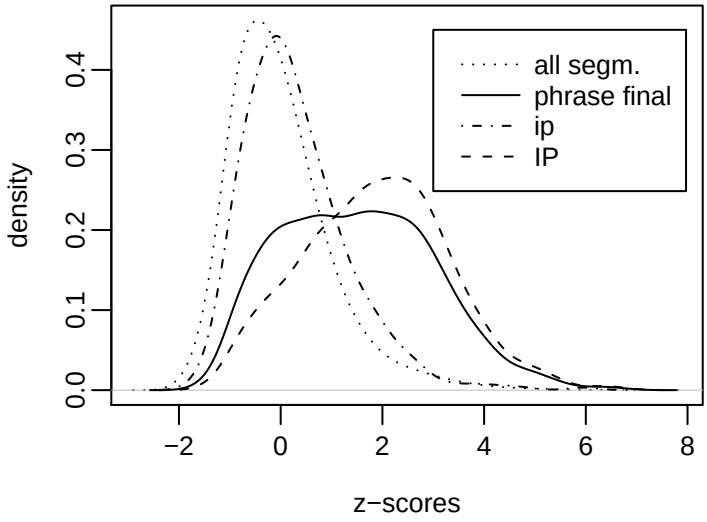

Figure 4.6: Z-score density function for phrase-final segments (solid line), repeated from figure 4.5. The bimodal distribution is due to different distributions for phrase-final segments in intermediate phrases (ip, dot-dashed) and intonation phrases (IP, dashed). For comparison, the z-score distribution for all segments is again indicated by the dotted line.

by the solid line. This is evidently due to the fact that there is less lengthening for phrase-final segments in intermediate phrases, whereas there is substantial lengthening in intonation phrases, as illustrated in figure 4.6. Here, the distributions for all segments (dotted line) and for phrase-final segments (solid line) are repeated from figure 4.5. Phrase-final segments have been separated into those from intermediate phrases (dot-dashed line) and those from intonation phrases (dashed line). The maxima of these two distributions lie where the two bumps occur in the overall distribution. The two means corresponding to the maxima are at 0.296 and 1.868, respectively, indicating that there is less lengthening for final phonemes in intermediate phrases than for accented nuclei (the mean for the latter was 0.547, see above). It should be noted that although this indicates that most intonation phrases have been realized with substantial lengthening, there are also cases where the lengthening is similar to the one typically observed for intermediate phrases: the distribution for intonation phrases has a negative skew; it overlaps with the distribution of intermediate phrases. Two pairwise Wilcoxon rank sum tests confirm that the values for intermediate phrases are still significantly different from those for all segments and those for accented nuclei, at the same significance level as above[7].

[7]The confidence intervals for the true differences in means between all segments and in-

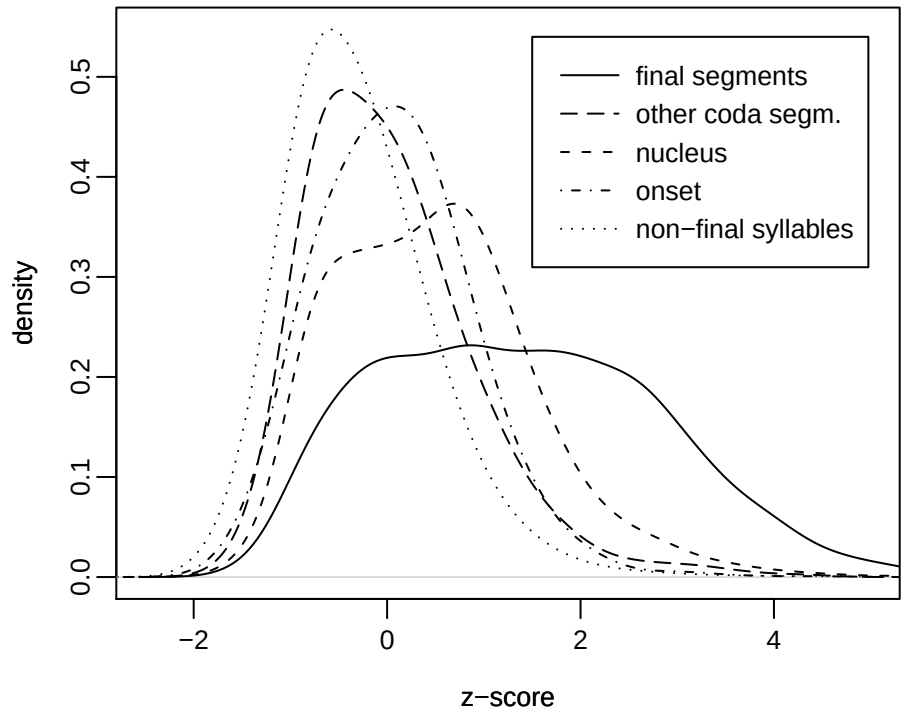

Figure 4.7: Z-score density functions of segments in different syllable positions in phrase-final unaccented syllables compared to segments in non-final unaccented syllables (dotted line). The lengthening effect is strongest for segments in syllable-final position (solid line), less strong for nuclei (short dashes), and only moderate for other non-final coda segments (long dashes) and onset segments (dot-dashed line).

The influence of prosodic context is not limited to single phonemes. For instance, phrase-final lengthening can be observed for all segments in phrase-final syllables. Figure 4.7 indicates the distribution of z-scores of segments in various positions in phrase-final unaccented syllables and compares them to the distribution of segments in non-final unaccented syllables[8] (dotted line). The effect is strongest for syllable-final segments (solid line). For these, the mean z-score is 1.371. Interestingly, the effect is also quite strong for nuclei in phrase-final syllables (short dashes, mean z-score 0.491), but less pronounced for onset segments (dot-dashed line, mean z-score 0.079) and for non-final coda segments (long dashes, mean z-score 0.036). The distribution for segments in syllables

termediate phrases and between intermediate phrases and pitch-accented nuclei are between 0.205 and 0.367 and between 0.191 and 0.379, respectively.

[8]Accented syllables are ignored here to separate lengthening effects caused by the accent from the lengthening effects caused by phrase-finality, which are of interest here.

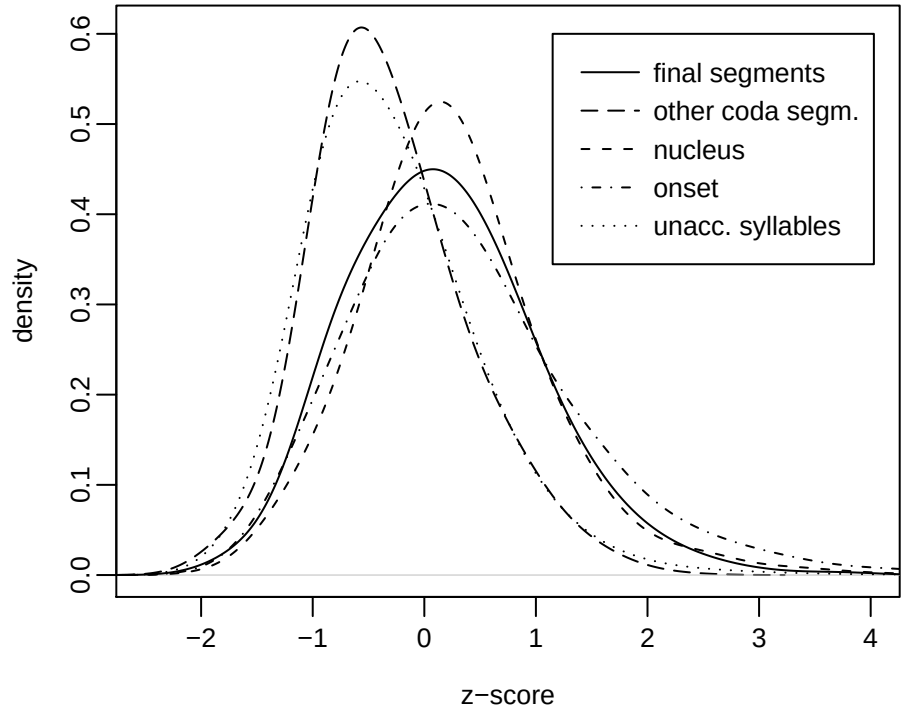

Figure 4.8: Z-score density functions of segments in different syllable positions in non-final accented syllables compared to segments in non-final unaccented syllables (dotted line). The lengthening effect is present and similarly strong for segments in syllable-final position (solid line), nuclei (short dashes), and onset segments (dot-dashed line). It is, however, not detectable for non-final coda segments (long dashes).

which are not phrase-final at all (dotted line) is indicated for comparison (mean z-score -0.286). Four Wilcoxon rank sum tests reveal that the differences are significant except for onset segments vs. non-final coda segments (long dashes vs. dot-dashed line)[9]. However, these latter distributions are still significantly different from the distribution of segments in non-phrase-final syllables (dotted line), confirming that the effects are detectable for all segments in phrase-final syllables.

Similarly, lengthening can be observed not only for accented nuclei, but for other segments in accented syllables. Figure 4.8 compares the z-score distribu-

[9]Calculation of the 99.9% confidence intervals for pairwise differences of the means reveals that the true differences in means lie between 0.733 and 0.962 for syllable-final segments compared to nuclei, between 0.297 and 0.450 for nuclei compared to onset segments, and between 0.188 and 0.372 for non-final coda segments compared to segments from non-final syllables. The difference between onset and non-final coda segments is not significant ($p \approx 0.004$).

tions of segments in various positions in accented non-final syllables to the z-score distribution of segments in unaccented non-final syllables (dotted line)[10]. It can be seen that the effect is similarly strong for onset segments (dot-dashed line, mean z-score 0.358), nuclei (dashed line, mean z-score 0.253), and final segments (solid line 0.176). Wilcoxon rank sum tests reveal that there is no significant difference between the former two ($p \approx 0.029$), and a small but significant difference between the latter two[11]. There is, however, still a significant difference between syllable-final segments in accented syllables (solid line) and segments in unaccented syllables (dotted line)[12]. Interestingly, the effect is not present for non-final segments in the coda of accented syllables (long dashes), as there is no significant difference ($p \approx 0.851$) between their distribution and the z-score distribution of segments in unaccented syllables (dotted line).

It is an astonishing finding that the strength of effect both in phrase-final lengthening and in lengthening related to pitch accents does not increase or decrease monotonically in the course of a syllable. One could have expected for instance that phrase-final lengthening increases towards the end of the phrase and is strongest for the very final segment in a phrase. Instead, figure 4.7 suggests that the effect is present in the onset segments already, increases for nuclei, then decreases for interceding coda segments, before reaching the maximum for the very final segment in the phrase. Similarly, one might have expected that nuclei are affected the most by lengthening related to pitch accents, but it is interesting that the effect is similarly strong for the final segment in the syllable, while it is not present at all for interceding coda segments. These observations will be investigated further in future work.

The influence of prosodic context can also be seen in syllable-level z-score distributions. When examining z-scores of syllables, the problem arises that some syllables are extremely rare. For instance, 1,650 of altogether 3,863 syllable types occur only once in the MS corpus; 2,947 syllables types occur 5 times or less, 3,322 types occur up to ten times. The smaller the number of instances of a particular syllable type, the less reliable is the z-score calculated on that basis, particularly because it is likely that there are still some segmentation errors in a corpus of this size even after manual checking. Therefore only the z-scores of those 340 types were examined for which there are at least 20 realizations in the corpus. These syllable types add up to 22,650 syllable tokens.

When looking at their z-score distributions, the same effects of prosodic context on the syllable level as on the segment level can be found. The dis-

[10]Phrase-final syllables are ignored here to separate phrase-final lengthening effects from the lengthening effects related to pitch accents, which are of interest here.

[11]Calculation of the 99.9% confidence interval for pairwise differences of the means yields true differences between 0.019 and 0.127 for nuclei compared to syllable-final segments.

[12]The true difference between means lies between 0.432 and 0.499 with a confidence of 99.9%.

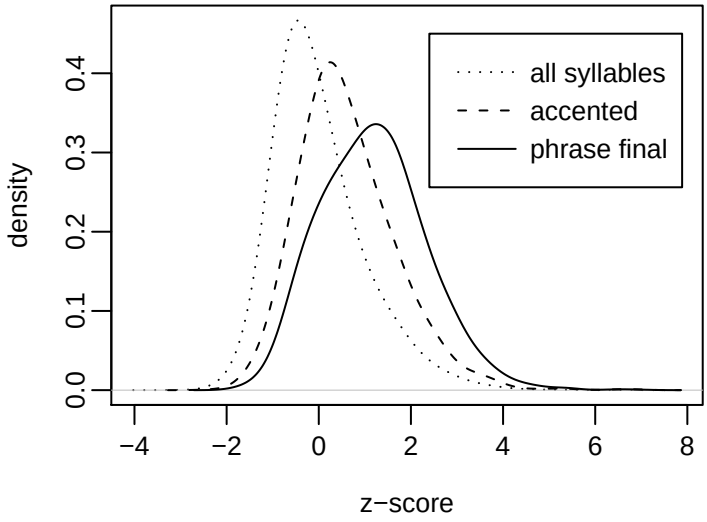

Figure 4.9: Lengthening effects on the syllable level: Z-score density functions for all syllables (dotted line), accented syllables (dashed), and phrase-final syllables (solid). Z-scores of accented syllables and phrase-final syllables are significantly higher than the average.

tributions are depicted in figure 4.9. Syllable-level z-scores are higher than average for pitch-accented syllables, and even higher for phrase-final syllables: the overall mean z-score is 0.000, while the means for pitch-accented syllables and for phrase-final syllables are 0.611 and 1.174 respectively. The differences in means are again highly significant[13].

As for the difference between intonation phrases and intermediate phrases observed on the segmental level in figure 4.6, the effect is less obvious in figure 4.9. There is only a very slight bump on the left slope of the distribution for phrase-final syllables. Nevertheless the underlying distributions of intermediate and intonation phrases are still different, as illustrated by figure 4.10: the distribution for phrase-final syllables (solid line) can be split up into intermediate phrases (dot-dashed line) and intonation phrases (dashed line). Again, the distribution for all syllables is depicted as a reference (dotted line). The bump is less obvious in this case because there is more overlap between the two distributions on the syllable level than on the segment level. Since there are also more realizations of boundary tones than of intermediate phrase boundaries, the peak corresponding to the maximum in the distribution for intonation

[13]Three pairwise Wilcoxon rank sum tests yield 99.9% confidence intervals for the differences in means between 0.565 and 0.675 for all syllables as compared to pitch-accented syllables, between 0.493 and 0.674 for accented syllables compared to phrase-final syllables, and between 1.134 and 1.274 for all syllables vs. phrase-final syllables.

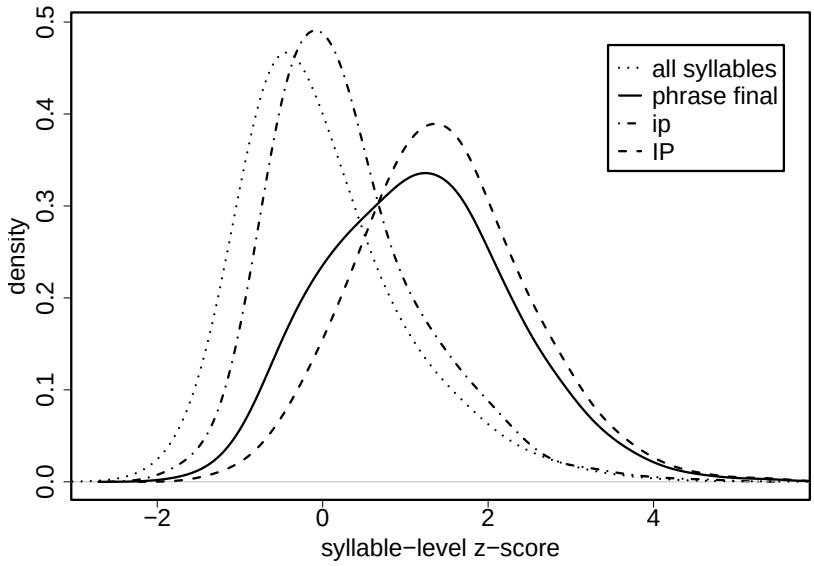

Figure 4.10: Z-score density function for phrase-final syllables (solid line), repeated from figure 4.9. The distributions of phrase-final syllables in intermediate phrases (ip, dot-dashed) and intonation phrases (IP, dashed) are significantly different. For comparison, the z-score distribution for all syllables is again indicated by the dotted line.

phrases is dominant even in the overall distribution for all phrase-final syllables.

To summarize the experimental results presented in this section, it can be stated that consistent effects of prosodic events on the duration z-scores have been observed both on the segment and on the syllable level. Different prosodic contexts produce significantly different z-score distributions.

Under the assumption that the z-scores can be interpreted as a relevant dimension in the perception of prosody, the z-score distributions for different prosodic events implicitly define overlapping, but significantly different target regions: they identify the range of acceptable z-scores for each prosodic event. Likewise, Keating (1990) proposed that the range of observed values along a dimension in production defines the window which serves as a target region for production.

Answering question (iii), which are the target regions corresponding to these events, I conclude that the z-score distributions for speech units related to prosodic events implicitly define the temporal target regions in the production of these events. Turning again to the question of the nature of the speech units—segments, syllables, or possibly higher-level units—I have shown that prosodic effects are not only detectable in the z-score distributions of segments but also in those of syllables, thus both segments and syllables may be the underlying units; however, I do not want to rule out that such effects could be

found on the word-level as well. The following section will shed more light on this issue, suggesting that both the segment and the syllable levels may play a role, depending on the frequency of the speech unit that is to be produced.

4.4 Temporal target regions and the syllabary

Levelt's model of speech production (Levelt 1992; Levelt and Wheeldon 1994; Levelt 1999) assumes that the basic unit in articulation is the syllable. Convincing evidence for this claim came from results of implicit priming experiments reported by Cholin et al. (2004). Levelt claims that gestural scores for the articulation of syllables are stored in a mental syllabary. It is left open, however, whether gestural scores for *all* syllables are stored, even for languages like English or Dutch, which are claimed to have approximately 12,000 syllables (Levelt 1999, p. 111), or whether scores for infrequent syllables are computed on-line. However Levelt notes that for English, 500 frequent syllable types would already cover 80% of the syllable tokens (Levelt 1999, p. 111), implicitly suggesting that only these frequent syllables might be stored in the syllabary.

Similar figures can be found in the MS corpus, even though the corpus has been designed to increase coverage of rare diphones. For instance, as mentioned in section 4.3, the corpus contains 3,863 different syllable types. 1,650 of them occur only once, but 588 syllables occur at least ten times each. These 588 types add up to 25,978 tokens and make up almost 77% of the entire corpus. Reasoning as Levelt above, this might already be a sufficient proportion to be stored in the syllabary. Moreover, the 1,650 syllables that occur only once can be said to be not only infrequent but extremely rare, considering that the corpus consists of 160 minutes of speech, and has been optimized for coverage, i.e. infrequent units should be overrepresented compared to randomly selected corpora of the same size. I would maintain that the average speaker produces significantly less than 160 minutes of speech per day, least of all with a vocabulary as diverse as that of the speech corpus investigated here. It is questionable whether gestural scores of extremely rare syllables that a speaker produces, say, once in a week can be expected to be stored in memory.

In any case, according to Levelt (1999), what is stored in the syllabary is the articulatory scores of syllables. In the sense of Guenther et al. (1998) and Perkell et al. (2000), however, these scores would be auditory rather than articulatory, and they can be thought of as trajectories through auditory target regions for subsequent speech segments. These trajectories are only later translated into motor commands by feed-forward mappings learned during speech acquisition (cf. section 2.2).

Assuming the exemplar-theoretic approach to target regions as outlined in the preceding section 4.3, according to which target regions are implicitly defined by accumulations of exemplars, it naturally follows that very infrequent

syllables might indeed not be stored in the syllabary but have to be computed on-line instead: Infrequent units are represented by considerably fewer exemplars; for the most infrequent units, there may be too few exemplars to qualify as an "accumulation" of exemplars over which we can abstract. This implies that there might be no or no reliable target regions available for infrequent units, and that the speaker has to resort to smaller and therefore more frequent units. I will illustrate this using some concrete figures from the MS corpus in the following section.

4.4.1 Frequent and infrequent syllables

In the MS corpus, the 340 most frequent syllable types, which occur, as mentioned in section 4.3 above, at least 20 times each, account for 22,650 syllable tokens and thus cover approximately 67% of the corpus. These figures give an impression of the order of magnitude of exemplars possibly stored in memory. They confirm that at least for very frequent syllables, there must be enough exemplars to be useful as a reference in speech production without resorting to the segment level.

As for determining which are the very infrequent syllables, the frequencies observed in the MS corpus cannot be relied on. Instead the frequency classification of the syllables was based on syllable probabilities induced from multivariate clustering (Müller et al. 2000), which allows estimation of the theoretical probability even for unseen syllables. In Müller et al. (2000), probabilities were obtained for a total of 41,711 German syllable types, ranging from $4.61*10^{-11}$ (for the syllable [R@sk]) to approximately 0.0259 (for the syllable [de:6]). The MS corpus contains 3,793 syllable types, which means that there could be almost 38,000 syllable types missing. Here the question arises how realistic the number of 41,711 is as an estimation of the number of different syllable types. It is worth noting that actually existing syllables can be found even among the least probable ones. On the other hand, some of them seem very unlikely, such as the syllable [R@sk] mentioned above for instance, since the consonant combination [sk] in the coda to my knowledge can only occur in the context of stressed syllables in morpheme final position in non-native German words (with the exception of the native German word *brüsk, [brYsk]*), and therefore should not occur in combination with the unstressed schwa [@], which usually only occurs in native German words.

For comparison, Celex (Baayen et al. 1995) contains only approximately 11,000 syllable types. But this is by no means the upper limit of different syllable types in German. For instance, the MS corpus contains syllables occurring in existing words that are not listed in Celex. Also, more words are used in German than can be expected to be listed in such a dictionary. German proper names for instance contain many types that are not listed in Celex. For ex-

ample, according to a cliché, the most popular German surname is *Schmidt*, pronounced [SmIt], but [SmIt] as a syllable does not occur in Celex. The syllables [ja:n] as in one possible pronunciation of *Jan*, [klaUs] as in *Klaus*, [jY6] as in *Jürgen*, all very frequent male first names, and the syllable [pOts], as in the German city *Potsdam*, are also missing. Other existing German syllables that are not in Celex comprise clitizations (e.g. [dU6Cs] as in *durchs*, which results from appending the clitic *'s* to the preposition *durch*, or [tsUm] as in *zum*, the popular short form of *zu dem*) and inflectional forms not listed in Celex (such as [hE6n] as in *Herrn*, the dative form of *Herr*). These are just some examples, more can be found easily. This leads me to conclude that realistically, there are many more than 11,000 syllable types in German, with the upper limit being approximately 41,000. This means that the number of syllables missing from the corpus is somewhere between many more than 7,000 and 38,000.

Summing up these considerations, it can be said that there are many existing syllable types that are not represented by even one token in a speech corpus of 160 minutes of speech. It is therefore likely that for very infrequent syllables, there are not enough exemplars stored in memory to serve as a reference in speech production, and that the respective segments must be used as targets instead.

4.4.2 Dual-route phonetic encoding

Assuming that not all syllables are stored in the syllabary but only the most frequent ones, it follows that gestural scores or trajectories through target regions have to be computed on-line for the less frequent syllables. Then, two different strategies would be used by speakers when producing frequent vs. infrequent syllables. This has been called dual-route phonetic encoding by Whiteside and Varley (1998a), with a direct route, and less computational demand, for frequent syllables, and an indirect route for infrequent syllables, which requires on-line computation of gestural scores.

Whiteside and Varley (1998a) hypothesize that the direct route encoding should be characterized by "higher syllable cohesion" and therefore "increased coarticulation or gestural overlap" (Whiteside and Varley 1998a, p. 3155) (in addition to shorter response latencies in production experiments). This hypothesis is supported by their view that in apraxia of speech "either the access to and/or storage of verbo-motor patterns are disrupted" (Whiteside and Varley 1998b, p. 223), and that patients suffering from apraxia of speech, who would then always have to use the indirect route, show indeed less coarticulation and inconsistent articulatory movements, among other things. They also claim (Whiteside and Varley 1998a, p. 3155) that the indirect route is probably used in careful speech, which again exhibits less coarticulation than normal spontaneous speech.

To substantiate their hypothesis, Whiteside and Varley (1998a) measure response latencies, utterance and word durations, and second formant frequency changes of frequent vs. infrequent English words. However, significant effects of word frequency are only confirmed for response latencies, supporting the hypothesis that different strategies are used in the production. As for the second formant changes as a measure of coarticulation, the effect is not significant. This is attributed to the small sample size of only 30 tokens in each condition and to inter-speaker variability.

Another reason for their failing to find significant differences other than response latencies may be the fact that they control for word frequency rather than syllable frequency, although the model proposed by Levelt and Wheeldon (1994), upon which they base their hypothesis, assumes syllables as the unit of articulation and not words.

An experiment aiming at differences in the degree of coarticulation in syllabary and non-syllabary syllables was conducted by Dogil, Ostry, and Schiller in 2000 (Dogil, personal communication) in an unpublished pilot study. They did find differences in the degree of coarticulation when looking at (labial) articulatory movements of syllables assumed to be in the syllabary vs. those of non-syllabary syllables. More precisely 12 two-syllable nonsense words, covering the four possible combinations of English syllabary and non-syllabary syllables, were embedded in an English carrier sentence and repeated 25 times each by a native (Canadian) English speaker. The classification of syllables as belonging to the syllabary was based on syllable frequencies obtained from the British National Corpus. For the non-syllabary syllables BST-ELITE lip tracings[14] of the 25 repetitions were more consistent than for the syllabary syllables, confirming that there was less coarticulation for the infrequent non-syllabary syllables.

4.4.3 The dual-route hypothesis in the temporal domain

With respect to temporal planning, the analogous assumption is that if, because of a lack of an appropriate syllable-level target, a very infrequent syllable is produced by concatenating segments, then the z-score of the resulting realization of the syllable should depend on the z-scores of the involved segments. There should be less dependency for very frequent syllables, because then the speaker does not access exemplars of the involved segments but directly uses exemplars of the syllable as a reference. For instance if a speaker intends to articulate a syllable lengthened by a z-score of 2, but does not have enough exemplars of the syllable to use as a reference, he will articulate the syllable using exemplars of the involved segments with a z-score of 2. Consequently, more variation can

[14]Two-dimensional movements of two reflector points on the upper and lower lip of the speaker are registered by two cameras, using a third reflector point on the nasion as a reference. The two camera tracings are combined to obtain a three-dimensional tracing.

be expected for frequent syllables than for infrequent syllables when looking at the relationship between syllable z-scores and the z-scores of the corresponding segments.

To assess the validity of the hypothesis, the specifics of the MS corpus can be exploited. Since the corpus was designed for unit selection synthesis (cf. section 4.1), one objective was to have a good coverage even of phonemes in infrequent contexts, and to have at least the same coverage as a diphone corpus. Therefore, after optimizing coverage for phoneme/context vectors, sentences containing diphone types that were not found in the corpus were manually added. As a consequence, the corpus differs from a randomly collected database in that it exhibits an unusual syllable frequency distribution with disproportionately many instances of some otherwise infrequent syllables. This makes it possible to compute z-scores for realizations of these syllables even though they should usually not be represented by a sufficient number of exemplars in a speaker's memory for him to have established stable target regions.

To test the hypothesis, two linear regression models were calculated, one for very frequent and one for very infrequent syllables. Both models predict the syllable z-score from the mean z-score of the involved segments. The criterion for identifying infrequent and frequent syllables was a probability of less than 0.00005 according to Müller et al. (2000) for very infrequent syllables, and of more than 0.001 for very frequent syllables (cf. section 4.4.1).

To obtain reliable z-scores, only those syllables were taken into account for which there were more than 20 realizations in the MS corpus. Some of the most frequent syllables obviously came from frequent function words which usually do not carry pitch accents. To avoid overrating the effects of typical prosodic contexts, such syllables were explicitly excluded, only taking types into account for which there was at least one instance carrying a pitch accent in the database. This left 108 very frequent and 16 very infrequent syllable types which met the requirements, adding up to 4,548 and 742 tokens, respectively. Figure 4.11 shows the mean z-scores of involved segments plotted against syllable z-scores for frequent (left panel) and infrequent (right panel) syllables, together with the regression lines.

For the two linear regression models, the residual standard errors were 0.391 for frequent and 0.340 for infrequent syllables. This indicates that indeed the model for frequent syllables is less accurate in predicting syllable z-scores from mean segment z-scores, confirming that there is a stronger linear dependency between the two values for infrequent syllables. To determine whether this difference is significant, the Bartlett test for homogeneity of variances was applied to the residuals. The test confirmed that the variances are significantly different ($p \ll 0.0001$) and thus supports the hypothesis: There is a stronger relationship between mean segment z-score and syllable z-score for infrequent than for frequent syllables.

This result confirms that mean segment z-scores better predict the syllable

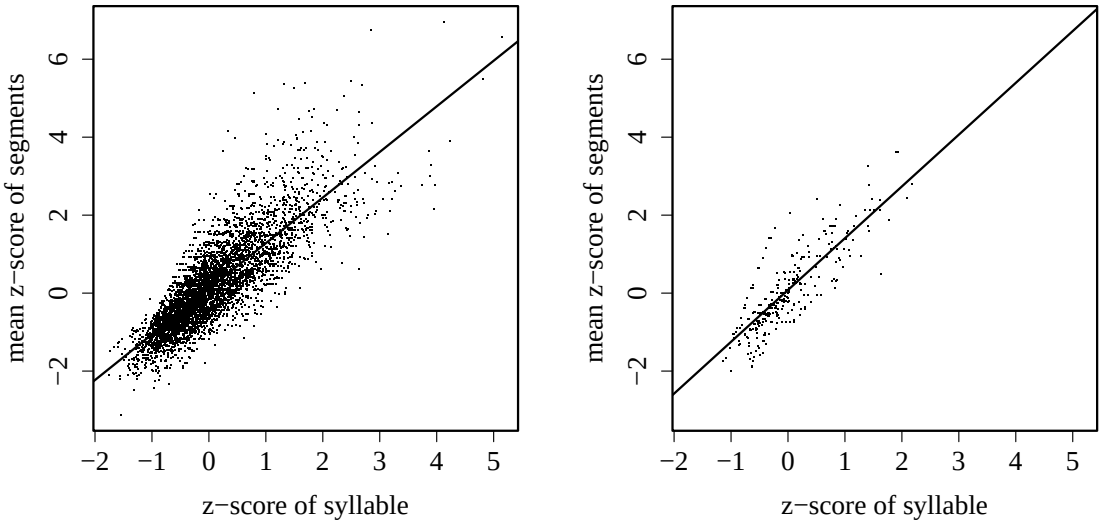

Figure 4.11: Mean z-scores of segments within a syllable plotted against z-score of the syllable for frequent (left panel) and infrequent (right panel) syllables.

z-score in the case of infrequent syllables. Note that the regression lines in figure 4.11 almost go through the origin and have a slope of approximately 1. This indicates that the mean segment z-scores in both cases not only predict the syllable z-score; they are even almost equal to the syllable z-score. However, the residuals are slightly greater in the case of frequent syllables, indicating that there was more variation in the relationship between segment and syllable z-score. Assuming that for frequent syllables, speakers establish a syllable z-score target, while for infrequent syllables, they establish a sequence of segment z-score targets would predict this outcome. For the segment targets, the durations of the constituent segments are given by the their z-scores. More variability is possible for syllable targets: here, only the z-score of the syllable is fixed, but there is no prescription how the duration should be distributed to the constituent segments.

The Multilevel exemplar model (Walsh et al. 2010) described in section 2.3.4 models production using such a dual-route approach. Specifically, it assumes that complete syllables ("units") are used as targets in production if their activation exceeds a certain threshold. Since strength of activation depends on the number of exemplars of that type, this will usually be the case for frequent syllables. If the activation level is not high enough (for instance for lack of exemplars of that type), which will typically be the case for infrequent syllables, the syllable is instead produced using "constituent" segments as production targets. Walsh et al. (2010) present a simulation experiment in which the model is seeded with typical durations form the MS corpus used here. It then iteratively produces syllables, adding the resulting syllables as exemplars after each

step. The model generates each syllable type *n* times, where *n* reflects the syllable's prior probability, i.e, frequent syllables are produced more often than infrequent syllables. Thus, the number of frequent syllable exemplars at some point becomes sufficient to allow for unit level production, i.e., for using a syllable exemplar as production target rather than exemplars of its constituent segments. The model obtained in this way then produces each syllable type 500 times (without storing the resulting syllables to avoid compromising the relative frequencies of syllable exemplars in memory). The syllable and segment durations generated that way exhibit the same behavior as the durations in the MS corpus: in case of infrequent syllables, mean segment z-scores better predict the syllable z-score than in case of frequent syllables.

4.5 Conclusion

In this chapter I have presented an exemplar-theoretic interpretation of Guenther and Perkell's speech production model (Guenther et al. 1998; Perkell et al. 2000) for the prosodic domain. I have suggested that z-scores of speech unit durations are the temporal dimension in the perception and production of prosody. From an exemplar-theoretic perspective, the z-score of a speech unit can be interpreted as its position in the exemplar cloud with respect to the duration dimension. Assuming that speech units are stored including their durations, which is in accordance with the exemplar-theoretic models discussed in section 2.3, speakers and listeners have access to the position of exemplars in the cloud and therefore to their temporal z-score.

The exact nature of the speech units is left open. Following Levelt (1999), I have assumed here that the basic unit in articulation is the syllable, and that speakers possibly have to resort to the segment level in the case of very infrequent syllables, as suggested by the dual-route hypothesis discussed in 4.4.3. Experimental results from a large speech corpus seem to corroborate this view. However, only 16 infrequent syllable types were frequent enough in the corpus to be considered in the experiment, which is probably not enough to be representative of the large number of infrequent syllables. On the other hand, simulation results presented by Walsh et al. (2010) confirm the effect. They do not face the data sparsity problems because in the simulation, any infrequent syllable can be produced as often as necessary. Still, I believe that carefully controlled production experiments would be valuable to add another methodology of assessing differences in variability between frequent and infrequent speech units.

In any case, I have suggested that temporal z-score distributions of speech units related to prosodic events implicitly define target regions in the production of temporal properties of prosodic events. As stated in section 2.3.5, this is similar to Keating's (1990) assumption that target regions are implicitly defined

by the range of observed values along a dimension. I have presented experimental results confirming that the z-score distributions of segments and syllables in different prosodic contexts are significantly different, i.e., from a theoretical perspective, they could serve to distinguish between different prosodic events. In other words, z-scores are an appropriate perceptual measure to make the target regions for different prosodic events more distinct from each other. Of course, there is still a high amount of overlap between the distributions. Obviously, the temporal z-scores alone will not make the regions sufficiently distinct from each other. I will turn to tonal aspects of prosody in the following chapter, and propose that when considering temporal and tonal dimensions together, the resulting multidimensional target regions may be sufficiently distinct.

Chapter 5

The tonal dimension of perceptual space

When turning to the tonal dimensions of the target regions of prosodic events, again three questions arise, just as for the temporal dimension discussed in chapter 4. These questions are, in analogy to the temporal domain: (i) what are the tonal dimensions of perceptual space, (ii) which are the relevant prosodic events, and (iii) which are the target regions corresponding to these events.

With regard to question (ii), as suggested in the introduction to chapter 4, I will assume for now that phrase boundaries and pitch accents according to GToBI(S) (Mayer 1995) are the relevant prosodic events. Since pitch accents and boundary tones are defined as tonally distinct units, it is self-evident that they will differ with respect to their tonal properties.

I will suggest to answer question (i) in this chapter by proposing the PaIntE parameters introduced in chapter 3 as perceptually relevant tonal dimensions. To motivate the PaIntE parameters as the perceptually relevant tonal dimensions, I will explain how they capture relevant tonal properties of the pitch accents and boundary tones of the GToBI(S) labeling system in section 5.1.

In analogy to the temporal domain discussed in chapter 4, I will then propose to answer question (iii) by assuming that the distributions of PaIntE parameters related to prosodic events can be regarded as target regions in the production and perception of tonal aspects of prosodic events, as explicated in section 5.2.

5.1 Interpretation of the PaIntE parameters

This section is intended to illustrate the relation between PaIntE parameters and the tonal implementation of pitch accent and boundary types. For four of the six parameters, I will explain which values of the parameters one would expect for which accents and boundaries, and show how these expectations

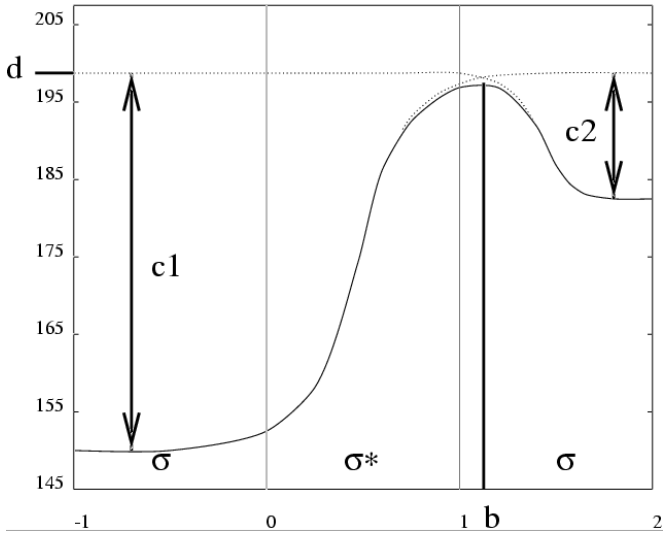

Figure 5.1: Schematic of the PaIntE approximation function repeated from figure 3.3.

are borne out using density plots of the parameter distributions of the different accent and boundary types found in the SWMS corpus.[1,2] I will also contrast their parameter distributions with those of unaccented syllables. Although I will only use data from one male speaker here, the experiments described in chapter 6 will show that prosodic classifiers trained using, among others, the PaIntE parameters generalize well to a similar corpus of a female speaker, which to some extent indicates that the tonal properties found in the PaIntE parameter distributions of the SWMS corpus are not speaker-specific but reflect general properties of the German intonation system.

The PaIntE approximation and the resulting parameters have been discussed in section 3.2 already. The graphical display of an example pitch contour as specified by the PaIntE parameters is repeated in figure 5.1. From the schematics of the pitch accents of GToBI(S), which had been discussed in section 3.1, and which is repeated in figure 5.2, it can be seen that almost all accents involve some kind of peak (L*H, H*L, H*) or at least high plateau (HH*L) in the pitch contour, with the exception of completely linked L*, in which case only the left ("valley") part of the contour given in figure 5.2 is realized. The peaks

[1]The SWMS corpus is described in some detail in section 6.1.2 because the exact specification is more important for the experiments presented in chapter 6 than it is here. Here, it suffices to say that the corpus consists of approximately 2 hours of speech of a male speaker, that the speaker was the same as in the MS corpus used in chapter 4, and that the corpus was originally recorded for unit selection speech synthesis.

[2]For the two remaining parameters, the expectations are less clear, and I will not discuss them here.

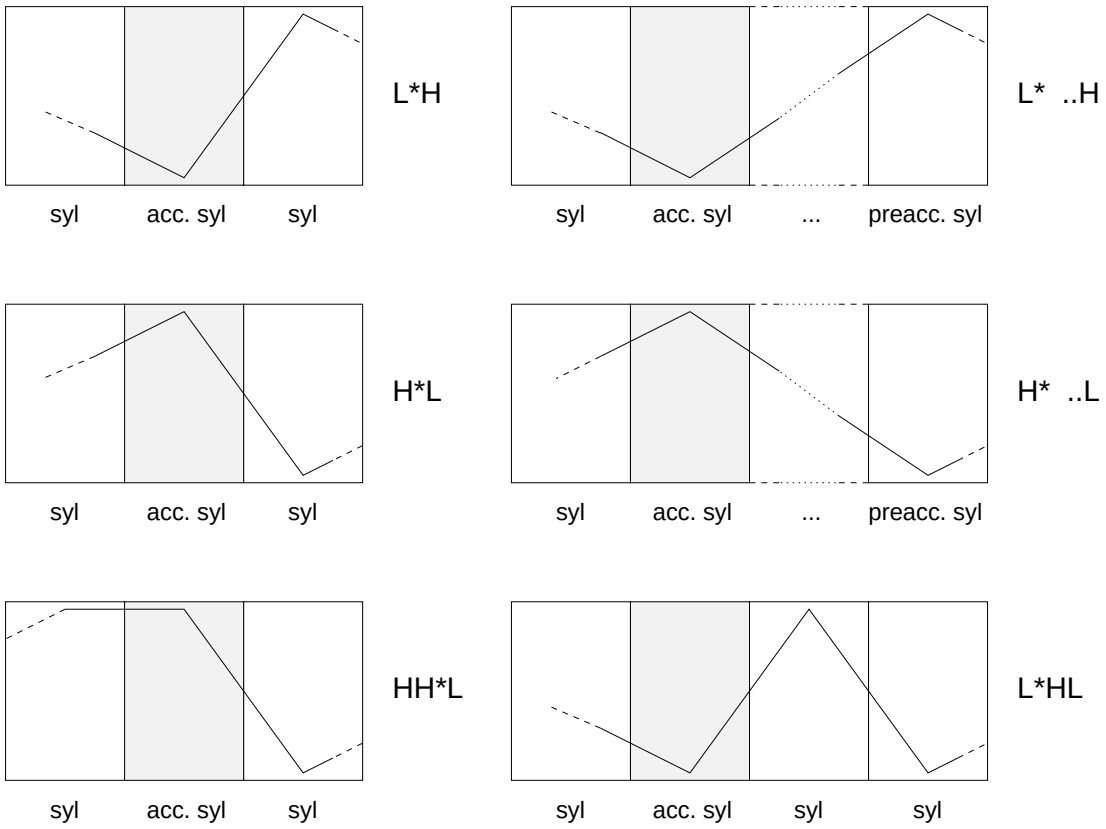

Figure 5.2: Schematic diagrams of the pitch contours of the GToBI(S) pitch accents, repeated from section 3.1. Boxes represent syllables, accented syllables are highlighted in gray color. Dotted lines indicate one or several interceding syllables. Dashed lines indicate that the actual contour arises from interpolation to targets of preceding or following accents.

can be modeled by the peak of the PaIntE function.

It may seem less promising at first glance to model L* using the PaIntE function, but I have described in chapter 3 that for syllables for which no F0 peak can be detected, the approximation is carried out using just the rising or just the falling part of the PaIntE function for the approximation, i.e., using just one of the single sigmoids indicated by the dotted lines in figure 5.1. In that case, the *b* parameter still determines where that sigmoid reaches its maximum, i.e., it specifies where the rise comes to an end or where the fall starts, and the corresponding *c* parameter (*c1* for a rise, *c2* for a fall) still represents the amplitude of the rise or fall. Thus, L* accents are expected to be modeled using just the rising sigmoid indicated as rising dotted line in figure 5.1. This may still be suboptimal, because it does not explicitly model a valley, but the start of the rise should still correspond to the low target for L*. Also, at least in my

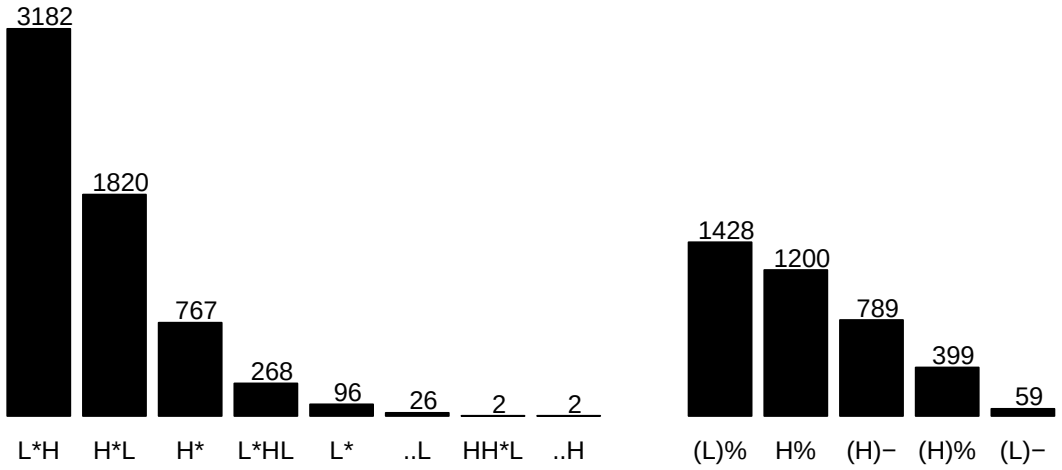

Figure 5.3: Histograms of pitch accent types (left panel) and boundary types (right panel) in the data used for the experiments presented in this chapter. The absolute numbers for each type are plotted just above the corresponding bar.

experiments, the question of whether the modeling is adequate for L* or not was of less importance because of the rarity of L* in our data (only 1.6% of the pitch accents in the SWMS corpus were L* accents).

For examining the PaIntE distributions of boundary tones, I have modified the inventory of boundary tones to disambiguate between tonally unspecified boundaries (%,-) that are high and those that are low. As mentioned in section 3.1, unspecified boundaries "inherit" their specification by way of spreading of the preceding trail tone. Therefore I differentiate between the following intonation phrase boundaries: (H)% boundaries, which are high because the preceding high trail tone is spread, H% boundaries, which are high inherently, and (L)% boundaries, which inherit the specification as low from the preceding low trail tone[3]. Analogously, for intermediate phrase boundaries, I distinguish between (H)- boundaries, which occur after high trail tones, and (L)- boundaries, which occur after low trail tones.

As for the frequencies of the prosodic events in the part of the SWMS corpus that I will be using in this chapter for illustrating the relation between PaIntE parameters and pitch accent and boundary types, there were approximately 6,200 pitch accents and 3,900 boundaries. Figure 5.3 shows a histogram of the pitch accent types and their frequencies. It can be seen that some of the pitch accents were quite rare in the data: L* occurs only 96 times, and HH*L

[3]Inherently low L% boundaries were not labeled in the data of this particular speaker because almost all full phrase boundaries are glottalized in his case, so that it was not possible to consistently distinguish between L% and low unspecified % boundaries.

occurs only twice. The linked trail tones ..L and ..H are also very infrequent, with only 26 and 2 occurrences, respectively. As for phrase boundaries, low intermediate phrase boundaries (L)- are also not frequent, with only 59 occurrences. Since I am going to examine the parametrization results by interpreting plots of estimated probability distributions (also referred to as "density plots" in the following),[4] I will for the sake of reliability only discuss the results of those prosodic events that are represented by at least 200 tokens. This leaves the four most frequent accents L*H, H*L, H*, and L*HL, as well as H%, (H)%, (L)%, and (H)-. I will also compare the probability distributions of the pitch-accented syllables with those of unaccented syllables, and those of phrase-final syllables with those of non-final syllables.

5.1.1 Peak alignment

As stated above, the peak of the PaIntE function corresponds to the peak that most pitch accents exhibit. Its alignment with the syllable structure is determined by the *b* parameter. If *b* is between 0 and 1, the peak is on the accented syllable. This is what one would expect for all accents that involve an H*, i.e., all accents that have a high target on the accented syllable. However, in the HH*L case, the peak could be realized almost as a plateau, with similar pitch across both the pre-accented and the accented syllable, giving rise to *b* values anywhere between -1 and 1, i.e., the peak will be either on the pre-accented syllable which is associated with the high target for the leading tone, or on the accented syllable associated with the high target for the starred tone, depending on which of the two syllables is realized with (possibly only slightly) higher pitch.[5] For values of *b* greater than 1, the peak is on the post-accented syllable or later. This should be the case for L*H and L*HL, where the trail tone causes a high target in the pitch contour.

Figure 5.4 shows density plots of the *b* parameter. Generally, in density plots, peaks appear at values that are more likely to occur for the underlying sample, whereas valleys appear at values that are less likely to occur. The density plots for different accent types can be compared because the density functions are normalized so that the area below the line is equal to 1. Thus, for instance, if the plots for two different pitch accent types have equally high peaks for a particular value of *b*, it can be inferred that for both accents, that particular value of *b* is equally likely. It can not be inferred, however, that given that particular value of *b*, the two accent types are equally likely, because of the

[4] All density plots were generated using the *density* function implemented in R (R Development Core Team 2009), which computes kernel density estimates for a sample of a population in order to estimate the probability distribution of the entire population.

[5] There are only two instances of HH*L in the SWMS database. For these, the expectation is confirmed by values of approx. -0.3 and 0.4, respectively. Distributions of HH*L accents will not be discussed in the following because of the sparsity of instances of this accent type.

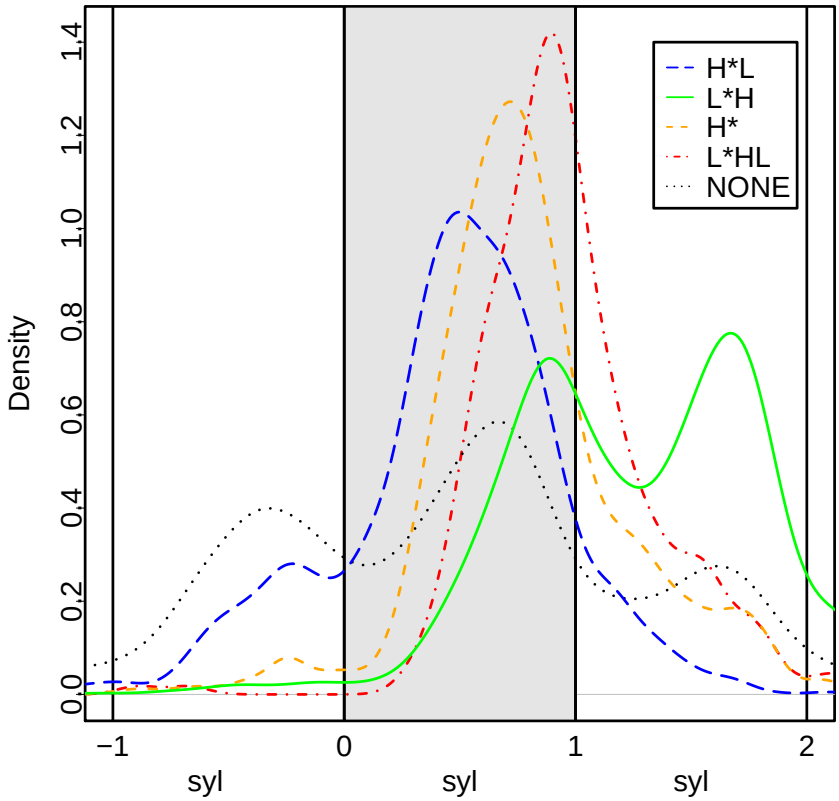

Figure 5.4: Density plots of the b *parameter for the most frequent accents and for unaccented syllables. The syllable for which the approximation was carried out is highlighted in gray. H*L accents (blue, long-dashed line) have their peak earlier in the accented syllable than H* accents (orange, dashed line) and L*HL accents (red, dot-dashed line). L*H accents (green, solid line) have their peak either on the accented or on the post-accented syllable. For unaccented syllables (black, dotted line) peaks are detected on any of the three syllables of the analysis window.*

differences in absolute frequencies of the accent types themselves.

The density plots for the *b* parameter in figure 5.4 show that the above expectations are borne out. The vertical lines indicate syllable boundaries. The middle syllable, which is highlighted in gray, is the syllable for which the approximation was carried out. In the distributions for pitch-accented syllables, this is the syllable associated with the accent. The blue, long-dashed line represents the probability distribution for H*L accents. The relatively broad peak in the middle of the accented syllable confirms that H*L accents have their peak somewhere in the course of the accented syllable, but not at the very beginning.

The peak is roughly between 0.3 and 0.8. I will discuss H*L accents in more detail below (cf. figure 5.6).

It can further be verified in this figure that H* accents, which are represented by the dashed orange line, also have their peak on the accented syllable; it is interesting, however, that for H* accents, the peak is more likely to be later in that syllable than for H*L accents.[6] The peak is also a little more narrow, indicating less variation of *b* for H* accents than for H*L accents.

For L*HL accents, which are indicated by the red, dot-dashed line, surprisingly, the peak is also on the accented syllable. From the description of L*HL accents given in section 3.1, one would expect that peak to be on the post-accented syllable. Still, compared to H*L and H*, the peak for L*HL accents is shifted further towards the syllable boundary. It is also a little more narrow than the two peaks of the H*L and H* distributions, indicating even less variation of *b* for L*HL accents.

Finally, the density for L*H accents (solid green line) is bimodal: L*H accents are almost equally likely to have their peak either right before the syllable boundary, just as L*HL accents did, or to have their peak in the later part of the post-accented syllable. One might interpret this as evidence for two distinct categories L+H* (with the peak on the accented syllable) and L*+H (with the peak on the post-accented syllable) as in the Saarbrücken dialect of GToBI (Grice and Baumann 2002; Grice et al. 2005); however, it turns out that L*H is realized differently on word-final syllables than on non-final syllables,[7] as demonstrated in figure 5.5. Here, the dashed line represents L*H accents that occurred on word-final syllables, and the dot-dashed line represents L*H accents that occurred on word-internal syllables. Obviously, these two contexts cause the bimodal distribution: L*H accents on word-final syllables almost always have their peak in the accented syllable, while word-internal L*H accents tend to have their peak on the post-accented syllable. In other words, the tonal movement on L*H accents usually does not cross word boundaries, instead it is timed to occur earlier before word boundaries.

As for the difference between accented and unaccented syllables, it is obvious from figure 5.4 that unaccented syllables (black, dotted line) vary much more in where the peak is detected. This is because peaks on unaccented syllables are often due to microprosodic variation in pitch rather than to conscious pitch movements and are thus almost equally likely to occur on any of the surrounding syllables. I claim that the reason why they are discovered a little more often on the middle syllable is that this is the place where the algorithm starts to look for a maximum in the implementation of the PaIntE approximation (cf. section 3.2.3).

Altogether, figures 5.4 and 5.5 show that peaks for H*L accents usually oc-

[6] All distributions will be compared using confidence tests at the end of this section.

[7] Thanks to Jörg Mayer for pointing me in the right direction.

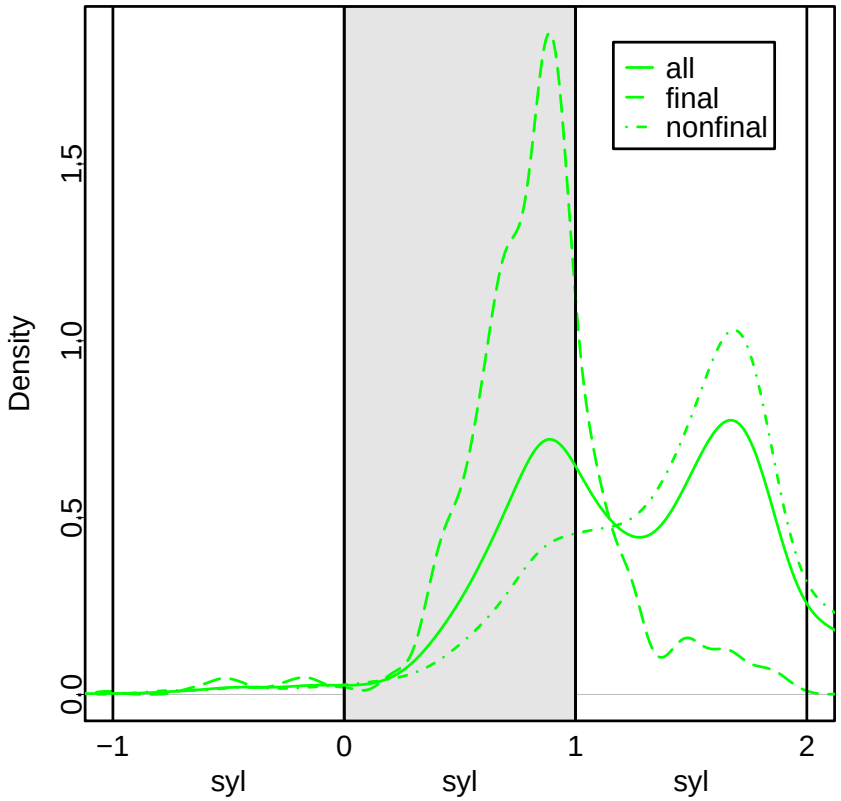

Figure 5.5: Density plots of the b *parameter for L*H accents in different contexts. The accented syllable, for which the approximation was carried out, is highlighted in gray. The bimodal distribution for all L*H accents (solid line, repeated from figure 5.4) obviously arises from the fact that L*H accents in word-final syllables (dashed line) have their peak on the accented syllable, and L*H accents in word-internal syllables (dot-dashed line) have their peak on the following syllable.*

cur earlier in the accented syllables than for H*, peaks for H* accents occur earlier than peaks for L*HL accents and word-final L*H accents, and that word-internal L*H accents have their peak on the post-accented syllable. Since the distributions overlap, the accent types can not generally be inferred from the value of the *b* parameter alone. Such an inference may be valid for peaks that occur on the pre-accented syllable: in this case, it is very likely that the syllable is unaccented. This is because it is very unlikely for the pitch accents discussed above to have their peak on the pre-accented syllable, but not unlikely at all for unaccented syllables, and because, in addition, unaccented syllables are in general more likely than accented syllables because of their dominance in fre-

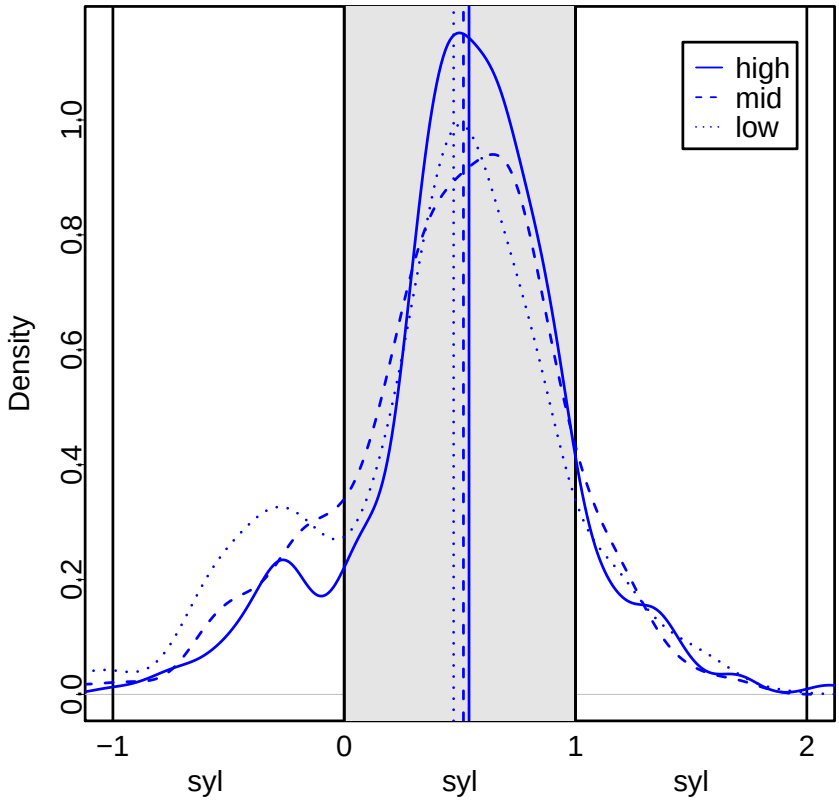

Figure 5.6: Density plots and mean values (vertical lines) of the b *parameter in H*L accented syllables with low (dotted lines), mid (dashed lines), and high vowels (solid lines).*

quency.

Another interesting fact that can be observed in figure 5.4 is that peaks are generally unlikely at the beginning of any syllable. This is in line with House's (1996) model of optimal tonal perception, which claims that tonal movements through areas of maximum new spectral information and intensity change (such as syllable onsets) are perceived as level tones rather than as tonal movements. The model states that gestures of tonal movement must be synchronized to occur after these areas in order to be perceived as such. The valleys occurring in the density plots in figure 5.4 for all pitch accent types and all syllable onsets confirm that our speaker avoided placing peaks (and thus, placing the beginning of the falling tonal movement) in the spectrally unstable onsets.

Returning to H*L accents, Jilka and Möbius (2007) observed using the MS

corpus (cf. section 4.1) that peak alignment in H*L accents depends on the height of the vowel in the accented syllable. They report that the point in the accented syllable where the maximal F0 value is reached is later for higher vowels: for low, mid, and high vowels, the maxima occur at 30.9%, 34.5% and 41.8% of the voiced part of the syllable, respectively (median values). They used pairwise t-tests to confirm that the alignment for the three vowel heights differ significantly, with $p<0.001$ when comparing high with non-high vowels, and $p<0.005$ when comparing mid vs. low vowels. The distributions of the *b* parameter observed here for the three cases confirm these differences. For high vs. low vowels, they are significantly different[8] with $p<0.001$; for high vs. mid vowels, they are significantly different with $p<0.01$, and for mid vs. low vowels, the difference is marginally significant ($p<0.05$). The means of *b* are 0.37 for low vowels, 0.48 for mid vowels, and 0.52 for high vowels. The distributions and the means are depicted in figure 5.6. The effect is significant in two of the three cases, but very subtle: the means, which are indicated by the vertical lines, are very close, in particular for mid vs. high vowels, even though the difference was significant with $p<0.01$.

Turning to phrase boundaries, there are systematic differences between the probability distributions of the *b* parameter for different boundary tones. Figure 5.7 shows the density plots for the most frequent boundary types. Here, four syllables are shown instead of just the three syllables in the analysis window, because in the case of low boundaries, the peak was sometimes found to be to the left, on the syllable just outside the window, i.e., two syllables before the actual phrase boundary. Therefore, the preceding syllable was included in the plot, and thus the penultimate syllable in figure 5.7, which is highlighted in gray, is the syllable that has been analyzed. In the distributions for the boundary tones, this is the phrase-final syllable.

For the high boundaries, viz. H% (blue, long-dashed line), (H)% (red, dashed line), and (H)- (orange, dot-dashed line), peaks are almost always detected late in the final syllable. However, the detected peaks only indirectly correspond to production targets: On the phrase-final syllable, the pitch contour ends at a high level in the speaker's register. The next syllable belongs to the next phrase and therefore is usually produced at a lower level. This does cause a peak in the contour, because of the subsequent drop in F0. But one would probably not want to interpret this drop in F0 as an intentional falling tonal movement, rather, one would assume two underlying targets: the rise to the high boundary, and, subsequently, a new, lower target for the next syllable.

Still, the PaIntE parameters capture valid, if not in every case intentional, F0 movements occurring at phrase boundaries. For instance, the distribution for H% boundaries shows that their peaks tend to occur even later in the phrase-

[8] All statistical tests reported in this chapter have been conducted using the R statistics package (R Development Core Team 2009).

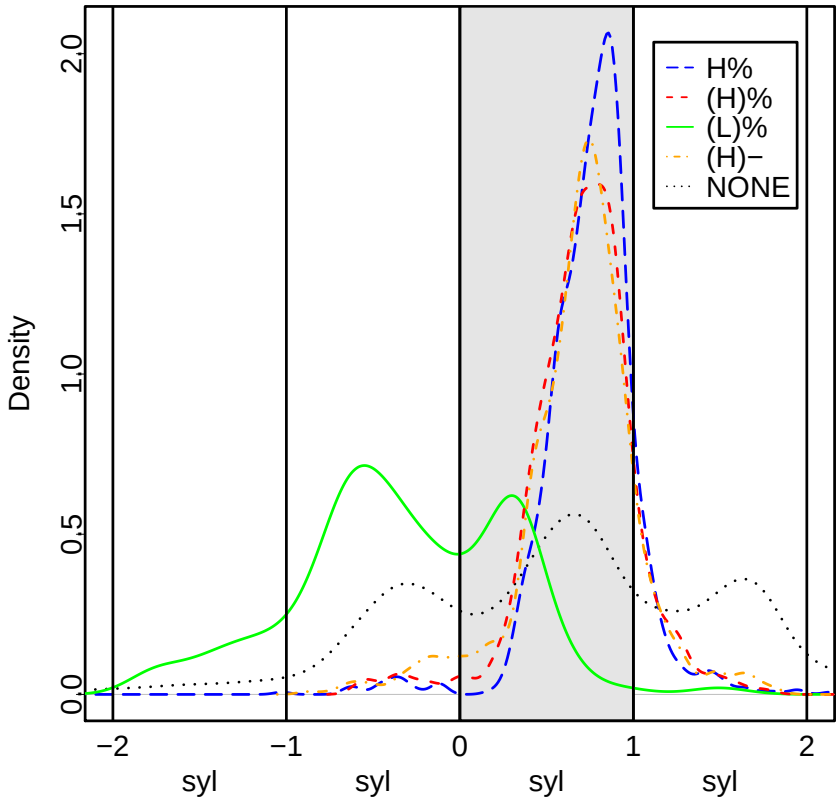

Figure 5.7: Density plots of the b *parameter for the most frequent boundary types and for non-final syllables. The syllable for which the approximation was carried out is highlighted in gray. For high boundaries (H%, blue, long-dashed line; (H)%, red, dashed line; (H)-, orange, dot-dashed line) peaks are usually detected late in the phrase-final syllable. For (L)% (green, solid line) there is a peak either on the phrase-final syllable or on the preceding syllables but almost never on the following syllable. For non-final syllables (black, dotted line), peaks are detected on any of the three syllables in the analysis window.*

final syllable than those for (H)% and (H)- boundaries. This is because H% boundaries are expected to exhibit a contour that is rising throughout the syllable and thus the maximum will be reached later; (H)% and (H)- boundaries, on the other hand, are expected to be associated with a level contour on the final syllable, thus the exact location where the peak is detected is to some extent subject to slight random variations in pitch on that syllable, as confirmed by the broader and slightly earlier peaks in the corresponding probability distributions.

Turning back to figure 5.7, the density plot for non-final syllables (dotted, black line) is depicted as well. As in the case of unaccented syllables discussed above (cf. figure 5.4), peaks are likely to be detected on any of the three syllables, with a preference for the middle syllable, but again, I claim that this is due to the fact that PaIntE starts to look for a maximum in the middle of that syllable.

(L)% phrase boundaries, finally, are usually found to have a peak in one of the preceding syllables. Their distribution is indicated by the green, solid line in figure 5.7. For (L)% boundaries, peaks are usually detected on either the final syllable or on the preceding one, which can be seen from the prevalence of values between -1 and 0.5. Occasionally the peak is even on the syllable before that, with values of *b* between -2 and -1. The peak is found on preceding syllables if there is no clear local maximum on the final syllable itself, but on one of the preceding syllables. This may be due to microprosodic variation, or to other prosodic events associated with these syllables. It can be noted here that if the peak is on the final syllable, it is early in the syllable, which is in line with the expectation that the tonal movement on that final syllable should be falling, making the location of a peak in the later part of the syllable unlikely. What is at first surprising, however, is that the peak is never detected on the following syllable: there are virtually no values of *b* greater than 1 in the case of (L)% boundaries. This is because in almost all cases of (L)% boundaries (96.4%), there is a silence following. Since the F0 contour is not interpolated across silences, there is a stretch of the speech signal for which no F0 values are present, and in these cases, the approximation window ends before that stretch. Thus, in almost all cases of (L)% boundaries, values of greater than 1 for *b* are less likely because the approximation is carried out in a window that ends on the syllable associated with (L)%. These values can only arise if the contour is best approximated by a PaIntE function which reaches its peak outside the window, which does not happen often, but is theoretically allowed if *b* is not more than half the window length beyond the window's end (cf. section 3.2.4). Also, in many cases, (L)% boundaries are utterance-final, so often there is no following syllable at all.

To conclude discussing the *b* parameter distributions, it is obvious that the *b* parameter captures meaningful aspects of peak alignment: the tendencies expected based on the phonetic description of the GToBI(S) events are present in the distributions. Also, effects observed by House (1996) and Jilka and Möbius (2007) can be found. It is also evident that the distributions for different types are usually different, even though the differences in some cases are subtle. For both accents and boundaries, pairwise Wilcoxon rank sum tests reveal that the differences in most cases are significant. Table 5.1 lists the corresponding p-values, rounded to 5 digits. Values of 0.00000 indicate that $p \ll 0.000005$. Al-

	NONE	H*L	H*	L*H	L*HL
NONE	–	0.05808	0.00000	0.00000	0.00000
H*L	0.05808	–	0.00000	0.00000	0.00000
H*	0.00000	0.00000	–	0.00000	0.00000
L*H	0.00000	0.00000	0.00000	–	0.00000
L*HL	0.00000	0.00000	0.00000	0.00000	–
	NONE	H %	(H)%	(H)-	(L)%
NONE	–	0.00000	0.00007	0.00000	0.00000
H%	0.00000	–	0.00072	0.00000	0.00000
(H)%	0.00007	0.00072	–	0.56745	0.00000
(H)-	0.00000	0.00000	0.56745	–	0.00000
(L)%	0.00000	0.00000	0.00000	0.00000	–

Table 5.1: p-values obtained by pairwise Wilcoxon rank sum tests comparing distributions of b *for the most frequent accent types (upper panel) and boundary types (lower panel). The distributions for unaccented syllables (NONE) vs. H*L accents and those for (H)% vs. (H)- boundaries are not significantly different at a confidence level of 0.999; all other distributions are significantly different. Values of 0.00000 indicate that $p \ll 0.000005$.*

most all distributions are significantly different at a confidence level of 0.999.[9] Exceptions are the distributions of unaccented syllables (NONE) vs. H*L accented syllables, and those of (H)% vs. (H)- boundaries.

5.1.2 Peak height

Parameter *d* of the PaIntE function determines the absolute height of the F0 peak in Hz. Again, the distributions differ between accent types and between accented vs. unaccented syllables. Density plots for parameter *d* are given in figure 5.8. Again, the distributions differ between accent types and between accented vs. unaccented syllables. Values for *d* in Hertz are indicated on the x axis: values to the right indicate higher peaks.

Figure 5.8 shows that syllables associated with H*L accents (blue, long-dashed line) and unaccented syllables (black, dotted line) exhibit the lowest values for *d*, i.e., their peaks are lower than peaks associated with the other accent types. In both cases, the peak in the distribution and thus the most likely value for F0 peak height is between 115 and 120 Hertz. In the case of

[9] This unusually high confidence level had been suggested in the preceding chapter in examining the duration z-score distributions of the GToBI(S) events to make up for the high number of confidence tests.

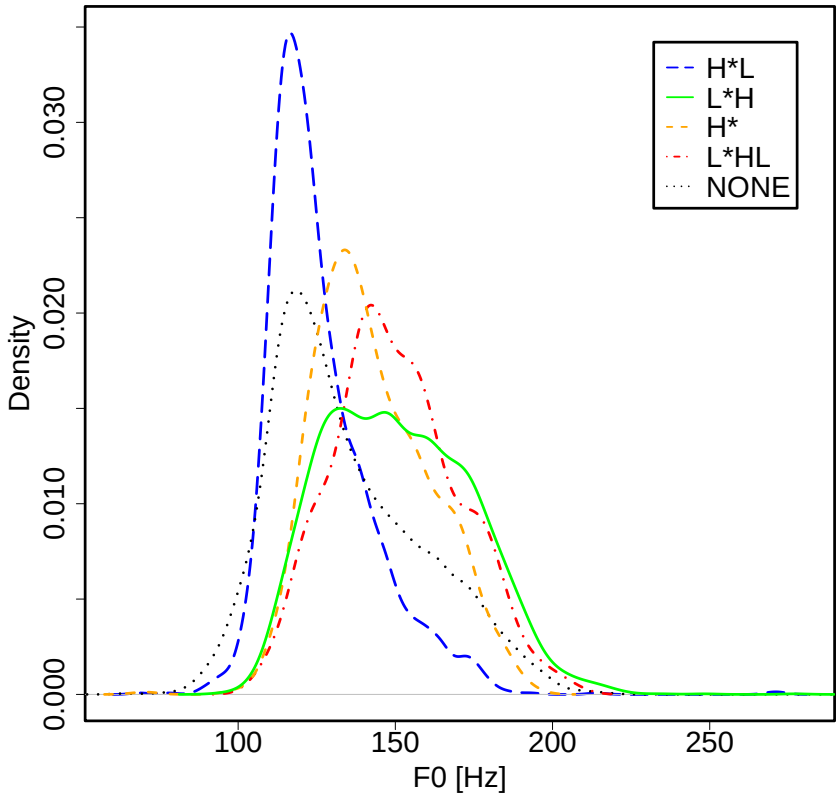

Figure 5.8: Density plots of the d *parameter for the most frequent accents and for unaccented syllables. H*L accents (blue, long-dashed line) and unaccented syllables (black, dotted line) exhibit the lowest values for* d*; H* accents (orange, dashed line), L*HL accents (red, dot-dashed line), and L*H accents (green, solid line) are characterized by higher and more variable values for* d*.*

H*L accents, I claim that this is due to the prevalence of nuclear, i.e., phrase-final, H*L accents over pre-nuclear H*L accents: 93% of the H*L accents are nuclear accents. For nuclear accents, I would expect lower peaks because of the declination. Indeed, the density plot for nuclear H*L accents only (not depicted here) looks almost identical to the one for H*L accents depicted here, with the exception that the peak in the distribution is a little more narrow because nuclear H*L accents are even more likely to have their F0 peak at that location. The density plot for non-nuclear H*L accents however (also not depicted here) is shifted to the right and broader, very similar to the distribution for L*H accents.

In the case of unaccented syllables, the dominance of lower peak heights

arises because peaks on unaccented syllables are often due to microprosodic variation and will thus exhibit very little rise and fall amplitudes (cf. section 5.1.3), so absolute peak heights depend more on where in the pitch range of the speaker the contour happens to be at that particular syllable than on the amplitude of the associated rise or fall. Thus, what is interpreted as a peak in the PaIntE approximation is often just a very local small "bump" rather than a real "peak", and one should probably speak of F0 maxima rather than of F0 peaks in these cases. So, although unaccented syllables most often exhibit F0 maxima of around 115 to 120 Hertz, there is more variation than for H*L accented syllables: higher F0 maxima are also possible, which can be seen from the broader shape of the distribution: values of up to 160 or 180 Hertz are not unlikely.

Figure 5.8 further demonstrates that peaks related to H* accents (orange, dashed line) are usually higher than those of H*L accents (blue, long-dashed line), which would follow, reasoning as above, from the fact that H* accents are always pre-nuclear accents and thus are expected to exhibit higher peaks than nuclear H*L accents because the declination effect is less strong for them than for nuclear H*L accents. Peaks of L*HL accents (red, dot-dashed line) are higher than those for H* (orange, dashed line), H*L (blue, long-dashed line) or unaccented syllables (black, dotted line). Since L*HL accents are claimed to be usually quite prominent and associated with greater amplitudes of rise and fall, this does not come as a surprise. Peaks of L*HL accents are higher even though they are usually nuclear accents in our data (84%).

The distribution for L*H accents (green, solid line) is very similar to that of L*HL accents (red, dot-dashed line), although there is again more variation for L*H accents. Figure 5.9 demonstrates an interesting observation: the effect of phrase-finality observed above for H*L accents is reversed for L*H accents: 60% of them are nuclear accents, i.e., they occur in phrase-final position, but compared to the overall distribution for L*H accents (green, solid line, repeated here from figure 5.8) their distribution (red long-dashed line) is shifted even a little more to the right, with the most likely value for peak height at around 150 Hertz, i.e., nuclear L*H accents tend to have a higher peak than non-nuclear L*H accents. For non-nuclear L*H accents, the effect is similar to that for H*L accents: the more accents follow an L*H within the same phrase, the higher the peak, or, in other words, the earlier in the phrase, the higher the peak: peaks are typically realized at 125 to 130 Hertz for L*H accents that are followed by only one more accent in the same phrase (blue, dashed line), at around 135 Hertz for L*H accents that are followed by exactly two more accents (black, dot-dashed line), and at around 145 Hertz for L*H accents that are followed by exactly three more accents (orange, dotted line).

Concerning nuclear L*H accents, they could exhibit higher peaks if they occur immediately before phrase boundaries, because then the boundary could reinforce the pitch movement associated with the L*H accent. Further analysis

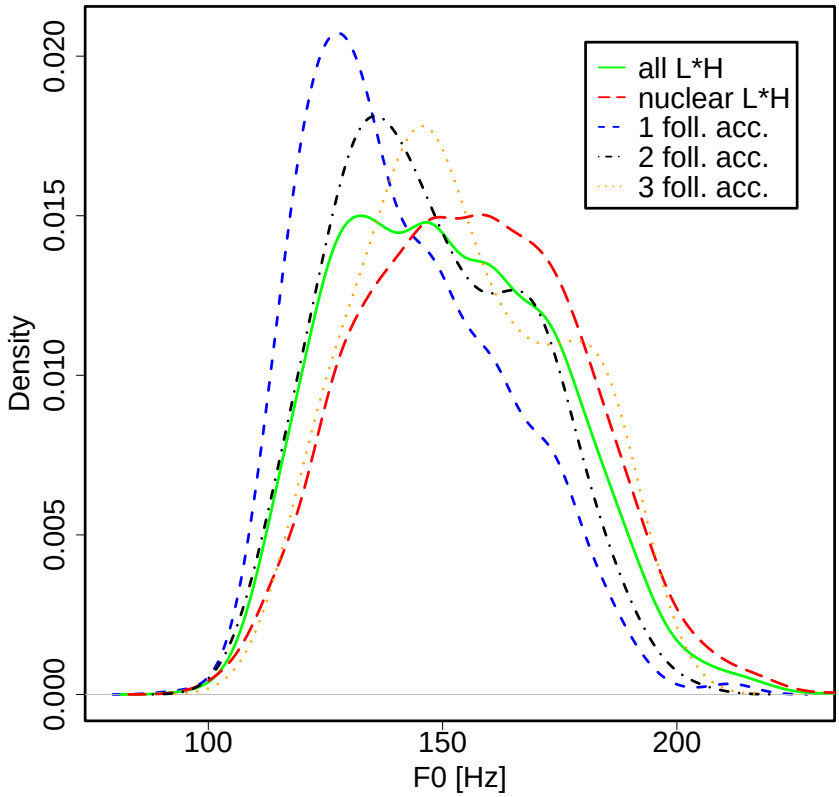

Figure 5.9: Density plots of the d *parameter for L*H accents in different positions in the phrase. Compared to L*H accents in general (green, solid line, repeated from figure 5.8), nuclear (i.e., phrase-final) L*H accents (red, long-dashed line) exhibit slightly higher values for d. For non-nuclear L*H accents, their position in the phrase seems to have an effect on the* d *parameter: L*H accents with only one more accent following in the same phrase (blue, dashed line) tend to have lower* d *values than L*H accents with two accents following (black, dot-dashed line), which tend to have lower* d *values than L*H accents with three accents following (orange, dotted line).*

reveals that this seems to be the case, but only to some extent: when looking at L*H accents for which the phrase boundary does not immediately follow (i.e., not on the syllable itself, and not on the following one), the distribution (not depicted here) is shifted to the left and more narrow, i.e., if the boundary does not immediately follow, lower peak heights are observed, as expected. However, the peak of the distribution is still to the right of the one observed for pre-nuclear L*H with only one accent following, i.e., peak height is still not

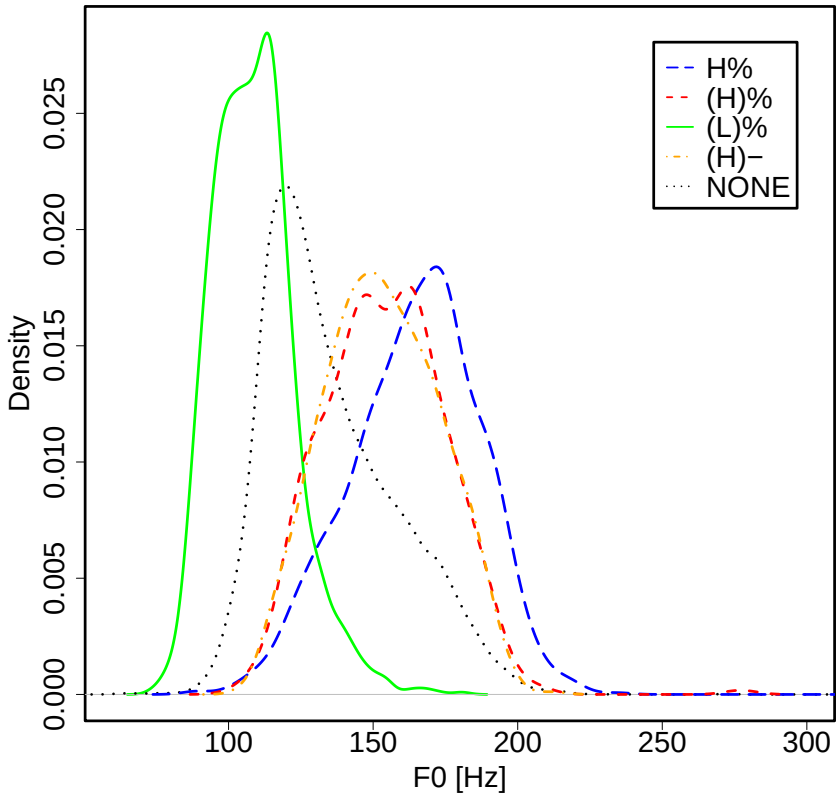

Figure 5.10: Density plots of the d *parameter for the most frequent boundary types and for non-final syllables. Low boundaries ((L)%, green, solid line) exhibits low values for d, while high boundaries (H%, blue, long-dashed line; (H)%, red, dashed line; (H)-, orange, dot-dashed line) are characterized by higher values. The* d *values for non-final syllables (black, dotted line) are in between.*

ordered according to the position of the L*H in the phrase, even when excluding syllables immediately before the phrase boundary.

Considering phrase boundaries, the *d* parameter also captures differences in the distributions of the boundary types.: Figure 5.10 confirms that low boundaries exhibit values that differ very clearly from those of high boundaries: The density plot for (L)% boundaries, which is given by the green solid line, documents a preference for relatively low values of *d* of around 100 Hertz, as would be expected, while (H)% and (H)- boundaries, whose probability distributions for *d* are indicated by the red dashed and the orange dot-dashed lines, respectively, typically exhibit higher *d* values of around 150 Hertz. Non-final syllables (black dotted line) are somewhere in between, with *d* mostly between 110 and

	NONE	H*L	H*	L*H	L*HL
NONE	–	0.00000	0.00000	0.00000	0.00000
H*L	0.00000	–	0.00000	0.00000	0.00000
H*	0.00000	0.00000	–	0.00000	0.0000
L*H	0.00000	0.00000	0.00000	–	0.92136
L*HL	0.00000	0.00000	0.00000	0.92136	–
	NONE	H %	(H)%	(H)-	(L)%
NONE	–	0.00000	0.00000	0.00000	0.00000
H%	0.00000	–	0.00000	0.00000	0.00000
(H)%	0.00000	0.00000	–	0.58325	0.00000
(H)-	0.00000	0.00000	0.58325	–	0.00000
(L)%	0.00000	0.00000	0.00000	0.00000	–

Table 5.2: p-values obtained by pairwise Wilcoxon rank sum tests comparing distributions of d *for the most frequent accent types. The distributions for L*H and L*HL accents and those of (H)% and (H)- boundaries are not significantly different; all other distributions are significantly different at a confidence level of 0.999. Values of 0.00000 indicate that $p \ll 0.000005$.*

150 Hertz, while H% boundaries, as could be expected, are realized with the highest pitch: their density plot (blue long-dashed line) has its peak at between 170 and 180 Hertz.

In concluding the discussion of the *d* parameter, it can be said that the distributions of this parameter confirm what could be expected from the phonetic description of the accents and boundary tones. Furthermore, there are clear differences in the distributions for different accent and boundary types. Wilcoxon rank sum tests confirm that all distributions are pairwise significantly different at a confidence level of 0.999, with the exception of L*H vs. L*HL accents, and of (H)% vs. (H)- boundaries, for which the distributions are not significantly different. Table 5.2 lists the p-values, rounded to 5 digits after the decimal point. Values of 0.00000 indicate that $p \ll 0.000005$.

5.1.3 Amplitudes of rise and fall

Parameters *c1* and *c2* determine the amplitude of the rise towards the peak (*c1*) and the amplitude of the fall after the peak (*c2*) in Hertz. As mentioned in chapter 3, before approximating the F0 contour, PaIntE looks for a local maximum in the contour. If such a maximum is found, the approximation is carried out using the actual PaIntE function, which is the sum of a rising and a falling sigmoid. If no such maximum is found, the approximation is carried out

using only the rising sigmoid for rising contours, and only the falling sigmoid for falling contours. In these cases, the *c* parameter that is undefined since it is not part of the term of the approximation function is set to 0. When using the actual PaIntE function with both sigmoids, values of 0 for either *c1* or *c2* do not arise. Thus, in the following plots, values of 0 mean that the parameter was undefined originally. However, since it can be deduced in these cases that the F0 contour can be best described as only a fall or only a rise, it is straightforward to interpret the amplitude of the opposite, "non-existing", movement as zero.

Figure 5.11 shows the distributions of *c1* (upper panel) and *c2* (lower panel) for different accent types. Looking at H*L accents first, which are indicated by the blue long-dashed line, there is little surprise. It is obvious that they tend to have low values of *c1*, but higher values of *c2*: their *c1* distribution shows a pronounced peak for *c1* values of around 0 to 10 Hertz, and although the distribution extends to the left with values of *c1* up to 60 to 80 Hertz, the higher values are much less likely. Their *c2* distribution, on the other hand, shows a clear dominance of moderately high *c2* values with values of around 0 being rather unlikely. There is a broad peak between 20 and 40 Hertz, indicating that these are typical values of *c2* for H*L accents. In short, H*L accents have small rise amplitudes but higher fall amplitudes, as expected for falling accents.

L*H accents (green solid line) show just the opposite behavior: their *c1* values are typically between 20 and 60 Hertz, while their *c2* values tend to be close to 0, as one would expect for rising accents. The distributions for L*HL accents are given by the red dot-dashed lines. They exhibit higher values for both *c1* and *c2*, reflecting their characterization as rise-fall accents.

H* accents should tend to have lower *c1* and *c2* values. These values should be greater than 0, since H* accents should exhibit a small peak, and they should be lower than the values observed for L*HL accents, since they do not necessarily involve pronounced falls or rises; rather, the course of the pitch contour depends on the preceding and following targets. These expectations are confirmed by the density plots for H* accents (orange dashed lines) in figure 5.11.

Unaccented syllables, which are represented by black dotted lines in both panels, tend to have *c1* and *c2* values close or equal to 0. This confirms that unaccented syllables usually do not exhibit significant F0 rises or falls. However, this is just a tendency, not an absolute rule, as can be seen from the fact that both distributions have non-zero densities for higher *c1* or *c2* values. In the case of unaccented syllables, I claim that these higher values for *c1* or *c2* can arise either (i) because of pitch accents or boundaries in the vicinity, in which case interpolation to tonal targets on adjacent syllables causes the F0 contour to rise or fall on the unaccented syllables themselves, or (ii) from the fact that tonal movements caused by accents associated with surrounding syllables are captured in the PaIntE parametrization. Figure 5.12 serves to illustrate these two causes.

The black dotted line in figure Figure 5.12 indicates the distribution of *c1*

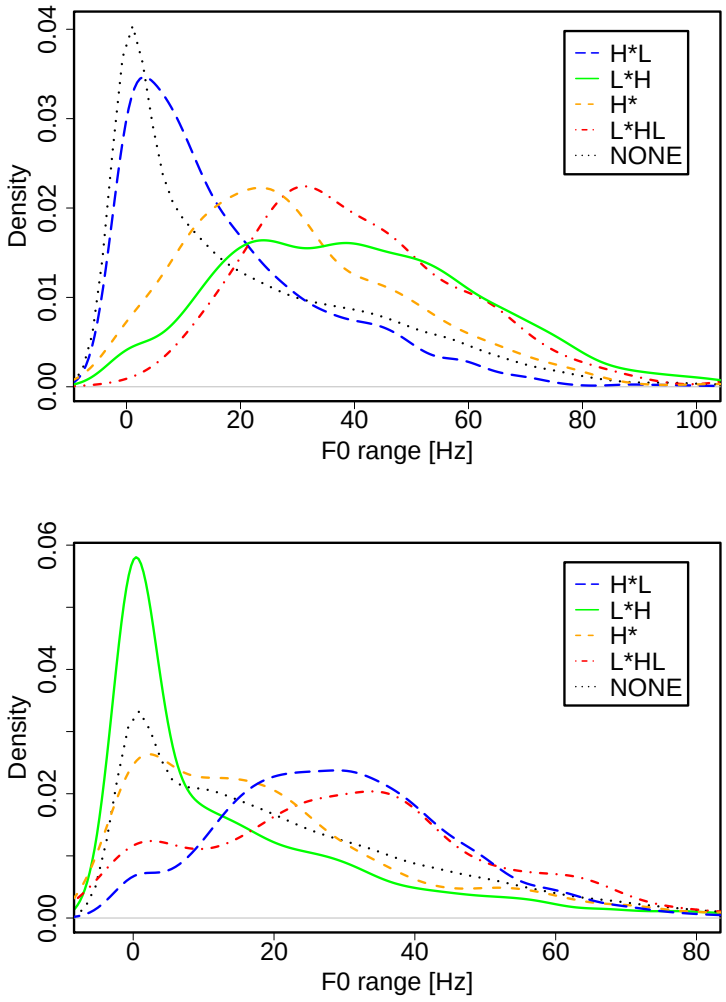

Figure 5.11: Density plots of the c1 *parameter (upper panel) and the* c2 *parameter (lower panel) for the most frequent accents and for unaccented syllables. H*L accents (blue, long-dashed line) usually exhibit low* c1 *and high* c2 *values; vice versa, L*H accents (green, solid line) exhibit low* c2 *and high* c1 *values. L*HL accents (red, dot-dashed line) are characterized by high values of both* c1 *and* c2*, and H* accents (orange, dashed line) by moderately high values of* c1 *and* c2*. For unaccented syllables (black, dotted line) low values for both* c1 *and* c2 *dominate, but higher values for either of the parameters can be observed as well.*

values for unaccented syllables in general and is just repeated from figure 5.11 to serve as a reference. The effect described in (i) above is visible in the distributions indicated by the blue, solid, and the green, long-dashed lines. The blue

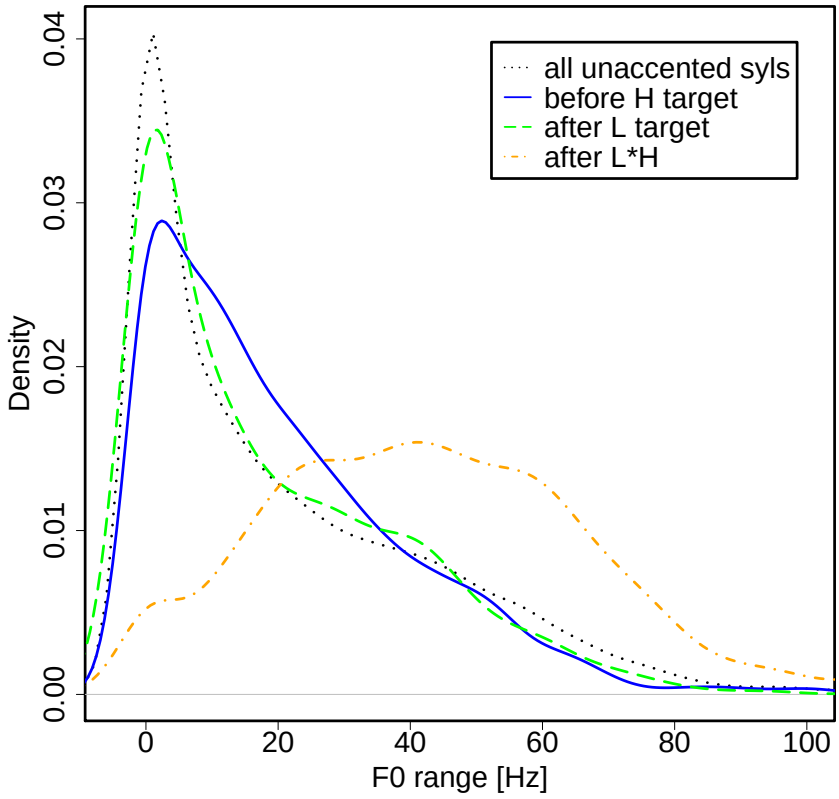

Figure 5.12: Density plots of the c1 *parameter for unaccented syllables in general (black, dashed line) and for unaccented syllables in the immediate neighborhood of pitch accented syllables. Unaccented syllables preceding high targets (blue, solid line) are more likely to exhibit non-zero* c1 *values than unaccented syllables in general. The effect is very subtle for unaccented syllables following low targets (green, long-dashed line). For unaccented syllables immediately following L*H accents (orange, dot-dashed line), values of around 40 Hz are the most likely.*

solid line represents the distribution of unaccented syllables that immediately precede a following high target, i.e., that immediately precede a H*L, H*, or HH*L accent. It can be seen that the peak is less pronounced: higher *c1* values of 10 to 40 Hertz are more likely than for unaccented syllables in general. This is because the F0 contour is expected to rise before high targets, and this is visible in the *c1* parameters of the preceding syllables. The symmetric effect, i.e., rising F0 contours after low targets is very subtle: the green long-dashed line represents the *c1* distribution for unaccented syllables immediately following low targets, i.e., following H*L or L*HL accents. Again, the peak is less

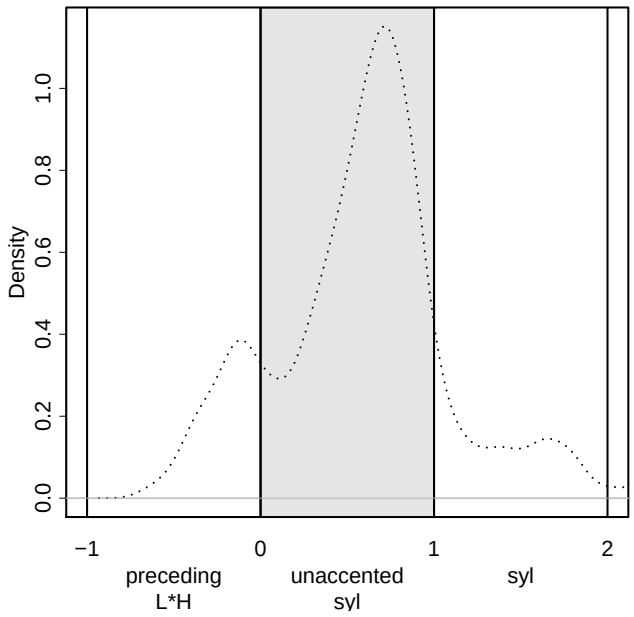

Figure 5.13: Density plot of the b *parameter for unaccented syllables with* c1>*20 which occur immediately after an L*H accented syllable. The unaccented syllable, for which the approximation was carried out, is highlighted in gray. The corresponding peak is usually detected on the unaccented syllable.*

pronounced, i.e., values around 0 are less likely than for unaccented syllables in general. However, this effect is not very strong and demonstrates that the F0 contour does not necessarily start to rise again immediately after low targets.

The distribution of *c1* for unaccented syllables immediately following L*H accents (orange, dot-dashed line) illustrates the effect described in (ii) above: the tonal movement related to the preceding L*H syllable is captured in the PaIntE parametrization of the following unaccented syllable. The distribution shows that the following unaccented syllables usually exhibit *c1* parameters of between 20 and 60 Hertz, with few occurrences of values equal to 0. The effect is due to the PaIntE approximation being carried out in a three-syllable window around the current (in this case unaccented) syllable, and detected peaks can be on any of the three syllables. The *c1* parameter gives the relative height of this peak, no matter on which syllable the peak is realized. So if in the approximation, a peak is found on a neighboring syllable, the *c1* parameter indicates the height of that peak and thus the amplitude of the rise on the neighboring syllable.

I will verify this for the cases with unexpectedly high values of *c1*, by looking at the distribution of unaccented syllables after L*H accents with values of

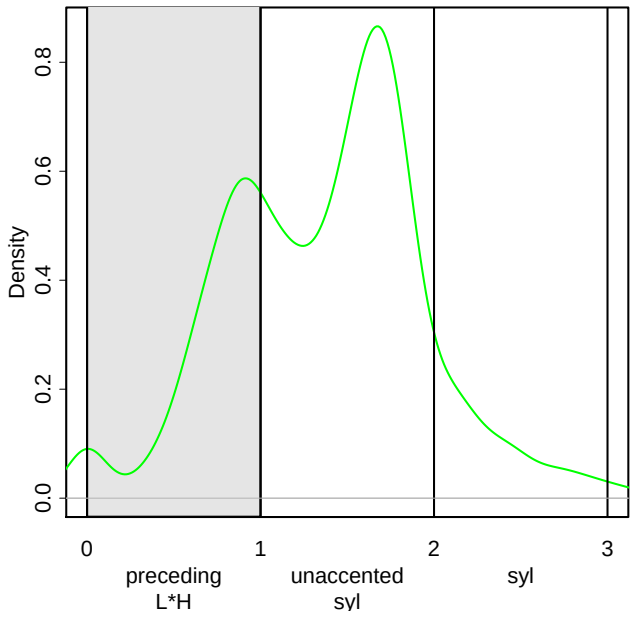

Figure 5.14: Density plot of the b *parameter for L*H accented syllables which occur immediately before unaccented syllables with* c1>*20. The L*H syllable, for which the approximation was carried out, is highlighted in gray. The peak is usually detected on the unaccented syllable in these cases.*

c1>20. The density plot of their *b* parameters, which is depicted in figure 5.13, confirms that the peaks are mostly detected on the unaccented syllable itself: *b* in most cases is between 0 and 1. More precisely, the peak is most likely to be detected in the later part of the unaccented syllable. This suggests that the peaks are actually the peaks of preceding L*H accents which are realized on the post-accented syllables and thus on the unaccented syllables that are under consideration here. As noted before in section 5.1.1, this is not unusual for L*H accents.

The hypothesis is borne out, as demonstrated by figure 5.14. It shows the density plot of the *b* parameters for only those L*H accented syllables that occur just before unaccented syllables, again restricted to syllables with *c1*>20, and confirms that their peaks are typically detected on the post-accented syllable: values between 1 and 2 are most likely. Thus, the high values of *c1* for unaccented syllables immediately after L*H accents can be attributed to the fact that the tonal movement associated with the L*H accent on the preceding syllable is partly realized on the unaccented syllable. This movement will also be captured

in the PaIntE parameters of the unaccented syllable.[10]

To summarize the discussion of the *c1* and *c2* parameters of unaccented syllables, it can be said that unaccented syllables tend to have low *c1* and *c2* parameters, indicating that there are usually no pronounced F0 rises or falls on unaccented syllables. Higher values of *c1* and *c2* can occur because of tonal targets in the vicinity of the unaccented syllables, which cause the F0 contour to rise or fall, or from the fact that tonal movements associated with the surrounding syllables are captured in the PaIntE parametrization of the unaccented syllables.

With respect to boundary tones, one would expect high values of *c1* for the high boundary tones, which should exhibit a rising contour, while low boundaries should have *c1* close to 0. This expectation is borne out as can be seen in figure 5.15. The probability distribution for (L)% (green solid line) shows a very strong preference for values close to or equal to 0, while the three distributions for high boundaries (orange dot-dashed line for (H)-, red dashed line for (H)%, and blue long-dashed line for H%) almost always have *c1* values of more than 20 Hertz, with the highest values a little more likely to belong to the inherently high H% boundaries. In light of the fact discussed at the end of section 5.1.1 that for (L)% boundaries, the peak is almost always detected on the final syllable or on one of the syllables preceding it, but almost never on the next syllable, it is not surprising that *c1* values close to 0 predominate for (L)% boundaries: Rises should only be expected from the final syllable with the (L)% boundary to the next phrase, but since the peak is always detected on one of the earlier syllables, the *c1* value always corresponds to the rise toward that earlier peak. I have suggested above that these earlier peaks are in most cases due to microprosodic variation, thus their amplitudes are usually low, yielding only small values for *c1*.

For the *c2* parameter (figure 5.16), non-zero amplitudes for all boundary types are observed. Although there are cases where *c2* is close to or equal to 0 for all boundary types, as can be seen from the small peaks or at least bumps at around 0 in the density plots (green solid line for (L)%, orange dot-dashed line for (H)-, red dashed line for (H)%, and blue long-dashed line for H%), the majority of boundaries is realized with fall amplitudes greater than 0. For (L)% boundaries, this is not surprising because they should be realized as falls, but it may be surprising at first in the case of high boundaries. However, I have discussed above (cf. 5.1.1) that for high boundaries, falling pitch movements

[10] It may be noted that this distribution of the *b* parameter looks different from the one for L*H depicted in figure 5.5. There, the peak was almost equally likely on the accented and on the post-accented syllable. The distribution depicted here is different because only L*H accents before unaccented syllables with values of *c1*$>$20 are taken into account. In cases where the peak is realized on the accented syllable, if some small peak is detected on the unaccented syllable, this peak will be approximated rather than the peak of the preceding accent, giving rise to lower *c1* values, and thus such cases are not included in the distribution depicted here.

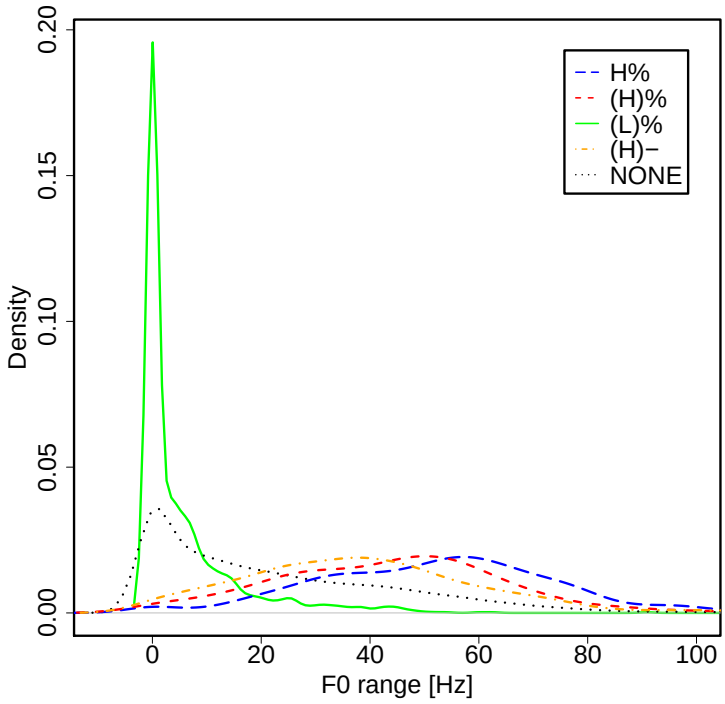

Figure 5.15: Density plots of the c1 *parameter for the most frequent boundary tones and for non-final syllables. (L)% boundaries (green, solid line) and unaccented syllables (black, dotted line) usually have low* c1 *parameters, while high boundaries exhibit higher values of* c1.

are detected because the reset of F0 for the next phrase causes a drop in F0 at the boundary to the next syllable. The amplitude of the drop is documented in the *c2* parameter for the high boundaries.

Again, I will conclude this section by confirming that the observed differences are significant. Table 5.3 lists the p-values obtained in pairwise Wilcoxon rank sum tests of the *c1* distributions for accents and boundaries. The distributions for unaccented syllables (NONE) vs. H*L accents are not significantly different, neither are those for L*H vs. L*HL; all other distributions are significantly different at a confidence level of 0.999. The p-values resulting from comparing the *c2* distributions are indicated in table 5.4. Here, the distributions for unaccented syllables (NONE) vs. H* accents are not significantly different, neither are those for H*L vs. L*HL; as for boundaries, there is no significant difference between H% and (H)%, between H% and (H)-, and between (H)% and (L)%; the other distributions are significantly different at a confidence level of 0.999.

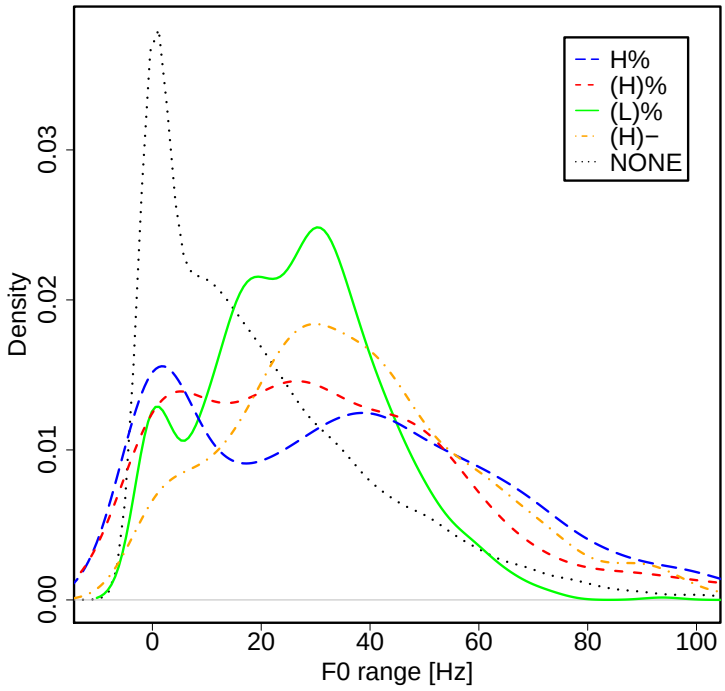

Figure 5.16: Density plots of the c2 *for the most frequent boundary tones and for non-final syllables. Unaccented syllables (black, dotted line) tend to have low values of* c2*, but higher values can also occur. For boundaries, values of* c2 *of around 0 do occur, but in the majority of cases, higher values are detected.*

Concluding this subsection, it can be stated that the PaIntE parameter distributions of GToBI(S) events reflect the expected tonal properties. I have not discussed parameters *a1* and *a2*, which determine the steepness of the related rise or fall, because GToBI(S) makes no prediction of what their probability distributions might look like. However, this does not necessarily speak against their perceptual relevance. I will leave open this question for now. However, I propose to view them as potentially relevant in perception. In the following, I will for the sake of simplicity speak of "the" PaIntE parameters as perceptually relevant, but one should keep in mind that the perceptual relevance of *a1* and *a2* is less well motivated than that of the remaining parameters.

	NONE	H*L	H*	L*H	L*HL
NONE	–	0.10361	0.00000	0.00000	0.00000
H*L	0.10361	–	0.00000	0.00000	0.00000
H*	0.00000	0.00000	–	0.00000	0.00000
L*H	0.00000	0.00000	0.00000	–	0.61154
L*HL	0.00000	0.00000	0.00000	0.61154	–
	NONE	H %	(H)%	(H)-	(L)%
NONE	–	0.00000	0.00000	0.00000	0.00000
H%	0.00000	–	0.00000	0.00000	0.00000
(H)%	0.00000	0.00000	–	0.00002	0.00000
(H)-	0.00000	0.00000	0.00002	–	0.00000
(L)%	0.00000	0.00000	0.00000	0.00000	–

Table 5.3: p-values obtained by pairwise Wilcoxon rank sum tests comparing distributions of c1 *for the most frequent accent types (upper panel) and boundary types (lower panel). The distributions for unaccented syllables (NONE) vs. H*L accents are not significantly different, neither are those for L*H vs. L*HL; all other distributions are significantly different at a confidence level of 0.999. Values of 0.00000 indicate that $p \ll 0.000005$.*

5.2 Target regions for intonation events

At the very beginning of this chapter, I have presented three questions, two of which I claim to answer here: (i) what are the tonal dimensions of perceptual space, (ii) which are the relevant prosodic events, and (iii) which are the target regions corresponding to these events. I have already stated above that in the present thesis, I will assume that the prosodic events as defined by GToBI(S) (Mayer 1995) are the relevant prosodic events.

As for question (i), I have argued in chapter 3 that the PaIntE parameters are linguistically motivated in that they are intended to model aspects of the realization of intonation events that are known to be meaningful. In this chapter, I have explained the relation between the PaIntE parameters and the pitch accents and boundary tones of the GToBI(S) labeling system. I have shown empirically that the PaIntE parameters capture the tonal properties attributed to the different intonation events posited by GToBI(S), and that rather fine-grained effects of prosodic and segmental context on the tonal implementation of these events are visible in the probability distributions of their PaIntE parameters.

Since these events are defined as tonally distinct units, it goes without saying that they will differ with respect to their tonal properties. I have explained

	NONE	H*L	H*	L*H	L*HL
NONE	–	0.00000	0.07034	0.00000	0.00000
H*L	0.00000	–	0.00000	0.00000	0.81739
H*	0.07034	0.00000	–	0.00000	0.00000
L*H	0.00000	0.00000	0.00000	–	0.00000
L*HL	0.00000	0.81739	0.00000	0.00000	–
NONE	–	0.00000	0.00000	0.00000	0.00000
H%	0.00000	–	0.06971	0.01203	0.00000
(H)%	0.00000	0.06971	–	0.00002	0.00761
(H)-	0.00000	0.01203	0.00002	–	0.00000
(L)%	0.00000	0.00000	0.00761	0.00000	–

Table 5.4: p-values obtained by pairwise Wilcoxon rank sum tests comparing distributions of c2 *for the most frequent accent types (upper panel) and boundary types (lower panel). In the upper panel, the distributions for unaccented syllables (NONE) vs. H* accents are not significantly different, neither are those for H*L vs. L*HL; in the lower panel, there is no significant difference between H% and (H)%, between H% and (H)-, and between (H)% and (L)%; all other distributions are significantly different at a confidence level of 0.999. Values of 0.00000 indicate that $p \ll 0.000005$.*

above how the PaIntE parameters serve to distinguish between the different intonation events: The probability distributions of the PaIntE parameters differ for different events. These differences were not always present in every parameter; rather, some parameters are similar for some intonation events, but useful in distinguishing between other intonation events, and others vice versa. For example, the distributions for the *b* parameter are quite different for H*L accents and L*H accents, but more similar for H* an L*HL accents (cf. figure 5.4), while for H* and L*HL accents the distributions for the *c2* parameter are different (cf. figure 5.11).

Building on these two aspects—that the PaIntE parameters model properties of the realization of intonation events that are meaningful and that they serve to distinguish different categories—I propose to answer question (i) by considering the PaIntE parameters as the relevant tonal dimensions in the perception and production of intonation events.

Regarding question (iii), which are the target regions corresponding to the prosodic events, in section 2.3.5 I have presented an exemplar-theoretic interpretation of Guenther and Perkell's speech production model (Guenther et al. 1998; Perkell et al. 2000, 2001) for the temporal domain. I have suggested that z-score distributions implicitly define target regions in the production of temporal properties of prosodic events. Reasoning the same way for the tonal

domain, I propose to regard the distributions of PaIntE parameters as target regions in the production of tonal properties of prosodic events. Thus, the abstract perceptual target regions posited by Guenther and Perkell's model for the segmental domain are implicitly defined by the distributions of PaIntE parameters of the exemplars stored in the memory of the speaker. This view is motivated by the fact that if phonetic details of the exemplars are stored in memory, speakers have access to the F0 contours of the stored exemplars, and their shapes in terms of peak alignment, peak height, rise and falls amplitudes, etc., as coded in the PaIntE parameters, and are likely to use them as a reference in production.

Moreover, I have presented experimental results confirming that realizations of syllables in different prosodic contexts show different distributions of PaIntE parameters, i.e., they are expected to be useful in distinguishing prosodic events. As with the z-score distributions examined in section 4.3, there is a lot of overlap between the distributions. However, I have only examined one parameter at a time; taking all parameters into account at once may be enough to make the regions sufficiently distinct from each other. This question will be addressed in the following chapter.

Chapter 6

Modeling categorization

I have shown in chapters 4 and 5 that the distributions of z-scores of segment durations in the temporal domain and of the PaIntE parameters in the tonal domain are significantly different for different prosodic events. With regard to perception, the question that immediately arises is whether these regions are sufficiently distinct to categorize new exemplars. From an exemplar-theoretic point of view, new instances are categorized based on their perceptual similarity to exemplars stored in memory (cf. section 2.3.3): they are assigned the category of the most similar exemplars. I will investigate in this chapter whether the perceptual dimensions introduced in the preceding chapters are suitable and sufficient for categorizing new instances.

In spite of the fact that most prosodic events exhibit significantly different duration z-score and PaIntE parameter distributions, there was a considerable amount of overlap in all distributions. However, I have only examined the perceptual dimensions one at a time. Because of the overlap in the distributions, few exemplars could have been categorized correctly based on one dimension only. The question that I am trying to answer in this present chapter is whether categorization is possible when all relevant dimensions contribute to categorization.

From a theoretical perspective, listeners are able to categorize new exemplars in speech perception. Thus, modeling a listener's memory by a large prosodically annotated speech corpus, it should be possible to categorize prosodic events if all perceptually relevant dimensions are known, i.e., it should be possible to predict prosodic labels for new exemplars based on the values that exemplars from the corpus exhibit for these dimensions. In the best case, one could even hope to detect prosodic categories by identifying accumulations of exemplars with similar perceptual properties, and then establishing which accumulations correspond to which category.

I will examine both approaches in this chapter. I will discuss clustering experiments as a way to detect prosodic categories in section 6.2. I will then model categorization of prosodic events using machine learning methods to

predict prosodic labels in section 6.3. But before turning to the actual experiments, I will describe the data used in both approaches in section 6.1.

6.1 Representing the data

In machine learning, whether the goal be clustering or classifying data, the first step is to represent the data as a set of *instances* of observations. In the experiments presented in this chapter, the data consist of instances of realizations of syllables. Each instance is described by some observed properties. These properties are called attributes, and thus each instance is characterized by the values that it exhibits for the attributes. For coherence with the preceding chapters, instances can be thought of as exemplars, and the attributes can be interpreted as their perceptual properties; however, not all attributes I will be using here have been confirmed as perceptually relevant.

In the following sections, I will first discuss the attributes used for the experiments described in this chapter; then, I will provide more details on the databases from which the data were extracted in section 6.1.2. Finally, a note of caution is in order (section 6.1.3) because the data are inevitably noisy.

6.1.1 Attributes

Most attributes used here are related to the duration z-scores and PaIntE parameters discussed in the preceding chapters. However, some more attributes that were thought to be useful in distinguishing accent types and boundaries were included, even though I have not examined in detail whether they can be claimed to be perceptually relevant or not. All attributes are discussed and illustrated using an example sentence from the database in the following. I will also explain for each attribute why it has been included, particularly for the attributes that are not directly related to the duration z-scores or PaIntE parameters.

Tables 6.1 to 6.5 list attributes and observed values of the first thirteen instances of syllables of the SWMS corpus, corresponding to the phrase “Das Zentrum blieb zumeist Cardoso vorbehalten” (*The center was mostly left to Cardoso*); each line represents an instance, and each column represents an attribute; the observed values are indicated in the table cells. For the experiments, all instances and attributes are collected in one huge table; this table is broken down here into several partial tables in order to have more space for discussion of the particular attributes in between tables.

Table 6.1 shows the first eight attributes. The first two specify which GToBI(S) accent (if any) and which GToBI(S) boundary tone (if any) have been

	accent	tone	a1'	a2'	b	c1'	c2'	d'
1	NONE	NONE	-0.443	-0.957	2.045	2.389	-0.976	2.637
2	L*HL	NONE	-0.306	0.624	0.876	2.178	2.695	2.471
3	NONE	(L)-	0.115	0.239	-0.206	1.389	2.987	2.549
4	NONE	NONE	-0.976	-0.849	0	-0.953	-0.976	-1.309
5	NONE	NONE	-0.57	0.007	0.833	-0.512	-0.646	-1.042
6	NONE	NONE	0.402	0.001	1.404	-0.61	-0.751	-0.91
7	NONE	NONE	0.126	-0.957	1.732	-0.438	-0.976	-0.498
8	H*L	NONE	-0.284	-0.441	0.997	-0.405	1.124	-0.44
9	NONE	NONE	-0.692	0.473	0.44	-0.105	0.101	-0.506
10	NONE	NONE	1.715	-0.184	0.475	-0.354	-0.55	-1.042
11	NONE	NONE	1.495	-0.392	-0.526	-0.353	-0.449	-1.071
12	NONE	NONE	2.638	-0.957	1.313	-0.687	-0.976	-1.446
13	NONE	(L)	1.833	0.237	0.344	-0.641	-0.195	-1.415

Table 6.1: Prosodic labels (first two columns, used for evaluation purposes) and first six attributes (remaining columns) with corresponding values for the first thirteen instances of the SWMS data. The six attributes are the PaIntE parameters determined for each instance. Values of parameters marked by ' have been z-scored.

realized on that particular syllable.[1] The values for these two attributes have been determined by manual prosodic labeling. They are not intended to be used in the clustering or classification experiments themselves, but are provided for evaluation purposes. In the clustering experiments, one hopes to find correlations between these manual labels and the observed clusters. In the classification experiments, they are the attributes whose values are to be predicted based on the values of all other attributes.

The **accent** attribute corresponds to the GToBI(S) accent label of the syllable, with downstepped accents mapped to their counterparts without downstep. Thus, the attribute can assume the following values: NONE for unaccented syllables, or any of the accents of the GToBI(S) annotation system discussed in section 3.1: L*H, H*L, L*, H*, ..H, ..L, L*HL, or HH*L. Please note that technically, ..H and ..L are not accents but trail tones related to an earlier accent. In contrast to the "real" accents, they may also occur on unstressed syllables. For the **tone** attribute, I have again mapped the underspecified boundary tones % and - of the GToBI(S) system to fully specified tones by integrating the preceding trail tone into the boundary tone labels (cf. chapter 5): to distinguish them from boundary tones that had been fully specified already, the preceding trail tones are specified in brackets. Thus, values of (H)%, for instance, occur

[1] GToBI(S) is discussed in more detail in section 3.1.

for syllables with an originally unspecified % boundary tone after a preceding H trail tone (which must have come from a preceding L*H accent), while H% occurs if the boundary tone had been fully specified in the manual prosodic annotation already. The full set of possible values for the tone attribute is NONE for non-final syllables, or (H)-, (L)-, (H)%, H%, (L)%, or L% for phrase-final syllables.

The remaining attributes in table 6.1 are attributes whose values are actually used for clustering or learning. They correspond to the six PaIntE parameters (cf. chapter 3). As I have substantiated in some detail in chapter 5, the PaIntE parameters are expected to be relevant in the perception of prosodic events. Since the experiments aim at modeling categorization, it is desirable to build models that are applicable to data of different speakers. However, at least three of the six PaIntE parameters are expected to be speaker-specific: *d* corresponds to the absolute height of the F0 peak in Hertz, which will be different between speakers, particularly between male and female speakers. *c1* and *c2* specify the amplitudes of the rising or falling F0 movement. It is often claimed that female speakers have a greater F0 range, therefore the amplitudes of the F0 movement in accented syllables are expected to be greater for female than for male speakers. Thus, *c1* and *c2* should be speaker-specific or at least gender-specific. To eliminate the speaker-specific aspects, *d*, *c1*, and *c2* were z-scored. Parameters *a1* and *a2* correspond to the amplitude-normalized steepness of the rising and falling movements, respectively. As the amplitudes (*c1* and *c2*) are likely to be speaker-specific, the amplitude normalization introduces some speaker-specificity, and thus, *a1* and *a2* have been z-scored as well.[2] The *b* parameter was left unchanged because there was no equally compelling reason to expect speaker-specific effects: parameter *b* corresponds to the alignment of the F0 peak with the syllable structure and there is no evidence that speakers might differ in that respect.[3] To the contrary, it is not advisable to z-score the b parameter since its values are normalized with respect to syllable structure, with integer values corresponding to syllable boundaries. Thus, z-scoring the *b* parameter would obscure the peak's exact position in the syllable.

In any case, z-scoring of five of the six PaIntE parameters seems to have been unproblematic, as I have applied most of the classification algorithms discussed in section 6.3 to both the absolute and z-scored values of these parameters, and there was no indication whatsoever that performance might have been lower in the z-scored case. The z-scored parameters are marked by the ' diacritic in the remainder of this chapter.

[2] I do not claim that listeners z-score or otherwise normalize the tonal parameters in perception; z-scoring here is just a technicality that is necessary to be able to apply the models and classifiers obtained on data of just one speaker to data of other speakers later.

[3] I do not want to deny that dialects might differ in the alignment of peaks, but dialectal variation is beyond the scope of this thesis. Here, I model categorization of speakers of the same dialect, viz. Standard German ("Hochdeutsch").

	maxc	c1'-c2'	prev syl			next syl		
			pc1'	pc2'	pd'	nc1'	nc2'	nd'
1	2.389	-3.364	-0.953	-0.976	-5.612	2.178	2.695	2.471
2	2.695	0.518	2.389	-0.976	2.637	1.389	2.987	2.549
3	2.987	1.598	2.178	2.695	2.471	-0.953	-0.976	-1.309
4	-0.953	-0.022	1.389	2.987	2.549	-0.512	-0.646	-1.042
5	-0.512	-0.133	-0.953	-0.976	-1.309	-0.61	-0.751	-0.91
6	-0.61	-0.141	-0.512	-0.646	-1.042	-0.438	-0.976	-0.498
7	-0.438	-0.538	-0.61	-0.751	-0.91	-0.405	1.124	-0.44
8	1.124	1.529	-0.438	-0.976	-0.498	-0.105	0.101	-0.506
9	0.101	0.206	-0.405	1.124	-0.44	-0.354	-0.55	-1.042
10	-0.354	-0.196	-0.105	0.101	-0.506	-0.353	-0.449	-1.071
11	-0.353	-0.096	-0.354	-0.55	-1.042	-0.687	-0.976	-1.446
12	-0.687	-0.289	-0.353	-0.449	-1.071	-0.641	-0.195	-1.415
13	-0.195	0.446	-0.687	-0.976	-1.446	0.809	-0.976	0.614

*Table 6.2: Attributes 7 to 14 and corresponding values for the first ten instances of the SWMS data. The first two attributes indicate the maximum amplitude on the current syllable as determined by max(*c1',c2'*) and the difference between* c1' *and* c2'*. The remaining attributes in this table correspond to the PaIntE parameters* c1*,* c2 *and* d *of the preceding (attributes* pc1', pc2', pd'*) and following syllables (attributes* nc1', nc2', nd'*).*

Table 6.2 lists further attributes used in the following sections. The first one, **maxc**, is the maximum of *c1'* and *c2'*. This attribute was included to capture the maximum amplitude of the F0 movement on the syllable in one parameter, irrespective of whether it is related to a rise or to a fall, and might be helpful in distinguishing unaccented syllables from accented ones, for instance. The second, **c1'-c2'**, codes the relative difference in F0 before and after the F0 movement by subtracting the amplitude of the falling movement, *c2*', from the amplitude of the rising movement, *c1'*. Thus, F0 movements for which *c2*' was greater than *c1'*, which can be interpreted as overall falling, are characterized by negative values, whereas F0 movements with greater *c1'* than *c2*', which can be described as overall rising, exhibit positive values for *c1'-c2'*.

The other attributes in table 6.2 are z-scored PaIntE parameters of the preceding (attributes **pc1'**, **pc2'**, **pd'**) and following syllables (attributes **nc1'**, **nc2'**, **nd'**). They are of interest because comparison of these parameters for adjacent syllables can indicate local changes in the slope of the F0 contour. Assume, for instance, that there is a longer rising stretch of F0 spanning several syllables. All syllables within that span should have moderately high, positive, and, most important, similar, values of *c1* (and *c2* parameters around 0). Here, the sim-

	syl before preceding syl			syl after next syl		
	ppc1'	ppc2'	ppd'	nnc1'	nnc2'	nnd'
1	-0.953	-0.976	-5.612	1.389	2.987	2.549
2	-0.953	-0.976	-5.612	-0.953	-0.976	-1.309
3	2.389	-0.976	2.637	-0.512	-0.646	-1.042
4	2.178	2.695	2.471	-0.61	-0.751	-0.91
5	1.389	2.987	2.549	-0.438	-0.976	-0.498
6	-0.953	-0.976	-1.309	-0.405	1.124	-0.44
7	-0.512	-0.646	-1.042	-0.105	0.101	-0.506
8	-0.61	-0.751	-0.91	-0.354	-0.55	-1.042
9	-0.438	-0.976	-0.498	-0.353	-0.449	-1.071
10	-0.405	1.124	-0.44	-0.687	-0.976	-1.446
11	-0.105	0.101	-0.506	-0.641	-0.195	-1.415
12	-0.354	-0.55	-1.042	0.809	-0.976	0.614
13	-0.353	-0.449	-1.071	0.457	-0.976	0.266

Table 6.3: Attributes 15 to 20 and corresponding values for the first thirteen instances of the SWMS data. The attributes indicate PaIntE parameters c1*,* c2 *and* d *of the syllables before the preceding (attributes* ppc1', ppc2', ppd'*) and after the following syllable (attributes* nnc1', nnc2', nnd'*).*

ilarity of the *c1* values is an indicator that this is not a local F0 rise caused by an accent but a more global F0 movement, possibly caused by the interpolation between a preceding low and a following high target. For the same reason, the parameters of the syllable before the preceding syllable (attributes **ppc1'**, **ppc2'**, **ppd'**) and the syllable after the next syllable (attributes **nnc1'**, **nnc2'**, **nnd'**) are also included. These are listed in table 6.3.[4]

The attributes cited so far are all related to the PaIntE parametrization. In short, all PaIntE parameters of the current syllable are used. Also, the maximum of *c1'* and *c2'* and the difference between *c1'* and *c2'* are provided. And finally,

[4]It should be noted that file boundaries are ignored when calculating the PaIntE parameters of the context syllables in order to avoid artifacts caused by the structure of the database. Usually, after running a PaIntE approximation for an utterance which corresponds to one speech file, for the first syllables, the PaIntE parameters of the left context syllables are not accessible, and for the last syllables, the PaIntE parameters of the right context syllables are missing. Thus, file boundaries, which virtually always coincide with phrase boundaries, might be easily detectible from the missing parameters. This could artificially increase the prediction accuracy of phrase boundaries for corpora like the SWMS corpus, which consist of rather short utterances in many separate files. Instead, the context syllables are taken from preceding or following files if necessary. This way, the only cases where the PaIntE parameters of the context are not accessible are the very first and the very last syllables of the corpus. In these few cases, the absolute parameters have been set to 0, resulting in negative z-scored values of -0.953, -0.976, and -5.612 for *c1'*, *c2'*, and d', respectively.

in addition to the PaIntE parameters of the current syllable, *c1'*, *c2'* and *d'* of the two preceding and the two following syllables are used; thus, these parameters are provided for a five-syllable window around the syllable in question.

Turning to temporal aspects, I have motivated in chapter 4 that phoneme-specific z-scores of segment durations are relevant in the perception of prosodic events. I have shown in section 4.3 that lengthening effects caused by accents are most prominent for onset segments and syllable nuclei (figure 4.8), while effects caused by phrase boundaries are most prominent for the last (phrase-final) segment (figure 4.7). Therefore, both the z-scores for nuclei[5] and the z-scores for final segments are provided. The former should be helpful in distinguishing accented syllables from unaccented syllables, the latter in recognizing phrase boundaries. Again, in order to facilitate comparison of values for the current syllable with those for context syllables, these attributes are provided for a three-syllable window around the current syllable. Comparing duration z-scores of the current syllable to that of context syllables should help to separate effects of prosodic events from effects of local speech rate: If lengthening is caused by a low speech rate, the z-scores of the context syllables should be similar, while, if lengthening is caused by prosodic events, context syllables should not be affected. Thus, six attributes are relevant: phoneme-specific duration z-scores of nuclei of the current (attribute **nucleus**), preceding (attribute **pnucleus**), and next syllables (attribute **nnucleus**) and phoneme-specific duration z-scores of final segments of the current (attribute **finalseg**), preceding (attribute **pfinalseg**), and next syllables (attribute **nfinalseg**). Again, the values of these attributes for the first thirteen instances are indicated in table 6.4.[6]

All attributes discussed so far are numeric. The remaining attributes are nominal attributes that are either derived from the annotation of the database or even from the underlying text. They are listed in table 6.5. The first three can be classified as phonological and are derived directly from the segment, syllable, and word label files. The **stress** attribute indicates whether a syllable is stressed or not, possible values are 1 (for stressed) and 0 (for unstressed). For clustering and classification of accents, it is clearly advantageous to know which syllables are stressed, since the accents can only occur on stressed syllables. I would hypothesize that even human listeners in the perception of accents are aware of the location of stressed syllables and use that information when categorizing accents, although I cannot confirm this hypothesis by formal ex-

[5]Since the lengthening effect is intended to be captured in just one attribute, it is straightforward to take the nucleus z-score instead of z-scores of possibly several segments, which would have to be condensed into one value.

[6]File boundaries are ignored when accessing the duration z-scores of the context syllables, for the same reason as in the case of the PaIntE parameters of the context syllables (cf. footnote 4). Again, the only cases where the z-score parameters of the context are not accessible are the very first and the very last syllables of the corpus. In these few cases, the absolute parameters have been set to 0.

	nucleus duration			final segment duration		
	nucleus	pnucleus	nnucleus	finalseg	pfinalseg	nfinalseg
1	-0.876	0	0.32	-0.009	0	0.32
2	-0.183	-0.009	1.237	0.32	-0.009	1.237
3	0.427	0.32	-1.953	1.237	0.32	-1.953
4	-0.601	1.237	-0.958	-1.953	1.237	-0.958
5	-0.958	-1.953	-1.248	-0.958	-1.953	-1.248
6	-0.399	-0.958	-0.679	-1.248	-0.958	-0.679
7	-0.896	-1.248	0.728	-0.679	-1.248	0.728
8	0.728	-0.679	-0.057	0.728	-0.679	-0.057
9	-0.057	0.728	-0.976	-0.057	0.728	-0.976
10	-0.509	-0.057	1.283	-0.976	-0.057	1.283
11	1.283	-0.976	0.48	1.283	-0.976	0.48
12	-0.165	1.283	3.783	0.48	1.283	3.783
13	-0.862	0.48	-1.139	3.783	0.48	-1.139

*Table 6.4: Attributes 21 to 26 and corresponding values for the first thirteen instances of the SWMS data. The attributes indicate phoneme-specific duration z-scores of nuclei of the current (*nucleus*), preceding (*pnucleus*), and next syllables (*nnucleus*), and phoneme-specific duration z-scores of final segments of the current (*finalseg*), preceding (*pfinalseg*), and next syllables (*nfinalseg*).*

periments currently. However, when trying to prosodically label speech based on the larynx signal only, where only the F0 information is present, I found that in some cases where there was a peak in the F0 contour derived from the larynx signal, I could not tell whether the peak was the high target of a rising L*H accent on a preceding syllable or whether it was a high target corresponding to an H* or H*L accent at the current location. The location of the word stress could have disambiguated the contour in these cases—if the high target co-occurs with a stressed syllable, it is likely to be the high target of an H* or H*L accent, and unlikely to be the high trail tone of an L*H, because that would rather be expected on the following syllable.

The **wordfin** attribute indicates whether the syllable is word-final or not; it can take values of 0 (word-internal) or 1 (word-final). Similar to the relevance of the *stress* attribute in clustering or classifying accents, the *wordfin* attribute is evidently helpful in recognizing boundaries since word-internal syllables at least in unmarked speech are never phrase-final and thus should never exhibit boundary tones. Again, I would hypothesize that human listeners make use of this information as well. Phrase boundaries, particularly intermediate phrase boundaries, are very hard to recognize when labeling speech based on the lar-

	phonological			text-based					
	stress	wordfin	silnext	pos	func	top	head	wght	punc
1	1	1	0	ART	func	Vf	-	1	0
2	1	0	0	NN	cont	VfEnd	+	1	0
3	0	1	0	NN	cont	VfEnd	+	1	0
4	1	1	0	VVFIN	cont	0	-	0	0
5	0	0	0	ADV	cont	0	-	0	0
6	1	1	0	ADV	cont	0	-	0	0
7	0	0	0	NN	cont	0	+	1	0
8	1	0	0	NN	cont	0	+	1	0
9	0	1	0	NN	cont	0	+	1	0
10	1	0	0	VVINF	cont	0	-	0	0
11	0	0	0	VVINF	cont	0	-	0	0
12	0	0	0	VVINF	cont	0	-	0	0
13	0	1	1	VVINF	cont	0	-	0	,

*Table 6.5: Attributes 27 to 35 and corresponding values for the first thirteen instances of the SWMS data. The phonological attributes indicate whether the syllable is stressed (*stress*), whether it is word-final (*wordfin*), and wether a silence follows (*silnext*). The text-based attributes specify the part-of-speech of the corresponding word (*pos*), whether it is a function word or a content word (*func*), whether the syllable belongs to the topological field "Vorfeld" (*top*), whether it belongs to the head of a noun chunk (*head*), the weight of the noun chunk in terms of contained number of content words (*wght*) if the syllable is part of a noun chunk, and which kind of punctuation symbol occurs at that word, if any (*punc*).*

ynx signal only.[7] This is to be expected since temporal cues play an important role in the perception of phrase boundaries. However, it is not possible to manipulate speech in a way that integrates temporal information in the larynx files, or at least to manipulate speech in a way that leaves only temporal information and takes away segmental information, thus it can not be experimentally verified whether temporal information alone or in addition to the larynx information is sufficient for human perception of boundary tones, or whether humans might use cues such as word finality in perception. However, since the hypothesis can be neither verified nor dismissed properly, at least not in this thesis, the attribute *wordfin* is provided with the other attributes. It will turn out later on that it does increase performance in the experiments presented in section 6.3.

[7] An exception are high boundary tones (H%, or % which inherit their high specification from a preceding trail tone); these are easier to identify because of the often quite pronounced rise in F0.

The last of the "phonological" attributes is **silnext**. It specifies whether a silence follows (in which case its value is 1, otherwise 0). It is evident that the *silnext* attribute should be a good indicator of phrase boundaries since silences, i.e. pauses, are very frequent at phrase boundaries.

The remaining attributes are "text-based" attributes which are derived from orthographic information and punctuation marks. These attributes are a subset of the attributes used by the prosody prediction module of the IMS Festival text-to-speech synthesis system (IMS Festival 2010). Thus, they are known to be predictive of prosodic structure, and they were included mainly to see if their presence would increase performance—given that they contribute in text-based prediction of prosodic events, they might be helpful in the (phonetically based) classification experiments as well. The attribute **pos** specifies the part-of-speech (POS) tag of the word that the syllable is related to. POS tags were obtained by the German Tree Tagger (Schmid 1994, 1995), which uses the STTS tag set (Schiller et al. 1995). The STTS tag set comprises 12 main classes of POS tags (adjectives, adverbs, adpositions, determiners, cardinals, interjection, conjunctions, nouns, pronouns, particles, verbs, and miscellaneous), which are further subdivided into altogether 54 different types of POS tags. For instance, verbs are subdivided into full verbs, modal verbs, and auxiliary verbs, and for each of these subclasses, STTS distinguishes between finite verbs, imperative forms, infinitives, and participles, yielding 12 different POS tags pertaining to verb forms. I have also included the attribute **func**, which maps the POS tags to just two classes for function and content word (values: func, cont) in the following way: adpositions, determiners, conjunctions, pronouns, particles, and auxiliary verb forms are function words, everything else is classified as content word.

The attribute **top** is also derived from the POS tags. It specifies whether the syllable belongs to the so-called Vorfeld, if it ends the Vorfeld, or if it is not part of the Vorfeld at all. Vorfeld is a term which was introduced by topological models of German syntax (e.g., Höhle (1986), or see Grewendorf et al. (1989) for a description of topological models in general). These models state that sentences are divided into topological fields: Vorfeld, left bracket, Mittelfeld, right bracket, and Nachfeld (note that the right bracket is called "Verbalkomplex" by Höhle (1986)). The Vorfeld can accommodate exactly one constituent, which is often the subject. The left bracket is reserved for either the finite verb (in main clauses) or for conjunctions (in subordinate clauses). However, in the latter case the Vorfeld must remain empty. Thus, the occurrence of the finite verb is a good indicator of the boundary of a preceding constituent of possibly high complexity, which is why our TTS prosody prediction module assigns prosodic phrase boundaries at the end of the Vorfeld if the length of the Vorfeld exceeds a certain threshold. The end of the Vorfeld is located by the occurrence of a POS tag indicating the finite verb. Possible values of the **top** attribute are VF (for words contained in the Vorfeld), VFEnd (for the final word in the Vorfeld), and 0 (else).

Another concept that is used for predicting prosodic events in our TTS system is the noun chunk. According to Abney (1995), chunks are those fragments of a parse tree that remain when detaching all elements whose attachment to the tree is not unambiguous, and chunk boundaries correspond to prosodic boundaries. Interpreting Abney in a simplistic way, noun chunks thus can only consist of possibly a determiner, followed by zero to several adverbs and adjectives, and one to several nouns. Modifications by prepositional phrases, for instance, would give rise to their own, separate noun chunk, because the attachment of prepositional phrases would be ambiguous. Thus, using the POS tags obtained from the Tree Tagger, noun chunks can be approximately identified by their typical patterns of sequences of POS tags. Noun chunks are not only predictive of prosodic boundaries; they also often carry pitch accents, usually on the final noun, because of the right-branching structure of German noun phrases. Thus, the final noun in a noun chunk not only signals a potential site for a prosodic boundary, it is also a candidate site for a pitch accent. In the IMS German Festival TTS system, final nouns in noun chunks are therefore referred to as head of the noun chunk. The classification of a word as being the head of a noun chunk, based on the identification of noun chunk boundaries by patterns of POS tags from the Tree Tagger, is captured in the **head** attribute (values +, -). Since the length of noun chunks also seems to be relevant for the assignment of prosodic boundaries or accents (very short noun chunks are less likely to be realized with prosodic boundaries, for instance), we also determine their weight in terms of number of content words in the noun chunk. This weight is indicated by the **wght** attribute; it can take positive integer numbers for syllables in noun chunks, or 0 for syllables which do not belong to noun chunks.

The last attribute, **punc**, specifies whether there was a punctuation symbol after the syllable in the text underlying the utterance (i.e., in the text the speaker was prompted with). Possible values are 0 (for no punctuation) or any of the following symbols: ,;:?!.)"

6.1.2 Databases for training and testing

The data (i.e., the attributes for each syllable instance) were extracted from two databases, viz. the SWMS database (2 hrs., male) and the SWRK database (3 hrs., female), which had been recorded in the course of the SmartWeb project (Wahlster 2004). The speakers are professional speakers of Standard German. For both databases, the utterances represent typical utterances of 5 different genres. They were read off a screen at recording time. Afterwards, the utterances were annotated on the segment, syllable, and word level, and prosodically labeled according to GToBI(S) (Mayer 1995). Prosodic labeling for each utterance was carried out by one of three human labelers, all supervised and

instructed by myself, without having the present experiments in mind. The SWMS database amounts to 72,000 segments, 28,000 syllables, and 14,000 words. The SWRK data contain 88,000 segments, 34,000 syllables, and 17,000 words.

The databases were originally used for unit selection speech synthesis in the SmartWeb project (Barbisch et al. 2007). The recording procedure and the prompts were identical for both databases. Utterances corresponding to different prompts were saved in separate speech files. The utterances represent typical utterances of the following genres: soccer reports, tourist information, newspaper articles, dialog system prompts, and numbers, and they usually consist of one or at most two short sentences, corresponding to several prosodic phrases.

The utterances were recorded in blocks of 30 or 20 utterances of one genre, switching to the next genre after each block if there were still unrecorded utterances left for that genre. Before recording, all utterances within a genre had been indexed according to their order of presentation during recordings. The file names of the speech files consist of a character identifying the genre and that index. The procedure and the prompts were identical for both the SWMS and the SWRK database, but more utterances were recorded for the SWRK database than for the SWMS database. Thus, the SWMS utterances are a subset of the SWRK utterances, and utterances with identical file names in the two databases represent utterances with identical prompts.

For the experiments discussed in the following sections, the data have been divided into test and training set according to the index contained in the file name: files with names ending on 1 are in the test set, files with names ending on 0 or 2 to 9 are in the training set. This ensures that there are approximately equal proportions of utterances of each genre in both sets, and that there are comparable amounts of utterances from early or late in the recording session in both sets. Another aspect is that by taking into account the file names in separating test and training data, one avoids having originally adjacent syllables from one utterance in test and training set. This was deliberate since it is assumed that adjacent syllables are less independent from each other than syllables originating from different prompts. Last but not least, it was intended to test results of the experiments gained on the SWMS data (which was spoken by a male speaker) on the SWRK data (which was spoken by a female speaker) in order to get a first estimation of how well the results generalize to other speakers. Since the prompts of the utterances were identical for the SWMS and the SWRK database, this means that the split in test and training set is the same for both databases—utterances that are in the test set of the SWMS database are in the test set of the SWRK database, and utterances that are in the training set of the SWMS database are in the training set of the SWRK database.[8]

[8]However, there are utterances in the SWRK test and training set that are not present in the

6.1.3 Noisy data

Before turning to the actual experiments, a note of caution is in order: One should keep in mind that almost all attributes discussed above are extracted automatically, and there is a lot of room for errors. Regarding the PaIntE parameters, errors can be introduced in estimating the F0 values from the speech signal. Then, smoothing the resulting F0 files may introduce undesired artifacts. The approximation itself is of course also prone to error—outliers are not uncommon, which is why a plausibility check is carried out after approximation (cf. section 3.2.4). One of the criteria used there is that the results are only trusted if the approximation window contained at least 10 frames, i.e., 10 F0 values. This helps to reduce outliers, however, it does occasionally reject correct parameters, and on the other hand, sometimes will still not be strict enough. Also, not all errors introduced in the approximation are caused by short approximation windows.

Turning to the duration z-scores, they are based on the segment labels in the database. These were generated automatically using forced alignment on the segment level, and corrected manually afterwards. However, even though the manual correction was done in a systematic way, by identifying possible outliers, there was no complete segment-by-segment check, and it is likely that some errors have gone undetected. This could involve erroneous start or end times for some segments, but also segment identity errors caused by incorrect lexical entries or by out-of-vocabulary words for which transcriptions had been generated automatically. Several of the higher-linguistic attributes are determined by simple, not always perfect, heuristics based on results of the part-of-speech tagger, which itself occasionally will misclassify words.

Thus, the data are inevitably noisy. Adding to that, the results of the experiments presented in this chapter are evaluated using manually annotated prosodic labels—but prosodic labeling is notorious for its subjectivity, and inter-labeler reliability is much lower than on the segment level. Thus, unexpected attribute values for an instance of a prosodic category may arise either because of errors in the extraction of the attributes, or because of subjectivity in prosodic labeling.

These considerations may sound pessimistic, but I have shown in chapters 4 and 5 that both duration z-scores and PaIntE parameters do capture well-known aspects of prosodic categories. The following two sections will show that even though the data are noisy, there are enough sound instances in the data to allow for automatic detection and prediction of prosodic categories. I will now address clustering as a means for automatically detecting these categories (section 6.2) before turning to prediction in section 6.3.

SWMS database at all because the SWMS utterances were a subset of the SWRK utterances, as mentioned above.

6.2 Detecting prosodic categories

The aim of this chapter is to model human categorization of prosodic events. From an exemplar-theoretic perspective, categories are represented by accumulations of exemplars in perceptual space (cf. section 2.3.2). According to exemplar theory, phonetic categories can be thought of as "clouds" of exemplars in perceptual space: each exemplar is located in perceptual space according to its perceptual properties. Exemplars of the same category form clouds because they are perceptually similar, and since similar exemplars are stored closely together, they should make up a cloud of exemplars with higher density.

Given that all relevant perceptual dimensions are known, it should be straightforward to detect such clouds in perceptual space using clustering techniques. For instance, Pierrehumbert (2003) reviews clustering results obtained by Kornai (1998) where clusters of F1/F2 data corresponded well to vowel categories. In this vein, she suggests that acquisition of phonetic categories may be guided by probability distributions of exemplars in perceptual space, claiming that "well-defined clusters or peaks in phonetic space support stable categories, and poor peaks do not" (Pierrehumbert 2003, p. 210).

Thus, a straightforward approach to detecting prosodic categories in perceptual space is to cluster prosodic data. The idea in clustering is to partition multidimensional data in a way that instances within each partition are similar to each other, and dissimilar from instances in all other partitions. Similarity is quantified using a distance measure. In other words, clustering tries to identify groups of instances in the data which are particularly close to each other. These groups or partitions are called clusters. Since instances in the clusters are particularly close, the clusters will correspond to regions with high population density, and thus, hopefully, to prosodic categories.

In this section, I will investigate the correspondence between clustering results and prosodic categories using pitch accents as an example. To this end, I will run several experiments on the SWMS database (cf. 6.1.2), using the attributes described in section 6.1.1 as dimensions for clustering. It will turn out that expecting a one-to-one correspondence between categories and clusters is too strong, at least for the data used here; however, I will show that there is a good correspondence between clusters and categories, and that for each cluster, one can identify a category which is particularly dominant in this cluster.

This section is organized as follows. I will first discuss some preliminaries in section 6.2.1 before turning to a first experiment, which is mainly provided as an introduction (section 6.2.2). Then, three cluster algorithms with varying numbers of clusters will be examined with regard to their performance in categorization using a new evaluation method, which will be introduced below (section 6.2.3). In section 6.2.4 I will use a more standard evaluation measure, viz. the v-measure suggested by Rosenberg and Hirschberg (2007), to determine the optimal number of clusters. Finally, in section 6.2.5 some problems

with the v-measure will be discussed, and I will return to the accuracy measure for determining the optimal number of clusters.

6.2.1 Preliminary remarks

In this section, using pitch accents as an example, I will consider clustering as a method to identify prosodic categories, by identifying clusters of very similar instances. To this end, I used the attributes discussed in the preceding section 6.1 as dimensions in clustering. Most of them are related to the duration z-scores and PaIntE parameters which have been argued to correspond to perceptually relevant properties of prosodic events in chapters 4 and 5. However it should be noted here that I used almost all attributes discussed in the preceding section, i.e., I did not only consider the duration z-scores and PaIntE parameters of the current syllable, but those of context syllables as well, as described above (section 6.1). Also, most other attributes described above were included. The only attributes that were left out were those that were expected to be useful in distinguishing boundaries only, i.e., while I did use the duration z-score parameters of nucleus segments, I did not use those of final segments (**finalseg**, **pfinalseg**, and **nfinalseg**), because they were intended for distinguishing phrase boundaries; the word finality attribute (**wordfin**) and the punctuation attributes (**punc**) were left out for the same reason, as well as the **silnext** attribute, which states whether a silence follows after the current syllable. Apart from these, all other attributes discussed in the preceding section 6.1 were used. I did try clustering the data using only the attributes motivated in chapters 4 and 5, however, the results were clearly better when using all attributes, thus I will only discuss results involving the full set of attributes here. I will get back to the issue of reducing the number of attributes in discussing the cluster results in section 6.2.6 and in the final discussion in chapter 7.

In theory, if all perceptually relevant parameters are provided as dimensions, prosodic categories should be sufficiently distinct in the resulting perceptual space. In this case, one can expect to find clusters in the data that represent the prosodic categories, or at least are related to them. In case of the experiments here, this can be verified using the manually determined prosodic labels for evaluating the observed clusters.

Several practical aspects have to be considered before actually turning to clustering experiments. The first issue is whether the data should be normalized before clustering or not. All clustering methods use some sort of distance function to establish how similar instances or clusters are. Many of the distance functions commonly used in clustering, such as the Euclidean distance,

the Manhattan distance,[9] or the Chebyshev distance,[10] are susceptible to differences in scaling between dimensions—changing the scaling of one dimension affects the distances between data points. Thus, scaling the values in a dimension will put more or less emphasis on this dimension because after scaling, the increased or decreased distance between data points along that dimension will also increase or decrease their overall distance.[11]

In many clustering applications, it is common to normalize all attributes to obtain values within a certain range, or with equal variance, in order to give each dimension equal weight. However, since in the experiments presented here, most numeric attributes have undergone z-scoring already, it was decided to not normalize these values further. One reason is that especially for PaIntE-related attributes, outliers are not uncommon.[12] In these cases, normalizing all values to lie in a certain range would cause the majority of values to be "crowded together" in some smaller interval. The extent of this crowding would be determined by the value of the most extreme outliers in that dimension, which in my opinion introduces a kind of uncontrolled and undesired scaling.

A second issue is how the cluster results can be evaluated. The most direct way of evaluating them in this particular setting is to compare the clusters to the "correct" classes, i.e., the manually labelled prosodic categories. Such an evaluation is called an external evaluation, and I will address this in more detail in section 6.2.4, where I will use the *v-measure* recently suggested by Rosenberg and Hirschberg (2007) to evaluate the clustering results in a more sophisticated way. However, to get a first idea of what can be expected from the clustering approach, I will calculate what I call *classification accuracy* of each clustering. To that end, each cluster is assigned a "predicted" class, which is the prosodic class most often observed in this cluster.[13] Then, the classification accuracy is the proportion of instances for which the manually labelled class matches the predicted class of its cluster. In other words, the classification accuracy is what one would get when using the clusterings to predict prosodic classes by assigning each cluster the class label with the highest likelihood in

[9] The Manhattan distance between two data points $p = (p_1, p_2, ..., p_n)$ and $q = (q_1, q_2, ..., q_n)$ is the sum of the absolute pairwise differences between their coordinates: $dist(p, q) = \sum_{i=1}^{n} |p_i - q_i|$

[10] The Chebyshev distance between two data points $p = (p_1, p_2, ..., p_n)$ and $q = (q_1, q_2, ..., q_n)$ is the maximal absolute pairwise difference between their coordinates: $dist(p, q) = \max_i(|p_i - q_i|)$

[11] In case of the Chebyshev distance, the overall distance is only affected in those cases where before or after scaling the difference along that dimension is the maximum distance.

[12] It is a wide-spread practice to remove outliers before clustering; however, as discussed in section 3.2.4, some cases of outliers or contexts that are likely to produce outliers are dismissed already in a check for plausibility in the approximation process. Thus, the remaining outliers may represent meaningful values, which is why I chose to not remove these as well.

[13] As described in section 6.1.1, the prosodic classes are taken to be GToBI(S) labels, with downstepped accents mapped to their counterparts without downstep.

that cluster. Or, put yet another way, in an exemplar-theoretic fashion, if one takes the clusters to be accumulations of similar instances stored in memory, and categorizes each instance based on the majority category label in that cluster, how many instances would have been categorized correctly?

The formal definition of the classification accuracy is as follows. Let $\mathcal{K} = \{K_1, K_2, ...K_N\}$ be the clusters and $\mathcal{C} = \{C_1, C_2, ...C_M\}$ the externally determined classes. To calculate the corresponding contingency table a, we determine for each pair of clusters and classes how many instances are both members of cluster i and of class j, i.e., the cells a_{ij} of the contingency table are calculated using

$$a_{ij} = |\{x | x \in K_i \wedge x \in C_j\}|, 1 \leq i \leq N, 1 \leq j \leq M$$

Then, the classification accuracy can be calculated from the contingency table:

$$class\_acc(a) = \sum_{i=1}^{N} \frac{\max_j a_{ij}}{\sum_{j=1}^{M} a_{ij}} \tag{6.1}$$

Explicating definition 6.1 in more detail, if each instance is predicted to belong to the class which is most frequent in its cluster, this prediction will be correct for all those instances which are in the most frequent class, i.e., looking at cluster K_i, the prediction will be correct for $\max_j a_{ij}$ instances, out of the total number of instances in cluster K_i, which is given by $\sum_{j=1}^{M} a_{ij}$. Summing over all clusters K_1 to K_N yields the overall accuracy.

Consider, for instance, the contingency table listed in table 6.6. It lists 15 clusters which have been found using the Expectation Maximization (EM) algorithm. The table indicates which manually labelled prosodic categories were located in which cluster. Each row in the table corresponds to a cluster, and the columns indicate for each class how many instances of that class were located in each cluster. For example, cluster K1 comprised, among others, 58 instances which had manually been labelled H*, 51 which had been labelled H*L, 147 which had been labelled L*H, and a majority of 1198 which had been labelled as unaccented. Indeed, unaccented syllables dominate all clusters but K13: here, 891 instances were L*H instances, while there were only 761 unaccented syllables. K13 is highlighted in table 6.6.

At first glance, these results look discouraging, as almost all clusters are dominated by the "NONE" class, which is of course also the overall most frequent class. According to the procedure suggested above, all instances in these clusters are predicted to be unaccented. Instances which are located in cluster K13 are predicted to be L*H accented since this is the most frequent prosodic class in K13. The classification accuracy then is the percentage of instances for which the prediction was correct. And indeed, the classification accuracy of

	H*	..H	HH*L	H*L	L*	..L	L*H	L*HL	NONE
K1	58	0	0	51	0	0	147	36	1198
K2	193	0	1	150	4	1	317	27	1544
K3	25	0	0	10	0	3	51	5	1600
K4	14	0	0	12	7	3	13	0	1785
K5	48	0	0	114	3	3	321	31	1119
K6	1	0	1	101	4	2	8	0	1288
K7	2	0	0	14	1	0	18	0	317
K8	25	1	0	201	23	2	77	3	2075
K9	18	0	0	50	1	0	128	9	1580
K10	7	0	0	474	2	0	10	1	1176
K11	70	0	0	280	3	1	417	66	1302
K12	119	0	0	114	14	1	288	22	1316
K13	41	0	0	14	0	5	891	32	761
K14	47	0	0	39	22	1	100	4	1854
K15	26	0	0	19	1	2	108	5	493

Table 6.6: Contingency table obtained in clustering pitch accents using the EM algorithm with 15 clusters. K13 (highlighted gray) is the only cluster for which unaccented ("NONE") syllables were not the most frequent class.

this clustering is only 78.18%, which is only marginally better than the baseline of 77.66%, which one would get by always predicting "NONE".

Looking more closely, it can at least be observed that the proportion of manual classes in the clusters varies. This is shown in table 6.7. Here, the absolute frequencies from table 6.6 have been replaced by relative frequencies: for each cluster, the table cells indicate the percentage of instances belonging to each pitch accent class. For comparison, the first line indicates percentages over all data. Looking at H*L, for instance, the proportion of H*L instances in the clusters ranges from 0.59% in K3 to 28.38% in K10. Comparing these numbers to the proportion of H*L instances in the whole data set, 6.57%, it can be said that H*L accents are much less frequent than expected in K3, while they are much more frequent than expected in K10. Similarly, the proportion of L*H instances ranges from 0.60% in K10 (note that this was the cluster with the highest proportion of H*L instances) to 51.09% in K13. It can be stated that for all accents, there was at least one cluster in which the accent was at least 3 times more frequent than expected based on the proportions observed in the whole data set. These values are highlighted in table 6.7.

Thus, despite the discouraging dominance of unaccented syllables in all clusters, this particular clustering does capture some tonal aspects of pitch accent realization. This is most obvious in cluster K13, which contains much

	H*	..H	HH*L	H*L	L*	..L	L*H	L*HL	NONE
all data	2.78	0.00	0.01	6.57	0.34	0.10	11.58	0.96	77.66
K1	3.89	0.00	0.00	3.42	0.00	0.00	9.87	2.42	80.40
K2	8.63	0.00	0.04	6.71	0.18	0.04	14.17	1.21	69.02
K3	1.48	0.00	0.00	0.59	0.00	0.18	3.01	0.30	94.45
K4	0.76	0.00	0.00	0.65	0.38	0.16	0.71	0.00	97.33
K5	2.93	0.00	0.00	6.96	0.18	0.18	19.59	1.89	68.27
K6	0.07	0.00	0.07	7.19	0.28	0.14	0.57	0.00	91.67
K7	0.57	0.00	0.00	3.98	0.28	0.00	5.11	0.00	90.06
K8	1.04	0.04	0.00	8.35	0.96	0.08	3.20	0.12	86.21
K9	1.01	0.00	0.00	2.80	0.06	0.00	7.17	0.50	88.47
K10	0.42	0.00	0.00	28.38	0.12	0.00	0.60	0.06	70.42
K11	3.27	0.00	0.00	13.09	0.14	0.05	19.50	3.09	60.87
K12	6.35	0.00	0.00	6.08	0.75	0.05	15.37	1.17	70.22
K13	2.35	0.00	0.00	0.80	0.00	0.29	51.09	1.83	43.64
K14	2.27	0.00	0.00	1.89	1.06	0.05	4.84	0.19	89.70
K15	3.98	0.00	0.00	2.91	0.15	0.31	16.51	0.76	75.38

Table 6.7: Contingency table for EM clustering pitch-accented syllables, using 15 clusters. Table cells indicate cluster-wise relative frequencies of manually labelled pitch accent classes. For comparison, relative frequencies over the whole data set are given in the first line. Highlighted cells indicate relative frequencies that are at least 3 times higher than on the whole set.

more L*H accents than expected, and much less H*L accents than expected (their proportion is only 0.80% instead of overall 6.57%). Also, K10 contains much more H*L accents than expected (28.38% instead of 6.57%), but much less L*H accents than expected (0.60% instead of 11.58%). That way, K10 and K13 nicely separate L*H and H*L accents; however, they are less successful in filtering out unaccented instances at the same time, particularly in the case of K10, which still contains 70.42% of unaccented syllables, but also in the case of K13. Even though there are more L*H accents than unaccented syllables in this cluster, the proportion of unaccented syllables is still high (43.64%).

Summing up the preliminary considerations, I have suggested classification accuracy as an external evaluation measure that mimics exemplar-theoretic categorization: Category labels of instances are determined by the class label of surrounding (i.e., similar), exemplars. Using this measure for evaluating a first clustering example yields results that are discouraging at first glance, but a more detailed examination of the clustering reveals that there is some correlation between clusters and prosodic classed. I take this as an indication that it may be worthwhile pursuing clustering experiments further.

6.2.2 A first experiment

As a first step, I have compared the classification accuracy in pitch accent clustering of six algorithms implemented in WEKA (Witten and Frank 2005), viz. Cobweb, DBScan, EM, FarthestFirst, OPTICS, and SimpleKMeans. WEKA's HierarchicalClusterer could not be applied successfully due to memory limits. In most cases, the default settings suggested by WEKA were used, which is to determine the appropriate number of clusters automatically. However, SimpleKMeans and FarthestFirst require specification of the desired number of clusters, and the default settings with only 2 clusters are clearly inappropriate. As a first, straightforward approach the number of clusters was specified to be the number of prosodic categories, assuming as a first approximation that it will be possible to identify one cluster for each prosodic category.[14] In identifying accents, there are 8 different categories present in the SWMS data (cf. figure 5.3), but two of them occur only twice, thus 6 clusters appeared to be the better choice.[15]

Figure 6.1 shows the results for the six algorithms in terms of classification accuracy. The baseline of 77.657% is indicated by the solid line. As stated above it is obtained by assigning each instance the overall most frequent prosodic class as its predicted label ("NONE" indicating no accent). Again, these results seem discouraging, as only two algorithms (Cobweb and DBScan) reach classification accuracy rates that are clearly above the baseline (85.687% and 82.142%, respectively). However, these two algorithms yield much higher numbers of clusters than the other algorithms: Cobweb finds 3150 clusters, and DBScan finds 237. The numbers of clusters are indicated inside the bars in figure 6.1. The other extreme is OPTICS clustering, with a degenerate single cluster containing all instances. For this case, the classification accuracy trivially corresponds to the baseline. EM clustering has identified 15 clusters, which seems more appropriate as the number of clusters is in the range of the number of manually determined prosodic classes.[16]

For FarthestFirst and SimpleKMeans clustering, the most frequent prosodic class in each cluster was "NONE". Even though it is likely that in some of these clusters some prosodic classes have made up a greater proportion of instances than on average, as was the case in the EM example above, there was no cluster where a single class was more frequent than the "NONE" class, thus the results in both cases are equal to the baseline in terms of classification accuracy.

[14] Having discussed the EM clustering indicated in tables 6.6 and 6.7, it can be suspected that this assumption is too simple.

[15] One might even consider aiming at 5 clusters, since it is disputable whether the 26 instances of category ..L are sufficient. However, later on in this section, the question of appropriate cluster numbers will be treated in more detail.

[16] The results of the EM algorithm in this experiment are actually the results discussed above in more detail (tables 6.6 and 6.7).

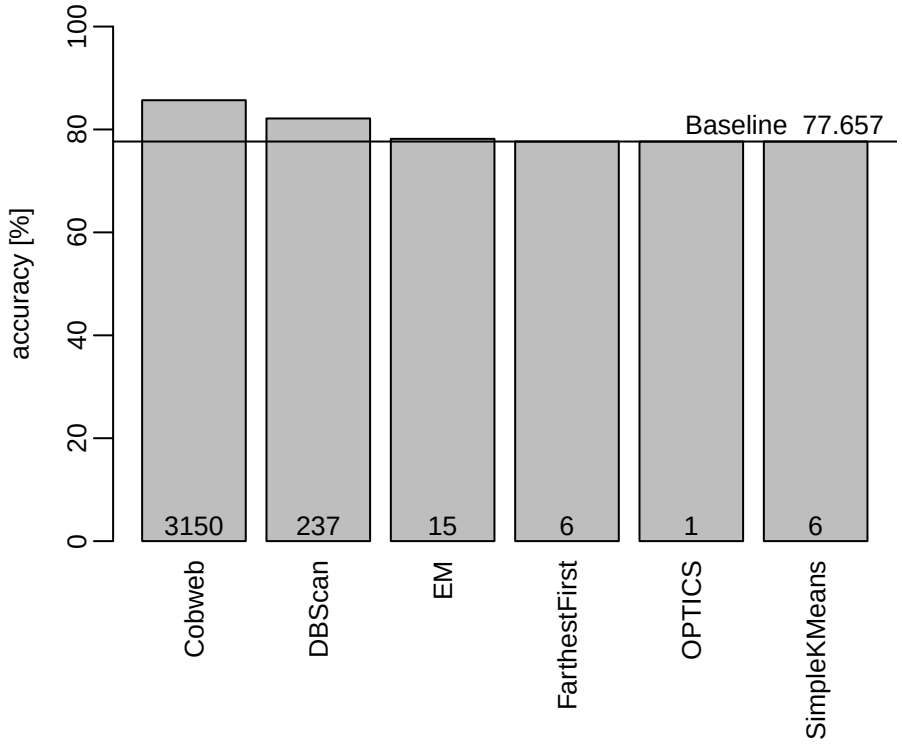

Figure 6.1: Classification accuracy rates for pitch accent clustering using several WEKA clustering algorithms with default settings (except for the number of clusters, see text). Only two algorithms reach rates clearly above the baseline (solid line). The number of clusters obtained is indicated inside the bars.

Two questions arise immediately: (i) Could the high numbers of clusters for Cobweb and DBScan indicate that the clusterings were over-adapted to the data? (ii) Which classification accuracy rates could be obtained by allowing FarthestFirst and SimpleKMeans to identify more clusters?

Addressing question (i), it must be considered that the decision of which hypothetical class label should be assigned to the instances in each cluster was based on the very same data on which the accuracy is now evaluated. To rule out over-adaptation to the data, I have applied the clusterings to the independent test data (cf. section 6.1.2), and repeated the evaluation, taking clusters-to-classes assignments from the clusterings of the training data. Applying a cluster model obtained on some data to new data is provisioned for by WEKA: it is possible to save clusterings to a cluster model and to load this model for clustering new data. In the case of k-means, clusters are represented by their cluster medoids, and thus, saving the model would effectively just save these medoids. Similarly, in the case of EM clustering, clusters are represented by multivariate Gaussian distributions, so saving a clustering obtained by the EM

algorithm would result in saving the distribution for each cluster. In any case, unseen instances from a new data set can be assigned to the clusters based on the saved model: in the case of k-means clustering for instance, the pertinent cluster is determined by finding the closest medoid. In the case of EM clustering, it is determined by finding the cluster to whose Gaussian distribution the instance is most likely to belong to.[17]

The idea in using the classification accuracy measure given in definition 6.1 above was to determine for each cluster a corresponding prosodic class. In the definition of the classification accuracy, the cluster-to-class correspondence was determined by the majority class in each cluster. However, if the clusters represent "real" categories, one would expect that the cluster-to-class correspondence should be the same on new data. In fact, this is a very desirable property in the context of my experiments — given that both clustering and cluster-to-class correspondence have been determined on independent data, applying the clustering to new data and evaluating it in terms of classification accuracy will assess its generalizability, and thus predict how useful the clusters are in categorization. In evaluating the clusterings on new data, I therefore suggest to not only keep the clustering, but also the cluster-to-class correspondence of the original data, and to apply these in classifying new data. To my knowledge, this is a new evaluation method, and it necessitates redefining the classification accuracy measure to allow for independently determining the cluster-to-class correspondence.

The formal definition of the classification accuracy given in definition 6.1 above can be replaced by definitions 6.2 and 6.3 below, which separate determining the cluster-to-class correspondence from calculating the actual classification accuracy. For each cluster K_i the corresponding class can be determined from contingency table a by finding the index of the most frequent class in that cluster, i.e., the index of the corresponding class is the index k for which a_{ik} is maximal. Since it could happen that several classes are equally frequent in a cluster, we take the minimal index for which the maximum occurs from the set of indices for which a_{ij} is maximal, L_i:

$$class(a,i) = min(L_i) \quad with \quad L_i = \{k | a_{ik} = \max_j a_{ij}\}, 1 \leq i \leq N \tag{6.2}$$

This way, the classification accuracy of a clustering applied to new data can be determined in the following way. Let b_{ij} be the cells of contingency table b obtained by applying the clustering to the new data, with b_{ij} calculated in analogy to a_{ij} for instances from the new data. The classification accuracy for the new data is then defined as

[17]Both algorithms are described in some more detail at the beginning of the following section.

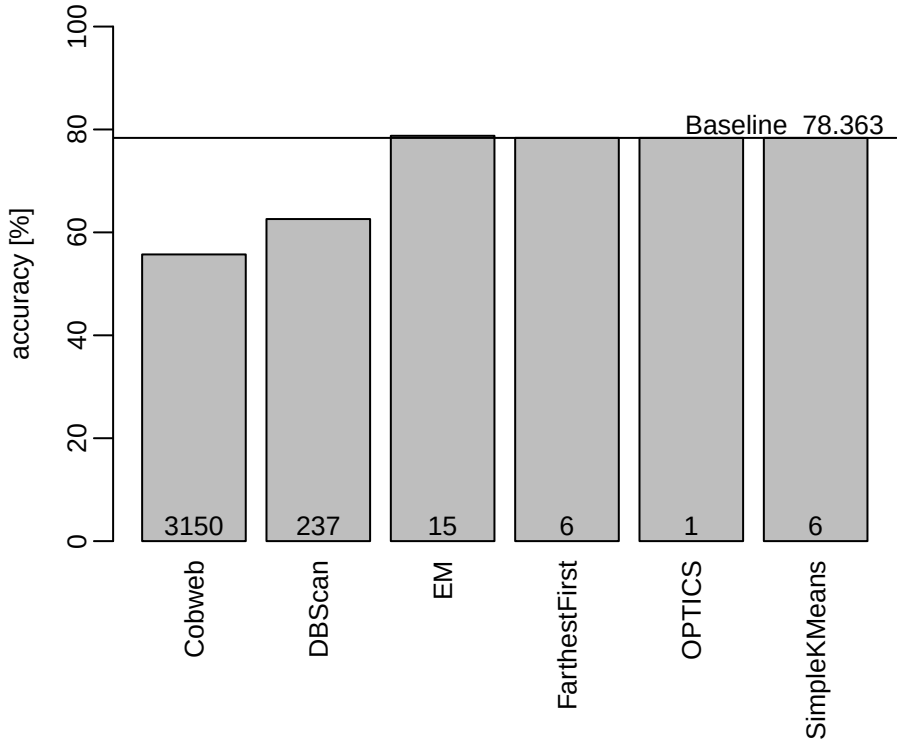

Figure 6.2: Classification accuracy rates for pitch accent clustering using several WEKA clustering algorithms with default settings, obtained on independent test data. Here, the rates for Cobweb and DBScan have dropped below the baseline (solid line); for EM clustering, the results are again very slightly above the baseline, and FarthestFirst and SimpleKMeans clustering are still at the baseline. Again, the number of clusters obtained is indicated inside the bars.

$$class\_acc(b) = \sum_{i=1}^{N} \frac{b_{i,class(a,i)}}{\sum_{j=1}^{M} b_{ij}} \tag{6.3}$$

Here, the number of instances which are classified correctly is the number of instances in the class determined by the clusters-to-class correspondence. The index of this class is $class(a, i)$, and thus the number of instances which are classified correctly is $b_{i,class(a,i)}$. Thus, the clusters-to-class correspondence is determined from contingency table a, and the classification accuracy is determined from contingency table b using this correspondence. If $b = a$, i.e., if the classification accuracy is determined on the original data, definition 6.3 is equivalent to definition 6.1.

Re-evaluating the clusterings above on the independent test data in this way yields the classification accuracy rates given in figure 6.2. Obviously, in the case

of Cobweb and DBScan, the clustering was indeed too specific to the training data—on the test data, the accuracy has dropped considerably below the baseline. This indicates that the clustering did not reveal "real" structure caused by regularities in the perceptual attributes of prosodic categories investigated here but rather structure specific to that particular selection of data, including noise.

These results are rather discouraging: OPTICS has found only one degenerate cluster, which yields baseline performance. Three algorithms (SimpleKMeans, FarthestFirst, and EM) perform only at or slightly above the baseline because in almost all clusters, unaccented syllables are more frequent than any of the pitch accents. And the two algorithms which had identified more clusters, Cobweb and DBScan, despite performing clearly above the baseline on the training data, have obviously failed to capture generalizable regularities in the realization of prosodic categories—they now perform dramatically below the baseline.

However, for EM, the minimal advantage over the baseline was stable when the cluster model was applied to new data. The results of the EM clustering are the results discussed in more detail in the previous section (cf. tables 6.6 and 6.7). As explicated there, this particular clustering did capture tonal aspects, despite the fact that it performed approximately at the baseline, and it can be suspected that the same is true for the FarthestFirst and SimpleKMeans clusterings.

Question (ii), which classification accuracy rates could be obtained by allowing FarthestFirst and SimpleKMeans to identify more clusters, is still relevant, even though it has turned out that the good classification accuracy rates of Cobweb and DBScan clustering did not carry over to new data. In the EM clustering example, the proportion of accents in the clusters varies, but the predominance in numbers of unaccented instances mostly prohibited clusters where accents dominate. When increasing the number of clusters, the average size of the clusters will decrease, and it is conceivable that then clusters can emerge in which the unaccented instances are less dominant.

Whether increasing the number of clusters will improve the clustering results will be addressed in the following section.

6.2.3 Varying numbers of clusters

To systematically assess the influence of the number of clusters on the clustering results, I have varied the number of clusters for FarthestFirst, SimpleKMeans, and EM clustering. These three algorithms are the only ones for which it was possible in WEKA to specify the number of clusters manually. I will briefly sketch how the three algorithms are implemented in WEKA (Witten and Frank 2005).

In SimpleKMeans clustering, k points from the data set are randomly chosen

as cluster centers. All remaining instances are then assigned to the cluster center which is nearest in terms of Euclidean distance, resulting in k initial clusters. Then, iterating until the assignment of instances to clusters does not change any more, the cluster centers are modified: for each cluster, its centroid, i.e., the mean over all instances in the cluster, is chosen as new cluster center, and all instances in the data set are in turn assigned to the nearest new cluster center. Choosing the means as new cluster centers causes accumulations of instances to "attract" the cluster centers because if there are many instances with similar values in a cluster, they all contribute to the mean, and thus the resulting centroid will usually be close to these similar instances. In this way, SimpleKMeans always finds a local optimum, but not necessarily a global one. The results usually strongly depend on the k initial randomly selected cluster centers.

FarthestFirst clustering is different in that it does not iterate. As SimpleKMeans does, it determines cluster centers, and instances are assigned to the nearest cluster center. However, once a cluster center has been determined, it is fixed and will not be revised later. The procedure is as follows: one first cluster center is chosen from the instances at random. The remaining k-1 cluster centers are each chosen in a way that they are farthest from the previously determined cluster centers, by maximizing the distance to the closest center. The effect of this procedure is that the clusters will be distributed across the space in which instances occur, and since placing clusters too close to each other is avoided, they will be distributed more or less evenly. But, in contrast to SimpleKMeans, the clusters do not necessarily occur in regions which are characterized by a higher density of instances, except that, once the space begins to get crowded with clusters and new cluster centers have to be placed in between already existing cluster centers, it may be more likely to find an optimal place between clusters in regions with higher density because there are more instances to select the new center from.

In contrast to SimpleKMeans and FarthestFirst, EM clustering does not take into account Euclidean distance. Also, EM produces soft clusters, i.e., instances may belong to several clusters, but with different probabilities. However, EM is similar to SimpleKMeans in that initially, k clusters are chosen at random, and these are modified iteratively. The difference is that clusters are not represented by cluster centers, but by multivariate Gaussian distributions over the attribute space. Then, iteratively, two steps are carried out. First, each instance is assigned to the cluster to which it most likely belongs to, given the cluster distributions. One can imagine that even if an instance is most likely to belong to a specific cluster, it may still exhibit values that are not very typical of this cluster. The next step can be thought of as counteracting such cases: the parameters of the Gaussian distribution of each cluster are modified in a way that the likelihood of the instances given their cluster is maximized. The iteration terminates if the change in the overall likelihood is below a predefined thresh-

old. In EM clustering, regions with higher density will correspond to peaks in the Gaussian distributions of the clusters, otherwise, their likelihood would not be maximal.

While for SimpleKMeans and FarthestFirst clustering, the clusters can be thought of as regions around cluster centers which stretch in each dimension half-way to the next cluster center, the clusters in EM clustering have "soft borders", and the parameters of their Gaussian distribution determine how far the cluster stretches in each dimension. Since in multivariate Gaussians, the variances along each dimension are specified independently of the other dimensions, clusters in EM clustering may stretch quite far in one dimension, but much less so in another dimension. I will come back to this in the discussion in section 6.2.6.

Turning to the experiments carried out using the three algorithms sketched above, I have systematically varied the number of clusters from 5 to 300 clusters. I will refer to this experiment as the 300 clusters experiment. Figure 6.3 gives the classification accuracies for all three algorithms, on training (upper panel) and test data (lower panel), respectively. The baselines are indicated by the dashed lines. As before, the baseline is obtained by assigning each instance the "NONE" label. The baseline is slightly higher on the test data because there are slightly more unaccented instances.

Clearly the best results are obtained by SimpleKMeans clustering (dark gray circles). Here, classification accuracy rates of more than 84% are obtained on both training data and test data. As may have been expected, the classification accuracy increases with the number of clusters. However, for SimpleKMeans clustering, the increase seems to level off at around 200 clusters on the training data, and at around 150 clusters on the test data. For FarthestFirst (light gray squares), a clear increase can be observed at close to 300 clusters; it is possible that this increase may continue beyond 300 clusters. However, given the results for up to 300 clusters, it seems unlikely that FarthestFirst approaches the classification accuracy of SimpleKMeans. The best rates for FarthestFirst are between 82 and 83% both on training data and test data. Regarding the EM algorithm (black diamonds), its results are not as good as the ones for SimpleKMeans and FarthestFirst, and it can be noted that they seem less stable than for the other two algorithms: they are more "jumpy" than the others even for higher numbers of clusters. Even for EM the results are clearly above the baseline now, approaching 81% on the training data when reaching cluster numbers of 300. This is better than the 78.18% obtained for 15 classes in the first experiment, where the number of clusters had been determined automatically by cross validation.[18] It should be noted however that for EM clustering, the results on the

[18]In the WEKA implementation of EM clustering, the optimal number of clusters is obtained in the following way: first, the model parameters are determined starting with only one cluster. Then, iteratively, the number of clusters is increased by one if this increases the cross-validated log-likelihood of the data, otherwise, the current number of clusters is regarded as optimal.

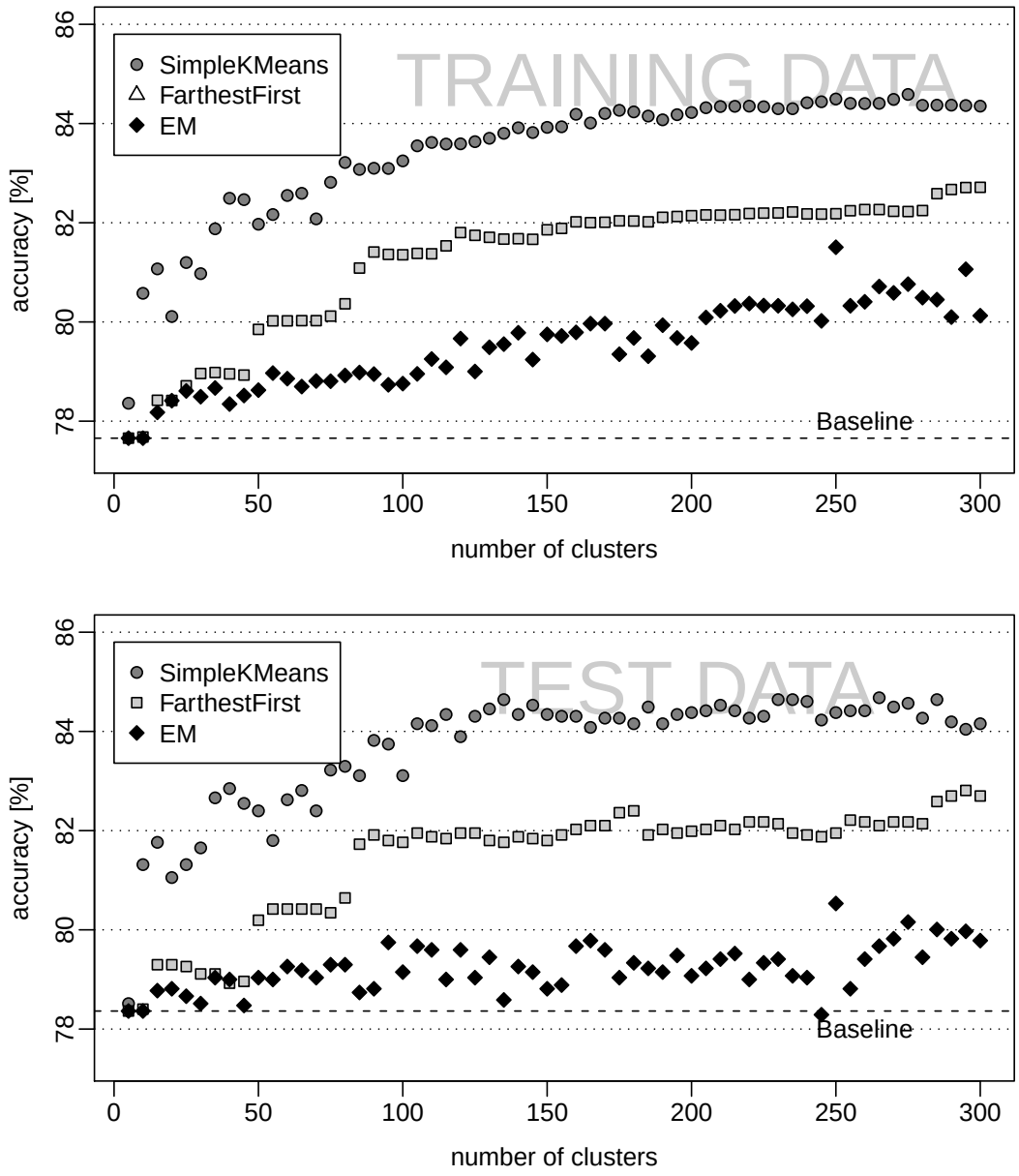

Figure 6.3: 300 clusters experiment. Classification accuracy rates for the SimpleKMeans, FarthestFirst, and EM algorithms with varying numbers of clusters between 5 and 300, evaluated on the training data (upper panel), and on independent test data (lower panel). The baselines are indicated by the dashed lines. For SimpleKMeans (dark gray circles), the accuracies seem to level off, but for FarthestFirst (light gray squares) and EM (black diamonds), the increase may well continue beyond 300 clusters.

The "jumpiness" observed here for classification accuracy may have been present in the log-likelihoods as well, causing WEKA to cease increasing the number of clusters with the first drop in the likelihoods.

test data are clearly below those reached on the training data.

In summary, the accuracies obtained in the 300 clusters experiment show that despite the first discouraging results, there is some potential in using clustering for detecting prosodic categories: if one takes the classification accuracy measure introduced above to simulate exemplar-theoretic categorization accuracy, it is a good sign that the results are clearly better than the baseline. Also, in contrast to the first experiment, all three algorithms have picked up regularities in the realization of prosodic categories that can be generalized to unseen instances. For SimpleKMeans and FarthestFirst, the classification accuracy scores on the test data are in the same range as those on the training data, indicating that the clusterings generalize very well, while in the case of EM, lower scores on the test data indicate that generalizability is limited. However, since the baseline is slightly higher on the test data, this may have shed a slightly too favorable light on the classification accuracy on the test data in all three cases.

Evaluating classification accuracy on new data is not a standard measure for external evaluation of clusterings, at least not to my knowledge. But particularly from an exemplar-theoretic perspective it is interesting because it resembles exemplar-theoretic categorization: there, new instances are categorized by comparing them to exemplars stored in memory, and assigning them to the category which is most frequent in the neighborhood (cf. section 2.3.3). Here, new instances are categorized by assigning them to the category which is most frequent in their cluster, i.e., if the cluster is taken to represent the neighborhood, the proposed procedure models exemplar-theoretic categorization, and the classification accuracy measures its performance.

A second aspect is that in the second part of this chapter (section 6.3), I will turn to predicting prosodic events using various classifiers. Against that background it is of course desirable to evaluate how the clusters obtained here would perform in prediction, compared to results that state-of-the-art classifiers would yield. In my opinion the most straightforward way of using the cluster results for prediction is the procedure suggested above.

Still, before turning to further experiments and to the question whether increasing cluster numbers further would deserve some attention, I will introduce another, more standard, evaluation measure than the classification accuracy used above.

6.2.4 Evaluation of clustering results

In many applications, there exists no manually or otherwise determined "correct" grouping or "gold standard" which could serve to evaluate the clustering. In these cases, the goodness of the resulting clusters has to be evaluated using internal criteria. For instance, the goodness of a cluster can be assessed by the sum of squared errors between the cluster centroid and the objects in the

cluster; or the goodness of classification of one particular object can be evaluated using its silhouette value (Kaufman and Rousseeuw 1990), which relates the distance of this object to other objects in the same cluster to its distance to objects of other clusters.

If on the other hand, the correct assignment of instances to classes is known, as in the present application, an external evaluation is viable. In this case, the clustering can be evaluated by comparing the clusters to the correct, externally given, classes. In the present experiments, the externally given classes are the prosodic categories, which have been manually labelled. A variety of external evaluation methods has been proposed in the literature. Recently, Rosenberg and Hirschberg (2007) have suggested the *v-measure* and demonstrated its applicability, among others, on a pitch accent type clustering task very similar to the present task. They claim that, contrary to many other external measures, v-measure is independent of the clustering algorithm or the data set, and independent of the number of clusters found. The latter property makes v-measure appealing to the present application in that it is feasible to determine the optimal number of clusters based on the v-measure scores. Also, given the independence from data sets and cluster algorithms, and the similarity of the two applications, the scores obtained here can be compared to Rosenberg and Hirschberg's (2007) results for pitch accent type clustering. The difference to the present task is that Rosenberg and Hirschberg (2007) classify only syllables that are known to be pitch-accented, while I want to obtain the classification as pitch-accented vs. unaccented from the clustering as well.

The v-measure is obtained by taking the harmonic mean of two scores which represent desirable criteria in clustering applications. One score measures homogeneity of the clustering. Homogeneity is satisfied if for each cluster, all members belong to the same class. The other score measures completeness, which is achieved if for each class, all instances that belong to this class are assigned to the same cluster. Homogeneity and completeness scores are implemented with reference to the conditional entropies of the class distribution given the clusters and that of the cluster distribution given the classes, respectively (Rosenberg and Hirschberg 2007, pp. 411–412).

For clustering pitch accents, Rosenberg and Hirschberg (2007) used data from the Boston Directions Corpus, which has been annotated according to the ToBI labeling conventions for Standard American English (Silverman et al. 1992). They mapped downstepped accents to their "normal" counterparts, obtaining 5 classes of accent types. Similarly, as discussed in section 6.1.1, in the present experiments, I have used GToBI(S) labels, with the same mapping of downstepped accents, obtaining 8 classes of accents.

To get an idea of how the classification accuracy results discussed so far would translate to v-measure results, I evaluated the first experiment again using v-measure for evaluation (cf. figures 6.1 and 6.2 for the classification accuracy rates of these clusterings). The results are indicated in figure 6.4. This

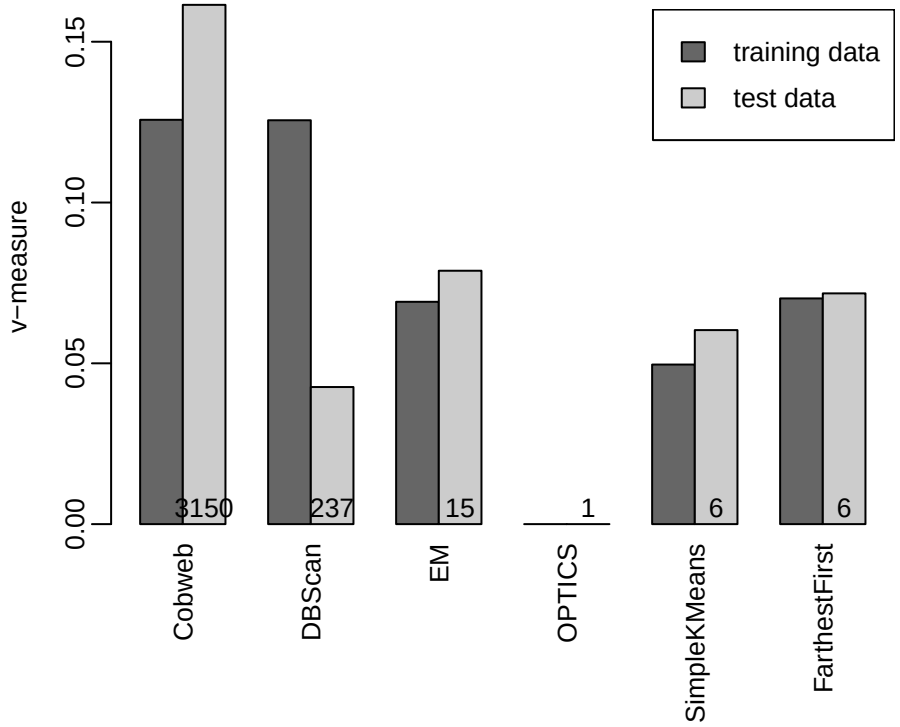

Figure 6.4: v-measure results for pitch accent clustering on training data (dark gray bars) and on test data (light gray bars) using several WEKA clustering algorithms with default settings. The number of clusters obtained is indicated inside the light gray bars.

time, results on training data and on test data are plotted next to each other in different shades of gray. The Cobweb algorithm yields the best results, as was the case for classification accuracy. Surprisingly for almost all algorithms the results are better on the independent test data, i.e., when applying the model obtained on the training data to the test data, v-measure is higher than on the training data itself. This effect is strongest for the Cobweb algorithm, even though a considerable drop had been observed in the classification accuracy scores for this algorithm when testing the clustering on the test data (cf. figure 6.2). A small increase in v-measure could be argued to be expected since the test data contained a slightly higher proportion of unaccented syllables. Recall that v-measure combines two scores which measure homogeneity and completeness of the clusterings. Admittedly, homogeneity must increase with a more skewed distribution of classes in the dataset — perfect homogeneity is achieved if the class distribution in each cluster is skewed to a single class (cf. Rosenberg and Hirschberg 2007, p. 411). However, it is less clear why completeness should improve. But since the effect is quite strong in the case of

Cobweb it is unlikely that it is only due to the improvement in homogeneity. I will discuss this effect in more detail below, arguing that it is an artificial effect which gets stronger with increasing number of clusters, and which arises when evaluating the clusterings using less data for testing than for training.

As for DBScan, it can be seen that v-measure drops considerably on the test data, confirming the effect found for classification accuracy, where the accuracy had dropped below the baseline. The results for EM, SimpleKMeans, and FarthestFirst are comparable to the classification accuracy results—the three algorithms perform roughly in the same range. OPTICS, with its single cluster, is successfully singled out by a v-measure of 0. These results demonstrate that v-measure is a more sensitive measure than the classification accuracy used in the first experiment, where there were almost no differences between EM, SimpleKMeans, FarthestFirst, and the baseline.

Turning to the 300 clusters experiment (cf. figure 6.3 for the results in terms of classification accuracy), the results in terms of v-measure are presented in figure 6.5. Results on the training data are presented in the upper panel, and results on the test data in the lower panel. Again, it can be observed that the results in general are better on the test data than on the training data: the best values of v-measure on the training data are around 0.12, while on the test data, the best values are in the range between 0.14 and 0.15. The values on the training data are slightly lower than the values of v-measure obtained for pitch accent clustering in Rosenberg and Hirschberg (2007), where values of approximately 0.125 to 0.13 were obtained for 100 clusters, and values of approximately 0.175 for 300 clusters.[19] It should be noted that their task was slightly different from the task presented here, since only pitch-accented syllables were clustered there, while I want the clustering to pick up not only differences in the realization of pitch accents, but also differences between unaccented and pitch-accented syllables.

Returning to figure 6.5, while in terms of classification accuracy SimpleKMeans was clearly better than FarthestFirst, in terms of v-measure, this depends on the number of clusters. For up to about 80 clusters, SimpleKMeans (dark gray circles) is indeed better than FarthestFirst (light gray squares), both on test data (upper panel) and training data (lower panel). Then, for up to about 230 clusters on the training data, and for up to about 180 clusters on the test data, SimpleKMeans and FarthestFirst are roughly comparable in terms of v-measure. For higher numbers of clusters, FarthestFirst performs better than SimpleKMeans. On both test and training data, SimpleKMeans reaches a maximum in v-measure at 150 clusters; for greater numbers of clusters, the values start to decrease, while v-measure for FarthestFirst both on training and on test data increases beyond 150 clusters, and possibly beyond 300 clusters. The same is true for EM (black diamonds)—while it does not reach the scores ob-

[19] v-measure results read off the diagram in Rosenberg and Hirschberg (2007, p. 418, fig. 5)

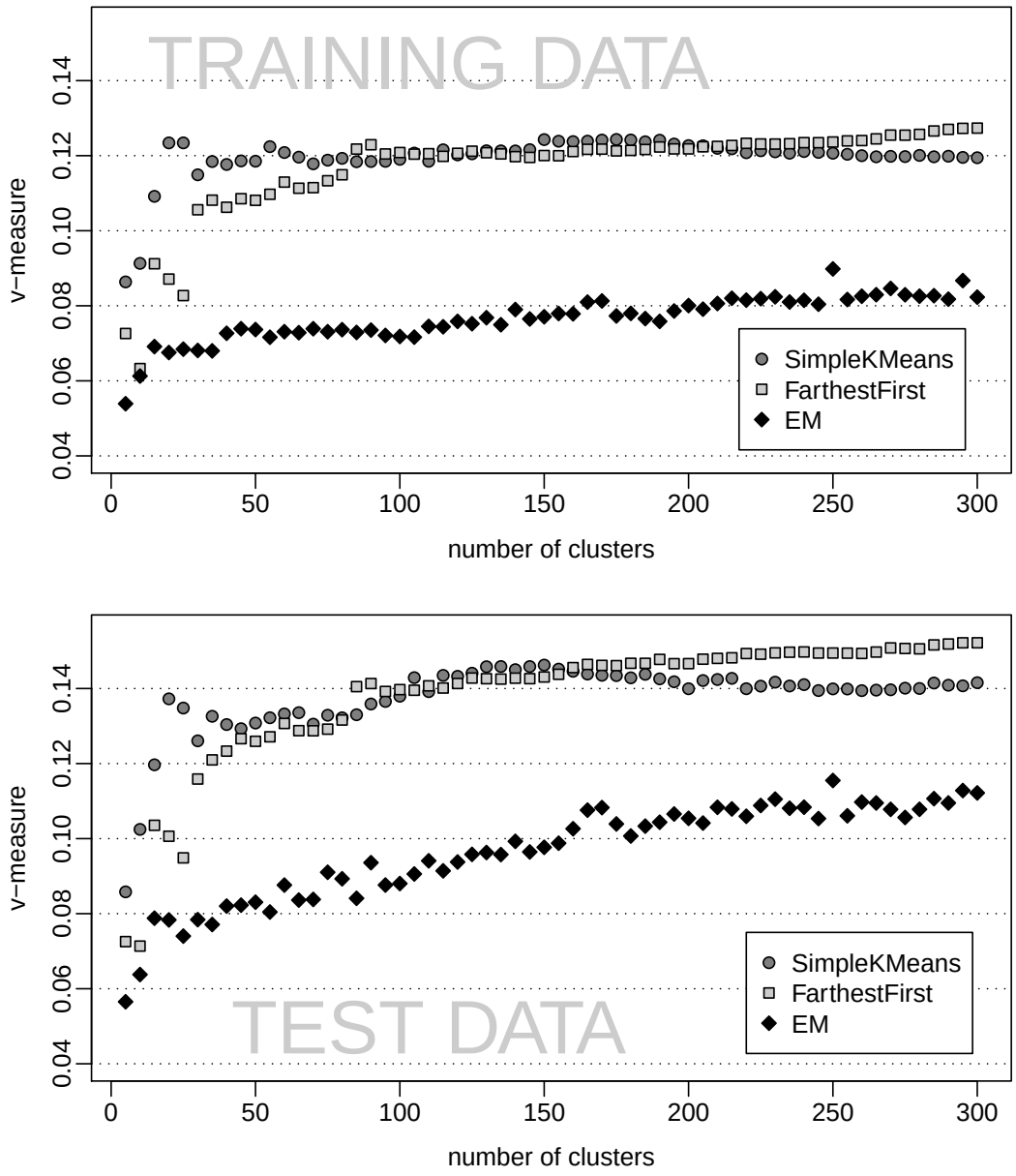

Figure 6.5: 300 clusters experiment. v-measure results for SimpleKMeans, FarthestFirst, and EM algorithms with varying numbers of clusters between 5 and 300, evaluated on the training data (upper panel), and on independent test data (lower panel). The v-measure scores seem to level off at around 150 clusters for SimpleKMeans (dark gray circles), but for FarthestFirst (light gray squares) and EM (black diamonds), the increase may well continue beyond 300 clusters.

tained for FarthestFirst and SimpleKMeans at this point, its scores increase with the number of clusters, and it cannot be said whether it could reach values of v-measure in the same range as the other two algorithms.

Summarizing the v-measure results of this experiment, it can be stated that for SimpleKMeans clustering, a number of no more than 150 clusters seems appropriate. However, for EM and FarthestFirst clustering, the results suggest that even higher numbers of clusters might be of interest.

I thus ran a third experiment aiming at up to 3200 clusters, which will be called the 3200 clusters experiment in the following. The results are presented in figure 6.6. The results on the training data (upper panel) indicate that for SimpleKMeans and FarthestFirst clustering, allowing more clusters does not increase v-measure further. For SimpleKMeans, the tendency observed already in the 300 clusters experiment is confirmed: for numbers of clusters beyond 150, there is no further increase in v-measure. For FarthestFirst clustering, the best values obtained in the 300 clusters experiment were at 300 clusters, and the results here show that this was an optimum even for up to 3200 clusters. In contrast, for EM clustering, the scores continuously but very slowly increase, possibly beyond 3200 clusters. The scores do no reach the scores obtained for SimpleKMeans and FarthestFirst, and since the increase seems to decelerate towards the end, I take this to indicate that EM clustering will not outperform the other two algorithms even for higher numbers of clusters.

For comparison, the v-measure for Cobweb clustering with its 3150 clusters (cf. figure 6.4) is indicated by the solid lines in figure 6.6. The lines are drawn from left to right across the whole diagram to allow for comparison to the results of the other three algorithms for all numbers of clusters, even though technically, the Cobweb results should be indicated by just one data point at 3150 clusters in each panel. For both FarthestFirst and SimpleKMeans, the optimal results on the training data are very close to the Cobweb result. However, the optimal v-measure scores are obtained at much smaller numbers of clusters, viz. at around 150 and 300 clusters, respectively.

The v-measure results on the training data are in line with the classification accuracy results of the 300 clusters experiment presented so far: there, the accuracy obtained on the test data had leveled off at approximately 150 clusters for SimpleKMeans, and there had been a small increase at approximately 300 clusters for FarthestFirst. However, so far I have only presented accuracy results up to 300 clusters; higher numbers of clusters may well yield better results in terms of classification accuracy.

But turning to the lower panel in figure 6.6, when applying the clusterings obtained on the training data to the test data, there is a continuous, slight increase up to 3200 clusters for all three algorithms. This is surprising, at least for SimpleKMeans and FarthestFirst—why should the results for the test data improve even beyond the point where they stagnate for the training data? Also, the difference in v-measure between test data and training data approaches

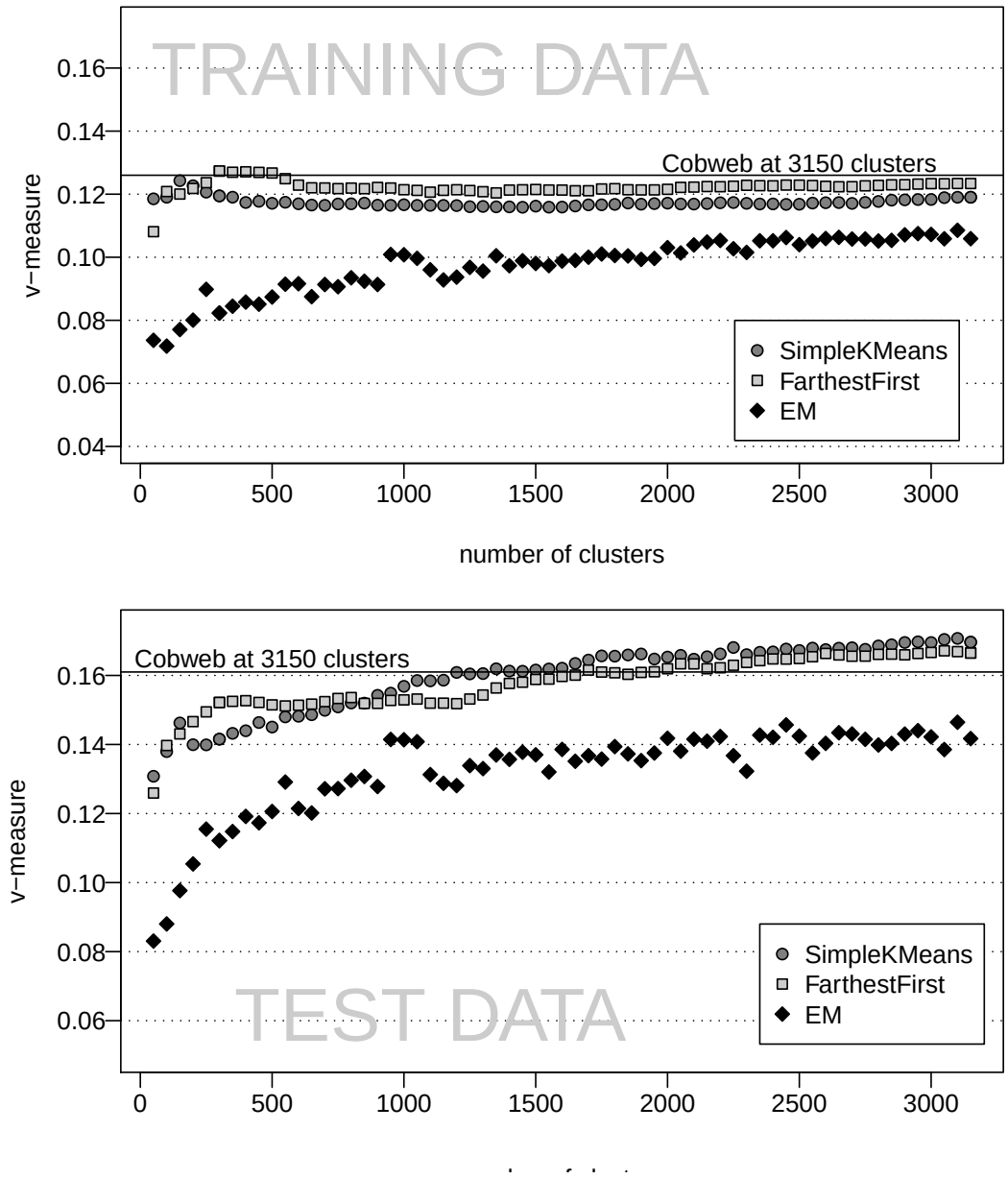

Figure 6.6: 3200 clusters experiment. v-measure for up to 3200 clusters, evaluated on the training data (upper panel), and on test data (lower panel). The horizontal solid lines indicate values obtained for Cobweb clustering with 3150 clusters. For SimpleKMeans (dark gray circles) and FarthestFirst (light gray squares), optimal results on the training data are at 150 and 300 clusters, respectively. On the test data, a continuous increase for all three algorithms can be observed.

0.04 for the highest numbers of clusters, while it was at roughly 0.02 in the 300 clusters experiment.

As explicated in the discussion of the v-measure results of the first experiment already, slight differences may have been attributed to slightly different class distributions in test and training data. But it is questionable whether differences in v-measure between training and test data in this order of magnitude can arise solely from slightly different distributions. I will address this issue in the following section.

6.2.5 Cross-validating the results

To rule out that the differences are only due to an unfortunate split into test and training data with different class distributions, I ran another experiment with up to 3000 clusters for SimpleKMeans and FarthestFirst clustering, using 10-fold cross validation. This will be referred to as the 3000X experiment. EM was not included because it takes a considerable amount of computational resources even without cross-validation. Since the main interest in this experiment was not to compare different algorithms to one another but to investigate the unexpected test vs. training data effect discovered above, leaving EM out here was considered unproblematic, in particular since its results both in terms of v-measure and in terms of classification accuracy were inferior to those of the other two algorithms.

The procedure for evaluating clusterings in the 3000X experiment was as follows. As described in section 6.1.2, the split into test and training data had been based on file names—files ending on 1 made up the test set. Basing the split on file names ensures that data from the same utterance never end up both in test and training set. I have kept to this principle and assigned the data to 10 different bins, based on the digit on which the file name ended. In clustering, I have used the data of 9 bins as training data, holding out the data from one bin. The obtained clustering was evaluated on the training data ("validation");[20] then, the clustering was applied to the data from the remaining bin and these results were evaluated as well (cross-validation). This procedure was repeated 10 times, holding out every bin once, resulting in an evaluation scheme that might be called 10-fold validation and cross-validation.[21]

The results are presented in figure 6.7. By and large, the results are similar to the results obtained in the simple validation procedure used above. Looking at the validation results from the training data (upper panel), FarthestFirst is superior to SimpleKMeans throughout. There is a little difference, however. In

[20] Usually, when using cross-validation, the results on the training data are not of interest; but they are relevant here to estimate the difference in performance on training vs. test data.

[21] It should be noted that since in each run, 9 out of 10 bins are used for training, the training sets in the 10 runs are never identical, but always overlapping. Only the test data are different and non-overlapping in each run.

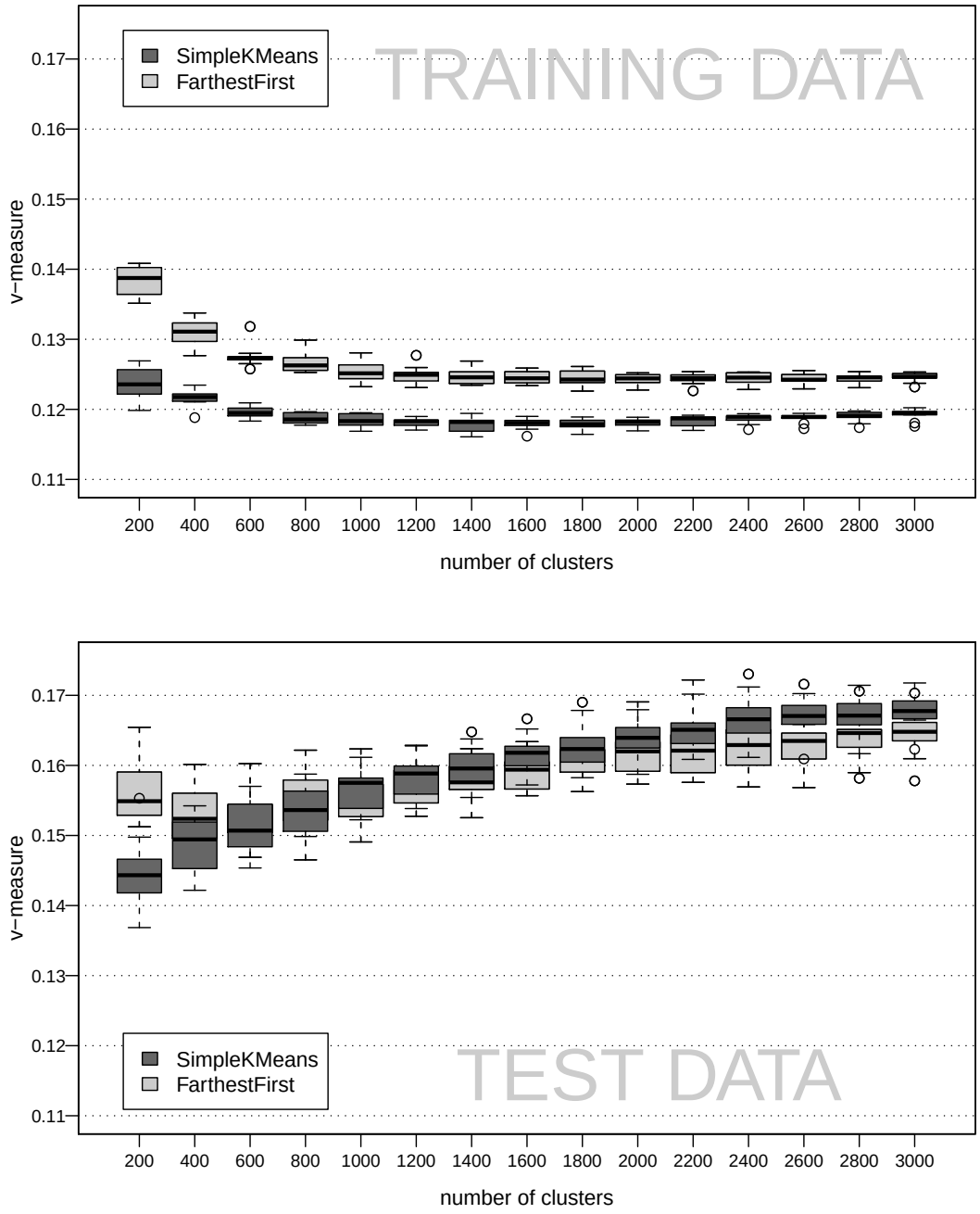

Figure 6.7: 3000X experiment. Cross-validated v-measure values obtained by 10-fold validation and cross validation using 200 to 3000 clusters, on the training data (upper panel), and on independent test data (lower panel). The boxes indicate the variation observed in each of the 10 folds. It can be seen that the difference in results on test and training data is present across all folds and increases with the number of clusters.

the simple validation case presented in figure 6.5, FarthestFirst and SimpleKMeans performed in the same range for cluster sizes between 150 and 250 clusters. This is not the case here, where in the 200 clusters case, FarthestFirst is better than SimpleKMeans for all 10 folds. The only difference in the clustering process was that the order of the input data has changed—due to sorting the instances according to the bins to which they belonged, which has changed the initial cluster centers. This demonstrates how sensitive FarthestFirst and SimpleKMeans can be to this initial choice. Anyway, in general the validation results on the training data in the upper panel are compatible with the earlier results—FarthestFirst is better than SimpleKMeans, and the values of v-measure decrease from 200 to 400 clusters, which would not contradict the earlier results which suggested that 300 clusters is optimal for FarthestFirst, and 150 clusters is optimal for SimpleKMeans.

But more importantly, the cross-validation results on independent test data (lower panel) confirm the earlier observation that the results are always better than the results on the training data: when comparing the boxes for the training data in the upper panel with the boxes for the test data in the lower panel, it can be observed that they never overlap: the effect is present across all folds.[22] The effect is quite strong for higher numbers of clusters: while on the training data, v-measure is very consistently slightly below 0.12 for SimpleKMeans and around 0.125 for FarthestFirst for numbers of clusters beyond 1000, on the test data, it is above 0.15 for cluster numbers beyond 1000 and even reaches 0.17 for some folds with increasing number of clusters for both algorithms. This confirms that the effect was not due to an unfortunate split into training and test data—rather, it persists in 10-fold cross-validation.

I suggest that the effect is caused by using different amounts of data for evaluating the clusterings. About 10% of the data are used for testing, while about 90% are used for training. This entails that the clusters, which have been determined on the training data, are less dense in the testing case, i.e., on average, there are fewer test instances in the clusters than training instances. Some clusters will not be populated by test instances at all. This disproportion is quite dramatic for numbers of clusters towards 3000: since the test set contains roughly 10% of altogether 28,000 instances, 2800 instances will be distributed to 3000 clusters, resulting in less than one instance per cluster on average. This should result in very good scores for homogeneity: in case of clusters with only one instance, homogeneity is trivially satisfied. On the other hand, completeness is expected to deteriorate when the instances are distributed across many different clusters. However, the completeness scores for the present data are very low already—simply said, they cannot get much worse anyway. Since v-

[22]The results are also always more variable on the test data than on the training data, but this is due to the fact that the training data are overlapping for the 10 folds, as mentioned above in footnote 21.

measure combines both scores, considerably increased homogeneity scores will more than compensate for the slightly decreased completeness scores.

In order to show that the higher v-measure values on the test data are indeed caused by evaluating the clusters on fewer instances than were in the training set, I repeated the experiment in two different settings. In one setting, I used only one bin for training, and the remaining 9 bins for testing, effectively reversing the training-test split from the 3000X experiment to have fewer training data than test data. This experiment is called the 10:90 experiment in the following. In the second setting, I have again used only one bin for training, and only one other bin for testing, resulting in about the same amount of training and test data. This will be referred to as the 10:10 experiment.

The results for the 10:90 experiment, with less training data than test data, are presented in figure 6.8. At first glance, it can be seen that the evaluation results seem exchanged. In evaluating the clusterings on the training data (upper panel), v-measure increases with increasing numbers of clusters, which is the trend that had been observed in evaluating on the test data in the 3000X experiment. When evaluating the clusterings on the test data (lower panel), v-measure is best for smaller numbers of clusters and then decreases with increasing numbers of clusters, as it had when evaluating on the training data in the 3000X experiment.[23]

It is further interesting to note that not only are the trends for increasing numbers of clusters exchanged when reversing the training-test split, but the scores are a little more extreme:[24] In the 10:90 experiment, the scores on the training data resemble those obtained on the test data in the 3000X experiment, but they are slightly better. Vice versa, the scores on the test data in the 10:90 experiment resemble those on the training data in the 3000X experiment, but they are slightly worse.

This observation indicates that, while v-measure may be independent of the clustering algorithm, the numbers of clusters, and the data set, as Rosenberg and Hirschberg (2007) claim, it is obviously not independent of the size of the data set. Thus, for the present application, in which testing on independent data is interesting because generalizability to new data is desirable, the setting for evaluation should be chosen in a way that the clusterings can be evaluated using the same amount of data for training and testing. This is what I have done in the 10:10 experiment, using one bin for training, and one bin for testing, resulting in roughly equal amounts of training and test data, each consisting of approximately 10% of the original data.

The results are given in figure 6.9. Evaluating on the training data (upper

[23]It can also be observed that the scores on the training data are more variable than those on the test data. This is because this time, there was no overlap between training data, but there was an approximately 90% overlap in the test data, cf. footnote 21.

[24]The axes in figures 6.7 and 6.8 are identical to allow for direct comparison of the two diagrams.

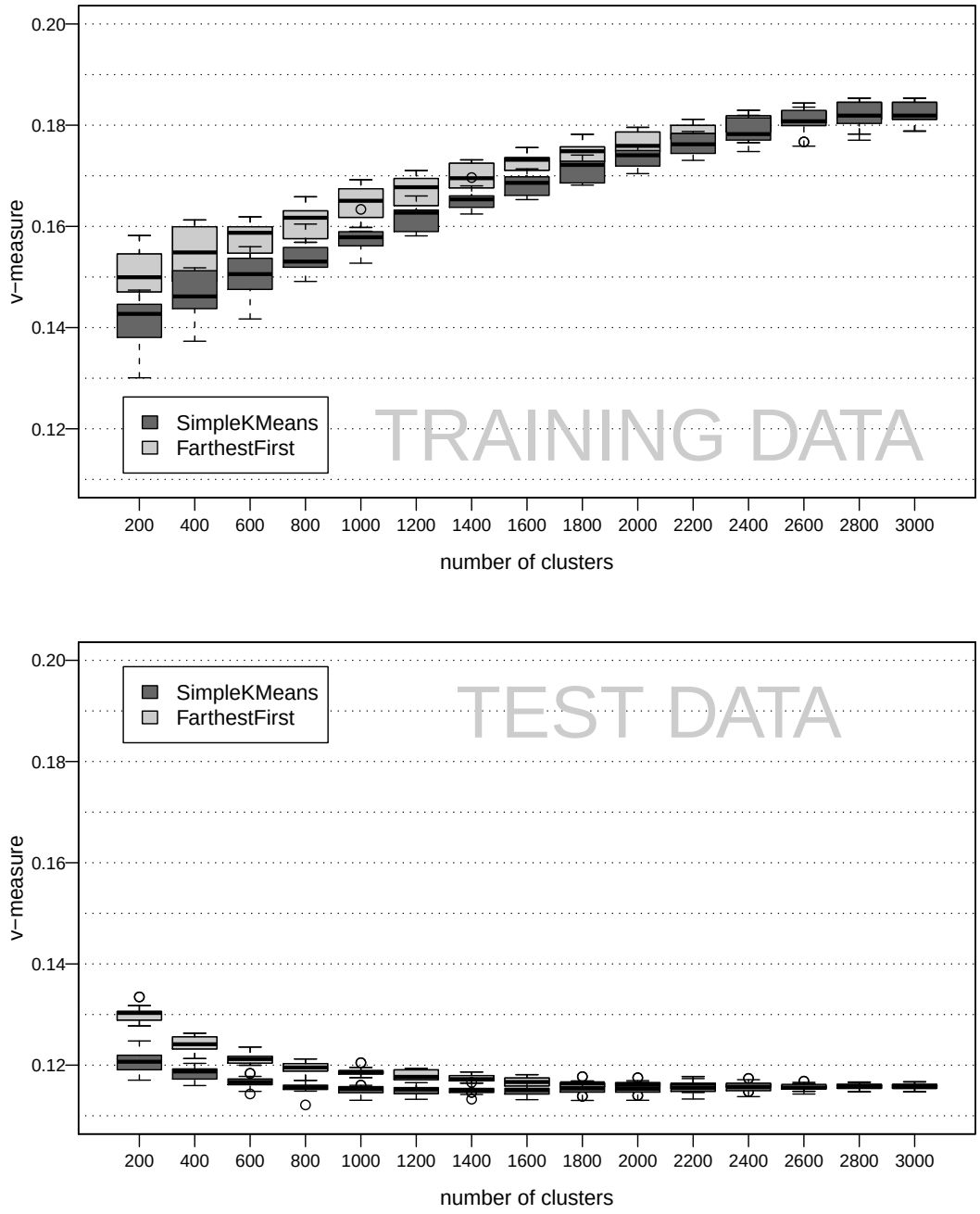

Figure 6.8: 10:90 experiment. Cross-validated v-measure values obtained by 10-fold validation and cross validation with 10% training and 90% test data. The results on the training data (upper panel) resemble those on the test data in the 3000X experiment; results on the test data (lower panel) resemble those on the training data in the 3000X experiment.

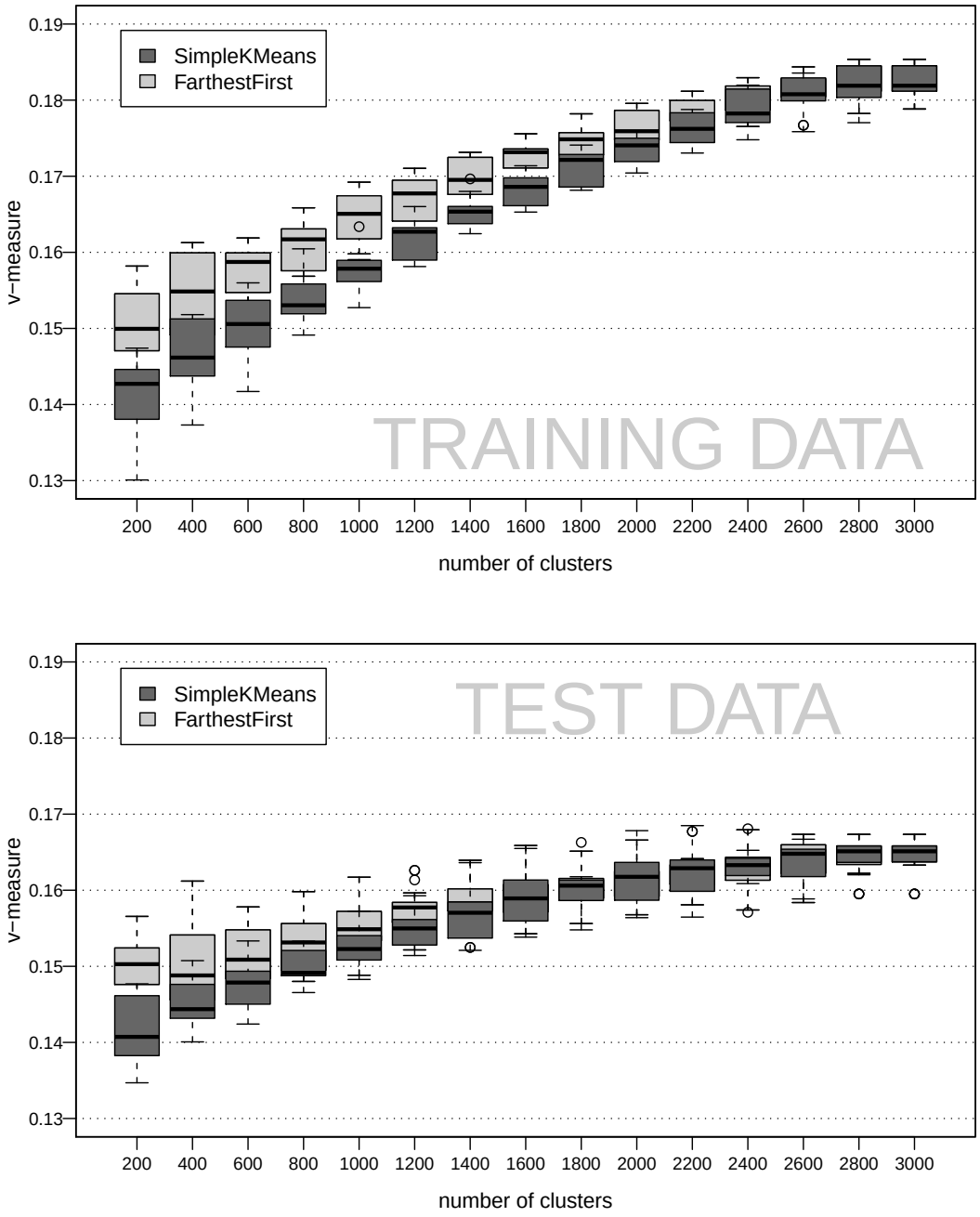

Figure 6.9: 10:10 experiment. Cross-validated v-measure obtained by 10-fold validation and cross validation with 10% training and 10% test data. Results on the training data (upper panel) and results on the test data (lower panel) indicate that higher numbers of clusters are better.

panel) yields results identical to the 10:90 experiment, since the training data were identical. As stated before, the scores increase with the numbers of clusters. Looking a little closer, the best score of slightly more than 0.18 on average is obtained at 2800 clusters, and there is no further increase for 3000 clusters. When evaluating the clusterings on the test data (lower panel), the results for 200 clusters are only very slightly lower, indicating that the clusterings generalize very well. Also, now the same trend as for the training data can be observed: the scores increase when increasing the number of clusters. However, the increase is smaller than on the training data, reaching an optimal value of about 0.165 on average for 2800 clusters. Again, there is no further increase for 3000 clusters. This indicates that the clusterings are not over-adapted to the training data for numbers of clusters up to 2800, but that up to these numbers of clusters, the clusterings have captured structure in the data that generalizes to new data. However, the degree of generalization decreases with the number of clusters.

These results obtained in the 10:10 setting indicate that up to 2800 clusters are optimal for both FarthestFirst and SimpleKMeans. The results in the 10:90 setting (cf. figure 6.8) suggest that 200 clusters is optimal for both algorithms. However, in both cases, training data and obtained clusterings are identical. If one decides not to consider the results on the test data and to evaluate just on the training data instead, the 10:90 experiment, with less training data, indicates that high numbers of clusters of up to 2800 are optimal, while the 90:10 experiment, with much more training data, suggests that 200 clusters are optimal. It is puzzling that for small amounts of training data, the best scores are obtained for high numbers of clusters, while for large amounts of training data, the fewer clusters yield the best scores.

Thus, the results obtained using v-measure for evaluation are inconclusive. If my claim is right and it is not viable to compare v-scores obtained for different amounts of data to each other, one can only determine the optimal number of clusters depending on the amount of data. Unfortunately, then, it is not possible to determine the optimal number of clusters for the large amount of training data here, because of the lack of the same amount of test data. However, it may well be that the small amount of training data is not sufficient to capture the general structure of prosodic categories.

To assess whether the amount of training data in the 10:10 setting was sufficient to capture the structure of prosodic categories, I have returned to the classification accuracy measure. This serves to get an idea of how well clusterings obtained on a small training set could perform in categorization. I have calculated classification accuracy rates for the 10:10 setting as well as for the 10:90 setting. The results for evaluating on the training data, which are identical in both settings, are depicted in figure 6.10. It can be seen that the accuracies increase almost linearly with increasing numbers of clusters and reach 100% for 2800 and 3000 clusters, which is suspiciously perfect and must

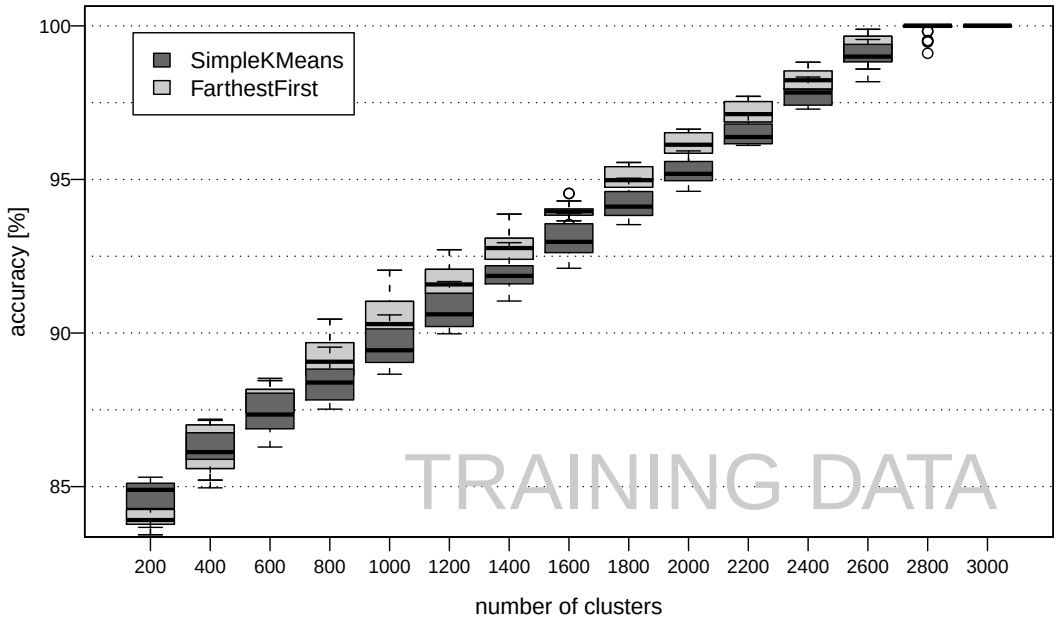

Figure 6.10: 10:10 experiment. Classfication accuracies obtained by 10-fold validation using 200 to 3000 clusters, on training data corresponding to about 10% of the original data. For higher numbers of clusters, accuracies of 100% or close to 100% are reached, which very likely indicates over-adaptation to the data.

certainly be caused by over-adaptation to the data.

Indeed, the classification accuracies for the test data in the 10:10 setting, which are given in figure 6.11, are much less perfect. Results for the two algorithms are given in two separate panels because there was too much overlap. The results for SimpleKMeans (upper panel) indicate that 400 or possibly 600 clusters are optimal in terms of classification accuracy, while for FarthestFirst (lower panel), 600, or possibly 1000 clusters are best. Comparing the two algorithms, SimpleKMeans is better than FarthestFirst. However, as can be seen, the results are quite variable, which makes it hard to tell which number of clusters is definitely better than the rest. This variability can be reduced by using more data for testing: for the 10:90 setting, compatible, but less variable results are obtained. These are presented in figure 6.12. Here, it can be said with more confidence that 600 clusters is optimal for both algorithms, and that the accuracies in this case are slightly above 83% for SimpleKMeans and slightly below 83% for FarthestFirst.

Recall from the beginning of this section that the accuracies obtained on the test data when training on the full training set, without cross-validation, were up to 84% and 83% for SimpleKMeans and FarthestFirst, respectively (cf. figure 6.3). Given the variability observed in evaluating the 10-fold clusterings, it is

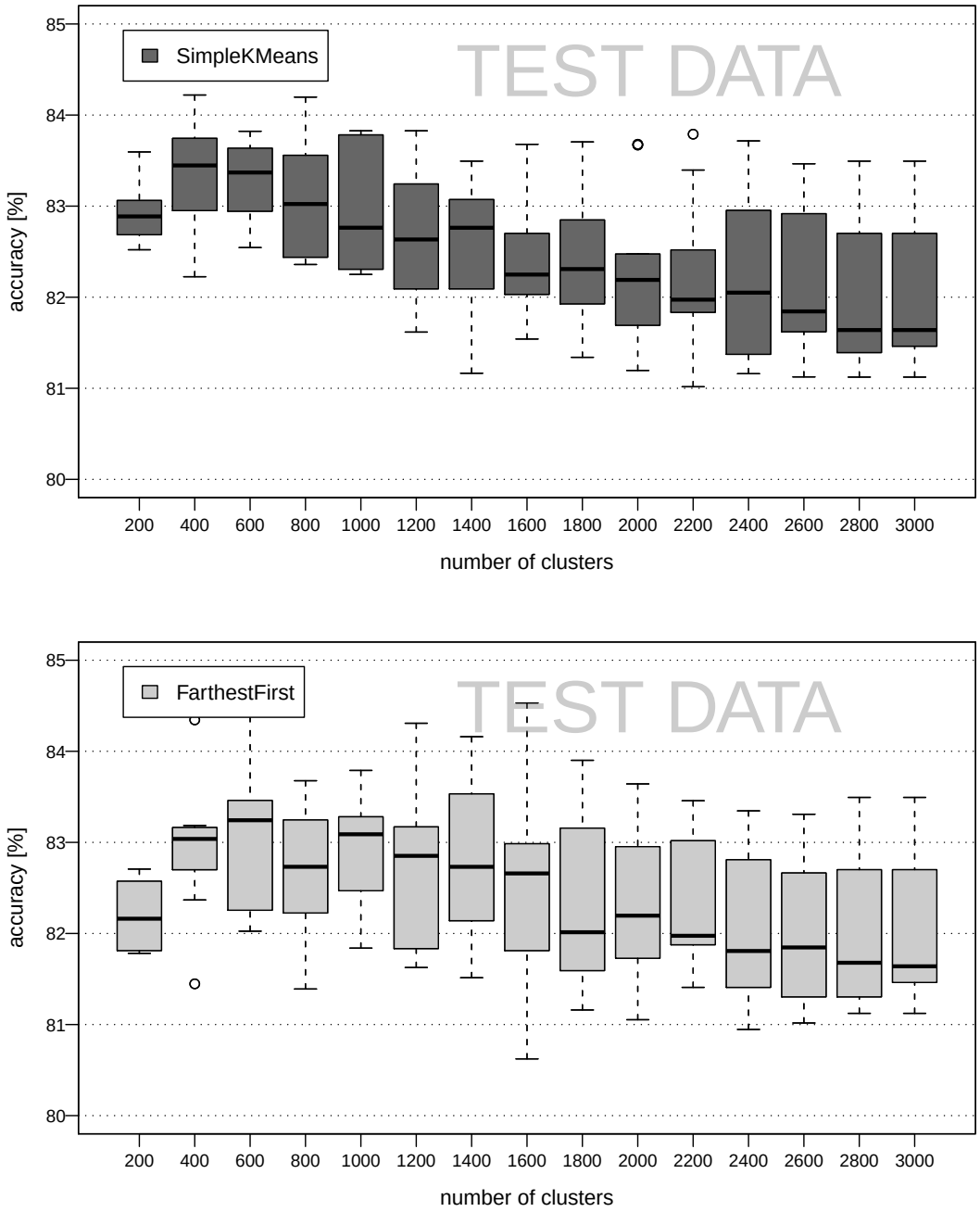

Figure 6.11: 10:10 setting. Classfication accuracies obtained on the test data by 10-fold cross-validation using 200 to 3000 clusters, for SimpleKMeans clustering (upper panel) and FarthestFirst clustering (lower panel). Comparable amounts of training and test data were used, each corresponding to about 10% of the original data. Best results are obtained for 400-600 clusters in case of SimpleKMeans, and at 600 or 1000 clusters in case of FarthestFirst.

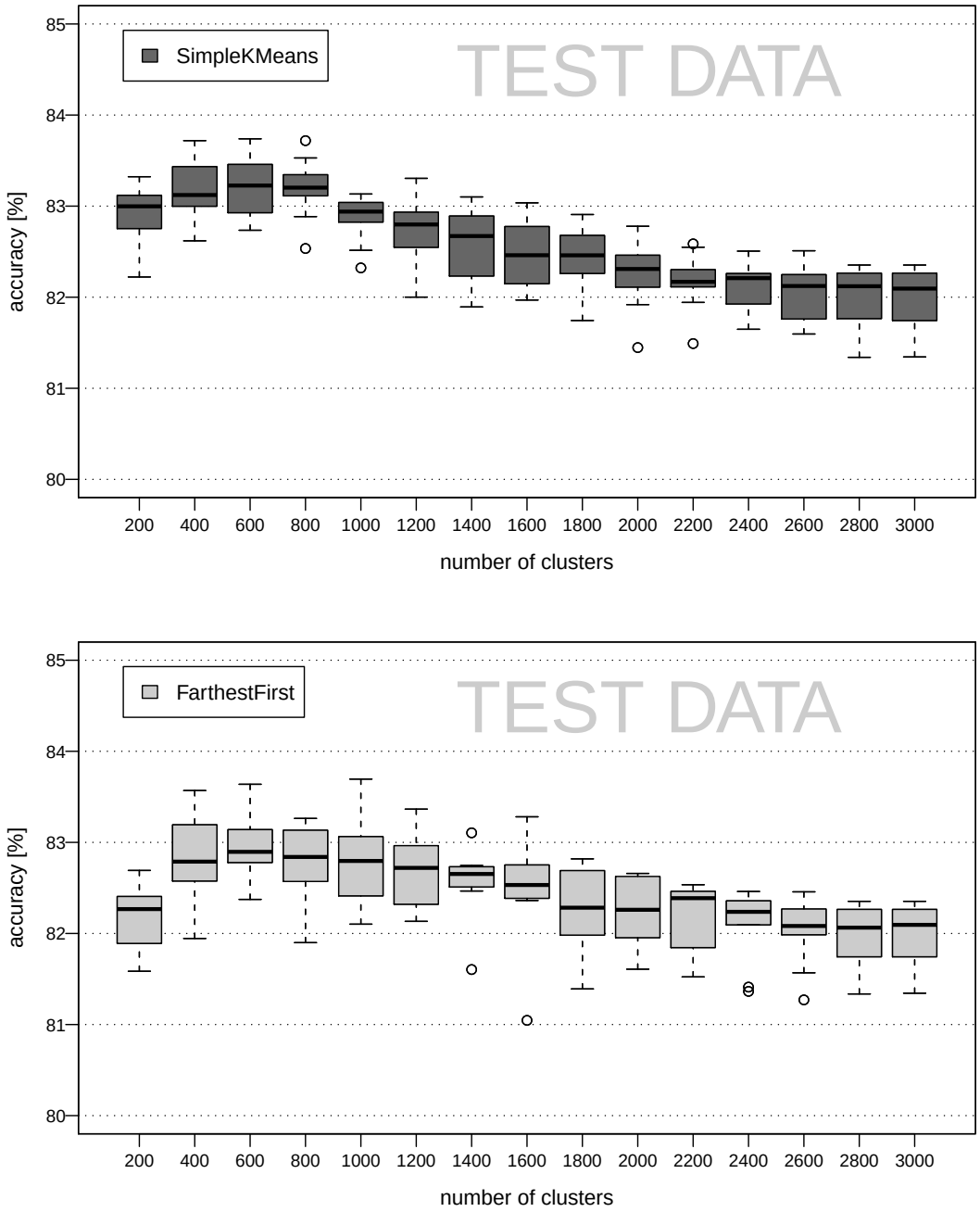

Figure 6.12: 10:90 setting. Classfication accuracies obtained on the test data by 10-fold cross-validation using 200 to 3000 clusters, for SimpleKMeans clustering (upper panel) and FarthestFirst clustering (lower panel). About 10% of the original data were used for training, and about 90% for testing. The results on the greater amount are compatible with the earlier results, but clearer due to reduced variability: 600 clusters are optimal for both algorithms.

Figure 6.13: 90:10 setting. Classfication accuracies obtained by 10-fold validation using 200 to 3000 clusters, on training data corresponding to about 90% of the original data. The accuracies increase with the number of clusters.

hard to tell whether the classification accuracies obtained here on the small training set are actually in the same range as the accuracies expected when training on the full set.

Since it is unproblematic to use different amounts of test and training data when using classification accuracy as an evaluation criterion, except for the greater variability when using smaller amounts of test data, I have calculated classification accuracy rates for the 3000X experiment. Recall that the training-test split in this case was 90:10. The results on the training data are presented in figure 6.13. In contrast to the 10:10 and 10:90 settings, the classification accuracies on the training data do not reach 100%. However, they increase continuously for higher numbers of clusters. The increase is not linear but diminishes with increasing numbers of clusters.

The results on the test data are given in figure 6.14. Again, the results for SimpleKMeans and FarthestFirst are given in two panels because they overlap. Both algorithms reach accuracy rates of more than 85%,[25] with slightly better scores for SimpleKMeans. Also, SimpleKMeans reaches these values for fewer numbers of clusters, viz. for numbers of clusters around 1600. The best values for FarthestFirst are obtained for around 2000 clusters.

[25] i.e., the median for the 10 folds is slightly higher than 85%

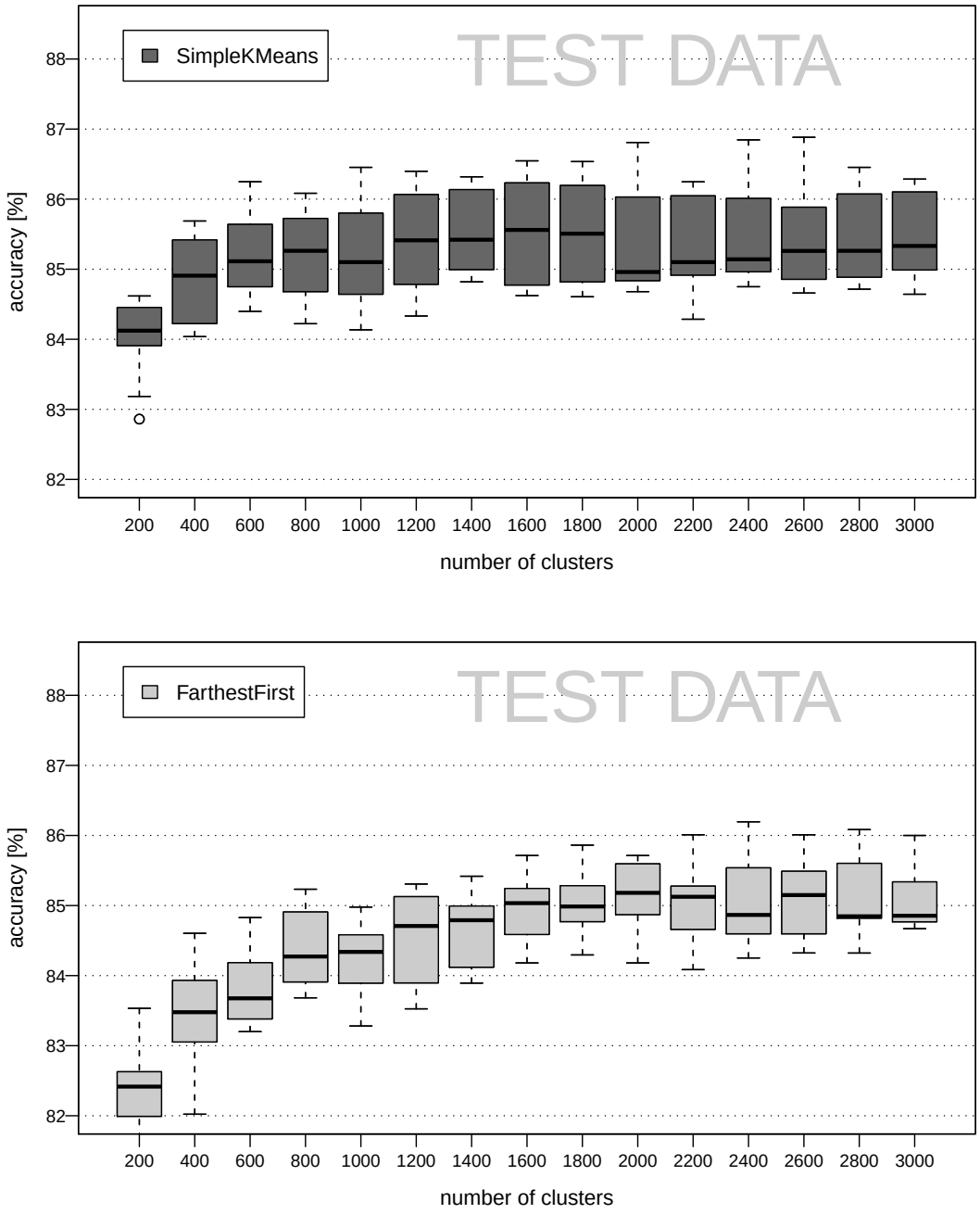

Figure 6.14: 90:10 setting. Classfication accuracies on the test data by 10-fold cross-validation using 200 to 3000 clusters, for SimpleKMeans clustering (upper panel) and FarthestFirst clustering (lower panel). About 90% of the original data were used for training, and about 10% for testing. Best results are obtained at 1600 clusters for SimpleKMeans, and at 2000 clusters for FarthestFirst.

Summarizing the results of this section on cross-validation experiments, according to the classification accuracy scores, SimpleKMeans is slightly better than FarthestFirst. Also, for the amount of training data available here, around 1600 clusters are optimal for SimpleKMeans, and around 2000 clusters are optimal for FarthestFirst. Using these settings, classification accuracy rates of around 85% can be reached for pitch accent clustering. Using a small amount of training data, viz. only 10% of the full data set instead of 90%, the classification accuracy rates are lower, but not dramatically so, reaching approximately 83%.

Using v-measure for evaluating the clusterings, the results are less homogenous. They indicate that for the larger amount of training data, around 200 clusters are optimal, while for the small amount of training data, up to 3000 clusters are best. Evaluating the clusterings using independent training data is problematic because the amount of test data should match the amount of training data. Thus, an evaluation on independent test data is not possible for the large amount of training data; however, for the small training set, evaluating on independent test data confirms that up to 3000 clusters are optimal.

6.2.6 Discussion and Outlook

In this section, I have investigated clustering as a means to automatically detect prosodic categories in perceptual space. Given that all relevant perceptual dimensions are known, it should be straightforward to detect clouds corresponding to phonetic categories using clustering techniques. In this vein, Pierrehumbert (2003) reviews clustering results obtained by Kornai (1998) where clusters of F1/F2 data corresponded well to vowel categories. She posits that stable categories are characterized by "well-defined clusters or peaks in phonetic space" (Pierrehumbert 2003, p. 210).

The experimental results presented in this section show that it is possible to identify clusters which correspond well to prosodic categories. Usually in clustering, the goodness of this correspondence is determined by an external evaluation measure which relates clusters to a gold standard, viz. the clusters to pitch accent classes in case of the experiments here. However, applying the v-measure (Rosenberg and Hirschberg 2007) as a standard measure was problematic because of its sensitivity to the amount of data, as elaborated in section 6.2.5. Instead, I have suggested to evaluate clusterings on independent test data using the classification accuracy measure, which models exemplar-theoretic categorization: it assumes that all instances in a cluster belong to the same pitch accent category, viz. the category with the highest likelihood in that cluster. This way, each cluster will be assigned a pitch accent category, and new instances can be categorized by determining to which cluster they belong.

Two clustering algorithms, namely SimpleKMeans and FarthestFirst, per-

form similarly well with respect to this measure, reaching classification accuracies of slightly more than 85% on independent test data. This is clearly above the baseline of around 78%. I will show in the following section using the same data as in this section that prediction accuracies of approximately 87.5% can be reached when classifiers are specifically trained to predict pitch accents. Given that in the clustering experiments presented here, the intention was to model human categorization by using classification accuracy, rather than finding the optimal classifier for prediction in general, the 85% classification accuracy obtained here compares well to the overall optimum of 87.5%.

As for the number of clusters, these scores are reached for approximately 1600 clusters in case of SimpleKMeans, and for approximately 2000 clusters in case of FarthestFirst, suggesting that these are appropriate numbers of clusters. Such high numbers of clusters may be unexpected at first. However, a one-to-one correspondence of clusters to phonetic categories cannot be expected. For instance, as Pierrehumbert (2003) points out, there may be overlap between distributions for different phonemes in different contexts, such as the overlap of devoiced /z/ realizations with "real" /s/ realizations. She states:

> Cases such as these mean that parametric distributions for phonemes are not always well distinguished from each other, if they are tabulated without regard to context. Within each given context, the distributions are much better distinguished. Thus, positional allophones appear to be a more viable level of abstraction for the phonetic encoding system than phonemes in the classic sense. (Pierrehumbert 2003, p. 211).

There are two solutions to this problem. Solution one is to cluster data from comparable contexts only (i.e., tabulated with regard to context, in Pierrehumbert's (2003) terms). Then, the clusters should be sufficiently well-defined to be detected automatically. However, this implies that the contexts that are supposed to be relevant have to be known beforehand. Solution two is to include contextual information as additional dimensions.[26] For instance, if voicing of subsequent phonemes is included as a dimension, this will help to separate /z/ realizations in devoicing contexts from /s/ realizations. By way of example from the prosodic data used here, it has been discussed in chapter 5 that the PaIntE parameter distributions for different prosodic events overlap even though they are often significantly different from one another. For instance, in figure 5.4 on page 81 the distribution of the *b* parameter (which quantifies F0 peak alignment) for L*H accents is indicated by the green solid line. The distribution is bimodal, indicating that the F0 peak is located either in the accented

[26]To some extent, I have done this here, since I have used more attributes than the perceptually motivated ones, even though doing so was motivated by the better results rather than by this theoretical consideration.

syllable (the first peak of the distribution) or on the next syllable (the second peak in the distribution). The distribution overlaps with the distribution for L*HL accents: if for L*H accents, the F0 peak is indeed on the accented syllable, it is located at the same point where it is usually located for L*HL accents (the peak of the L*HL distribution is at the same point as the first peak of the L*H distribution). As discussed in section 5.1.1, the bimodal distribution arises because L*H accents in word-final syllables have their peak at the earlier point, and word-internal L*H accents have their peak at the later point (cf. figure 5.5)—one could say that there are two positional variants of L*H. Including the position of the syllable in the word as a further dimension would help to keep the two distributions separate, hopefully resulting in two clusters for these two positional variants of L*H.

The attributes that I have used for clustering are described in section 6.1.1. There are altogether 33 attributes, of which 29 are used for pitch accent clustering here. Thus, the clusters are detected in a 29-dimensional space. If each dimension served to split each lower-dimensional cluster in two, one would get 2^{29}, i.e., more than 500 million, clusters. This is of course grossly exaggerated and unrealistic, not least because training and test data together contain only 28,000 syllables, which entails that it is simply impossible to obtain more than 28,000 clusters. But above all, one would not expect that each dimension is relevant for each cluster. For instance, the position of the syllable in the word was only relevant for L*H accents, thus this dimension should only serve to split L*H clusters in two and thus just add one cluster instead of doubling the number of detected clusters. Still, relative to the high dimensionality of the clustering space, 1600 to 2000 clusters seems more appropriate than it seemed at first.

The experiments presented in this chapter are a first step towards detecting prosodic categories using the parameters suggested above. There are many questions left open. Maybe the most immediate one of these is which of the dimensions really contribute to detecting the clusters. The three algorithms which were systematically used in this section are SimpleKMeans, FarthestFirst, and EM clustering. The former two use the Euclidean distance to find clusters of instances. However, all dimensions contribute to this distance with equal weight. Thus, adding dimensions which do not serve to distinguish clusters may even have the opposite effect in that it adds to the Euclidean distance, weighting down the contribution of the other dimensions. This could obscure meaningful clusters in the other dimensions. The EM algorithm on the other hand represents clusters as probability distributions over the attribute space. These distributions are multivariate Gaussians, thus the clusters are characterized by their means and variances along each dimension. Theoretically, EM clustering should be better suited for data sets which include irrelevant dimensions, because dimensions which are not important for a cluster could be characterized by high variance in that dimension, covering all possible values. In

practice, however, it seems that these parameters are not correctly detected, as attested by the consistently lower classification accuracy scores observed for EM clustering as compared to SimpleKMeans and FarthestFirst.[27] Thus, irrelevant dimensions are expected to be problematic, and possibly more problematic for EM clustering than for the other two algorithms. Identifying which dimensions are irrelevant can be addressed using attribute selection methods. This is an issue which is also relevant for the classification experiments described in the following section. I will therefore defer it to the general discussion in chapter 7.

A related issue is whether values in more important dimensions should be scaled to have greater impact in determining the similarity of instances. This idea is not implausible; in the exemplar-theoretic framework, its effect can be interpreted as increasing the attentional weights to certain aspects of the exemplars. However, this would require that the relevance of each dimension must be quantified beforehand, possibly using attribute selection as suggested above.

A third question is how the clustering results could improve with cleaner data. As mentioned in section 6.1.3, the data are noisy, and some improvements in the PaIntE approximation process envisioned for the future, in particular an improved smoothing algorithm, are expected to reduce this noise. I assume that this will also have a positive effect on the clustering results.

Not an open question, but a point worth mentioning, concerns the distribution of pitch accent categories in the data. As stated before, it is heavily skewed, especially so because of the overwhelming dominance of unaccented syllables. Such skewed distributions are often seen as disadvantageous for machine learning schemes, and a common solution to this problem is to select only a subset of the data with a more homogenous distribution. In this vein, Rosenberg and Hirschberg (2007) have taken only a subset of the most frequent accents in their data for clustering. However, not taking this option here was deliberate because the experiments aim at modeling human categorization, thus the data should reflect "real-life" distributions rather than distributions that are optimal for clustering. After all, humans manage to detect pitch accent categories despite the dominance of unaccented syllables.

Concluding this section on clustering pitch accents, the results in terms of classification accuracy are promising. For SimpleKMeans and 1600 clusters, cross-validated classification accuracies of between 84.6% and 86.5% are reached,[28] or 85.5% on average. These are clearly above the baseline of between 76.9% and 78.4%, or 77.7% on average. The baseline is the classification accuracy which would be obtained by predicting all instances to be "un-

[27] As mentioned above, I have used the default parameters for all algorithms tested here, except for the number of clusters, which was varied systematically. In case of EM clustering, there is a parameter restricting the number of iterations, and it defaults to 100 iterations. It is possible that the EM results could be improved when allowing more iterations.

[28] Classfication accuracies and baselines vary for the ten folds.

accented". Furthermore, the accuracies compare well to those achieved in the following section when training various classifiers to predict pitch accent categories; for the best classifier, the averaged cross-validated prediction accuracy was 87.5% (cf. section 6.3.5), i.e., the best result was only 2% above the averaged accuracy obtained in SimpleKMeans clustering. In my opinion, these results are quite satisfactory, given that training classifiers for accent prediction is a supervised process, i.e. the correct labels are provided to guide construction of the classifiers, while in the cluster-based prediction discussed here, the labels are only used to determine the clusters-to-classes correspondence after the clusters have been detected.

The aim of clustering the data was to detect prosodic categories. This ambition is dropped in training classifiers for prediction—even though the resulting classifiers can be interpreted as partitioning the data space in some way, these partitions do not necessarily correspond to accumulations of instances, or to regions with higher density of instances. The following section will reveal how much can be gained in terms of accuracy in this case.

6.3 Prediction of prosodic events

The aim of this section is to model human categorization of prosodic events using machine learning algorithms for classification.[29] I have built classifiers for both syllable-based accent and boundary prediction using the attributes discussed above (cf. section 6.1.1). These attributes have been used for clustering in the preceding section as well. The main interest in the prediction experiments is in how well automatically built classifiers can possibly perform this task, particularly compared to human classification. The underlying idea is that if the classifiers perform well, this corroborates my claim that the attributes, which include PaIntE parameters and duration z-scores, capture the most important perceptual aspects of prosodic events. Performance of the classifiers is measured in terms of prediction accuracy. As in the previous section, the accuracy rate is the proportion of instances which are correctly classified by a classifier. Thus, beyond evaluating the obtained classifiers, the results of the prediction experiments can be used to assess the performance of the clustering experiments presented in the previous section. A second, related question is whether any learning scheme is more suitable for this problem than others, i.e., whether classifiers built using a particular scheme classify syllable instances with better accuracy than other learning schemes. Third, the resulting classifiers are to be evaluated on data of another speaker in order to assess how well speaker-specific classifiers generalize to other speakers. Generalizability would

[29] The experiments discussed in this section have been published in Schweitzer and Möbius (2009).

not only confirm that the classifiers capture aspects of prosodic events that are relevant in perception in general, it would also allow the use of the classifiers for automatic prosodic annotation.

This section is organized as follows: I will sketch the procedure in section 6.3.1 and the evaluation method in section 6.3.2. The evaluation results are presented in section 6.3.3. Section 6.3.4 will address how the resulting classifiers generalize to data from another speaker. The classifiers will be compared to results of other studies and to human labeling in sections 6.3.5 and 6.3.6, and finally their application to automatic prosodic labeling is illustrated on examples from the database in section 6.3.7. I will conclude this section on the prediction of prosodic events by a discussion including an outlook on future work in section 6.3.8.

6.3.1 Procedure

Using the training part of the SWMS database as training data (cf. section 6.1.2), I have applied various machine learning schemes implemented in the WEKA software (Witten and Frank 2005) to build classifiers for both syllable-based accent and boundary prediction, i.e. to build classifiers that decide on the value of the **accent** or **boundary tone** attribute of a syllable instance based on the values observed for the remaining attributes (cf. section 6.1.1).

In order to compare the results of the experiments presented here to results of studies which just predict two classes of accent (no accent vs. accented) and two classes of boundaries (boundary vs. no boundary), classifiers for these two-class problems were trained in addition to the classifiers predicting the full set of pitch accents and boundaries.[30] Except for one case (for IBk instance-based learning), the default parameters suggested by WEKA were used in building the classifiers.[31]

For the experiments, I used the training set of the SWMS data. WEKA provides a more sophisticated 10-fold stratified cross-validation for evaluating its classifiers than the one I have used in clustering in the previous section: the data in each run is split into 10 folds which each contain approximately equal

[30] Even though the aim was to build classifiers that predict the full set of accents, many resulting classifiers effectively predict just the most frequent classes: obviously, the less frequent classes cannot be predicted with enough confidence to make up for their lower a-priori probabilities. Still, in evaluating these classifiers, classification results are compared to the full set of manual labels.

[31] The default settings of the IBk learning scheme implemented in WEKA set the k parameter to 1. The k parameter determines the number of neighbors considered in classification of new instances, and with k=1 this learning scheme is identical to the separate IB1 learning scheme in WEKA. Therefore, k=30 was used instead. Also, per default attribute values are normalized with the IBk learning scheme. Since the z-scoring of the PaIntE and duration parameters already introduces normalization, I have run the IBk scheme once using normalization and once without any normalization.

proportions of instances of all classes. In each run, 10 classifiers are built using the data of all but one fold, and each classifier is evaluated on the data of the remaining fold. The split into folds is randomized which means that it is different for different runs, reducing the risk of incidentally choosing a split that is advantageous for the evaluation. However, WEKA has no access to the genre that a particular syllable belongs to nor to the order of recording (cf. 6.1.2), which means that the resulting split may be unbalanced in that respect. Also, originally adjacent syllables may end up in test and training set. Nevertheless, in order to have objective estimates of the accuracy of the resulting classifiers, it was better to have WEKA execute several runs using different splits. By evaluating the cross-validated classifiers once more afterwards on the test data, it can be verified that the prediction accuracies reached on the randomized splits is comparable to that reached on the test set, which confirms that the randomized splits do not spuriously improve prediction accuracies.

The SWMS test set had to be held out from the experiments completely because its utterances are identical to the test set of the SWRK data, on which the classifiers were to be evaluated later when assessing their generalizability. Using utterances from this set in building the classifiers would have violated the assumption of independence of the test set.

6.3.2 Syllable-based vs. word-based evaluation

When evaluating the classifiers, one question that arises is whether the evaluation should be syllable-based or word-based. Looking at the data, which is composed of instances of syllables, the most straightforward way is to calculate the prediction accuracy syllable by syllable, as in evaluating the clustering results in the preceding section. However, as described in section 6.1.1 above, two attributes that are available in classification are the attributes **stress**, which indicates whether the syllable is lexically stressed, and **wordfin**, which indicates whether the syllable is word-final. It is theoretically impossible for an unstressed syllable to carry a pitch accent—theory claims that pitch accents are realized on lexically stressed syllables. The only exceptions in our data are (i) the very rare trail tones ..H and ..L, which are not pitch accents strictly speaking, but which count as accents here because they are possible values of the **accent** attribute, and (ii) syllables that have not been correctly labeled for word stress. Thus, the simple rule that syllables with values of 0 for the **stress** attribute should be classified as unaccented is expected to be captured by most classifiers, and successful application of this rule will increase accuracy rates. The same can be said of the **wordfin** attribute: phrase boundaries in fluent speech without hesitations do not occur word-internally, and thus, the classifiers are expected to classify syllables with values of 0 for the **wordfin** attribute as not phrase-final and thus belonging to class NONE. Again, accuracy rates should

benefit from this simple rule.

This can be confirmed by running WEKA's attribute selection schemes: *wordfin* is ranked among the 7 most predictive attributes when predicting boundaries, and so is *stress* when predicting accents. The latter is even ranked first by several selection methods.[32] Thus it is undeniable that the two attributes *wordfin* and *stress* contribute to the accuracy of the resulting classifiers. However, they are less interesting from a perceptual point of view in that they are determined by the linguistic context and reflect its lexical properties rather than perceptually relevant acoustic aspects of the particular rendition of this context.

In order to assess the performance of the classifiers beyond these two simple rules, I have trained classifiers on data from which, for boundary classification, I have eliminated all word-internal syllable instances, and, for accent classification, I have eliminated all syllable instances which were lexically unstressed. Since in the annotation of the database, only one syllable per word is marked as stressed, this leaves only one syllable per word—the final syllable for boundary classification, and the stressed syllable for accent classification. Thus, accuracy rates obtained for classifiers on the original data can be interpreted as syllable-based accuracies, and accuracy rates obtained on the transformed data sets can be interpreted as word-based accuracies. Accordingly, the underlying data sets are termed "syllable-based" data set and "word-based" data set in the following. This also allows for comparing the present results to many other studies which only report word-based measures of performance.

6.3.3 Results

All classification algorithms for which I am reporting results in this section have been evaluated in three runs using 10-fold cross-validation in each run. Thus, the accuracy rates correspond to averaged accuracy rates of 30 different classifiers built from various splits of the SWMS training data. Separate classifiers were trained for classifying accents and boundaries, and for syllable-based and word-based data sets, thus, for each algorithm, 120 classifiers were built altogether.

Figure 6.15 is intended to give an impression of the performance of all the learning schemes that I have experimented with. It presents the word-based accuracy rates of all classification schemes that are implemented in WEKA (version 3.4) and that were applicable to the present classification problem (to be more exact, it presents the averaged accuracy rates of the classifiers built in the

[32] In assessing the contribution of an attribute or a subset of attributes, I have applied several attribute selection methods implemented in WEKA including *CfsSubsetEval*, which selects attributes based on their predictive ability taking into account possible redundancies between attributes, *InfoGainAttributeEval*, which measures the attributes' information gain with respect to the class, and *ChiSquaredAttributeEval*, which assesses the attributes based on their chi-squared statistic with respect to the class.

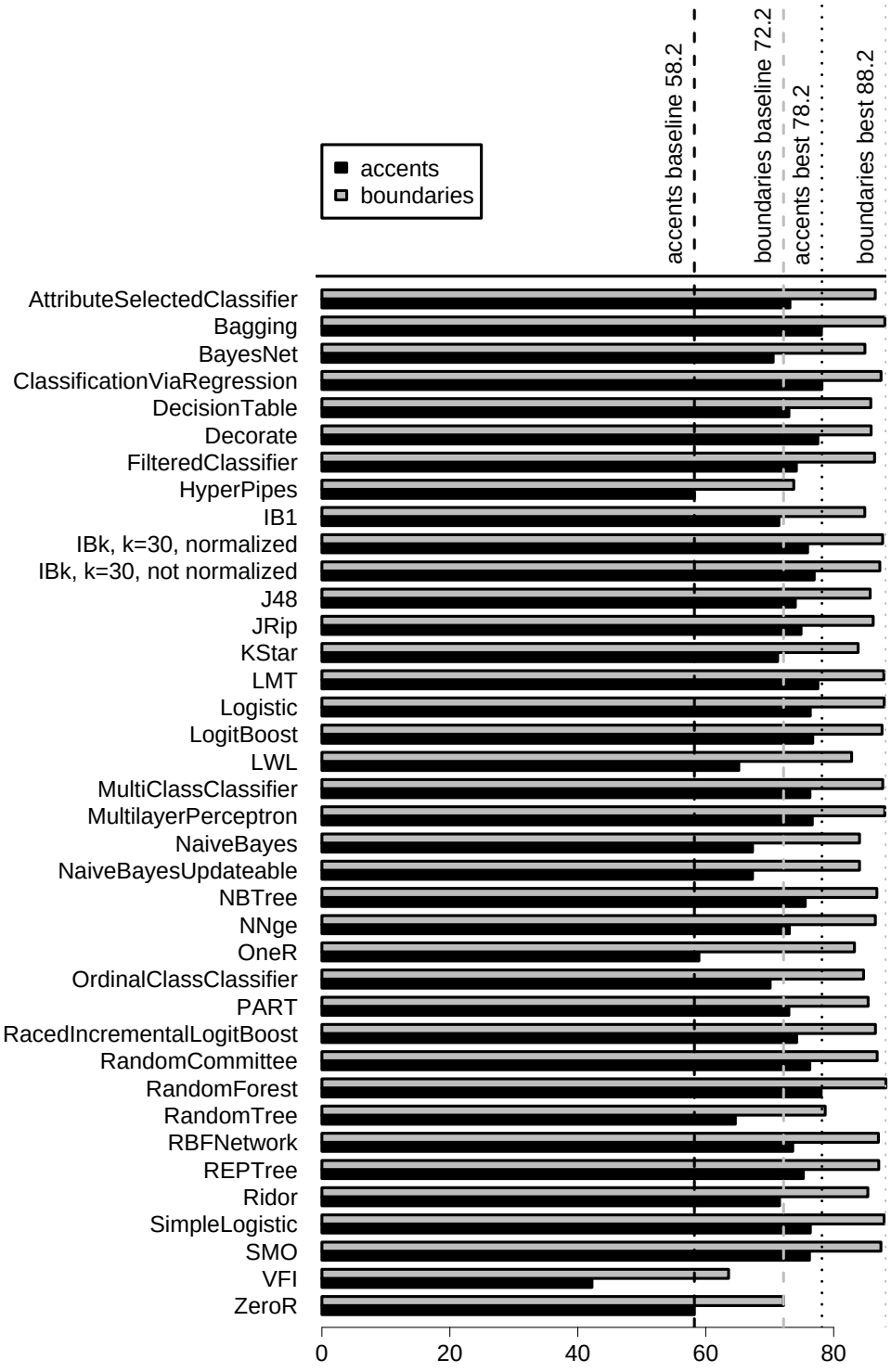

Figure 6.15: Overview of various machine learning algorithms implemented in WEKA and their word-based accuracy rates in accent classification (black bars) and boundary classification (gray bars). Several algorithms with accuracy rates equal to the baseline have been omitted. The rates indicated here pertain to the classifiers predicting the full set of prosodic event. See text for further details.

three runs mentioned above and thus gives an estimate of the accuracy rate that can be expected when building a classifier using this particular scheme). The accuracy rates of the classifiers for accent classification are indicated by black bars, those of the classifiers for boundary classification are indicated by gray bars. The rates pertain to the classifiers trained to predict the full set of accents or boundaries. The classifiers are listed in alphabetical order. To avoid confusion, I have kept the original names of the algorithms from the WEKA implementation, even if they are admittedly somewhat cryptic. I will discuss some of the schemes in more detail below; a short description of all schemes can be found in the WEKA book (Witten and Frank 2005).

In figure 6.15, the vertical dashed lines represent the baselines (black for accent classification, and gray for boundary classification). The baselines are determined as the word-based accuracy rates achieved by the ZeroR learning algorithm, which are indicated at the very bottom of 6.15. The ZeroR classifier does not infer any "rules", as indicated by its name, but just assigns all instances the most frequent class found in the training set and thus in a way can be interpreted as providing the chance level. On average, the most frequent accent was NONE with a proportion of approximately 58.2%, and the class NONE was also the most frequent boundary tone class with approximately 72.2%. Thus, on average, one can reach accuracy rates of 58.2% and 72.2% for accents and boundary tones, respectively, by just assigning each stressed syllable the class NONE in case of accents, and by assigning each word-final syllable the class NONE in case of boundaries. Several learning schemes achieved the same accuracy rates as the ZeroR classifiers, and these were left out in figure 6.15 in order to save space.[33]

The two dotted lines indicate the best results obtained in terms of word-based accuracy—the black dotted line shows the best word-based accuracy in accent classification, which was at 78.2%, and was obtained using the ClassificationViaRegression scheme; the gray dotted line indicates the best word-based accuracy for boundary tone classification, which was at 88.2%, and was obtained by the RandomForest scheme.

Figure 6.15 illustrates that the best results are not due to one or few outstanding learning algorithms that are particularly suitable for the present data; rather, when providing the information coded in the attributes discussed above, one can reliably reach quite high accuracy rates using any of a variety of different learning algorithms.

The exact word-based accuracy rates of the best learning algorithms from figure 6.15 are listed in table 6.8. Accuracy rates are given both for the two-class problem and for predicting the full set of GToBI(S) labels. To assess

[33]The following learning schemes were left out: AdaBoostM1, ConjunctiveRule, CVParameterSelection, DecisionStump, Grading, HyperPipes, MultiBoostAB, MultiScheme, OneR, Stacking, StackingC, Vote

*Table 6.8: Word-based accuracy rates for the best algorithms and the baseline (ZeroR). Most accuracy rates are comparable to the rates obtained by the RandomForest algorithm: only accuracy rates marked by * are statistically significantly worse than the corresponding rate for the RandomForest algorithm.*

Algorithm	accents		boundaries	
	2-class	full set	2-class	full set
Bagging	86.19	78.08	93.33	88.00
ClassificationViaRegression	85.49	78.17	*92.29	*87.41
LMT	86.24	77.54	93.37	87.84
RandomForest	86.17	78.04	93.31	88.16
ZeroR	*58.30	*58.23	*72.16	*72.16

whether any of the algorithms is better than the rest, the asterisks indicate which classifiers performed significantly worse than the RandomForest classifiers. The RandomForest classifiers were chosen as a reference here because they performed best for the boundary tones and were also among the best classifiers for accents. It can be seen that these classifiers perform similarly well. When predicting the full set of boundary tones, RandomForest classifiers with an accuracy of 88.16% were slightly ahead of the Bagging (88.00%) and LMT (87.84%) classifiers, while, when predicting the 2-class set, LMT performed best (93.37% vs. 93.33% for Bagging and 93.31% for RandomForest). However, all three performed similarly well, with no significant differences in accuracy rates between RandomForest classifiers and those obtained using the other two schemes. The ClassificationViaRegression classifiers, however, were slightly, but significantly, worse than the RandomForest ones. When predicting the full set of accents, the accuracy rates ranged between 78.17% (ClassificationViaRegression) and 77.54% (LMT), with no significant difference between RandomForest classifiers and those built using any of the other three learning algorithms. In predicting the two-class set of accents, the order was almost reversed, with LMT best at 86.24%, and Classification via Regression worst (among these top four) at 85.49%. Again, there were no significant differences between RandomForest classifiers and those trained using any of the other three learning algorithms.

It is not feasible in the scope of this thesis to investigate what exactly makes the four best schemes best for the data here. However, I will briefly discuss some thoughts on this in the following paragraphs. Two of these best four schemes above are actually very similar, viz. the Bagging and RandomForest schemes. The idea in bagging is to train an ensemble of several classifiers on random samples of the training data, and to let them "vote" for the predicted class, taking the majority vote as the prediction of the ensemble. The Random-

Forest classifier implemented in WEKA bags several RandomTree classifiers, i.e., it is actually an instance of bagging as well. In contrast, the Bagging classifier in WEKA by default bags REPTree classifiers. Thus what is referred to as Bagging in the tables and figures above corresponds to bagged REPTree classifiers, and what is referred to as RandomForest corresponds to bagged RandomTree classifiers. The WEKA book (Witten and Frank 2005, pp. 316–317) notes that inducing decision trees is an unstable process which is very susceptible to small changes in the training data, thus the single trees trained on different random samples of the data are expected to be very different. According to Witten and Frank (2005), this has the effect that the ensemble classifier is usually more accurate than a single one trained on the full training data. This is confirmed in case of Bagging and RandomForest here: both a single REPTree and a single RandomTree perform clearly worse for the current tasks, as can be seen in figure 6.15. Intuitively, the benefit in bagging is that each classifier in the ensemble will focus on different aspects of the data with more or less emphasis, due to differently sampled training data, resulting in several "specialist" classifiers instead of one that has to cover everything. In RandomForest classification, additional randomness is introduced in building the RandomTree classifiers which are to be combined; here, only a random subset of attributes is considered when splitting the tree, and no pruning is performed; in contrast, the REPTree classifiers which are combined in the default Bagging variant in WEKA are pruned, and all attributes are considered when building them. Despite the differences in building the classifiers, the difference in performance between Bagging and RandomForest was negligible in the experiments here.

LMT and ClassificationViaRegression are similar in that they both apply regression models to a classification problem. LMT, which is abbreviated for Logistic Model Trees, builds a model tree with additive linear logistic models at the leaves. Similar to bagging, the models at the leaves combine several logistic regression models to contribute in prediction. However, classification is not based on a majority vote; instead, the results of all models in the ensemble are added in calculating the final classification result. The contributing models are built iteratively, with each new model maximizing the likelihood of the training data given the ensemble classifier. ClassificationViaRegression transforms the classes into binary attributes and builds a model tree for each of them. However, these trees have linear regression models at their leaves instead of logistic ones.

It is not surprising that several of the best models here are ensembles of classifiers, since these are currently state-of-the-art and have often proven to be superior to traditional, single classifier models. There are several techniques to create ensemble classifiers, for instance bagging, boosting, and randomization. As stated in section 6.1.3, a characteristic of the data is that they are quite noisy. Dietterich (2000) claims that bagging schemes are particularly suitable for noisy data, in that bagging is superior to boosting and often to randomiza-

Table 6.9: Word-based accuracy rates for instance-based learning with and without normalization, using 30 neighbors compared to the best algorithms. Gray cells indicate which version of IBk turned out better for the particular problem.

Algorithm	accents		boundaries	
	2-class	full set	2-class	full set
IBk no norm	84.57	76.96	91.77	87.22
IBk norm	84.44	75.92	93.23	87.65
Bagging	86.19	78.08	93.33	88.00
ClassificationViaRegression	85.49	78.17	92.29	87.41
LMT	86.24	77.54	93.37	87.84
RandomForest	86.17	78.04	93.31	88.16

tion for such data. Obviously, this claim is supported here.

As for the good results of LMT, I suggest that LMT can be interpreted as partitioning the instance space into regions, as any decision tree does—leaves correspond to regions defined by the attribute values along the path from the root to the leave. However, classical decision trees then always predict the same class for all instances within the regions defined by the leave. LMT, on the other hand, models the class distribution in each of these regions separately, allowing for much finer distinctions.

A further learning scheme worth mentioning here is IBk, which is short for instance-based learning. Instance-based learning is a so-called "lazy learning" scheme because no model is induced in training; rather, the data itself represents the model. For prediction, new instances are compared to instances from the training data using a distance metric, and the predicted class is the class observed most often among the k closest neighbors of the new instance. I have experimented using 10, 20, and 30 neighbors for classification, with no dramatic differences in performance between the three. Table 6.9 lists the results obtained for IBk using 30 neighbors, in comparison to the results of the best algorithms, repeated from table 6.8. In WEKA's implementation of IBk, it is possible to specify whether the instances should be normalized or not. An interesting outcome of my experiments is that normalization proved to be beneficial in boundary classification, while using unnormalized data was better for pitch accent classification, as indicated by the gray cells in table 6.9. As can be seen, the best results for IBk come close to the overall best results, however, there is a small difference. IBk is particularly interesting from an exemplar-theoretic perspective because human categorization is claimed to work exactly that way (cf. section 2.3.3).

Summarizing this section, the classification results are very promising: when predicting the full set of accents or boundaries, prediction accuracies

of the best classifiers are around 78% for pitch accents, which is 20% above the baseline, and around 88% for boundaries, i.e., 16% above the baseline. For the two-class problem, the accuracies are even higher, reaching 86% accuracy for pitch accents (i.e., 28% above the baseline), and 93% for boundaries (i.e., 21% above the baseline). These best results are reached by several classifiers, with no significant differences between them. Also, many other learning schemes work almost as well, as illustrated by figure 6.15. Among these is the IBk learning scheme, which can be interpreted as modeling exemplar-theoretic categorization. The next step in assessing the usefulness of the classifiers is to see whether they generalize to data of other speakers. I will present first results on this in the following section.

6.3.4 Generalizability

As a first step towards assessing the generalizability of the classifiers to other data, the best classifiers have been applied to the test set of the (female) SWRK database. The results are comparable for boundaries: for instance, a RandomForest classifier trained on the full SWMS training data reaches accuracy rates of 88.9% on the SWMS test data and of 88.6% on the SWRK test data. Thus, applying classifiers built on the male data to the female data yields a drop in averaged word-based accuracy of only 0.3%.

For pitch accents, performance is lower on the SWRK data: for instance, a RandomForest classifier trained on the full SWMS training data reaches accuracy rates of 78.9% on the SWMS test data and of only 74.7% on the SWRK test data corresponding to a drop in word-based accuracy of 4.2%. However, when building the classifier directly on the SWRK training data, the same accuracy of 74.7% is reached on the SWRK test set. Thus, the RandomForest classifiers built from the male SWMS training data are just as good in classifying the SWRK test data as their SWRK counterparts. This demonstrates that the lower accuracy in applying the SWMS pitch accent classifier is not due to unsatisfactory generalizability but rather is inherent to the SWRK data. On the contrary, the SWMS classifiers perform just as well as the SWRK classifiers.

These results are very encouraging. Even though the two databases are very similar in that they were recorded from mostly identical text material, the two underlying voices are not particularly similar—after all, one is a male voice, and one a female voice. Most attributes that were used for classification were related to F0 and temporal aspects, and the few attributes that were text-based were just rather abstract attributes derived from the text, but not, for instance, word identities, which would have been much more specific of a genre. The next step, in any case, will be to apply the classifiers to data of other speakers, with different material, which will hopefully confirm these expectations.

6.3.5 Comparison with other studies

There are early studies on automatic classification of prosodic events (e.g. Wightman and Ostendorf 1994; Ross and Ostendorf 1996), which do not reach the accuracy rates of more recent studies. Among the recent studies, Sridhar et al. (2008) obtain word accuracy rates of 86.0% for the two-class problem of predicting pitch accents, and of 93.1% for predicting two classes of boundary tones, obtained on data from the Boston Radio News Corpus. These accuracies are slightly, but probably not significantly, lower than the best rates achieved in my experiments, which for the two-class problem were obtained by the LMT algorithm (86.24% and 93.37%, respectively). Similar results are reported by Hasegawa-Johnson et al. (2005), who obtain word-based accuracy rates of 84.2% and 93.0% for two-class pitch accent and boundary prediction on similar data. Comparing the results presented here to the results of these two studies is relatively unproblematic because the data are similar: the two classes of pitch accents and boundary tones are derived from ToBI labels, and the corpora both consist of news-style read speech by professional speakers. However, the corpora are from two different languages with different ToBI systems.

Turning to German data, Zeißler et al. (2006) report accuracy rates of 77.0% for (two-class) pitch accent detection and of 88.6% for (two-class) boundary detection for German data. This is significantly lower than the rates obtained here, but they classify spontaneous user interactions with a wizard-of-Oz system and thus their data is very different from the data used here.

Another study reporting German results, which also predicts the full set of (English or German) ToBI labels instead of just two classes, is Braunschweiler (2006). He reports syllable-based accuracy rates of 65% and 60% for German and American English pitch accent classification, respectively, and syllable-based accuracy rates of 71% and 68% for German and American English boundary tone classification. I have only provided word-based accuracies in table 6.8 above, because for the syllable-based data, I have not systematically evaluated all algorithms in several runs. However, taking the RandomForest algorithm, for instance, 10-fold cross-validation in one run yields estimated accuracy rates of 87.5% and 93.9% on the syllable level for pitch accent and boundary tone classification, respectively, using the full set in both cases. Thus, the present results are significantly better than the results reported by Braunschweiler (2006).

6.3.6 Comparison with human prosodic labeling

In order to assess the performance of the classifiers with regard to what could maximally be expected, it must be stated that manual prosodic labeling is notorious for its subjectivity. Even human prosodic labelers are not perfect: if several labelers label the same data prosodically, they usually do not perfectly

agree with each other.

Several studies (Pitrelli et al. 1994; Grice et al. 1996; Syrdal and McGory 2000) have assessed inter-labeler reliability in the ToBI labeling framework. Consistency between labelers in these studies is measured as the percentage of equal transcriber-word pairs, with no claim as for which of the transcriptions is correct, i.e., these studies do not compare labeling results to a gold standard, as I do in the present experiments, instead they assess in how many instances the labelers agreed. This makes it hard to relate the results to the current context, in which there are just two competing variants, viz. the gold standard and the automatically generated prosodic transcription. In case of just two transcribers, the percentage of equal transcriber-word pairs is equivalent to the percentage of words for which the prediction was correct. However, the above measure of inter-labeler reliability is intended to compare more than just two transcribers. For instance, Grice et al. (1996) compare results from 13 transcribers. If 12 of them agree, the percentage of equal transcriber-word pairs is only 84%, even though 12 out of 13 corresponds to a percentage of 92%. Thus, the inter-labeler agreement is a very stringent measure: if 92% of the labelers in a 13-labeler experiment agree on a word, this yields an inter-labeler agreement of only 84%. If 11 out of 13 agree, i.e., 85%, and the two others agree with each other, the inter-labeler agreement is only 73%.

For German GToBI(S), Grice et al. (1996) report inter-labeler consistencies of 70% for pitch accents, and of 86% for boundaries, and similar consistencies have been reported for English (Pitrelli et al. 1994; Syrdal and McGory 2000). In comparison, the best results were around 78% accuracy for pitch accents, and around 93% for boundaries. While these numbers look better at first glance, it should be kept in mind that a direct comparison is not valid because the inter-labeler consistencies depend on the number of participants. Even though the present evaluation can be interpreted as comparing two labelers, a human one and an automatic one, consistency experiments are intended to include more than just two participants, but including more participants will always decrease the consistency score.

6.3.7 Illustrating the results

I have implemented a prototype of an automatic prosodic labeling tool by combining the RandomForest classifiers obtained above with our German extension (IMS Festival 2010) of the Festival TTS system (Festival 2010). Festival can not only synthesize utterances, it also provides tools to build utterance structures automatically from speech label files, and these utterance structures can be read and used by Festival. The prototype prosodic labeler thus reads in utterances from the database, derives the attributes needed for predicting pitch accents and boundaries, and then lets WEKA predict them. The result is read

into Festival for further processing; for instance, to generate prosodic label files for the utterance.

In order to give an impression of the performance of the classifiers for accent and boundary prediction, I have used the prototype to generate prosodic label files for the first five utterances from the test data. The results are displayed in screenshots of F0 contours and automatically generated label files in figures 6.16 to 6.19. Each example is accompanied by a table listing correct and predicted prosodic events word by word. The classifiers were the word-based RandomForest classifiers trained to predict the full set of pitch accents and boundary tones. Even though the full set of boundaries was predicted, distinguishing between (H)% and H%, and between (L)% and L%, in generating the label files, (L)% boundaries were mapped to L% boundaries, and (H)% boundaries were mapped to H% boundaries, thus, manual label files and predicted label files in figures 6.16 to 6.19 differ in that respect. However, the unmapped predicted labels are listed in the corresponding tables.

There is one aspect in which the prototype's results differ from the results obtained directly from the classifiers: IMS German Festival assumes that monosyllabic function words are never stressed. However, word-based pitch accent prediction as described above is only carried out for the stressed syllable of each word; since there is no stressed syllable in monosyllabic function words, no prediction is generated for them. This is why in tables 6.10 to 6.14 monosyllabic function words are left out. Given that there is not even one monosyllabic function word that is pitch-accented in our data, this is a reasonable procedure. However, the syllable-based classifiers might have predicted pitch accents for these function words had they been stressed, and thus, the results presented here differ in this little detail from the results discussed above.

Results for the first utterance from the test data ("Das Zentrum blieb zumeist Cardoso vorbehalten", *The center was mostly left to Cardoso*) are presented in table 6.10, and a screenshot of the smoothed F0 contour and the corresponding label files is displayed in figure 6.16. The manually labeled accents and boundaries are indicated in the top tier, above the word tier. Predicted pitch accents and boundaries are in two separate label files, which are displayed in the two tiers below the word tier. The corresponding table 6.10 lists correct and predicted accents and boundaries, with prediction errors highlighted in gray. In this first example, there were only two incorrect predictions: the pitch accent on *rechts* was missed; also, the intermediate phrase boundary after *Zentrum* was not detected. It should be noted that the L*HL on *Zentrum* has been predicted correctly even though it is one of the less frequent accents. Thus, the example demonstrates that occasionally the less frequent accents can be correctly predicted as well.

Figure 6.17 displays the results for the second utterance from the test set, f011 ("Beide Teams waren sehr motiviert und ständig im Vorwärtsgang, es war viel Tempo in der Partie", *Both teams were very motivated and always pushing*

	accents		boundaries	
word	correct	predicted	correct	predicted
Das			NONE	NONE
Zentrum	L*HL	L*HL	(L)-	NONE
blieb	NONE	NONE	NONE	NONE
zumeist	NONE	NONE	NONE	NONE
Cardoso	H*L	H*L	NONE	NONE
vorbehalten	NONE	NONE	(L)%	(L)%
weil			NONE	NONE
Spörl	L*H	L*H	NONE	NONE
sich			NONE	NONE
deutlich	NONE	NONE	NONE	NONE
rechts	H*L	NONE	NONE	NONE
orientierte	NONE	NONE	(L)%	(L)%

Table 6.10: Correct and predicted pitch accents and boundaries in utterance f001, corresponding to the phrase "Das Zentrum blieb zumeist Cardoso vorbehalten" ("The center was mostly left to Cardoso"). For monosyllabic function words no accents were predicted because they contain no stressed syllable by default in IMS Festival. Incorrect predictions are highlighted in gray.

forwards, there was much tempo to the match). Correct and predicted prosodic events are indicated in table 6.11. There are several problems in this utterance, and it is certainly one of the more problematic cases from our database. Looking at the smoothed F0 contour in the middle of figure 6.17, the second word in the utterance, monosyllabic *Teams*, is characterized by a strong rise followed by a fall. It is classified as an H*L accent, which visually seems quite plausible: the syllable ends where the intermediate phrase boundary is labeled in the manual label tier. Indeed, the contour is lower at that point than at the beginning of the word. However, this is only because smoothing the F0 contour has had the undesired effect of smoothing a clear rise into a peak: originally, the contour rose throughout the word *Teams* to the voiceless /s/ at the end. On the next word, *waren*, the contour is much lower. This can be verified by looking at the raw F0 contour, which is displayed above the smoothed F0 contour. In smoothing, interpolation across the voiceless /s/, where no F0 frames were present, turns the rise into a peak which is located where the voiceless part started. Thus, smoothing here causes an H*L to be predicted where there is only a rise towards a high boundary. The boundary tone on this word has also been incorrectly predicted—however, here, the predicted high intonation phrase boundary is much more plausible than the manually labeled intermediate boundary, so this is a manual labeling error rather than a prediction error. The next two errors are on the word *motiviert*, and there is no obvious rea-

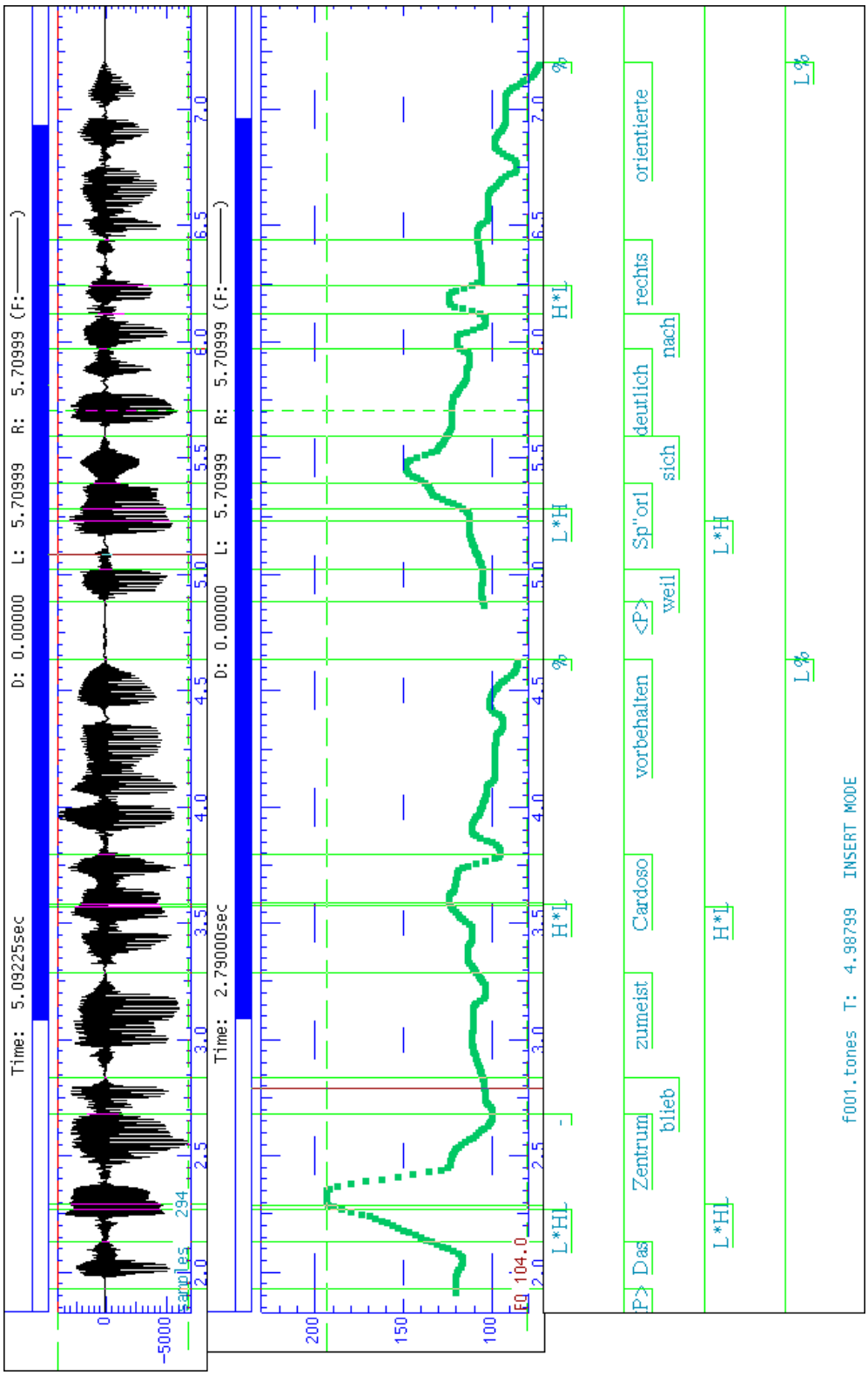

Figure 6.16: Prediction result for utterance f001 from the test set. Speech signal and smoothed F0 contour are indicated at the top; the label tiers indicate manually labeled intonation (top tier), words (2nd tier), predicted accents (3rd tier), and predicted boundary tones (4th tier)

	accents		boundaries	
word	correct	predicted	correct	predicted
Beide	L*H	L*H	NONE	NONE
Teams	NONE	H*L	(H)-	H%
waren	NONE	NONE	NONE	NONE
sehr	NONE	NONE	NONE	NONE
motiviert	L*H	H*	H%	NONE
und			NONE	NONE
ständig	NONE	NONE	NONE	NONE
im			NONE	NONE
Vorwärtsgang	H*L	H*L	(L)%	(L)%
es			NONE	NONE
war			NONE	NONE
viel	H*		NONE	NONE
Tempo	H*L	NONE	NONE	NONE
in			NONE	NONE
der			NONE	NONE
Partie	NONE	H*L	(L)%	(L)%

Table 6.11: Correct and predicted pitch accents and boundaries in utterance f011, corresponding to the phrase "Beide Teams waren sehr motiviert und ständig im Vorwärtsgang, es war viel Tempo in der Partie" ("Both teams were very motivated and always pushing forwards, there was much tempo to the match"). For monosyllabic function words no accents were predicted because they contain no stressed syllable by default in IMS Festival. Incorrect predictions are highlighted in gray.

son why these have been wrongly predicted. However, the next accent on the word *viel* is missed in prediction because *viel* has been mistakenly labeled as unstressed (in spite of being accented) and thus no prediction is made, which is of course wrong here. The H*L accent on *Tempo* is also missed, which is probably because the amplitude of the related peak is not too high—it can be seen that there are a lot of similarly vague "bumps" in the contour which indeed do not correspond to accents. Finally, an H*L accent is predicted on the very last word, which is not surprising because the contour is clearly falling there; however, this is due to the following boundary and not to a pitch accent. In my experience from manual prosodic labeling, pitch accents at the very end of utterances are often not associated with clear F0 movements; it is easier to judge their accent status based on other parameters such as loudness, and then just decide on the type of accent by the contour shape. Thus, adding parameters related to loudness might help in automatic classification as well in such cases.

Turning to the third utterance ("Bei Kaiserslautern agierte im Vergleich zur Champions League gegen Eindhoven Ramzy für Koch als Manndecker", *In con-*

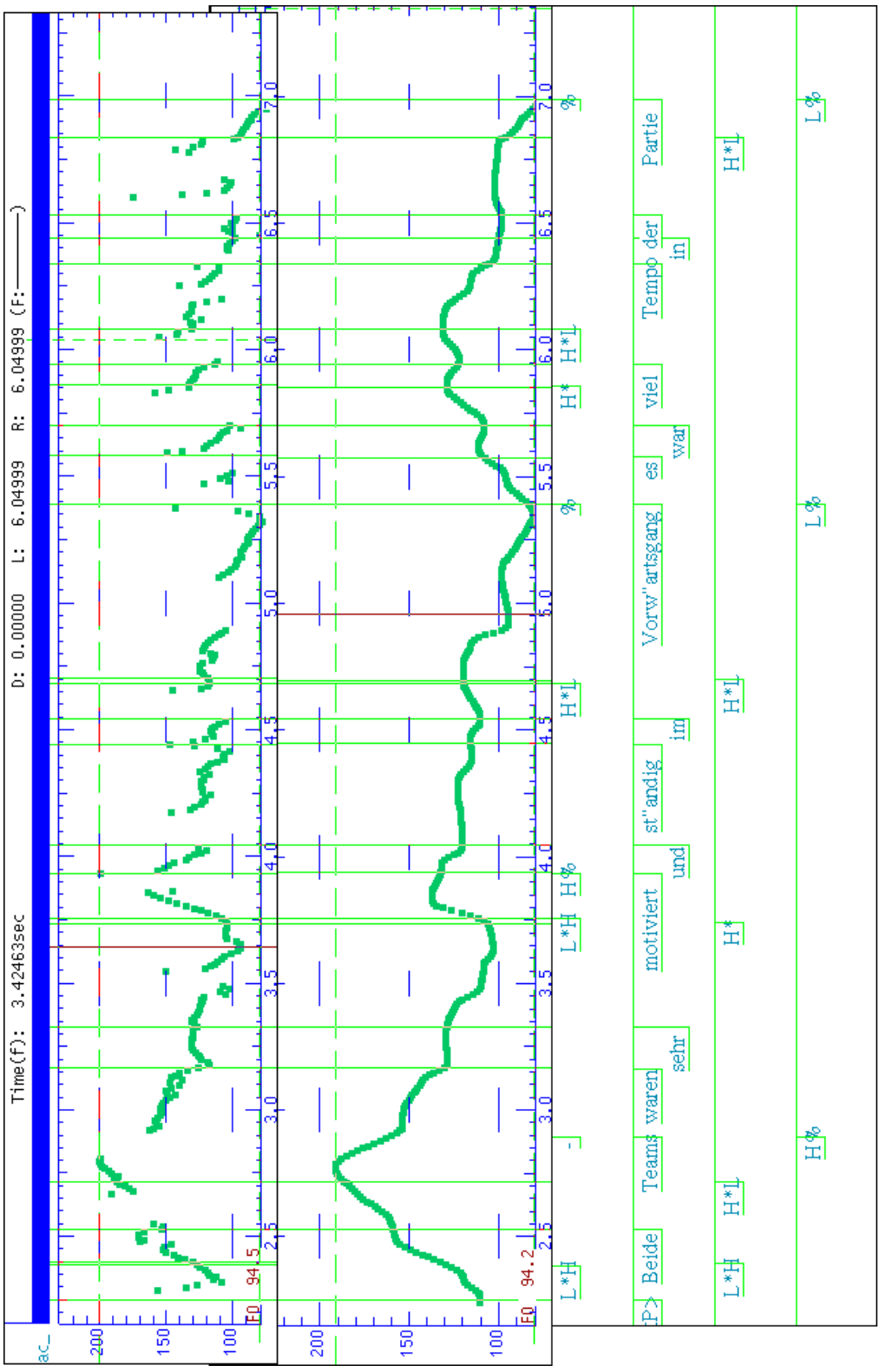

Figure 6.17: Prediction result for utterance f011 from the test set. Original and smoothed F0 contours are indicated at the top; the label tiers indicate manually labeled intonation (top tier), words (2nd tier), predicted accents (3rd tier), and predicted boundary tones (4th tier)

	accents		boundaries	
word	correct	predicted	correct	predicted
Bei			NONE	NONE
Kaiserslautern	L*H	L*H	H%	H%
agierte	L*H?	NONE	NONE	NONE
im			NONE	NONE
Vergleich	L*H	NONE	NONE	NONE
zur			NONE	NONE
Champions	L*HL?	L*H	NONE	NONE
League	NONE	NONE	(H)-	NONE
gegen	NONE	NONE	NONE	NONE
Eindhoven	L*H	L*H	(H)%	H%
Ramzy	L*H	L*H	NONE	(H)-
für			NONE	NONE
Koch	NONE	NONE	NONE	NONE
als			NONE	NONE
Manndecker	H*L	H*L	(L)%	(L)%

Table 6.12: Correct and predicted pitch accents and boundaries in utterance f021, corresponding to the phrase "Bei Kaiserslautern agierte im Vergleich zur Champions League gegen Eindhoven Ramzy für Koch als Manndecker" ("In contrast to the Champions League match against Eindhoven, Ramzy was positioned to play man-for-man marking for Kaiserslautern"). For monosyllabic function words no accents were predicted because they contain no stressed syllable by default in IMS Festival. Incorrect predictions are highlighted in gray.

trast to the Champions League match against Eindhoven, Ramzy was positioned to play man-for-man marking for Kaiserslautern), in table 6.12 and figure 6.18, for two of the incorrectly predicted pitch accents, the labeler had indicated uncertainty by the ? diacritic, indicating that these accents were problematic even in manual labeling. Indeed, the amplitude of the pitch movement on *agierte*, which had been labeled L*H?, is not very pronounced, and even more so on *Vergleich*—here, one may wonder if it was justified to manually label an L*H accent given that there is almost no rise associated with the accent. Similarly, the amplitude of the movement on *Champions*, which should have been predicted L*HL, is smaller than usually observed for L*HL accents, which explains why it has been mistaken for L*H. As for phrase boundaries, there were two mistakes involving intermediate boundaries; one was missed (on *League*), and one was inserted (on *Ramzy*). However, the H% boundary predicted for *Eindhoven* may even be more justified than the (H)% boundary that had been labeled manually.

For the fourth utterance of the training data, the results are given in table 6.13, but no screenshot is displayed because the utterance consists of a single

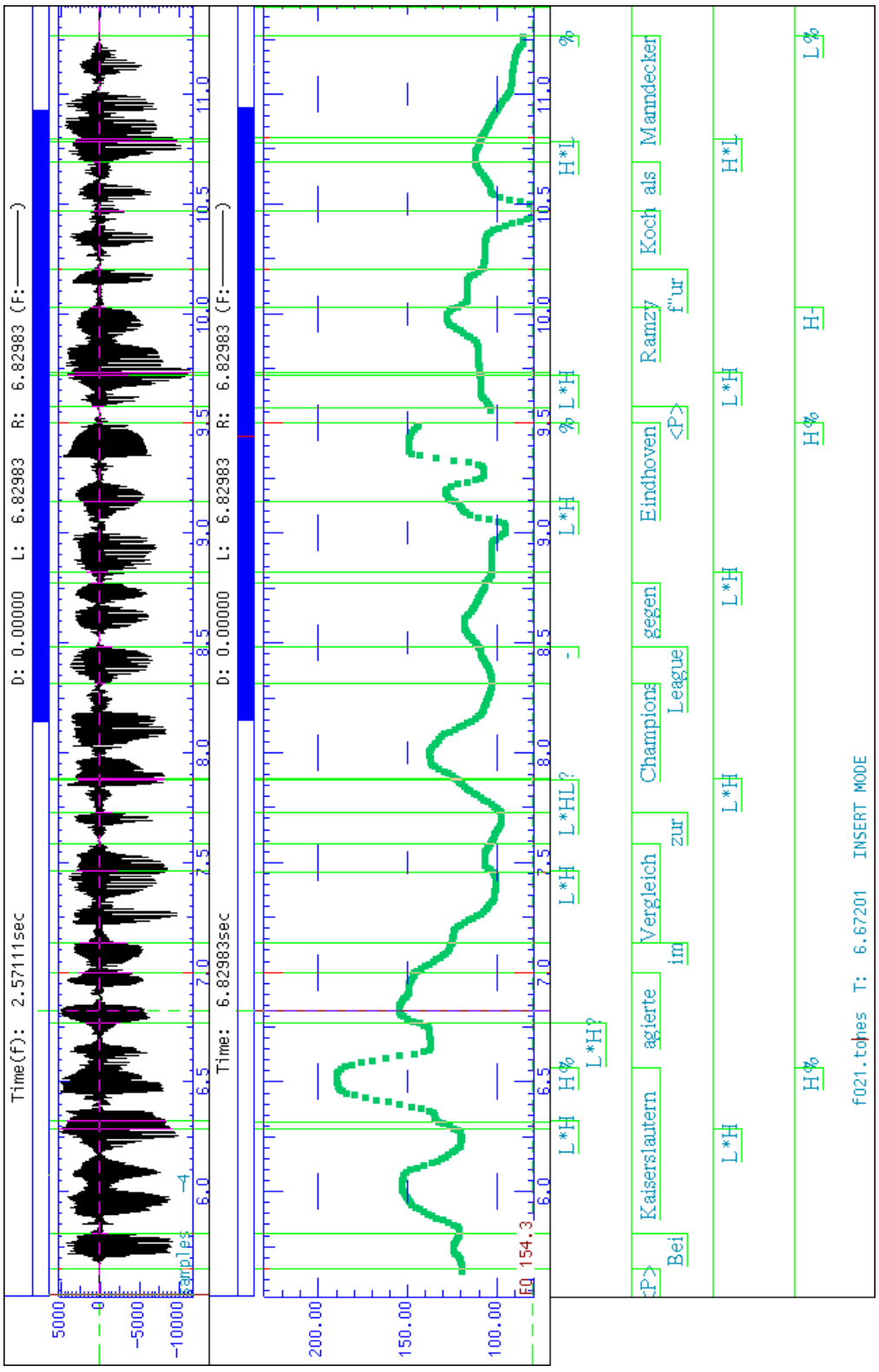

Figure 6.18: Prediction result for utterance f021 from the test set. Speech signal and smoothed F0 contour are indicated at the top; the label tiers indicate manually labeled intonation (top tier), words (2nd tier), predicted accents (3rd tier), and predicted boundary tones (4th tier)

word	accents		boundaries	
	correct	predicted	correct	predicted
Schlusspfiff	H*L	H*L	(L)%	(L)%

Table 6.13: Correct and predicted pitch accents and boundaries in utterance f031, corresponding to the single-word phrase "Schlusspfiff" ("Final whistle").

word	accents		boundaries	
	correct	predicted	correct	predicted
Nach			NONE	NONE
einer	NONE	NONE	NONE	NONE
Flanke	L*H	L*H	NONE	(H)-
von			NONE	NONE
der			NONE	NONE
rechten	NONE	NONE	NONE	NONE
Seite	L*H	L*H	(H)%	H%
köpft	NONE	NONE	NONE	NONE
Juskowiak	L*H	NONE	NONE	(H)%
nur	NONE	NONE	NONE	NONE
Zentimeter	L*H	L*H	NONE	NONE
am			NONE	NONE
FCN	NONE	NONE	NONE	NONE
Tor	NONE	NONE	NONE	NONE
vorbei	H*L	H*L	(L)%	(L)%

Table 6.14: Correct and predicted pitch accents and boundaries in utterance f041, corresponding to the phrase "Nach einer Flanke von der rechten Seite köpft Juskowiak nur Zentimeter am FCN Tor vorbei" ("Following a cross-path from the right side, Juskowiak's header missed the FCN's goal only by centimeters"). For monosyllabic function words no accents were predicted because they contain no stressed syllable by default in IMS Festival. Incorrect predictions are highlighted in gray.

word ("Schlusspfiff", *Final whistle*), and pitch accent as well as boundary tone have been correctly predicted.

Finally, results for the fifth utterance ("Nach einer Flanke von der rechten Seite köpft Juskowiak nur Zentimeter am FCN Tor vorbei", *Following a cross-pass from the right side, Juskowiak's header missed the FCN's goal only by centimeters*) are indicated in table 6.14 and figure 6.19. There is only one error in pitch accent prediction: the L*H accent on *Juskowiak* is missed. Turning to the boundaries, in the first two cases, on the words *Flanke* and *Seite*, I would

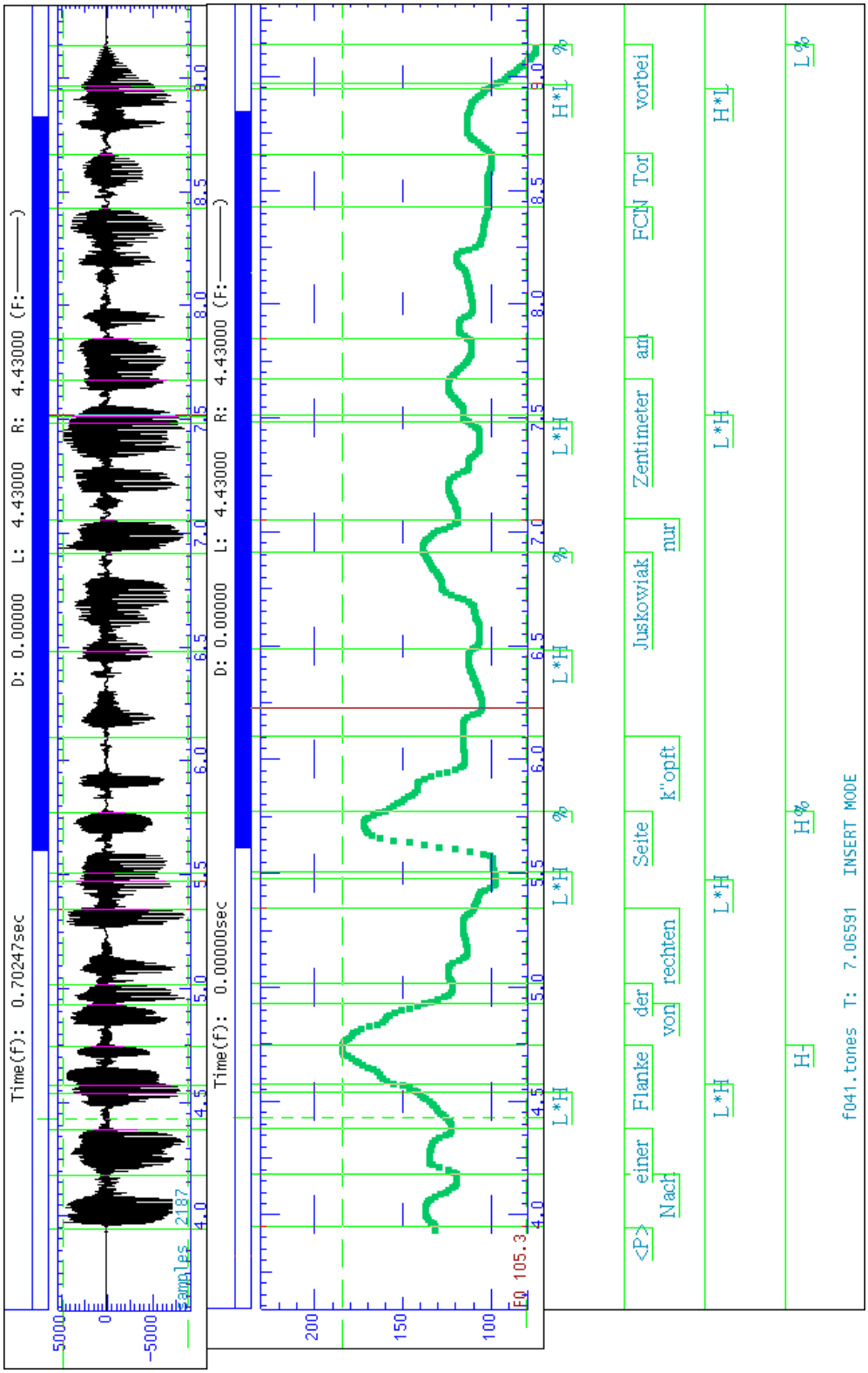

Figure 6.19: Prediction result for utterance f041 from the test set. Speech signal and smoothed F0 contour are indicated at the top; the label tiers indicate manually labeled intonation (top tier), words (2nd tier), predicted accents (3rd tier), and predicted boundary tones (4th tier)

again claim that the predicted versions are probably better than the manual labels: on *Flanke*, the contour rises throughout the word, continuing on the second, unstressed syllable. If there were no boundary following *Flanke*, one would expect that the contour starts to fall on this second syllable already. On *Seite*, the rise is characterized by a relatively high amplitude—this can be seen by comparing the rise on *Seite* to the rise on *Juskowiak*, for instance, which is much more typical of a (H)% boundary.

Summarizing these few examples, it can be said that prediction errors were sometimes caused by external factors and in these cases do not display weaknesses of the classifiers themselves. For instance, in one case, F0 smoothing had the undesired effect of turning a clear rise to the boundary into a peak, resulting in predicting an H*L accent. In one case, incorrectly labeled word stress in the database has prevented prediction of a pitch accent for one word. In several cases, one might argue that the predicted labels are actually more suitable than the manually labeled ones. These cases make up 8 out of 19 errors. They show that a considerable proportion of errors could be avoided if the input data were more consistent. However, in "real-life" application, one cannot expect perfect input data, so such errors are almost certain to occur when automatically labeling new data.

Overall, the results of this prototype for automatic prosodic labeling are promising. Even though there are some errors, many predictions are correct, and the errors are usually not blatant errors but in my experience are similar to errors human labelers might make. In its current state, the prototype is certainly useful in a bootstrapping approach when labeling new data, and I believe that with some future improvement, it may even replace manual prosodic labeling in some application scenarios.

6.3.8 Discussion and Outlook

In this section I have presented results on simulating prosodic categorization using machine learning methods to classify new exemplars. The results compare very well to results reported in other recent studies, particularly to results on German. In contrast to most other studies, the classifiers predict the full set of GToBI(S) labels rather than just two classes. As illustrated in the preceding section, they are good enough to be useful in automatic prosodic labeling.

Various learning schemes implemented in WEKA (Witten and Frank 2005) proved to be equally suitable for prosodic classification showing that good performance in classification is not necessarily due to one outstanding learning algorithm that is particularly suitable for the data; instead, this can be interpreted as showing that the information provided was sufficient to reliably reach quite high accuracy rates using various learning algorithms.

From an exemplar-theoretic point of view, it is interesting to note that

instance-based learning performs almost as well as the best learning schemes. In instance-based learning, new instances are categorized based on their similarity to stored instances, assigning new instances the label which is most frequent among their neighbors in instance space. Exemplar-theoretic categorization has been claimed to work exactly this way (Lacerda 1995; Pierrehumbert 2003). Johnson (1997) models exemplar-theoretic categorization slightly differently, but he himself notes that the process can be interpreted as "a sort of K nearest-neighbors classification" (Johnson 1997, p. 148).

I have also begun to assess the generalizability of classifiers trained on data of one speaker to other speakers' data. Results showed that the classifiers generalize very well to similar data of another speaker in that they yield the same accuracy rates as classifiers trained directly on data of that speaker. Generalizability is not only desirable in the context of automatic prosodic classification; I claim that the classifiers model human perception of prosodic categories by encoding perceptually relevant structure of prosodic categories, and in that respect, they are necessarily expected to generalize to other speakers: one would not want to assume that different speakers encode prosodic categories differently.

To further pursue the assessment of generalizability, the best classifiers will be applied to data of more speakers in the future, in particular to data which do not match the training data so closely with respect to speech style or content. In the very near future, I will further develop the prototype to obtain a first version of an automatic labeling tool. The first challenge will be to apply this tool to spontaneous speech data in a new project on phonetic convergence in spontaneous speech. If on these data, similar accuracies can be reached as in the experiments presented here, the tool would constitute a very valuable contribution to research on spontaneous speech. But even in its current state, the prototype is expected to be very useful in prosodically labeling new speech data in a bootstrapping approach.

Chapter 7

Conclusion and Outlook

One aim of this thesis was to identify perceptual targets for prosodic events, extending Guenther and Perkell's model (Guenther et al. 1998; Perkell et al. 2001) from the segmental to the prosodic domain as explicated in section 2.3.5. This involves, first, identifying perceptually relevant prosodic dimensions, and, second, identifying which regions in the space spanned by these dimensions can be considered as target regions for the different prosodic categories.

I suggest that temporal aspects in the realization of prosodic events are captured in duration z-scores of segments and syllables. I have argued that the z-scores are a measure of "local" speech rate which is granular enough to capture lengthening effects related to pitch accents and boundary tones. From an exemplar-theoretic perspective, the z-score of a speech unit duration can be interpreted as the location of the exemplar in the temporal dimension of its exemplar cloud. For instance, exemplars which lie at the center of their cloud in the temporal dimension exhibit a z-score of 0. As for the tonal aspects of prosodic events, I suggest that properties such as peak alignment, peak height, and fall and rise amplitudes are relevant in perception, and that these properties can be quantified by the PaIntE parameters. Thus, the PaIntE parameters can be thought of as relevant dimensions in tonal perception.

Following Keating's (1990) window model of coarticulation, it can be assumed that target windows for prosodic categories (the target regions, in Guenther and Perkell's terms) are implicitly defined by the values observed for each category. In this vein, I have shown in chapters 4 and 5 that for different prosodic events the distributions of the parameters suggested above are often significantly distinct from each other. However, it was also evident that there is much overlap between prosodic categories. The fact that the distributions are significantly different does not necessarily imply that they are different enough for discrimination. As Pierrehumbert (2003) states in discussing the distributions of two categories:

> Discrimination is not the same as statistical significance. As the quantity of data contributing to the distributions [...] increases toward infinity, the difference in the means of these distributions can become as statistically significant as one could wish. That is, [...] we can become more and more sure, scientifically speaking, that the distributions [...] are not the same. But no amount of certainty on this point improves the situation with respect to discrimination of the categories when classifying an unknown incoming token" (Pierrehumbert 2003, pp. 208–209).

In order to test whether the distributions are distinct enough for robust discrimination, I have carried out two sets of experiments: first, I have used clustering as a means to automatically detect accumulations of similar syllable instances corresponding to pitch accent categories, and second, I have built classifiers for predicting prosodic events. In both cases, PaIntE parameters as well as duration z-scores were used as dimensions, but further attributes, which were mostly derived from these parameters, have been included. For instance, the PaIntE parameters and z-scores of surrounding syllables were included to distinguish local phenomena from more global phenomena. Local phenomena are expected to be caused by prosodic events, while more global phenomena may have no immediate linguistic function. For instance, high boundary tones usually cause F0 rises that are confined to one or few syllables, i.e. they cause a local rise, while more extended "global" rises in F0 are often due to interpolation between events. Similarly, boundaries or pitch accents are expected to locally lengthen syllable nuclei or syllable-final segments, while a slower global speech rate would affect more than just one syllable. Also, some higher linguistic properties were included to model speakers' expectations, which are based on the linguistic context in which a syllable occurs. For instance, speakers would expect pitch accents on content words rather than on function words, or phrase boundaries at syntactic boundaries rather than within constituents.

The results indicate that the proposed dimensions do capture perceptual aspects of prosodic events. For instance, when evaluating the clusterings by assessing their prediction accuracy, scores of 85.5% can be reached for pitch accents. Even higher accuracies are obtained in the prediction experiments; here, scores of 87.5% for pitch accents and of 93.9% for boundary tones can be reached when evaluating the classifiers syllable by syllable.[1] Given that even human labelers do not reach 100%, as is evident from studies on labeler consistency (Pitrelli et al. 1994; Grice et al. 1996; Syrdal and McGory 2000), these scores are encouraging.

Among the classifiers which yield the best scores there are several classifiers which are based on decision trees. Similar to clustering, decision trees can

[1] The corresponding word-based accuracies are lower, at approx. 78% for pitch accents and at approx. 88% for boundaries when attempting to predict the full set of accents.

be thought of as partitioning the instance space: at each decision node, the instance space is split in two, and instances in the two parts are treated by the two subtrees. The leaves then correspond to regions in instance space. When using standard decision trees, all instances in these regions are predicted to belong to the same class.[2] In this way, standard decision trees identify regions in instance space which correspond to the categories. Clustering, on the other hand, detects categories by identifying regions in instance space with higher instance density.

Thus, the clustering and prediction experiments can be seen as constituting the second step in applying Guenther and Perkell's model (Guenther et al. 1998; Perkell et al. 2001) to the prosodic domain, by identifying target regions corresponding to the prosodic categories in the proposed perceptual space. However, the results of the clustering and prediction experiments indicate that there are many target regions for each prosodic category: the clustering results indicate that 1600 to 2000 clusters are appropriate for pitch accents. Similarly, in the prediction experiments, the best decision-tree based results were achieved by bagged decision trees. If standard decision trees constitute partitionings of the instance space, then bagged decision trees are sets of trees which constitute different partitionings, i.e., they identify no unique partitioning which relates regions to classes, instead, they identify different partitionings which relate different regions to different classes. In contrast, for standard decision trees (for instance, J48, RandomTree, or REPTree in the WEKA implementation) there is a 1-1 correspondence between the regions defined by the leaves and the classes. However, these standard decision trees have been found to be inferior to the meta-trees cited above with respect to classification accuracy. Also, the trees that are obtained for these algorithms can be quite complex as well. For instance, a REPTree decision tree for pitch accent prediction with 10-fold cross-validated accuracy of 86.6%, i.e., 1% lower accuracy than the best algorithms, contained more than 500 leaves. A J48 tree with even lower accuracy contained almost 3000 leaves.

Clustering results and prediction results indicate that it is not possible to identify one unique target region, in the sense of Guenther and Perkell, for each prosodic category, at least not with the many dimensions currently used. However, a convincing argument for their model is that disjoint target regions in articulator space are replaced by single convex target regions in perceptual space. Assuming many different target regions for each prosodic category would take much of the elegance of the model away. This perspective is also in opposition to my earlier claim that the z-score and PaIntE parameter distributions can be

[2]This is not the case for some meta-learning algorithms which combine decision trees with other models. For instance, LMT and ClassificationViaRegression are decision trees with regression models at their leaves, as explained in section 6.3.3, i.e., they do not assume that all instances in a particular region belong to the same class, however they assume that they can be treated using the same model.

seen as implicitly defining the target regions for the prosodic categories, as I had suggested in the concluding remarks in chapters 4 and 5.

Guenther et al. (1998) illustrate the concept of a single convex perceptual target region on the example of American English /r/ using F1 and F2 values as dimensions. However, to my knowledge the issue of discriminability between different segment categories in categorization is not addressed in their work. Indeed, the mean F1 and F2 values of American English /U/ vowels produced by female speakers and those of /3:/ vowels produced by either female or male speakers found in a study by Hillenbrand et al. (1995) lie roughly in the target region posited for /r/. This indicates that when using only F1 and F2 as dimensions, the target region for /r/ will not be distinct enough from the target regions of these vowels to allow for discrimination, just as the target regions for prosodic events are not distinct enough when using only duration z-scores and PaIntE parameters of the current syllable as dimensions.

I therefore suggest to regard the perceptual target regions in the sense of Guenther et al. (1998) and Perkell et al. (2001) as very coarse, abstract perceptual descriptions of the underlying targets of speech events. For the prosodic domain, I maintain the view explicated in chapters 4 and 5: duration z-scores and PaIntE parameters span a perceptual space in which these target regions are located, and the distributions of values observed for each prosodic event define its target region in this space. However, these target regions are overlapping and therefore not fine-grained enough for discrimination.

To allow for discrimination, context must be considered as well. As demonstrated by the clustering and prediction experiments in chapter 6, including parameters which are mostly derived from the PaIntE and z-score parameters, such as the parameters of neighboring syllables, but also some higher-linguistic properties, is sufficient to obtain very encouraging prediction accuracies, particularly in relation to human labeler consistencies. This demonstrates that the parameters do capture most perceptually relevant properties.

However, some attributes may be irrelevant, and could artificially increase the number of clusters or the complexity of the classifiers more than necessary. One way to gain some insight into the perceptual relevance of the attributes is to experiment with attribute selection methods. Such methods are provided with WEKA (Witten and Frank 2005). I have not yet systematically looked at this issue, but I have experimented somewhat with attribute selection when training the classifiers, reducing the number of attributes according to the relevance they were attested by WEKA's attribute selection methods. However, different selection methods often yield different results, and the classification results were always worse when excluding attributes which had been judged less relevant by some selection method. Still, systematic experiments could serve to identify the most relevant dimensions. This could in turn then reduce the number of clusters or the complexity of the classifiers. However, I doubt that the reduction in complexity will be enough to yield single convex target

regions for the prosodic events.

There are other factors which in my opinion could help to reduce the complexity of the clusterings. As stated in section 6.1.3, the data are noisy, and I hope to eliminate some of this noise in the future. One issue is that the F0 smoothing often has undesired effects, as discussed in sections 3.2.5 and 6.3.7. A better F0 smoothing method, such as the one suggested by Reichel and Winkelmann (2010), could alleviate this problem. Some noise is also introduced by incorrect manual labelings – in illustrating the examples of automatically generated prosodic label files in 6.3.7, some instances were discussed for which I claimed that the predicted label may have been more adequate than the manual label. Using more consistently labeled data, for instance data where several labelers agreed, might help to assess the influence of such labeling errors.

To conclude, I will come back to the main contributions of this thesis, as they were stated in the introduction. First, I have proposed that Guenther and Perkell's speech production model for the segmental domain (Guenther 1995; Guenther et al. 1998; Perkell et al. 2001) can be applied to the prosodic domain, claiming that tonal and temporal aspects of the implementation of intonation categories are phonemic settings. I have also pointed out that their model is compatible with exemplar theory if it is assumed that the target regions are implicitly defined by the exemplar clouds. Second, I have suggested that the PaIntE parameters as well as duration z-scores as a measure of local speech rate are the perceptual dimensions in applying Guenther and Perkell's model to the prosodic domain. Third, I have extensively investigated realizations of GToBI(S) events in terms of these parameters, confirming that known effects on prosodic realization are visible in the parameter distributions. Fourth, I have conducted clustering experiments, in order to show that the suggested dimensions are sufficient to detect prosodic categories. For evaluating the clusterings, I have suggested an evaluation procedure which to my knowledge has not been used before, which aims at assessing the generalizability of clusterings. Finally, I have built classifiers to show that the suggested dimensions are good enough to reach state-of-the art classification accuracies using supervised methods. Thus, a fifth, very valuable contribution of this thesis is a prototype of a tool for automatic prosodic labeling, which will be improved further in the future.

Bibliography

Abney 1995 ABNEY, Steven P.: Chunks and dependencies: bringing processing evidence to bear on syntax. In: *Computational Linguistics and the Foundations of Linguistic Theory*. Stanford : CSLI, 1995

Baayen et al. 1995 BAAYEN, H. ; PIEPENBROCK, R. ; GULIKERS, L.: *The CELEX lexical database—Release 2*. CD-ROM. 1995. – Centre for Lexical Information, Max Planck Institute for Psycholinguistics, Nijmegen

Baayen 2001 BAAYEN, Harald: *Word Frequency Distributions*. Dordrecht : Kluwer, 2001

Barbisch et al. 2007 BARBISCH, Martin ; DOGIL, Grzegorz ; MÖBIUS, Bernd ; SÄUBERLICH, Bettina ; SCHWEITZER, Antje: Unit selection synthesis in the SmartWeb project. In: *Proceedings of the 6th ISCA Workshop on Speech Synthesis (SSW-6, Bonn)*, 2007, pp. 304–309

Beckman and Ayers 1994 BECKMAN, Mary E. ; AYERS, Gayle M.: *Guidelines for ToBI labelling, version 2.0*. February 1994

Braunschweiler 2006 BRAUNSCHWEILER, Norbert: The Prosodizer — Automatic Prosodic Annotations of Speech Synthesis Databases. In: *Proceedings of Speech Prosody 2006 (Dresden)*, 2006

Browman and Goldstein 1986 BROWMAN, Catherine ; GOLDSTEIN, Louis: Towards an articulatory phonology. In: *Phonology Yearbook* 3 (1986), pp. 219–252

Browman and Goldstein 1992 BROWMAN, Catherine ; GOLDSTEIN, Louis: Articulatory phonology: an overview. In: *Phonetica* 49 (1992), pp. 155–180

Byrd 1996 BYRD, Dani: A phase window framework for articulatory timing. In: *Phonology* 13 (1996), pp. 139–169

Calhoun and Schweitzer accepted CALHOUN, Sasha ; SCHWEITZER, Antje: Can intonation contours be lexicalised? Implications for Discourse Meanings. In: *Prosody and Meaning (Trends in Linguistics)*. Mouton De Gruyter, accepted

Campbell 1992 CAMPBELL, W. N.: Syllable-based segmental duration. In: BAILLY, G. (ed.) ; BENOÎT, C. (ed.) ; SAWALLIS, T.R. (ed.): *Talking Machines: Theories, Models, and Designs*. Amsterdam : Elsevier, 1992, pp. 211–224

Campbell and Isard 1991 CAMPBELL, W. N. ; ISARD, S. D.: Segment durations in a syllable frame. In: *Journal of Phonetics* 19 (1991), pp. 37–47

Cholin et al. 2004 CHOLIN, Joana ; SCHILLER, Niels O. ; LEVELT, Willem J. M.: The preparation of syllables in speech production. In: *Journal of Memory and Language* 50 (2004), pp. 47–61

Dietterich 2000 DIETTERICH, Thomas G.: An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. In: *Machine Learning* 40 (2000), pp. 139–157

Dogil and Möbius 2001 DOGIL, G. ; MÖBIUS, B.: Towards a model of target oriented production of prosody. In: *Proceedings of the European Conference on Speech Communication and Technology (Aalborg, Denmark)* vol. 1, 2001, pp. 665–668

Edinburgh Speech Tools Library 1999 TAYLOR, Paul ; CALEY, Richard ; BLACK, Alan W. ; KING, Simon: *Edinburgh Speech Tools Library*. [http://festvox.org/docs/speech_tools-1.2.0/]. 1999. – System Documentation Edition 1.2, for 1.2.0 15th June 1999

Féry 1993 FÉRY, Caroline: *The meaning of German intonational patterns*. Tübingen : Max Niemeyer Verlag, 1993

Festival 2010 CENTRE FOR SPEECH TECHNOLOGY RESEARCH, UNIVERSITY OF EDINBURGH: *The Festival text-to-speech synthesis system*. [http://www.cstr.ed.ac.uk/projects/festival/]

Fujisaki and Hirose 1984 FUJISAKI, Hiroya ; HIROSE, Keikichi: Analysis of voice fundamental contours for declarative sentences of Japanese. In: *Journal of the Acoustical Society of Japan (E)* 5 (1984), pp. 233–242

Goldinger 1996 GOLDINGER, Stephen D.: Words and voices: Episodic traces in spoken word identification and recognition memory. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22 (1996), pp. 1166–1183

Goldinger 1997 GOLDINGER, Stephen D.: Words and voices—Perception and production in an episodic lexicon. In: JOHNSON, Keith (ed.) ; MULLENNIX, John W. (ed.): *Talker Variability in Speech Processing*. San Diego : Academic Press, 1997, pp. 33–66

Goldinger 1998 GOLDINGER, Stephen D.: Echoes of echoes? An episodic theory of lexical access. In: *Psychological Review* 105 (1998), pp. 251–279

Goldinger 2000 GOLDINGER, Stephen D.: The role of perceptual episodes in lexical processing. In: *Proceedings of the Workshop on Spoken Word Access Processes*. Nijmegen, The Netherlands : Max-Planck Institute for Psycholinguistics, 2000, pp. 155–158

Grewendorf et al. 1989 GREWENDORF, Günther ; HAMM, Fritz ; STERNEFELD, Wolfgang: *Sprachliches Wissen*. Frankfurt : Suhrkamp, 1989

Grice and Baumann 2002 GRICE, Martine ; BAUMANN, Stefan: Deutsche Intonation und GToBI. In: *Linguistische Berichte* 191 (2002), pp. 267–298

Grice et al. 2005 GRICE, Martine ; BAUMANN, Stefan ; BENZMÜLLER, Ralf: *German Intonation in Autosegmental-Metrical Phonology*. In: JUN, Sun-Ah (ed.): *Prosodic Typology: The Phonology of Intonation and Phrasing*, Oxford University Press, 2005

Grice et al. 1996 GRICE, Martine ; REYELT, Matthias ; BENZMÜLLER, Ralf ; MAYER, Jörg ; BATLINER, Anton: Consistency in Transcription and Labelling of German Intonation with GToBI. In: *Proceedings of the Fourth International Conference on Spoken Language Processing, Philadelphia*, 1996, pp. 1716–1719

Guenther et al. 1998 GUENTHER, F. H. ; HAMPSON, M. ; JOHNSON, D.: A theoretical investigation of reference frames for the planning of speech movements. In: *Psychological Review* 105 (1998), pp. 611–633

Guenther 1995 GUENTHER, Frank H.: A modeling framework for speech motor development and kinematic articulator control. In: *Proceedings of the XIIIth International Congress of Phonetic Sciences*, 1995, pp. 92–99

Hasegawa-Johnson et al. 2005 HASEGAWA-JOHNSON, Mark ; CHEN, Ken ; COLE, Jennifer ; BORYS, Sarah ; KIM, Sung-Suk ; COHEN, Aaron ; ZHANG, Tong ; CHOI, Jeung-Yoon ; KIM, Heejin ; YOON, Taejin: Simultaneous recognition of words and prosody in the Boston University Radio Speech Corpus. In: *Speech Communication* 46 (2005), no. 3-4, pp. 418–439

Heid 1998 HEID, Sebastian: Phonetische Variation: Untersuchungen anhand des PhonDat2-Korpus. In: *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München, FIPKM* (1998), pp. 193–368

Hermes and van Gestel 1991 HERMES, D. J. ; GESTEL, J. C. van: The frequency scale of speech intonation. In: *Journal of the Acoustical Society of America* 90 (1991), pp. 97–102

Hillenbrand et al. 1995 HILLENBRAND, James ; GETTY, Laura A. ; CLARK, Michael J. ; WHEELER, Kimberley: Acoustic characteristics of American English vowels. In: *Journal of the Acoustical Society of America* (1995), Mai, no. 5 Pt. 1

Hirst and Espesser 1993 HIRST, Daniel ; ESPESSER, R.: Automatic modelling of fundamental frequency using a quadratic spline algorithm. In: *Travaux de l'Institut de Phonétique d'Aix-en-Provence* 15 (1993), pp. 75–85

Höhle 1986 HÖHLE, Tilman N.: Der Begriff "Mittelfeld". Anmerkungen über die Theorie der topologischen Felder. In: SCHÖNE, Albrecht (ed.): *Kontroversen, alte und neue. Akten des VII. Germanisten-Kongresses, Göttingen 1985*. Tübingen : Niemeyer, 1986, pp. 329–340

House 1996 HOUSE, David: Differential perception of tonal contours through the syllable. In: *Proceedings of the International Conference on Spoken Language Processing (Philadelphia, PA)* vol. 1, 1996, pp. 2048–2051

IMS Festival 2010 INSITUT FÜR MASCHINELLE SPRACHVERARBEITUNG: *IMS German Festival home page*. [www.ims.uni-stuttgart.de/phonetik/synthesis/]

Jilka and Möbius 2007 JILKA, Matthias ; MÖBIUS, Bernd: The influence of vowel quality features on peak alignment. In: *Proceedings of Interspeech 2007 (Antwerpen)*, 2007, pp. 2621–2624

Jilka et al. 1999 JILKA, Matthias ; MÖHLER, Gregor ; DOGIL, Grzegorz: Rules for the generation of ToBI-based American English intonation. In: *Speech Communication* 28 (1999), pp. 83–108

Johnson 1997 JOHNSON, K.: Speech perception without speaker normalization: An exemplar model. In: JOHNSON, K. (ed.) ; MULLENNIX, J. W. (ed.): *Talker Variability in Speech Processing*. San Diego : Academic Press, 1997, pp. 145–165

Kamp and Reyle 1993 KAMP, Hans ; REYLE, Uwe: *From Discourse to Logic*. Dordrecht : Kluwer, 1993

Kaufman and Rousseeuw 1990 KAUFMAN, Leonard ; ROUSSEEUW, Peter J.: *Finding groups in data: an introduction to cluster analysis*. John Wiley & sons, 1990

Keating 1990 KEATING, Patricia A.: The window model of coarticulation: articulatory evidence. In: KINGSTON, J. (ed.) ; BECKMAN, Mary (ed.): *Papers in Laboratory Phonology I*. Cambridge University Press, 1990, pp. 451–470

Kohler 1991 KOHLER, Klaus J.: Prosody in speech synthesis: the interplay between basic research and TTS application. In: *Journal of Phonetics* 19 (1991), pp. 121–138

Kornai 1998 KORNAI, A.: Analytic models in phonology. In: DURAND, J. (ed.) ; LAKS, B. (ed.): *The organization of phonology: Constraints, levels and representations*. Oxford, U.K. : Oxford University Press, 1998, pp. 395–418

Kruschke 1992 KRUSCHKE, John K.: ALCOVE: An exemplar-based connectionist model of category learning. In: *Psychological Review* 99 (1992), no. 1, pp. 22–44

Kuhl 1991 KUHL, Patricia K.: Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not. In: *Perception and Psychophysics* 50 (1991), pp. 93–107

Lacerda 1995 LACERDA, F.: The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory. In: *Proc. 13th Internat. Congr. Phonet. Sci. (Stockholm)* vol. 2, 1995, pp. 140–147

Ladd 1983 LADD, D. R.: Phonological features of intonational peaks. In: *Language* 59 (1983), pp. 721–759

Levelt 1992 LEVELT, Willem J. M.: Accessing words in speech production: Stages, processes and representations. In: *Cognition* 42 (1992), pp. 1–22

Levelt 1999 LEVELT, Willem J. M.: Producing spoken language: a blueprint of the speaker. In: BROWN, C. M. (ed.) ; HAGOORT, P. (ed.): *The Neurocognition of Language*. Oxford, UK : Oxford University Press, 1999, pp. 83–122

Levelt and Wheeldon 1994 LEVELT, Willem J. M. ; WHEELDON, L.: Do speakers have access to a mental syllabary? In: *Cognition* 50 (1994), pp. 239–269

Lindblom 1990 LINDBLOM, Björn: Explaining phonetic vatiation: A sketch of the H & H theory. In: HARDCASTLE, William J. (ed.) ; MARCHAL, Alain (ed.): *Speech Production and Speech Modelling*. Dordrecht : Kluwer, 1990, pp. 403–439

Maye and Gerken 2000 MAYE, J. ; GERKEN, L.: Learning phonemes without minimal pairs. In: *Proceedings of the 24th Annual Boston University Conference on Language Development*. Somerville, Mass. : Cascadilla Press, 2000, pp. 522–533

Maye et al. 2002 MAYE, J. ; WERKER, J. F. ; GERKEN, L.: Infant sensitivity to distributional information can affect phonetic discrimination. In: *Cognition* 82 (2002), no. 3, pp. B101–B111

Mayer 1995 MAYER, Jörg: Transcription of German intonation—The Stuttgart system / Institute of Natural Language Processing, University of Stuttgart. 1995. – Forschungsbericht

Möhler 2001 MÖHLER, Gregor: *Improvements of the PaIntE model for F0 parametrization*. Manuscript. 2001

Möhler and Conkie 1998 MÖHLER, Gregor ; CONKIE, Alistair: Parametric modeling of intonation using vector quantization. In: *Proceedings of the Third International Workshop on Speech Synthesis (Jenolan Caves, Australia)*, 1998, pp. 311–316

Müller et al. 2000 MÜLLER, K. ; MÖBIUS, B. ; PRESCHER, D.: Inducing probabilistic syllable classes using multivariate clustering. In: *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (Hong Kong)*, 2000, pp. 225–232

Nielsen 2008 NIELSEN, Kuniko: *Word-level and Feature-level Effects in Phonetic Imitation*. Los Angeles, University of California, Ph.D. thesis, 2008

Nosofsky 1988 NOSOFSKY, Robert M.: Exemplar-based accounts of relations between classification, recognition, and typicality. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14 (1988), pp. 700–708

Palmeri et al. 1993 PALMERI, Thomas J. ; GOLDINGER, Stephen D. ; PISONI, David B.: Episodic Encoding of Voice Attributes and Recognition Memory for Spoken Words. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19 (1993), no. 2, pp. 309–328

Perkell et al. 2001 PERKELL, J. ; GUENTHER, F. ; LANE, H. ; MATTHIES, M. ; VICK, J. ; ZANDIPOUR, M.: Planning and auditory feedback in speech production. In: *4th Internat. Speech Motor Conf. (Nijmegen)*, 2001, pp. 5–11

Perkell et al. 2000 PERKELL, J. S. ; GUENTHER, F. H. ; LANE, H. ; MATTHIES, M. L. ; PERRIER, P. ; VICK, J. ; WILHELMS-TRICARICO, R. ; ZANDIPOUR, M.: A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. In: *Journal of Phonetics* 28 (2000), no. 3, pp. 233–272

Pfitzinger 1996 PFITZINGER, Hartmut R.: Two approaches to speech rate estimation. In: *Proceedings of the 6th Australian International Conference on Speech Science and Technology (SST, Adelaide)*, 1996, pp. 421–426

Pfitzinger 1998 PFITZINGER, Hartmut R.: Local speech rate as a combination of syllable and phone rate. In: *Proceedings of the International Conference on Spoken Language Processing (Sydney)* vol. 3, 1998, pp. 1087–1090

Pfitzinger 2001 PFITZINGER, Hartmut R.: Phonetische Analyse der Sprechgeschwindigkeit. In: *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München, FIPKM* (2001), pp. 117–264

Pierrehumbert 2001 PIERREHUMBERT, J.: Exemplar dynamics: Word frequency, lenition and contrast. In: BYBEE, J. (ed.) ; HOPPER, P. (ed.): *Frequency and the Emergence of Linguistic Structure*. Amsterdam : Benjamins, 2001, pp. 137–157

Pierrehumbert 1980 PIERREHUMBERT, Janet: *The phonology and phonetics of English intonation*. Cambridge, MA, MIT, Ph.D. thesis, 1980

Pierrehumbert 1981 PIERREHUMBERT, Janet: Synthesizing intonation. In: *Journal of the Acoustical Society of America* 70 (1981), pp. 985–995

Pierrehumbert 2003 PIERREHUMBERT, Janet: Probabilistic phonology: Discrimation and robustness. In: BOD, Rens (ed.) ; HAY, Jennifer (ed.) ; JANNEDY, Stefanie (ed.): *Probability Theory in Linguistics*. The MIT Press, 2003, pp. 177–228

Pierrehumbert and Steele 1989 PIERREHUMBERT, Janet B. ; STEELE, Shirley A.: Categories of Tonal Alignment in English. In: *Phonetica* 46 (1989), pp. 181–196

Pitrelli et al. 1994 PITRELLI, J. ; BECKMAN, Mary ; HIRSCHBERG, J.: Evaluation of Prosodic Transcription Labeling Reliability in the ToBI Framework. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP, Yokohama)*, 1994, pp. 123–126

R Development Core Team 2009 R DEVELOPMENT CORE TEAM: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing (Veranst.), 2009. – URL `http://www.R-project.org`. – ISBN 3-900051-07-0

Redi 2003 REDI, Laura: Categorical effects in the production of pitch contours in English. In: *Proceedings of the 15th International Congress of the Phonetic Sciences, Barcelona*, 2003, pp. 2921–2924

Reichel and Winkelmann 2010 REICHEL, Uwe D. ; WINKELMANN, Raphael: Removing micromelody from fundamental frequency contours. In: *Proceedings of Speech Prosody 2010*, 2010

Rosenberg and Hirschberg 2007 ROSENBERG, Andrew ; HIRSCHBERG, Julia: V-Measure: A conditional entropy-based external cluster evaluation measure.

In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, June 2007, pp. 410–420

Ross and Ostendorf 1996 ROSS, K. ; OSTENDORF, M.: Prediction of abstract prosodic labels for speech synthesis. In: *Computer Speech and Language* 10 (1996), October, pp. 155–185

Saltzman and Munhall 1989 SALTZMAN, Elliot L. ; MUNHALL, Kevin G.: A dynamical approach to gestural patterning in speech production. In: *Ecological Psychology* 1 (1989), no. 4, pp. 333–382

Savariaux et al. 1995 SAVARIAUX, C. ; PERRIER, P. ; ORLIAGUET, J. P.: Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production. In: *Jornal of the Acoustical Society of America* (1995), pp. 2428–2442

Schiller et al. 1995 SCHILLER, Anne ; TEUFEL, S. ; THIELEN, Christine: *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Institut für maschinelle Sprachverarbeitung, Univ. Stuttgart. 1995

Schmid 1994 SCHMID, Helmut: Probabilistic part-of-speech tagging using decision trees. In: *Proceedings of the International Conference on New Methods in Language Processing (Manchester, UK)*, 1994, pp. 44–49

Schmid 1995 SCHMID, Helmut: Improvements in part-of-speech tagging with an application to German. In: *From text to tags—Issues in multilingual language analysis. Proceedings of the EACL SIGDAT Workshop (University College, Belfield, Dublin, Ireland)*, 1995, pp. 47–50

Schneider et al. 2009 SCHNEIDER, Katrin ; DOGIL, Grzegorz ; MÖBIUS, Bernd: German boundary tones show categorical perception and a perceptual magnet effect when presented in different contexts. In: *Proceedings of Interspeech 2009 (Brighton)*, 2009, pp. 2519–2522

Schweitzer and Möbius 2003 SCHWEITZER, Antje ; MÖBIUS, Bernd: On the structure of internal prosodic models. In: *Proceedings of the 15th International Congress of Phonetic Sciences (Barcelona)*, 2003, pp. 1301–1304

Schweitzer and Möbius 2004 SCHWEITZER, Antje ; MÖBIUS, Bernd: Exemplar-based production of prosody: Evidence from segment and syllable durations. In: *Speech Prosody 2004 (Nara, Japan)*, 2004, pp. 459–462

Schweitzer and Möbius 2009 SCHWEITZER, Antje ; MÖBIUS, Bernd: Experiments on Automatic Prosodic Labeling. In: *Proceedings of Interspeech 2009*, 2009, pp. 2515–2518

Schweitzer et al. accepted SCHWEITZER, Katrin ; CALHOUN, Sasha ; SCHÜTZE, Hinrich ; SCHWEITZER, Antje ; WALSH, Michael: Relative Frequency Affects Pitch Accent Realisation: Evidence for Exemplar Storage of Prosody. In: *Proceedings of the 12th Australasian International Conference on Speech Science and Technology (Melbourne)*, accepted

Schweitzer et al. 2009 SCHWEITZER, Katrin ; WALSH, Michael ; MÖBIUS, Bernd ; RIESTER, Arndt ; SCHWEITZER, Antje ; SCHÜTZE, Hinrich: Frequency Matters: Pitch accents and Information Status. In: *Proceedings of EACL-09 (Athens, Greece)*, 2009, pp. 728–736

Shockley et al. 2004 SHOCKLEY, Kevin ; SABADINI, Laura ; FOWLER, Carol A.: Imitation in shadowing words. In: *Perception & Psychophysics* 66 (2004), no. 3, pp. 422–429

Silverman et al. 1992 SILVERMAN, Kim ; BECKMAN, Mary ; PITRELLI, John ; OSTENDORF, Mari ; WIGHTMAN, Colin ; PRICE, Patti ; PIERREHUMBERT, Janet ; HIRSCHBERG, Julia: ToBI - a standard for Labeling English Prosody. In: *Proceedings of the International Conference on Spoken Language processing (ICSLP, Banff)*, 1992, pp. 867–870

Sridhar et al. 2008 SRIDHAR, Vivek Kumar R. ; BANGALORE, Srinivas ; NARAYANAN, Shrikanth: Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework. In: *IEEE Transactions on Audio, Speech, and Language Processing* 16 (2008), no. 4

Syrdal and McGory 2000 SYRDAL, Ann K. ; MCGORY, Julia: Inter-transcriber reliability of ToBI prosodic labeling. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP), vol.3*, 2000, pp. 235–238

't Hart et al. 1990 'T HART, Johan ; COLLIER, René ; COHEN, Antonie: *A Perceptual Study of Intonation—An Experimental-Phonetic Approach to Speech Melody*. Cambridge, UK : Cambridge University Press, 1990

Taylor 1998 TAYLOR, Paul: The Tilt Intonation Model. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 1998, pp. 1383–1386

van Santen et al. 2004 VAN SANTEN, Jan ; MISHRA, Taniya ; KLABBERS, Esther: Estimating Phrase Curves in the General Superpositional Intonation Model. In: *Proceedings of the ISCA Speech Synthesis Workshop 04*, 2004, pp. 61–66

Wade et al. 2010 WADE, Travis ; DOGIL, Grzegorz ; SCHÜTZE, Hinrich ; WALSH, Michael ; MÖBIUS, Bernd: Syllable frequency effects in a context-sensitive segment production model. In: *Journal of Phonetics* 38 (2010), no. 2, pp. 227–239

Wahlster 2004 WAHLSTER, Wolfgang: SmartWeb: Mobile Applications of the Semantic Web. In: BIUNDO, Susanne (ed.) ; FRÜHWIRTH, Thom (ed.) ; PALM, Günther (ed.): *KI 2004: Advances in Artificial Intelligence*. Berlin/Heidelberg : Springer, 2004, pp. 50–51

Walsh et al. 2010 WALSH, Michael ; MÖBIUS, Bernd ; WADE, Travis ; SCHÜTZE, Hinrich: Multilevel Exemplar Theory. In: *Cognitive Science* 34 (2010), pp. 537–582

Walsh et al. 2008 WALSH, Michael ; SCHWEITZER, Katrin ; MÖBIUS, Bernd ; SCHÜTZE, Hinrich: Examining pitch-accent variability from an exemplar-theoretic perspective. In: *Proceedings of Interspeech 2008 (Brisbane)*, 2008, pp. 877–880

Whiteside and Varley 1998a WHITESIDE, S. P. ; VARLEY, R. A.: Dual-Route Phonetic Encoding: Some Acoustic Evidence. In: *Proceedings of the 5th International Conference on Spoken Language Processing (Sydney)* vol. 7, 1998, pp. 3155–3158

Whiteside and Varley 1998b WHITESIDE, S. P. ; VARLEY, R. A.: A reconceptualisation of apraxia of speech: a synthesis of evidence. In: *Cortex* 34 (1998), pp. 221–231

Wightman and Ostendorf 1994 WIGHTMAN, Colin W. ; OSTENDORF, Mari: Automatic labeling of prosodic patterns. In: *IEEE Transactions on Speech and Audio Processing* 2 (1994), no. 3, pp. 469–481

Witten and Frank 2005 WITTEN, Ian H. ; FRANK, Eibe: *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd edition. San Francisco, USA : Morgan Kaufman, 2005

Zeißler et al. 2006 ZEISSLER, Viktor ; ADELHARDT, Johann ; BATLINER, Anton ; FRANK, Carmen ; NÖTH, Elmar ; SHI, Rui P. ; NIEMANN, Heinrich: *The Prosody Module*. pp. 139–152. In: WAHLSTER, Wolfgang (ed.): *SmartKom: Foundations of Multimodal Dialogue Systems*. Berlin : Springer, 2006