Reports in Phonetics, University of the Saarland

Berichte zur Phonetik, Universität des Saarlandes

PHONUS

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

No. 14, September 2008

Institut für Phonetik
Universität des Saarlandes
D-66041 Saarbrücken, September 2008

ISSN 0949-1791

FOREWORD

Volume 14 of the PHONUS series presents a doctoral thesis which was written within the framework of the International Postgraduate College in the department of Computational Linguistics and Phonetics and was accepted by the Faculty of Arts, Saarland University in 2007. The work brings together three areas of research at our institute which have been documented in previous issues of PHONUS: speech technology, Slavonic languages and prosody. The thesis is of inherent interest to speech technologically and phonetically interested scientists: It deals with a Slavonic language that has previously been subject to limited instrumental phonetic analysis; it combines phonetic and engineering methods to model the intonation of Polish for use in Speech Synthesis. The thesis illustrates firstly the possibility and problems of 'cleaning' prosodically and segmentally unlabelled recordings to derive stylized intonation contours of sufficient data in order to undertake a statistically robust modelling procedure. Secondly it demonstrates a two-level modelling process to predict the location of accented words in an utterance and to generate subsequently the intonation contour.

William J. Barry and Jürgen Trouvain   Saarbrücken, September 2008

VORWORT

Band 14 der Reihe PHONUS präsentiert eine Dissertation, die im Rahmen des Internationalen Studienkollegs in der Fachrichtung Computerlinguistik und Phonetik erarbeitet und im Jahre 2007 von den Philosophischen Fakultäten der Universität des Saarlandes angenommen wurde. Die vorliegende Arbeit verknüpft drei Forschungsgebiete unseres Instituts, die auch in vorhergehenden Ausgaben von PHONUS dokumentiert worden sind: Sprachtechnologie, slavische Sprachen und Prosodie. Die Arbeit ist aus theoretischer Sicht für sprachtechnologisch und phonetisch Interessierte wichtig: Sie behandelt eine slavische Sprache, die bisher nur begrenzt instrumentalphonetisch untersucht worden ist; sie bringt phonetische und ingenieurwissenschaftliche Methoden zusammen, um die Intonation des Polnischen im Hinblick auf ihre Verwendung in der Sprachsynthese zu modellieren. Die Arbeit zeigt zunächst die Möglichkeiten und Probleme der 'Säuberung' von prosodisch nicht annotierten Aufnahmen, um stilisierte Intonationskonturen in ausreichender Menge zu erhalten, dass eine robuste Modellierung möglich ist. Zweitens demonstriert sie einen zweistufigen Modellierprozess, der die akzentuierten Wörter einer Äußerung voraussagt und danach die Intonationskontur generiert.

William J. Barry and Jürgen Trouvain   Saarbrücken, September 2008

MODELLING POLISH INTONATION FOR SPEECH SYNTHESIS

Dissertation
zur Erlangung des Grades eines Doktors der Philosophie
der Philosophischen Fakultäten der Universität des Saarlandes

vorgelegt von

DOMINIKA OLIVER

Saarbrücken, im Januar 2007

Dekan: Prof. Dr. Michael Hüttenhoff

Berichterstatter: Prof. Dr. William J. Barry und Prof. Dr. D. Robert Ladd

Tag der letzten Prüfungsleistung: 17.07. 2007

Abstract

This thesis describes a Polish prosody modelling module for speech synthesis. The module uses classification and regression trees for pitch accent type and place prediction and a linear regression technique for F0 contour generation. We show how improvements were achieved by the use of perceptually equivalent stylised F0 curve, clustering of accent types using self–organising maps, and language specific features, and describe techniques used to attempt to overcome problems with the available speech data.

In order to get high quality speech synthesis for Polish, a good prosodic modelling needed to be developed. Modelling prosody for diphone-based concatenative speech synthesis involves generating pitch contours based on a linguistic description of a language. The training of a model to do this requires a viable linguistic description of the intonation of the language and a suitably annotated speech database. We outline the characteristics of Polish prosody and describe the linguistic resources available in Polish.

Modelling of prosodic events is made possible when the acoustic parameters associated with them can be clustered into distinct groups. We use methods not presupposing a number of such classes to derive acoustic description of prototypical pitch contours found in Polish. Precise parameters of contour classes are generated from an automatically stylised F0 curve, made using a revised MOMEL algorithm for boundary locations, where it is found to be erroneous. The acoustic basis for most of these errors seems to be a too heavy smoothing of the intonation contour, with many too high initial F0 values and missing peaks and final falls.

We then discuss two models built for use with the Polish voice, a CART model for accent prediction and a Linear Regression model for F0 contour generation. We pay particular attention to the language specific

details of these models and the techniques used to overcome problems with the resources that are available in Polish.

Finally, we present the results of a formal evaluation process designed to test the approach taken in the thesis.

Zusammenfassung

Die vorliegende Arbeit beschreibt ein Modul für die Prosodiemodellierung polnischer Sprachsynthese. Das Modul verwendet Klassifikations- und Regressionsbäume für die Vorhersage von Akzent- und Grenztönen und eine lineare Regressionsanalyse für die Generierung der Grundfrequenz. Durch die perzeptionsnah stilisierte F0-Kurve, das Clustering von Akzenttypen mit Hilfe von Self-Organising Maps und die Verwendung sprachspezifischer Merkmale werden Verbesserungen der Synthese erreicht. Die Arbeit beschreibt außerdem Möglichkeiten zur Behebung eventueller Probleme im vorhandenen Sprachmaterial.

Eine gute Prosodiemodellierung ist die Voraussetzung für eine qualitativ hochwertige polnische Sprachsynthese. Prosodie in einer diphonbasierten konkatenativen Sprachsynthese zu modellieren verlangt nach einer linguistisch basierten Sprachbeschreibung, von der ausgehend der Tonhöhenverlauf generiert wird. Um ein Modell zu trainieren, das dieses vermag, muss eine umfassende linguistische Beschreibung der Intonation sowie eine entsprechend annotierte Sprachdatenbasis vorliegen. Die Eigenschaften polnischer Prosodie werden hier erläutert und allgemein verfügbare polnische Sprachressourcen vorgestellt.

Prosodische Ereignisse zu modellieren wird möglich, wenn die mit ihnen assoziierten akustischen Parameter unterschiedlichen Gruppen zugeordnet werden können. Die in der vorliegenden Arbeit verwendeten Methoden setzen jedoch keine genaue Anzahl solcher Klassen voraus, um eine akustische Beschreibung prototypischer polnischer Intonationskonturen zu erreichen. Bestimmte Parameter mehrerer Konturklassen wurden von einer automatisch stilisierten F0-Kurve erzeugt. Hierzu wurde eine Abwandlung des MOMEL-Algorithmus verwendet, da der ursprüngliche Algorithmus anfällig für das Setzen falscher Grenzmarkierungen ist. Die

akustische Grundlage vieler dieser Fehler scheint ein zu starkes Glätten der Intonationskontur zu sein, wodurch zu hohe initiale F0 Werte und zuviele fehlende Gipfel und finale Tonhöhenabfälle berechnet werden. Zwei Modelle, die speziell für die polnische Sprache entwickelt wurden, werden vorgestellt: das CART-Modell für die Vorhersage des Tonakzents und ein lineares Regressionsmodell für die Generierung der Grundfrequenzkontur. Besonderes Augenmerk wird dabei auf die sprachspezifischen Aspekte der jeweiligen Modelle, sowie auf die Techniken zur Behebung von Fehlern in den verwendeten polnischen Sprachressourcen gelegt. Abschließend werden die Ergebnisse eines formalen Evaluationsprozesses präsentiert, der den für die vorliegende Arbeit verwendeten Ansatz überprüft.

Die vorliegende Arbeit ist wie folgt strukturiert: Kapitel 1 beschreibt den Forschungsgegenstand und den Aufbau der vorliegenden Arbeit. In Kapitel 2 werden die Merkmale polnischer Intonation anhand eines Überblicks über die polnische Intonationsliteratur dargelegt, mit Fokus auf den in der Literatur beschriebenen verschiedenen Klassifikationen für Intonationskonturen. Da frühere Untersuchungen eine freie Wortbetonung für polnische Verben postuliert haben, wurden verschiedene Studien zur Phonetik und Phonologie von Betonung und Akzentuierung durchgeführt, deren Ergebnisse in Kapitel 3 vorgestellt werden. Im Hinblick auf eine mögliche Implementierung in einer polnischen Sprachsynthese wurde untersucht, inwieweit Wortbetonung verlässlich wahrgenommen werden kann und wie sich die akustischen Korrelate dieses perzeptiven Eindrucks manifestieren. In einem Perzeptionsexperiment wurden naiven Hörern Verben mit alternativen Wortbetonungen zur Evaluierung vorgelegt. In Kapitel 4 folgt eine Studie zum Alignment von Gipfeln in der Intonationskontur. Die Motivation dieser Studie war es, ein deskriptives Modell der Platzierung von F0-Gipfeln zu erstellen und die Beziehung zwischen Gipfel-Lokalisierung und segmentaler Verankerung. Die in dieser Studie gestellten Fragen betreffen des Weiteren die phonetische Realisierung des phonologisch spezifizierten F0-Gipfel-Alignment im Polnischen als Funktion der folgenden Faktoren:

Sprechtempo, Informationsstruktur und Struktur innerhalb der Äußerung. Auf der Basis experimenteller Daten untersuchen wir, ob in Bezug auf Peak-Alignment die Informationsstruktur die Wahl und Realisierung des nuklearen Tonakzentes im Polnischen beeinflusst. Kapitel 5 beschreibt die gesprochen-sprachlichen Quellen, die für das Polnische vorhanden sind. Es bietet eine eingehende Analyse zweier Korpora, die bereits für die Analyse und Modellierung polnischer Intonation benutzt wurden. Wir präsentieren Details der akustischen Analyse von Sprachbeispielen aus diesen Korpora und zeigen, wie das Sprachmaterial innerhalb der Arbeit genutzt wird. Dies beinhaltet auch die Details aller vorgenommenen Modifizierungen der Annotationsschemata und ihrer Zusätze.

Kapitel 6 behandelt den Entwicklungsablauf eines Sprachsynthesesystems mit Schwerpunkt auf dem Modul zur Prosodie-Modellierung. Zunächst führen wir die Hauptkomponenten eines Text-to-Speech-Systems ein und gehen dann näher auf die Komponenten eines prosodischen Modells ein, einschließlich Dauermodellierung, Modellierung der Phrasengrenzen, Tonakzent-Voraussage und -Zuweisung und Techniken der Konturgenerierung. Es wird sowohl eine Übersicht über die Studien, die sich der polnischen Prosodie-Modellierung widmen, als auch ein Überblick über vorhandene polnische Synthese-Systeme geboten. Schließlich stellen wir das Forschungsinstrument vor, das in dieser Studie benutzt wurde und zeigen, welche Methoden der Intonationsmodellierung mit ihm möglich sind.

Kapitel 7 befasst sich mit der phonetischen Analyse der F0-Kontur, mit dem Ziel, genaue Parameter prosodischer Ereignisse zu spezifizieren. Eine automatische F0-Stilisierungs-Methode wird präsentiert, die eine präzise akustische Repräsentation der zugrundeliegenden F0-Kontur bei polnischen Tonakzentklassen gewinnen soll. Der Ausdruck wird hier benutzt, um einen Vorgang anschaulich darzustellen, der die gemessene F0-Kontur einer Äußerung in eine einfachere aber funktional äquivalente Form umwandelt, unter Bewahrung der perzeptuellen Identität des ursprünglichen Signals.

Der Beweggrund für die in Kapitel 8 durchgeführte Studie ist das Erfordernis, sowohl eine optimale Zahl an Tonakzentklassen als auch ein automatisches Annotations-Schema für prosodische Ereignisse zu ermitteln. Es wird eine Methode vorgestellt, die mittels nicht überwachter Clustering-Methoden automatisch die Tonakzentklassen für die polnische Prosodie-Modellierung ableitet. Die gewählten Methoden streben an, die systematischen Fehler und die Kosten der manuellen Annotation von Intonations-Datenbanken zu reduzieren.

Kapitel 9 beschreibt die Methodik zur und die Ergebnisse von der Konstruktion von Intonationsmodellen und ihrer Implementierung in ein Text-to-Speech-System. Der Prozess setzt sich aus drei Schritten zusammen: Vorhersage des Akzentortes, Vorhersage des Akzenttyps und Vorhersage der F0-Werte für die Kurvengenerierung. Das implementierte Prosodie-Modul benutzt Klassifikations- und Regressionsbäume für die Vorhersage des Akzentortes und -typs und eine lineare Regressionstechnik für die Generierung der F0-Kontur der entsprechenden Konturen. Beide Vorhersage-Modelle werden durch objektive Tests evaluiert.

In Kapitel 10 schließlich werten wir unter Verwendung der implementierten prosodischen Modelle die generierten F0-Konturen aus. Wir setzen Methoden zur Bewertung von Sprachsynthese mit Schwerpunkt auf Prosodie ein und beschreiben die subjektiv durchgeführten Bewertungstests. Sie bestehen aus einem dreiteiligen Perzeptionsexperiment mit polnischen Muttersprachlern, die Sätze mit einem Default Prosodie-Modell mit solchen mit dem neuen Prosodie-Modell aus Kapitel 9 im Vergleich beurteilen mussten. Schlussfolgerungen, inklusive der Hauptergebnisse der Arbeit, Grenzen und Ausblicke werden in Kapitel 11 beschrieben.

Acknowledgements

First of all, I would like to thank my supervisor Bill Barry for directing my research on intonation and for all his support, suggestions and patience over the years. I owe many thanks to Bob Ladd, on the Edinburgh side, whose critical comments and insights greatly strengthened this work. I also wish to thank my second supervisor Rob Clark who made work with Festival less painful and always had a solution to my technical queries.

The work on this thesis was funded by the International Post-Graduate College Language Technology and Cognitive Systems, a program of the Deutsche Forschungsgemeinschaft (DFG). Id like to say thank you to the organisers Matthew Crocker, Sabine Schulte im Walde, Claudia Verburg, and to my colleagues in the IGK including (in alphabetical order) Núria Bertomeu, Bettina Braun, Irene Cramer, Ralph Debusmann, Péter Dienes, Amit Dubey, Malte Gabsdil, Kerstin Hadelich, Dominik Heckmann, Pia Knöferle, Michael Kruppa, Ulrike Padó, Ingmar Steiner, Kristina Striegnitz, Victor Trón and Olga Uryupina.

I spent a part of my PhD study at the University of Edinburgh and my thanks go to all the staff members and students at CSTR including Sasha Calhoun, Steve Isard, Simon King, Cassie Mayo, Korin Richmond and Mirjam Wester. I have also profited from the collaboration with members from the Psycholinguistics department including Alissa Melinger and Andrea Weber. Many thanks to Marc Schröder from DFKI for adapting his Rating Test program, which I extensively used for my perception experiments.

For more than three years my second home was the department of Phonetics and without the support of everyone there this work would never have reached its fruition. My heartfelt thanks go to Bistra Andreeva, Stefan Baumann, Caren Brinckmann, Martine Grice, Cordula Klein, Jacques Ko-

reman, Andrew Morris, Manfred Pützer, Gudrun Schuchmann and Jürgen Trouvain.

Thanks also go to all the very patient participants (students, colleagues and friends) taking part in my production and perceptual experiments.

I had the pleasure to visit Polish-Japanese Institute of Information Technology, where I was invited by Prof. Krzysztof Marasek, who together with Krzyszof Szklanny have been kindly collaborating with me in the quest for the development of a better Polish text-to-speech synthesiser.

The backbone to the thesis were speech resources on which my prosody models have been built. Special thanks to Maciej Karpiński for valuable discussions and making the Polish Intonation database PoInt available to me. I am grateful to Prof. Ryszard Gubrynowicz for his phonetic insights and supplying the Babel database.

Finally, I must thank my family and close friends who had to put up with me during these years. Especially, I would like to thank my husband Steve and my mother for their endless love, support, and patience.

Contents

List of Figures

List of Tables

1. Introduction

Introduction

Speech Synthesis is a powerful tool for testing fundamental claims about speech. It is a good platform for testing prosodic theories, as they remain abstractions until they have been incorporated into speech synthesisers.

Within the realm of rendering natural sounding speech from raw text, one of the many tasks is generating natural sounding and adequate intonation. The task of generating adequate intonation remains one of the most difficult in text–to–speech synthesis systems to date. As the degree to which intonation is perceived as natural affects the quality of synthetic speech, generating appropriate prosody remains central in the research carried out within many text–to–speech (TTS) systems. As the quality of speech synthesis improves, a greater demand is put on the system to produce more varied intonation contours. A number of intonation theories have been utilised in various systems to try to accomplish this task. My research goal is to generate a model of Polish intonation which can be directly implemented in a speech synthesis system. Moreover, to test the model suitability, the aim is to generate intonation contours by integrating the model into a test–to–speech system and performing a formal perceptual evaluation on the resulting synthesised signal.

1.1. Motivation

1.1.1. Aims and Methods

Intonation generation in speech synthesis can be viewed as a two–stage process; the first aimed at representing grammatical structures and referential relations on a symbolic level and the second at rendering acoustic signals that convey the structural and intentional properties of the message.

After reviewing methods of modelling intonation, the system requirements for building intonation models can be summed up in the three components outlined as follows.

First, a language–specific intonation model needs to be created. Such a model should contain a phonological description of language specific accent types and a phonetic specification of particular pitch accents. While the existing intonation theories can be utilised to build a theoretical framework, additional studies and speech data are required to pinpoint specific aspects which are of importance to the current project, to resolve theoretical issues debated in the literature and to investigate points which have not received much attention till now.

The particular aspects chosen to be investigated in greater detail in the case of Polish are :

- Determining the number and types of pitch accents present in the language in relation to a choice of linguistic theory deemed best to describe the intonation phenomena in Polish,

- Lexical stress assignment and its state of flux in modern Polish in the subset of past tense verb forms,

- F0 peak alignment in different focus and position in a sentence, with aim to provide a descriptive model of peak placement and to analyse the relationship between F0 peak location and its segmental anchors.

In order to carry out a phonetic analysis, speech data is required. The form, style, and content depends on the topic being investigated and in the current study a speech corpus in the form of recordings is essential. The phonetic analysis with the aim of specifying precise parameters of prosodic events will yield reliable results when the speech signal's fundamental frequency is accurately determined by a pitch extraction program, its curve stylised, resulting in a smooth stylised pitch contour. The stylisation of the F0 curve performed on a speech corpus needs to be perceptually meaningful in order to create a parametric description of F0 contour properties characterising each accent type present.

Second, once the language specific components for intonation modelling are in place, the individual intonation events have to be labelled. If a prosodically annotated corpus is not available, an automatic annotation of prosodic events needs to be carried out. Machine learning methods can be used to do this task. Using parameters derived from stylised F0 contours, pitch accent classes can be derived and clustered according to their similarity and distribution in the corpus.

Third, the prosodically annotated data serves as input to the next steps in intonation modelling, namely, building an accent type and placement prediction model, and an F0 contour values estimation model.

In view of natural–sounding prosodic effects achieved through corpus based prosody modelling reported in literature (Dusterhoff et al., 1999; Syrdal et al., 1998), corpus based methods have been chosen to build such models in the current study.

Fourth, the intonation models can be implemented into a test system, where they will be used to predict and generate a synthetic speech signal. Within such a framework, the actual F0 contour generation consists of two stages, each using a different statistical method:

- Prediction of pitch accent place and types,

- Prediction of F0 contour values.

Finally, the quality of a synthesised signal needs to be evaluated. To evaluate the synthetic intonation, both objective and subjective methods are found to be useful. The objective methods can calculate the distance between the synthetic and the natural signal. On the other hand, the subjective or perceptual methods not only give us insight into the perceivable quality of synthetic intonation, but also allow us to directly compare different models of intonation and establish if the speech data used to build the model is adequate in its coverage of intonation contours.

1.2. Overview of the Thesis

1.2.1. Linguistic Aspects of Polish Prosody

The thesis is organised according to the requirements outlined in the above section. In the first part of the dissertation, spanning Chapter 2 to Chapter 5, we concentrate on the linguistic aspects of Polish intonation with aim to establish a language–specific intonation model.

Based on a literature review, in Chapter 2 we outline the characteristics of Polish prosody, presenting a review of research on Polish intonation, and concentrate on different classifications of intonation contours found in literature.

In Chapter 3 we present results of our studies on stress and accent in Polish concerning the phonetics and phonology of lexical stress in Polish verbs. Polish has a strong tendency towards fixed stress on the penultimate syllable. Exceptions are 1st and 2nd person plural past tense and conditional verb forms, to which Polish grammars assign antepenultimate stress. Although optionality has been reported for these forms, the antepenultimate stress is classed as conservative. We investigated how reliably lexical stress can be perceived and, with a view to implementation in speech synthesis, the nature of the acoustic correlates of perceived stress in these verbs. In a perception task naive listeners evaluated the synthesised texts for acceptability of the alternative stress placement.

The lexical stress placement study is followed by a tonal peak alignment study in Chapter 4. The purpose of this study was to produce a descriptive model of F0 peak placement and to analyse the relationship between F0 peak location and its segmental anchors. Questions posed in this study further pertain to the phonetic realisation of the phonologically specified peak alignment in Polish as a function of the following factors: speech tempo, information structure and position within utterance. On the basis of experimental data, we investigate whether the information structure affects the choice and realisation of the nuclear pitch accents in Polish with respect to peak alignment and whether the phenomenon of segmental anchoring, cf. (Arvaniti et al., 1998), (the idea that there exist specific points in segmental structure to which the tones are anchored) can be observed.

The following two hypotheses regarding the factors affecting the variability vs. stability of nuclear peaks are investigated:

1. Different focus types are associated with specific nuclear pitch accents.

2. The peak of the falling vs. rising pitch accent is consistently anchored to specific points in the segmental structure.

Chapter 5 describes the speech resources available for Polish and provides a closer analysis of two corpora which have been used in the analysis and modelling of Polish intonation. We present details of the acoustic analysis of speech samples selected from them and show how the speech material is used throughout the thesis, together with details of any necessary modifications and additions to the corpora annotation schemes.

1.2.2. Prosody in Speech Synthesis

The second part of the dissertation is devoted to prosodic modelling within a speech synthesis system.

Chapter 6 outlines the development cycle of a speech synthesis system with emphasis on a prosodic modelling module. We first introduce the main parts of a text–to–speech system and go into detail about the components of a prosodic model, including duration modelling, phrase break location modelling, pitch accent prediction and assignment, and techniques of F0 contour generation. A survey of studies devoted to prosodic modelling for Polish as well as an overview of existing Polish synthesis systems is presented. Finally, we introduce a research tool used for this study, and present methods of intonation generation available within it.

Chapter 7 deals with the phonetic analysis of F0 contour with aim to specify precise parameters of prosodic events. An automatic F0 stylisation method for obtaining precise acoustic representation of underlying F0 contour within pitch accent classes in Polish is presented. The term is used here to depict a procedure that modifies the measured F0 contour of an utterance into a more simple but functionally equivalent form, preserving the perceptual identity of the original signal.

The motivation behind the study conducted in Chapter 8 is the need to determine an optimal number of pitch accent classes as well as an automatic annotation scheme of prosodic events. A method for automatically deriving pitch accent classes for prosodic modelling in Polish by means of unsupervised clustering methods is presented. The methods chosen aspire to reduce the bias and cost of manual prosodic database annotation. Moreover, it remains theory independent by not predetermining the number or type of accents to be used.

Chapter 9 describes the methodology and results of building intonation models and implementing them into a text–to–speech system. The process is composed of three parts: accent place and type prediction, and prediction of F0 values for curve generation. The implemented prosody module uses classification and regression trees for accent place and type prediction and a linear regression technique for F0 contour generation for these contours. Both of the prediction models are evaluated by use of objective tests.

Finally, in Chapter 10 we evaluate the pitch contours generated using the implemented prosodic models. We introduce methods of speech synthesis assessment with an emphasis on prosody and describe the subjective assessment tests carried out. It consists of three-part perception study involving Polish native speakers who judge sentences using a default prosodic model against sentences produced with the new prosodic model discussed in Chapter 9.

Conclusions, including the thesis main findings, limitations and future directions are presented in Chapter 11.

Part I.

Linguistic Aspects of Polish Intonation

2. Background

Introduction

In this chapter we outline the characteristics of Polish intonation, presenting a review of research on Polish intonation, and concentrate on different classifications of intonation contours presented by different authors. In Section 2.1.1 we present classic Polish intonation description by Jassem and Steffen-Batog, which is followed by revisions of these theories by means of computer assisted methods in Section 2.1.2.

2.1. Previous studies

2.1.1. Intonation description

The analysis and description of Polish intonation has a long history with the most notable analysis by Jassem (1961).

Jassem's description of accent and intonation at the linguistic level is based on the main features of a British-English system developed essentially by Palmer (1922). Jassem (1952, 1954), Kingdon (1958), O'Connor & Arnold (1973) and Cruttenden (1997) followed the tradition, where an intonational phrase (referred to as Tone Unit) is defined in terms of a sequence of (optional) pre-nuclear, (constitutive) nuclear, and (optional) post-nuclear accents. Syllables within an intonational phrase can be stressed or accented. Accented syllables are stressed and pitch prominent. The final accent of an Intonation Unit referred to as the nucleus.

O'Connor and Arnold's constituents of an intonation phrase in English are presented as follows:

[prehead [head [[nucleus] tail]]]

while Jassem's (1952) definition includes the following:

[anacrusis] [[prenuclear intonation [nuclear intonation]]]

In relation to Polish, Jassem states that the intonation phrase consists of an optional prenuclear accent, an obligatory nuclear accent and an optional postnuclear accent. According to this view a Polish phrase includes only one ictic accent, which is the also referred to as nuclear accent, which is phrase final.

The prominent syllable of the nuclear accent and the prominent syllable of the strong prenuclear accent bear a pitch accent. The accent in the pre-nuclear accent is considered to be secondary, whereas the accent in the nuclear accent is called a primary accent. The pre-nuclear accent is also referred to as pre-ictic and post-nuclear accents are called post-ictic accents. The pre-nuclear and the nuclear accents are mainly determined by specific pitch relations, whilst the post-nuclear accent (if present) is essentially durational.

Jassem's theoretical model (Jassem, 1961) assumes the existence of three tone heights, low (L), mid (M), and high (H). This intonation system is divided into six main nuclear tones and four other context dependent ones.

Here is the inventory of nuclear accents (phrase final):

I Low rising,

II High rising,

III Low falling,

IV High falling,

V Low level,

VI High level.

Accent type characteristics in accented syllables in the middle of phrases :

VIIa peak higher than end of previous and beginning of the next syllable,

VIIb follows syllable with a lower height, precedes a syllable with an equally high tone, after which there is a fall.

Accent types on accented syllables (phrase initial) :

VIIIa precedes a syllable with a lower tone (at the beginning),

VIIIb precedes a syllable with an equally high tone, after which there is a fall.

The study has been updated (Jassem, 1987) in an experiment, carried out to explore acoustically the different F0 contours in relation to linguistic and perceptual differences in the intonation patterns. The author recorded eight versions of a simple Polish phrase ”Dobrze”, corresponding to the English ”OK”. The eight versions were realisations of the same phrase with eight different intonation contours : Low Rise (LR), Full Rise (FR), High Rise (HR), Low Fall (LF), Full Fall (FF), Level (L), Low Rise-Fall (LRF), and Full Rise-Fall (FRF). The subjects taking part in the study had to reproduce each of the versions, imitating the intonation of the phrase. The imitations were elicited in three sessions from fifteen listeners who recorded ten versions of each contour, resulting in 1200 tokens. The recordings were analysed by constructing eight element vectors based on eight F0 points per utterance. Contour classification was carried out using quadratic and linear discriminant functions. Over 80% correct classifications revealed strong similarities between HR and FR, between L and LR, and a weaker similarity between the Low and Full Falls and the corresponding Rise-Falls.

The analysis also revealed a strong dissimilarity between Rise-Falls and the Rises, as well as between Full Falls and Level contours.

Another in-depth study of Polish intonation was carried out by Steffen-Batog (1996). The author describes intonation segment classes where an intonation segment is defined as

> a smallest, subfunctional intonation unit which spans an F0 course within a phonetic syllable containing at least one voiced segment.

Each intonation segment, from a perceptual point of view, can be thought of as a building block of a bigger functional unit, an accent group. In the study the division is made into seven intonation segment classes:

I Level, where F0 within the segment is perceptually a constant function of time,

II Rising, where F0 within the segment is perceptually a non-decreasing and non-constant function of time,

III Falling, where F0 within the segment is perceptually a non-increasing and non-constant function of time,

IV Falling-Rising (weak), the course of the segment can be divided into two parts where in the first part

- F0 within the segment is perceptually a function of time which is

a) non-increasing and non-constant,

b) in the second part becomes non-decreasing and non-constant and

c) whose maximum value in part one is lower than maximum value on part two.

V Falling-Rising (strong), the course of the segment can be divided into two parts where in the first part

- F0 within the segment is perceptually a function of time which

a) is non-increasing and non-constant,

b) in the second part becomes non-decreasing and non-constant and

c) whose maximum value in part one is higher than maximum value on part two.

VI Rising-Falling (weak), the course of the segment can be divided into two parts where in the first part

- F0 within the segment is perceptually a function of time which

a) is non-decreasing and non-constant,

b) in the second part becomes non-increasing and non-constant and

c) whose minimum value in part one is higher than minimum value on part two.

VII Rising-Falling (strong), the course of the segment can be divided into two parts where in the first part

- F0 within the segment is perceptually a function of time which

a) is non-decreasing and non-constant,

b) in the second part becomes non-increasing and non-constant and

c) whose minimum value in part one is lower than minimum value on part two.

From a phonological perspective, Steffen-Batog distinguishes intonemes, functional intonation units which are realised in an intonation phrase by way of intonation segments, the same way as phoneme is realised as a phone.

What follows is a division of ’intonemes’ into seven functional units, as above, plus division into

- High,
- Full,
- Low,
- Neutral.

This means that each of the seven intonation segments can additionally be realised as High, Full, Low or Neutral. As a result, Steffen-Batog (1996) distinguishes 28 types of intonation unit types.

2.1.2. Modern modifications of the theory

More recent studies, e.g. (Demenko et al., 1993; Demenko, 1999), using modern computation methods and new speech corpora, have provided a statistical analysis of the above models. The analyses concerned pitch movements in a read and spontaneous Polish text.

The basic claim of the theory is that there exists a finite number of melodic patterns, each pattern forming an intonation phrase. This phrase includes, in Polish, exactly one nuclear accent, which is normally in a sentence final position. It may be preceded by one or more strong prenuclear tunes. The first, or only syllable of the nuclear tune and the strong prenuclear tune bears pitch accent. A prenuclear tune may be preceded by one or more syllables that do not form a strong prenuclear tune, and are therefore unaccented. Such unaccented syllables may also precede a nuclear tune directly. Nuclear and prenuclear tunes can be described in terms of the course of F0 and the alignment of F0 variation with syllabic cores.

The description, based on descriptions within the British tradition of Intonation analysis (Jassem, 1952, 1954), consists of five distinct tone heights. They are marked by xL, L, M, H, xH ranging from extra low to extra high respectively and define acoustic parameters for two prenuclear (H, L) and nine nuclear accents (HL, ML, HM, xL, LM, MH, LH, LHL(MHL), MM).

In her study, Demenko (1999), carried out an acoustic analysis of prenuclear and nuclear contours in Polish as part of a bigger project concerned with analysis of suprasegmental features of Polish for use in speech technology.

The nuclear contours (ictic) have been defined in Table 2.1.

HL	full falling
ML	low falling
HM	high falling
xL	extra low falling
LM	low rising
MH	high rising
LH	full rising
LHL or MHL	rising-falling

Table 2.1.: *Nuclear contours (ictic) (Demenko, 1999)*

The main characteristics of the nuclear accents are: (Demenko, 1999)

HL global maximum at the beginning and minimum at the end, F0 reduced from highest level to lowest,

ML global minimum at the end, F0 reduced from mid to lowest,

HM global maximum at the beginning, F0 reduced from high to mid level,

xL global minimum, F0 reduced from lower mid to below minimal,

LM global minimum at the beginning, F0 increases till mid level,

MH global maximum at the end, F0 increases from mid to highest level, LH global minimum at the beginning and maximum at the end, F0 increases from lowest level to highest,

LHL global maximum and minimum, F0 increases to global maximum and then is reduced to global minimum,

MM level frequency continuum, in mid level, change of frequency before nucleus, no change afterwards.

The main characteristics of the pre-nuclear accents are: (Demenko, 1999)

H is situated in the top of the frequency range

- on the vowel there is an extreme point, containing a tone higher than preceding and following vowel,
- precedes a vowel equally high, after which there is a lowering, and follows a vowel with a higher tone,
- at the beginning of phrase accented vowel precedes a vowel with an equally high tone, after which there is a frequency reduction,
- at the beginning of phase vowel precedes a vowel with a lower tone.

L is situated in the bottom of the frequency range,

- on the vowel there is a local minimum; preceding and following syllables are situated higher,
- at the beginning of phrase L type vowel precedes a vowel with a higher tone.

This classification can be summarised as:

5 tone heights xL, L, M, H, xH,

2 classes of pre-nuclear accents: H (high) and L (low) ,

9 classes of nuclear accents: HL, ML, xL, HM, LM, MH, LH, MM, and LHL(MHL) -
relative to the particular speakers average pitch range.

Demenko's model of intonation for Polish was based on the analysis of isolated utterances as well as excerpts of continuous speech. Concurrent with Jassem's findings, according to this theory, an intonational phrase consists of an optional prenuclear accent and an obligatory nuclear accent. The nine nuclear accents observed in isolated utterances, correspond to four observed in continuous speech : R (rising), F (falling), MM (level), LHL (rising-falling). The practical applications of the study will be described in the context of prosody modelling for text to speech in Chapter 6 Section 6.4.

The interesting point from the acoustic modelling is the right correlate of a nuclear accent. Jassem's acoustic measurements point out that out of duration, intensity, spectral structure and F0, only the last one is linked to an accent.

Dogil (1995a) verified Jassem's results and found that the only correlate of an accent (what he refers to as a primary stress) was the point of highest F0 with a sharp F0 slope. Intensity was not found to be a significant factor and duration correlated only with the secondary stress.

Language	Polish
Pitch Accents	H*L(fall) L*H(rise) L*HL
Boundary Tones	H% %

Table 2.2.: *ToBI based Polish Intonation description, (Bruce et al., 1996)*

Durand et al. (2002) in their study of Yes/No questions in Polish also concluded that no correlation existed between primary stress and intensity.

As for duration, because of the significant F0 shift on the final foot, bearing the main accent and composed of the penultimate stressed and final unstressed syllable, there is a significant final lengthening present.

From the above studies, only Dogil in Bruce et al. (1996) tries to present the Polish intonation system within autosegmental-metrical theory. Table 2.2 shows the ToBI categories found for Polish and table 2.3 proposes a comparison of traditional British descriptions of Polish intonation.

Author	Jassem	Steffen–Batog	Demenko
Pitch	[I,LR] low rise	[II] rise [IV] fall-rise (w)	LM low rise
Accents	[II,HR] high rise	[II]rise	MH high rise
	[III,LF] low fall	[III] fall	ML low fall
	[IV] high fall	[III] fall	HM high fall
	[V,L] low level	[I] level (w)	xL extra low fall
	[VI,L] high level	[I] level (s)	MM level
	[LRF] low rise-fall	[VI] rise-fall (w)	MHL rise-fall
	[FRF] full rise-fall	[VII] rise-fall (s)	LHL rise-fall
	Full Rise (FR)	[II] rise [V] fall-rise (s)	LH full rise
	Full Fall (FF)	[III] fall	HL full fall

Table 2.3.: *Comparison of Polish Intonation descriptions*

2.2. Conclusions

As Demenko's analysis was carried out with speech technology applications in mind, and its acoustically motivated findings are consistent with computer assisted classification of basic Polish contours (Jassem, 1987) and earlier impressionistic findings (Steffen-Batog, 1996; Dłuska, 1964), we take it as a reliable starting point for the validation of the classes of nuclear patterns and for the prosodic modelling undertaken in this study. The study presented here is not going to choose between the British tradition and the autosegmental-metrical framework and any subsequent labels applied to data should be treated as theory-independent.

3. Stress and Accent in Polish

Introduction

In this chapter we present results of our studies on stress and accent in Polish, concerning the phonetics and phonology of lexical stress in Polish verbs, first reported in Oliver & Grice (2003).

Polish has a strong tendency towards fixed lexical stress on the penultimate syllable of a word. Exceptions are 1st and 2nd person plural past tense and conditional verb forms, for which Polish grammars assign antepenultimate stress. Although optionality has been reported for these forms, the antepenultimate stress is classed as conservative. Since these studies are based on introspection or the auditory analysis of few speakers, we first verify these claims, (section 3.1.2), taking the database recordings of 40 speakers of Polish reading a narrative text. This involves investigating how reliably stress can be perceived and, with a view to implementation in speech synthesis, the nature of the acoustic correlates of perceived stress in these verbs.

A second experiment, (section 3.1.3), is carried out in order to shed light on the potential acceptability of using penultimate stress in these verb forms. In a perception task naive listeners evaluate the synthesised texts for acceptability of the alternative stress placement.

3.1. Phonetics and phonology of lexical stress in Polish verbs

3.1.1. Theoretical background

As stated above, Polish has a strong tendency towards a regular primary stress on the penultimate syllable, and a secondary stress on the initial syllable in words containing more than three syllables (Spencer, 1991; Dogil, 1995b). Notable exceptions are 1st and 2nd person plural forms of verbs in the indicative past. According to Polish prescriptive grammars, antepenultimate stress is assigned to these forms (Bąk, 1995), which are built by adding to the past stem elements that are remnants of the auxiliary być ”to be” (Kenstowicz & Kisseberth, 1977).

(1) sprowadzili-śmy
 moved.MPL-aux1.PL
 ”we moved”

In example (1) above, the element ”-śmy” and its equivalent for the 2nd person past tense plural ”-ście” are treated as verbal auxiliary clitics (Franks & King, 2000). Evidence in favour of their clitic status is the fact that they can appear on hosts other than the verb (Dogil, 1987), as in example (2):

(2) Kogo-ście zobaczyli?
 who-aux.2.PL saw.MPL
 ”who did you see?”

Despite a consensus in the prescriptive grammars, there is apparently some optionality in stress placement in the verb forms mentioned above, in that they may receive penultimate instead of antepenultimate stress.

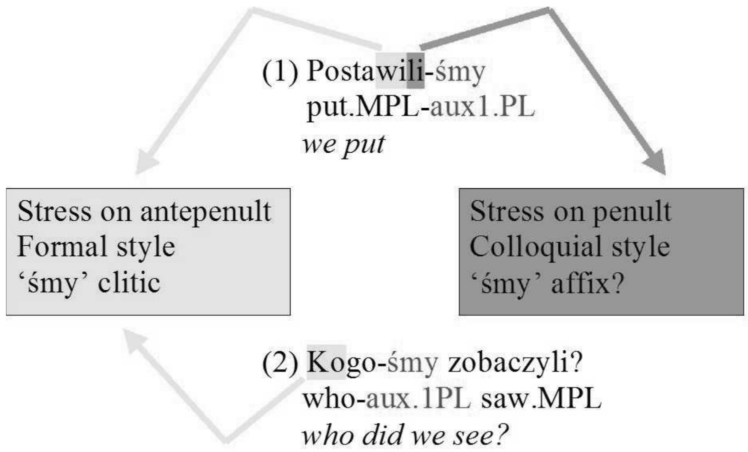


Figure 3.1.: *Stress assignment of 'śmy'*

According to Booij & Rubach (1987), this variation is conditioned both socially and stylistically.

The authors claim that antepenultimate stress is used in the "cultivated form" and penultimate stress in less cultivated speech (Booij & Rubach, 1987, 4:41). They point out that even highly educated speakers use penultimate stress in informal situations. The paradigm is illustrated in Fig. 3.1.

A phonological explanation for the variation is the changing status of the person-number agreement marker (Booij & Rubach, 1987; Kenstowicz & Kisseberth, 1977). For some speakers it might no longer be treated as a clitic, thus not affecting the stress pattern of the word as in example (2), but instead be incorporated into the word as an affix. In the latter case, the person-number marker would count towards the regular penultimate stress rules, shifting the stress one syllable to the right. If, on the other hand, it is an auxiliary clitic, it is considered extrametrical (Franks, 1985), and does not affect the stress pattern of the verb, hence antepenultimate stress.

In the experiments described below we first investigate the extent to which antepenultimate and penultimate stress is used in the verb forms of the type given in example (1) across a number of speakers when reading a literary text, cf. Section 3.1.2. Secondly, we collect listeners' judgements as to the acceptability of penultimate stress across different text registers and investigate whether antepenultimate stress can be taken as evidence for the clitic status of the person-number marker cf. Section 3.1.3.

3.1.2. Stress perception test

Considering that the above mentioned studies are based on introspection, or auditory analysis of few speakers, our first task was to investigate whether speakers place stress in verb forms of the type given in example (1) on the antepenultimate or penultimate syllable and to what extent optionality is found.

Since the perception of stress cannot be taken for granted, especially in a language where stress is not used contrastively, we approached the problem from two angles. First we collected judgements on stress placement from a number of subjects and measured inter-transcriber consistency. Second, we investigated acoustic correlates of stress in terms of duration of and F0 peak alignment on the antepenultimate and penultimate syllables of these verb forms.

We additionally investigated F0 peak alignment relative to the onset of the antepenultimate and penultimate syllables. It is important to point out here that in the corpus we used, nearly all cases of these verb forms had a pitch accent on the stressed syllable, so that in this study we are mainly investigating perceptual and acoustic aspects of accent at the phrasal level rather than of lexical stress alone. It is important to note that Polish is considered a weak stress language and the acoustic correlates of primary stress are highest F0 and a sharp F0 slope on syllables carrying primary stress, in relation to unstressed syllable, as reported by Dogil (1995b).

Method

The experiment made use of a part of Polish Intonational Database (Karpiński & Kleśta, 2001). The set included recordings of 40 speakers of Polish, aged between 19 and 38, all students and academics, reading a literary narrative text. Three sentences containing four verbs in the past tense 1st person plural masculine were selected for analysis. The four verbs were as follows: (3)

a) sprowadzili-śmy
moved.MPL-aux1.PL
”we moved”

b) kupili-śmy
bought.MPL-aux1.PL
”we bought”

c) postawili-śmy
put.MPL-aux1.PL
”we put”

d) wieźli-śmy
transported.MPL-aux1.PL
”we transported”

The corpus contained one rendition of each verb by each of the 40 speakers, constituting 160 tokens. Figure 3.2 shows an example of the recorded material which has been additionally labelled syllabically and prosodically.

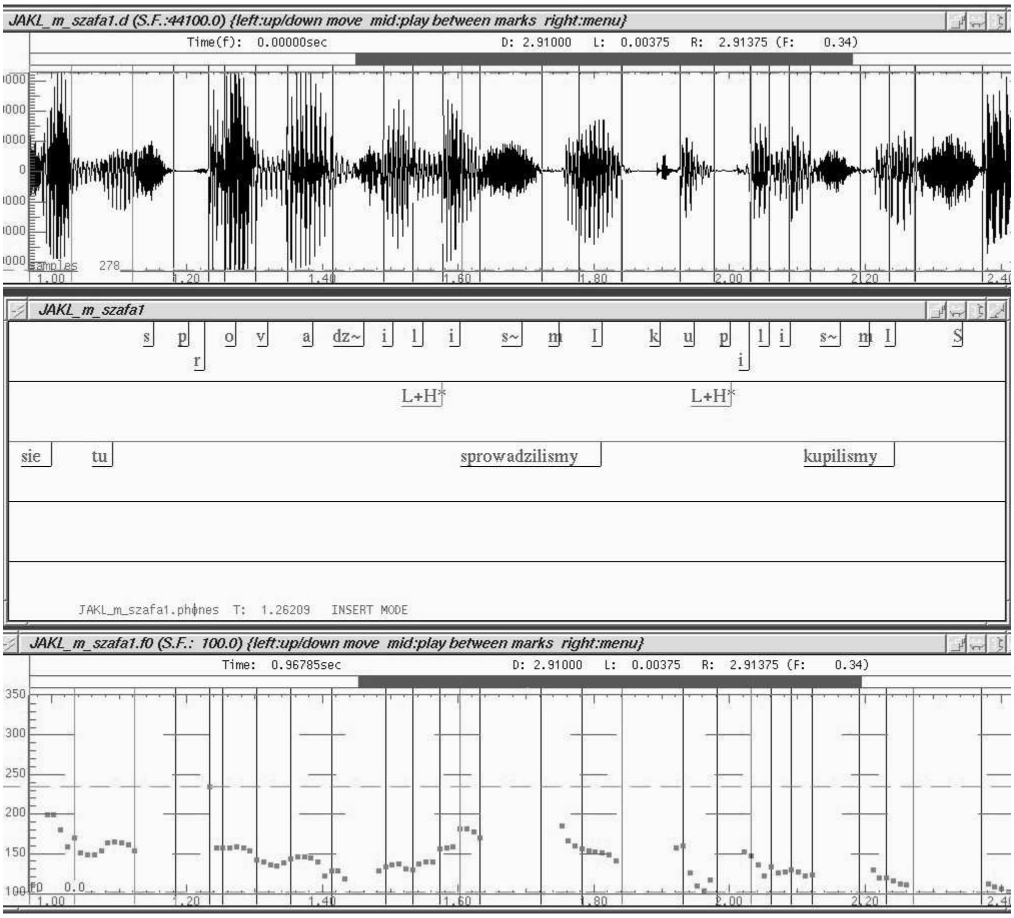

JAKL_m_szafa1.d (S.F.:44100.0) {left:up/down move mid:play between marks right:menu}
Time(f): 0.00000sec
D: 2.91000 L: 0.00375 R: 2.91375 (F: 0.34)
Samples 278
1.00
1.20
1.40
1.60
1.80
2.00
2.20
2.40
JAKL_m_szafa1
s
p
r
o
v
a
dz~
i
l
i
s~
m
I
k
u
p
i
l
i
s~
m
I
S
L+H*
L+H*
sie
tu
sprowadzilismy
kupilismy
JAKL_m_szafa1.phones T: 1.26209 INSERT MODE
JAKL_m_szafa1.f0 (S.F.: 100.0) {left:up/down move mid:play between marks right:menu}
Time: 0.96785sec
D: 2.91000 L: 0.00375 R: 2.91375 (F: 0.34)
350
300
250
200
150
100
F0 0.0

Figure 3.2.: *Example sentence including target verbs: 'sprowadziliśmy' (we moved) and 'kupiliśmy' (we bought).*

Six native speakers of Polish, aged between 22 and 36, all students and academics, participated in the test. Subjects listened to recordings of one sentence at a time and were instructed to indicate the syllable on which they heard stress in the selected verb forms. Three subjects performed the test in the Saarbrücken phonetics lab and three over the Internet.

The following sentences from the database were used for the listening test:

1. Kiedy się tu sprowadziliśmy, kupiliśmy szafę.
   ”when we moved here, we bought a wardrobe.”

2. Najpierw postawiliśmy ją w korytarzu.
   ”first, we put in in the hall.”

3. Miała dwoje drzwi ozdobionych roślinnym ornamentem, a trzecie były oszklone i w szybie odbijało się całe miasto, gdy wieźliśmy ją wynajętą bagażówką.
   ”it had two doors with floral ornaments, and the third was made of glass in which the whole city was reflected while we were transporting it in a hired van.”

The stimuli for which there was total inter-transcriber agreement were used for acoustic analysis. They were divided into two groups, those with antepenultimate and those with penultimate stress.

Since syllabification is not clear-cut (Rubach & Booij, 1990), the sequence ’li-śmy’ was syllabified in two alternative ways: one where the alveolo-palatal fricative [ɕ] is a part of the final syllable ’śmy’, and one where it is part of the penultimate ’li’.

Duration measurements were taken for all stimuli of the whole verb, and its antepenultimate and penultimate syllables (according to both syllabifications). Syllable durations were then expressed as a percentage of the total word duration in each case. The duration values of the same syllable (antepenultimate or penultimate) in stressed and unstressed cases were directly compared.

The F0 peak alignment was measured in relation to the onset of the antepenultimate syllable, whether stressed or not. Additionally, the position of the F0 peak was measured as a percentage value relative to the duration of the stressed syllable.

Results

Auditory analysis by six independent transcribers showed that less than a quarter of the database recordings (9/40) had antepenultimate stress in the 1st person plural past tense verb forms. Speakers were consistent as to which form they used. Stress assignment was affected by neither age, sex nor educational qualifications (to the limited degree these could be investigated, given the homogeneity of the corpus, cf. Chapter 5.2).

Inter-transcriber agreement was high. There was complete agreement across all six transcribers in 93% of cases. Agreement was even higher (95%) when word stress was manifested as a nuclear peak accent (L+H*)(Silverman et al., 1992) (Fig. 3.4 and Fig. 3.5). Transcriber disagreement occurred mainly in prenuclear cases.

In cases of total transcriber agreement, differences in F0 peak alignment in relation to the onset of the antepenultimate syllable were found to be highly significant ($p < 0.001$). The F0 peak was aligned late in the stressed syllable. On average, it was aligned at 93% of the duration of the stressed syllable (87% for antepenultimate syllable, 99% for penultimate syllable, although the measurements for the penultimate syllable are less reliable due to the microprosodic influence of the fricative). For this analysis the syllabification was ' li.śmy'.

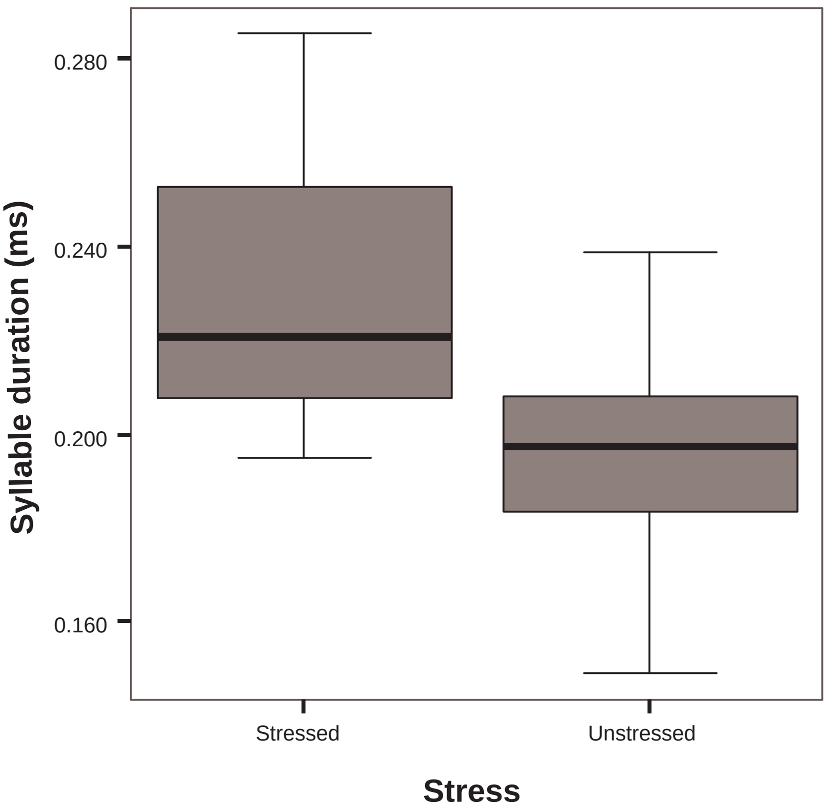


Figure 3.3.: *Mean syllable duration in stressed and unstressed antepenultimate syllables*

Additionally, both antepenultimate and penultimate syllable durations were significantly different in the stressed and unstressed cases (($p < 0.001$) and ($p < 0.05$) respectively). The results are equally significant regardless of whether [ɕ] is treated as belonging to the penultimate syllable or the final one. Taking the syllabification 'li.śmy', stressed antepenultimate syllables were on average 15% longer than unstressed ones, see Fig. 3.3, and stressed penultimate syllables were on average 5% longer than unstressed ones. With the syllabification 'liś.my', the stressed penult was 9% longer.

Figures 3.4 and 3.5 show examples of F0 traces of the sentence containing the verb 'postawiliśmy' (see example 3c) with different stress assignment spoken by two different speakers. Syllable boundaries are indicated by vertical lines and shaded areas represent the stressed (accented) syllable: penultimate in Fig. 3.4 and antepenultimate in Fig. 3.5.

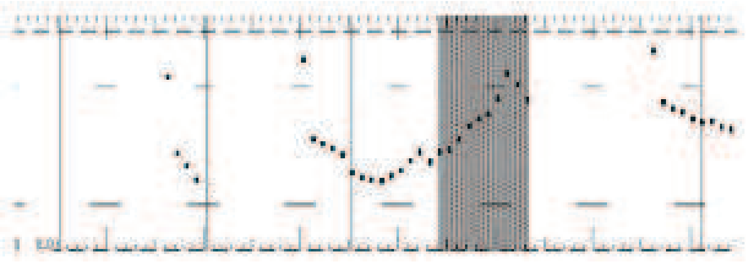


Figure 3.4.: *F0 trace for an example of penultimate stress on postawiliśmy "we put".*

3.1.3. Stress placement acceptability test

In the second part of the study we attempt to verify the claim in Booij & Rubach (1987) that stress placement is inter alia conditioned by stylistic factors. To do this we chose two text types: excerpts from literary texts, representing a formal text style, and sports reports, representing the more colloquial end of the scale. We investigate how far text register influences acceptability judgements when we change the stress assignment in these sentences.

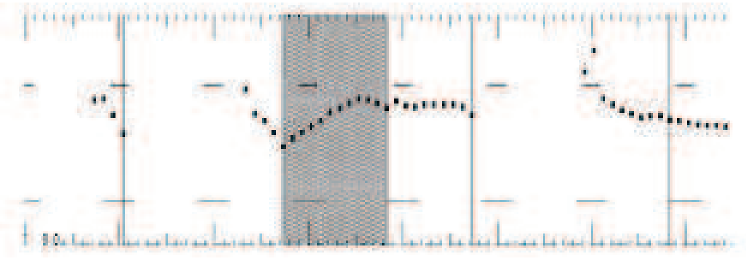


Figure 3.5.: *F0 trace for an example of antepenultimate stress on postawiliśmy "we put".*

An additional experiment investigates the status of the past tense person-number marker. We test the hypothesis formulated in Kenstowicz & Kisseberth (1977) that the difference in stress assignment (antepenultimate or penultimate) implies that the person-number marker is either treated as a clitic or not, respectively.

We tested these hypotheses by constructing synthetic speech material. Although the use of speech synthesis brings with it a considerable reduction in naturalness, it has the advantage of ensuring that only one aspect of the speech is manipulated, namely stress placement.

Method

We constructed stimuli using the Polish module (Oliver, 1998) of the Festival speech synthesis platform (Black & Taylor, 1998). Festival is a general multilingual speech synthesis system developed at the Centre for Speech Technology Research at the University of Edinburgh. The Polish module uses concatenative diphone based synthesis. The system offers a full and flexible environment for the development of language specific modules. By default, Polish stress rules in Festival always assign stress on the penultimate syllable. For the purpose of this experiment, a user dictionary was modified by hand to allow a change in the stress assignment on the verbs of interest.

Ten native speakers of Polish took part in the experiment. The Magnitude Estimation paradigm (Lodge, 1981; Cowart, 1997; Keller, 1998, 1999), was used to elicit acceptability judgements. This method has proved a reliable measure of judgements producing statistically significant results (Bard et al., 1996). For the presentation of stimuli the WebExp software was used (Keller et al., 1998).

In the first part of the test, sentences from two text types (literary and sport) were presented. Stimuli varied as to the stress assignment on past tense 1st person plural verb forms, i.e. antepenultimate or penultimate. For the second part, sentences presented had a person-number marker attached as a clitic to a clause initial constituent of the type given in section 3.1.1, example (2), and the rest of the sentence varied as to whether the

1st person plural verb was stressed on the antepenultimate or penultimate syllable. This means that by attaching the person-number marker to a host other than a verb it was clearly treated as a clitic in the first part of the sentence. The data for both parts of the experiment together consisted of a set of 60 sentences. Two pseudo randomisations, including 60 fillers, were used, each for half of the subjects.

The experiment was administered as follows: first, a training phase was carried out, in which subjects were familiarised with the magnitude estimation task. This was followed by a practice phase, in which subjects heard synthesised stimuli, and finally the experiment proper. No modulus was provided i.e. subjects did not evaluate each sentence in relation to one model token (modulus) but always in relation to the previous item. In this latter phase, subjects evaluated the synthesised stimuli in a self-paced experiment. The subjects were also instructed not to pay attention to the segmental quality of the synthesis nor to word choice, but instead were asked to judge the acceptability of the pronunciation of the sentences. They were not explicitly asked about the acceptability of the stress placement.

In order to compensate for individual differences in scaling across participants the data was normalised. This was done by first log-transforming the scores and then calculating z-scores to create a common scale for all subjects. The result is that the participants' scores are all on the same scale with a mean of 0, Standard Deviation of 1 and the remaining between-condition variance is symmetrically distributed around zero, without affecting differences between conditions.

Results

Taking all the stimuli together, penultimate stress was preferred and the differences in acceptability scores between antepenultimate and penultimate stress were significant ($p < 0.01$).

An ANOVA analysis revealed a significant interaction between text type (literary/sport) and stress (antepenultimate/penultimate) ($p < 0.01$). No significant order effect was found for stress preference in either text type. Further tests were carried out to explore effects within each text type. In the text type 'literature', penultimate stress was significantly preferred ($p < 0.01$) over antepenultimate stress. For text type 'sport' there was a tendency to prefer antepenultimate stress, although it was not significant.

In the second part of the stress placement acceptability test, an ANOVA analysis was performed to investigate the relationship between acceptability of tokens containing a fronted person-number marker (with a clear clitic status), followed in the next clause by a verb with either antepenultimate or penultimate stress. In this part of the experiment no significance was found for the preference of either stress pattern.

3.1.4. Discussion

The experimental findings from the first experiment show that despite the fact that all speakers had at least university level education and were reading a formal style text, penultimate stress was the predominant stress pattern produced on the 1st person plural past tense verb forms.

Where lexical stress was reliably perceived by all transcribers, it corresponded to phrase level accent, with significant peak alignment and duration differences between stressed and unstressed syllables.

Although subjects' acceptability judgements were sensitive to text register, this sensitivity showed, contrary to indications in the literature, that the preference for penultimate stress was stronger on formal literary texts. The tendency for subjects to prefer antepenultimate stress on sports reports, albeit insignificant, is left unexplained.

Despite the limited data analysed, the study indicates that not only do the majority of speakers assign penultimate stress in these verb forms when reading aloud, but that subjects also found penultimate stress acceptable when it was implemented in speech synthesis, even on formal texts. These results thus have implications for speech synthesis stress placement rules. It appears, at least in formal texts, that penultimate stress can be assigned to these verb forms across the board. Nevertheless, further research is needed to investigate how synthesis systems should deal with the more colloquial texts.

3.2. Conclusions

This chapter explored the usage of antepenultimate and penultimate stress in 1st person plural past tense verb forms in Polish. Although Polish generally has penultimate stress, prescriptive grammars treat these verb forms as an exception, assigning antepenultimate stress. It has been argued elsewhere that penultimate stress is possible in these forms, albeit in colloquial speech.

Data from 40 native speakers confirm that penultimate stress is used, but reveal that it is not restricted to colloquial speech: more than three quarters of speakers used penultimate stress when reading a literary text. Stress was auditorily transcribed by six independent judges. High inter-transcriber agreement was achieved when stress coincided with accent, the acoustic correlates of which were found to be F0 peak and syllable duration. A second experiment, using speech synthesis, revealed that the predominant stress pattern for reading was also the preferred pattern for speech synthesis, even on formal literary texts.

Moving the stress to the penultimate syllable in these verbs forms seems to be a strong trend among Polish speakers, and needs to be tested empirically before drawing conclusions about peak placement being a stronger cue than duration. Further tests are also needed before decisions about changing stress assignment in speech synthesis systems can be made.

In the next chapter, we continue the investigation of stress and particularly pitch accent types and the relationship between F0 peak location and its segmental anchors. Questions posed pertain to the phonetic realisation of phonologically specified peak alignment in Polish as a function of the following factors: speech tempo, information structure, and position within utterance.

4. Accent Types and F0 Peak Alignment

Introduction

In this chapter we present the results of our studies on F0 Peak Alignment, Pitch Accent types and Information structure. The purpose of study was to produce a descriptive model of peak placement and to analyse the relationship between peak location and its segmental anchors. Questions posed further pertain to the phonetic realisation of phonologically specified peak alignment in Polish as a function of the following factors: speech tempo, information structure, and position within utterance. We investigate whether information structure affects the choice and realisation of nuclear pitch accents.

4.1. Accent Types and F0 Peak Alignment

For over thirty years the levels-vs.-configuration debate has been a controversial issue for research in intonational phonology. Earlier traditions, such as that of the British school, e.g.,(Crystal, 1969; O'Connor & Arnold, 1973) describe the distinctive units of intonation in a holistic way in terms of complex tonal movements of the contour. These configurations or movements (fall, rise, rise-fall etc.) constitute the principal intonational feature of the intonation unit.

The autosegmental-metrical approach to intonation (see Ladd (1996) for an overview) argues against configurations as primitives of linguistic analysis and analyses the intonation contour as a sequence of phonological level tones such as H(igh) and L(ow) or a combination of the two, occurring at specific structural positions in the utterance. For example, a rising F0 movement is, in this view, taken merely as a transition from its beginning point (F0 minimum value) to its ending point (F0 maximum value). It is usual to refer to these points as tonal targets which can be defined in terms of two dimensions, alignment and scaling.

A number of recent studies have reported that pragmatic information such as information structure and sentence mode play a crucial role in the timing of tonal alignment (Kohler, 1987; Misheva, 1991; Frota, 2000), among others. In these studies the tonal targets are claimed to appear at different locations with respect to the segmental references, depending on the sentence's communicative meaning (e.g., broad vs. narrow focus, contrastive vs. non-contrastive focus, statement vs. question). Other studies, (Silverman & Pierrehumbert, 1990; Arvaniti et al., 1998; Ladd et al., 2000), among others, have suggested that the specification for the alignment of tonal targets is a function of speech tempo, phonological vowel length, syllabic structure and segmental effects (intrinsic vowel duration, consonant voicing etc.), adjacency to word and intonational boundaries, as well as proximity to other tones.

Previous studies have suggested that timing of F0 peaks is a function of: speech tempo, durations of segments in the accented syllable, distance in syllables from the accent to the word boundary and the next stressed syllable, focal structure, as well as sentence mode (Steele, 1986; Steele & Altom, 1986; Silverman & Pierrehumbert, 1990; Prieto et al., 1995; Rietveld & Gussenhoven, 1995; Prieto, 2004; Misheva, 1991; Ladd et al., 2000). Theories of the phonological structure of intonation suggest that there exist "segmental anchors" to which the tones are aligned. The idea of the "segmental anchoring" of F0 movements, is based on the notion that both the beginning (L) and the end (H) of a rising pitch accent are temporally anchored to specific points in segmental structure (Arvaniti et al., 1998). Similar studies were carried out for languages other than Greek, e.g. English (Ladd et al., 1999), Dutch (Ladd et al., 2000). The results indicate that the temporal alignment of both the start and the end of an F0 movement is defined with respect to landmarks in the segmental string and maintains an invariant F0 excursion size.

Therefore the anchors would be closer together at faster speaking rates, and the movement should therefore be shorter and steeper. In other words, the distances from the valleys and peaks to syllable onset and offset respectively would not differ across tempi.

Although this would be the case both if absolute or relative distances to the 'anchor points' remained constant, studies comparing these two measures have shown that, in fact, the relative distances remain more constant. This is tantamount to an orientation to the syllable as a whole.

Within this view our hypothesis is that, the peak of the falling and rising pitch accent would be jointly anchored to particular points in the segmental structure. The peak alignment, measured as an actual proportion of the syllable/rhyme length (relative alignment), should not be affected by increasing speech rate and the resulting shorter duration of the accented word. On the other hand, the absolute distance of the peak from the syllable/rhyme onset/offset should differ significantly with changing speech rate.

4.1.1. Aim

This study was conducted within the framework of a cross-language study which compared Polish and Bulgarian with respect to inter- and intra–language differences in F0 peak alignment. The details of the results are included in Oliver & Andreeva (2008) and Andreeva & Oliver (2005).

Using experimental data, we investigate whether information structure affects the choice and realisation of nuclear pitch with respect to peak alignment and whether segmental anchoring can be observed in Polish.

The following two hypotheses regarding the factors affecting the variability vs. stability of nuclear peaks are investigated:

1. Different focus types are associated with specific nuclear pitch accents.
2. The peak of the falling vs. rising pitch accent is consistently anchored to specific points in the segmental structure.

4.1.2. Experimental design

Material and Procedure Since we are primarily interested in the contribution of intonation for signalling focus, canonical word order was used for the test sentences, i.e.,

subject < verb < direct object < indirect object < oblique

This increases the role of intonation as an information-structuring factor, allowing us to analyse the realisation of focus-associated accent patterns statements with respect to the F0 peak alignment, independent of syntactic structure. Moreover, we designed the material (cf. the four test sentences below in Section Speech material) in such a way that there are one to four unaccented syllables between the metrically strong syllables, with the same maximally sonorant segmental structure ('ma') in order to avoid micro-prosodic effects.

Speech material The following four sentences were recorded six times per focus condition in a random order at normal and fast speech rate in a sound treated studio at the Institute of Phonetics (University of Saarland).

1. 'mama 'ma te'maty.
   mother has topics.
   'Mother has topics.'

2. ’mama wy’maga te’matu.
   mother requires topic.
   ’Mother requires a topic.’

3. a’mator nas na’mawiał do te’matu.
   amateur us urged to topic.
   ’The amateur urged us to accept the topic.’

4. a’mator nam niedo’magał przy Kaza’matach.
   amateur us felt unwell in Kazamaty.
   ’The amateur felt unwell in Kazamaty.’

In total there were 576 stimuli, 288 utterances per speaker set out as follows:

6 target word per focus x 4 test sentences x 2 tempi x 6 repetitions x 2 speakers

Speakers In this experiment, the subjects for the production experiment were two tertiary-level educated female speakers of standard Polish. One was selected from among the Erasmus/Socrates students at Saarland University and the other was a university employee.

Procedure We used two methods for eliciting different focus types. In one, the test sentences were embedded in dialogue sequences as replies to wh-queries uttered by the instructor (a) about the entire utterance, resulting in broad focus, and (b) directed towards the initial or final word, resulting in a narrow non-contrastive focus in the respective position. In the other, narrow contrastive focus condition (c) we embedded the test sentences in dialogues consisting of yes/no-query uttered by the instructor towards the initial, middle or final word, and a correcting reply by the subject. Thus we span broad, narrow and narrow contrastive focus respectively, e.g.:

1. Co dzisiaj nowego?
   'What's new today'?
   [F Mama ma TEMATY]

2. Co ma mama?
   'What has mother got?'
   Mama ma [F TEMATY]

3. Czy mama ma streszczenia?
   'Has mother got summaries?'
   Nie, mama ma [F TEMATY]

The subjects were not informed about the purpose of the study. No explicit instructions regarding accentuation were given to the subjects. In order to elicit broad, narrow non-contrastive, and narrow contrastive focus, the test sentences were embedded in dialogue exchanges as replies to wh-queries uttered by the instructor and directed towards the first, second, or last content word cf. Table 4.1. The subjects only had the test sentences on a sheet of paper and had to read them out in a manner that most suited the instructor's query, i.e. their reaction was elicited by the way the authors presented the material.

The subjects produced the sentences six times in random order at a normal and fast speech rate in a sound-treated studio at the Institute of Phonetics at Saarland University. The recordings were digitised at a sampling frequency of 16 kHz and with an amplitude resolution of 12 bits, and signals F0 values were obtained using the Advanced Speech Signal Processing Tool (xassp) (IPDS, 1997). Using xassp, the signal was manually segmented at the level of the speech sounds for target words using the synchronised microphone signal and a spectrogram, and labelled according to a slightly modified SAMPA (Wells, 1997) system. (cf.Fig. 4.2).

	broad	narrow initial	narrow medial	narrow final
statements [-contrast]	x	x		x
statements [+contrast]		x	x	x

Table 4.1.: *Realised focus conditions for four sentence modes (x indicates used focus positions)*

In addition to the segmental labelling, the pitch accents, phrase accents, and boundary tones were annotated (Silverman et al., 1992) following ToBI labelling instructions (Beckman & Ayers, 1994), with the peak alignment of the L(ow) and H(igh) targets explicitly specified, cf. Fig. 4.2. The positions of the F0 maxima and minima were double-checked by an automatic procedure for which the Praat pitch tracker was used. The window for the automatic procedure was the target word, and the autocorrelation Pitch extraction method was used (time step 0.01 s, frequency range 75 - 550 Hz). In cases where during the manual labelling, the F0 peak was found outside the word e.g. the manual annotation was always selected. The peak delay was calculated (a) as the absolute distance in time from the F0 peak to syllable onset, syllable offset, and rhyme onset, and (b) as the proportion of the rise/fall duration relative to the syllable or rhyme duration. To verify the labelling, the authors of Polish and Bulgarian language parts analysed random utterances from both languages and compared each other's analyses.

Measurements A number of studies have indicated that the following measurements may be significant in terms of segmental anchor points for F0 peak position (Atterer & Ladd, 2004; Prieto et al., 1995; Schepman et al., 2006). Thus, we calculated peak delay as a distance measure from the F0 peak to:

- Syllable onset,
- Syllable offset,
- Rhyme onset.

The peak delay was calculated as the absolute distance in time from the F0 peak to syllable onset (H - C1), syllable offset (H - C2) and rhyme onset (H - V1). Due to the possible effect of the varying segmental durations on peak delay, the above absolute measures were also converted to relative, taken as a proportion of syllable (C2 - C1) and rhyme duration(C2 - V1) (cf. Fig. 4.1). Additionally, the maximum F0 value of the pitch target was measured.

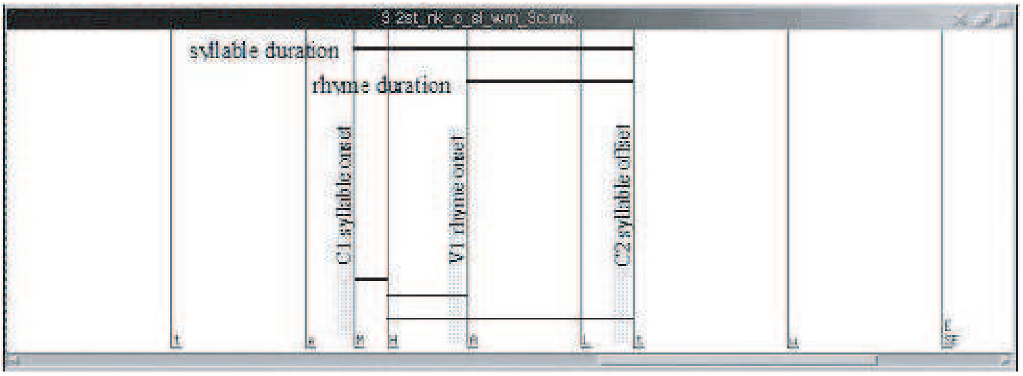


Figure 4.1.: *Example of distance measures taken*

The following points were labelled in each sentence:

- C0 - the beginning of the word initial consonant,
- V0 - the beginning of the word initial vowel,
- C1 - the beginning of the initial consonant of the target syllable,
- V1 - the beginning of the initial vowel of the target syllable,

- C2 - the beginning of the initial consonant of the syllable following the target syllable,
- V2 - the beginning of the initial vowel of the syllable following the target syllable,
- E - end of target word,
- SA - beginning of target utterance,
- SE - end of target utterance,]
- H - point of F0 maximum in the target word,
- L - point of F0 minimum in the target word,

In all our utterances C1 = /m/, V1 = /a/, and were marked with capital letters 'M' and 'A' as seen in Fig. 4.2.

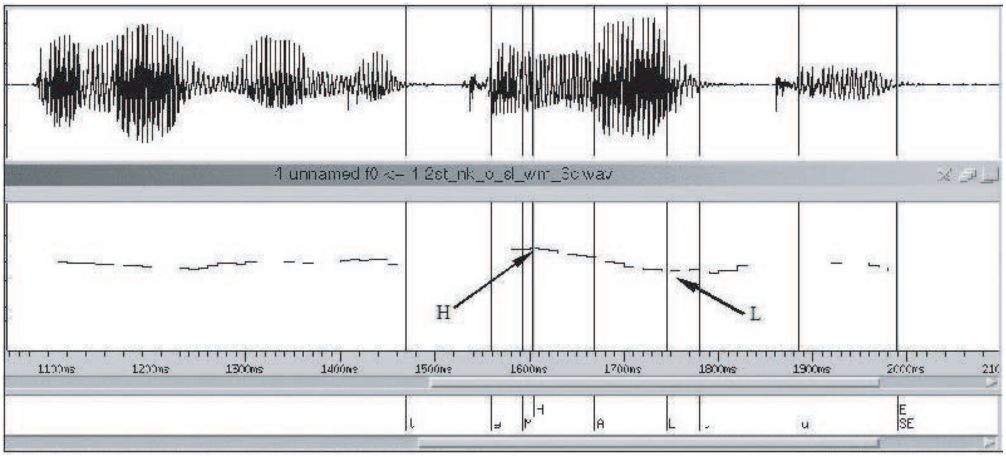


Figure 4.2.: *Example of a labelled target word 'tematu' together with its waveform and F0 contour*

4.1.3. Focus-driven peak alignment

The framework adopted in the present study is Pierrehumbert's autosegmental-metrical model of intonational phonology (Pierrehumbert, 1980). The phonological correlate of focus is a pitch accent which is associated with one of the prominent syllables. Before presenting the results, we wish to call attention to the different strategies used by the subjects in the realisation of the sentences under different focus conditions. The number of pitch accent types used in the different test conditions by the speakers is summarised in Table 4.2. The boundary tones in the test sentences are realised as L-L%.

With respect to the acoustic properties of H* and L+H*, there are conflicting views in the intonational research literature concerning whether these accents are categorically different or just two extremes of a simple accent type. Contrary to claims by Pierrehumbert (1980) and Pierrehumbert & Hirschberg (1990) that only L+H* can be preceded by a low target, Ladd & Schepman (2003) provide statistical evidence that this is also true for H*. A related issue is whether these two accents are associated with different meanings. With regard to the Bulgarian data we can argue that the domain of interpretation of H* and L+H* overlap. Both accent types can signal either new information or a presence of contrast.

Speaker	Focus	Accent type							
		!H+L*		H+L*		H*+L		L+H*	
		norm	fast	norm	fast	norm	fast	norm	fast
WM	broad	12	15	7	6	5	3	0	0
	non-contr	3	0	7	14	23	20	15	14
	contrastive	0	0	0	0	47	46	25	26
KA	broad	0	0	0	0	12	12	0	0
	non-contr	0	0	0	0	12	12	12	12
	contrastive	0	0	0	0	12	12	23	23

Table 4.2.: *Accent types used by speakers WM and KA in different focus conditions*

In the data fro Polish we observe four different pitch accent types: !H+L*, H+L*, H*+L, and L+H*. The first three are phonetically realised as a fall with an early peak aligned at different positions with respect to the accented syllable. The fourth one (L+H*) represents a rising movement within the vocalic nuclei. The two Polish subjects differed in their choice of pitch accent type across focus conditions. For example, in the broad-focus condition at both speech rates, only speaker WM used !H+L* and H+L* accents. These accent types were realised as a fall from a high target in the preceding syllable to a low target situated just after the rhyme onset. The difference between the two pitch accents is that the peak in the down-stepped one is perceived as lower in comparison to the preceding high target in the utterance cf. Fig. 4.3 and Fig. 4.4.

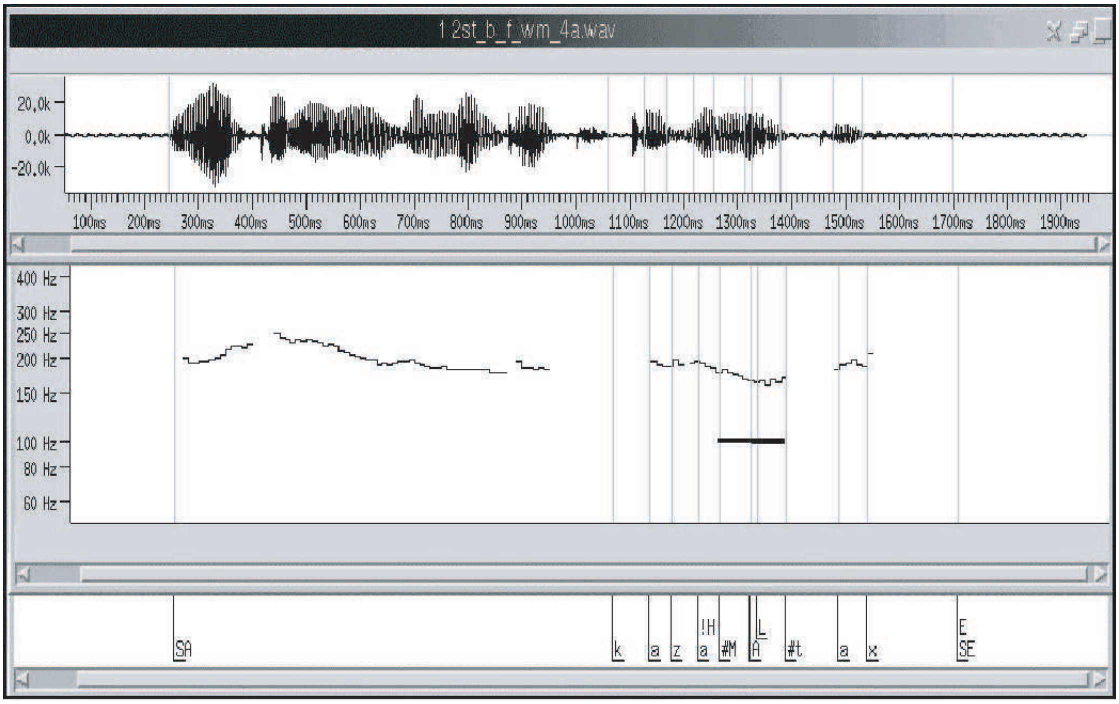


Figure 4.3.: *Realisation of !H+L* by speaker WM*

In contrast to speaker WM, speaker KA used H*+L in broad-focus condition. This accent type was also used by both speakers in the narrow non-contrastive focus condition (there were also seven realisations of !H+L* by speaker WM). In comparison to H+L*, the high target of H*+L is aligned later, just after the rhyme onset (cf. Fig. 4.5 and Fig. 4.6).

When narrow non-contrastive focus is on the final content word in an utterance, the speakers had to disambiguate between this condition and a broad focus with the focus exponent in the same position. While

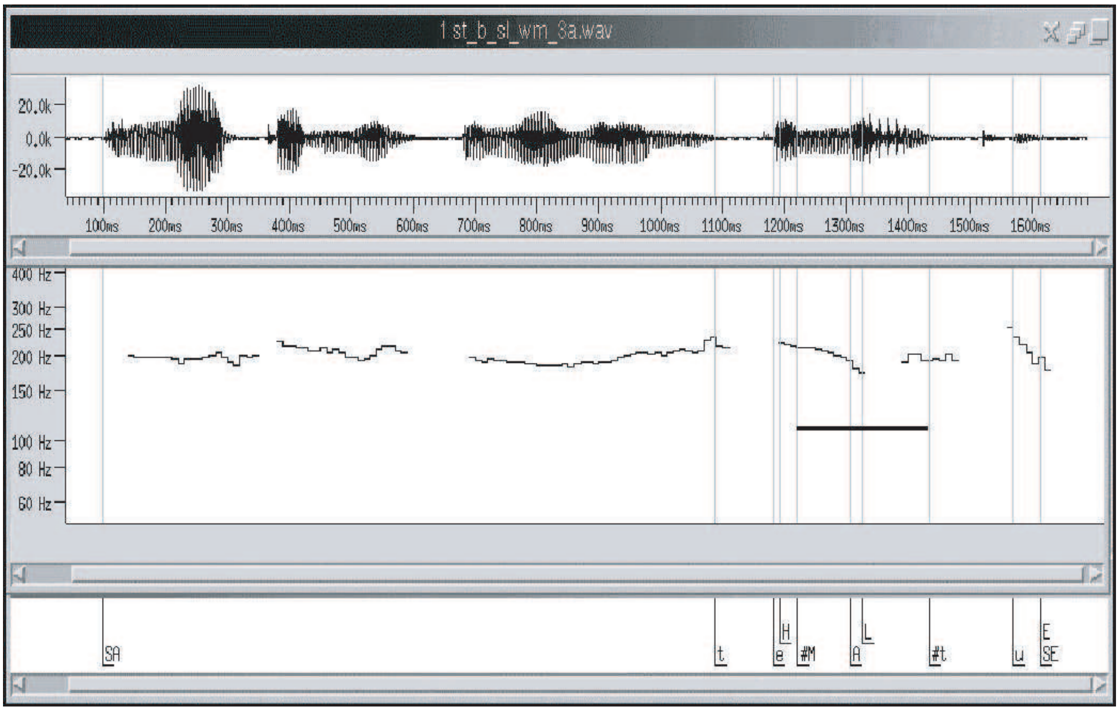


Figure 4.4.: *Realisation of H+L* by speaker WM*

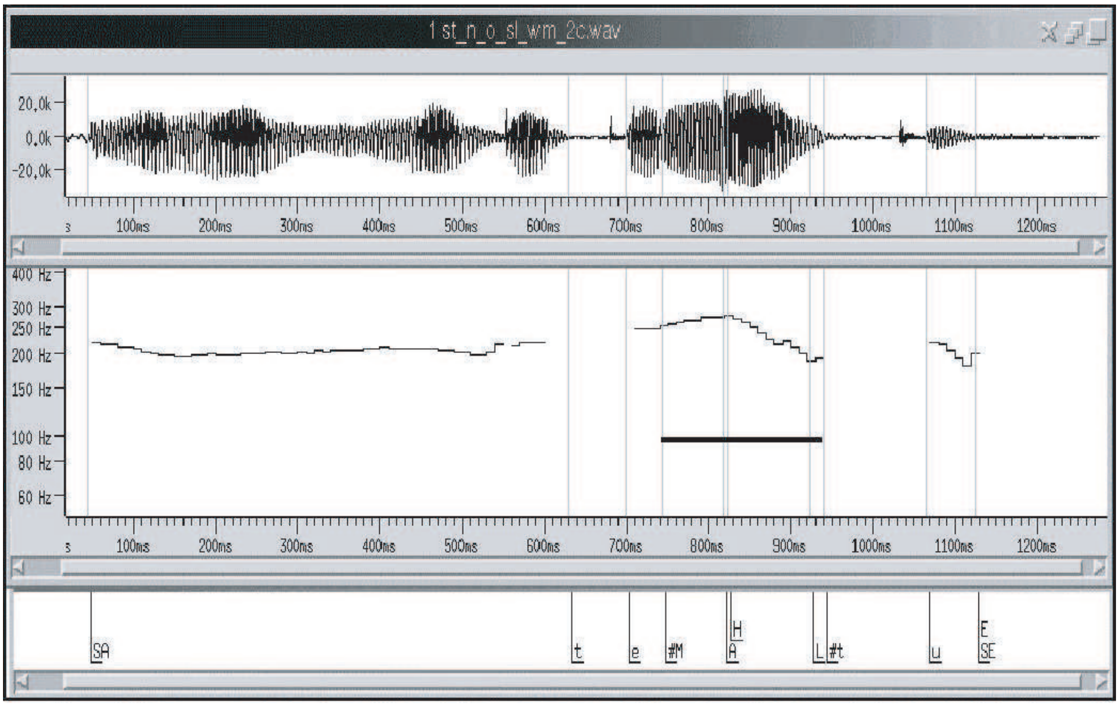


Figure 4.5.: *Realisation of H*+L by speaker WM*

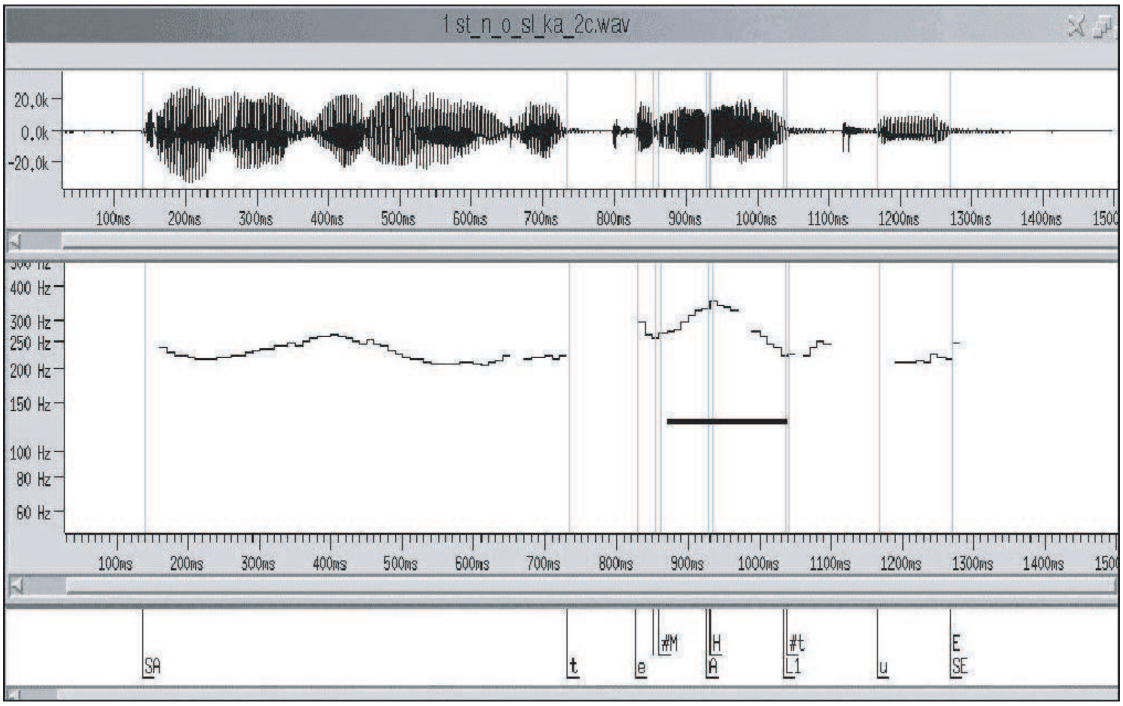


Figure 4.6.: *Realisation of H*+L by speaker KA*

speaker WM achieved this by using two different accent types, (!)H+L* vs. H*+L, for broad and narrow non-contrastive respectively, speaker KA used the frequency domain. For this speaker we find peak F0 values for narrow non-contrastive to be significantly higher than the broad focus F0 values. When narrow non-contrastive focus was in sentence-initial position (subject in focus) both speakers used L+H*. They placed the low target of the L+H* accent just before or at the beginning of the accented syllable. The high target was placed at the end of the accented syllable or at the beginning of the next syllable (cf. Fig. 4.7 and Fig. 4.8).

In narrow contrastive focus speaker WM used both L+H* (51 occurrences) as well as H*+L accents (93 occurrences). In the same condition speaker KA used just L+H* accent type. Because both speakers used L+H* on the sentence-initial word in narrow non-contrastive as well as in contrastive conditions, they needed to disambiguate them. Speaker WM achieved this in the time domain by varying F0 peak alignment, reaching the peak significantly later in the contrastive condition. Speaker KA on

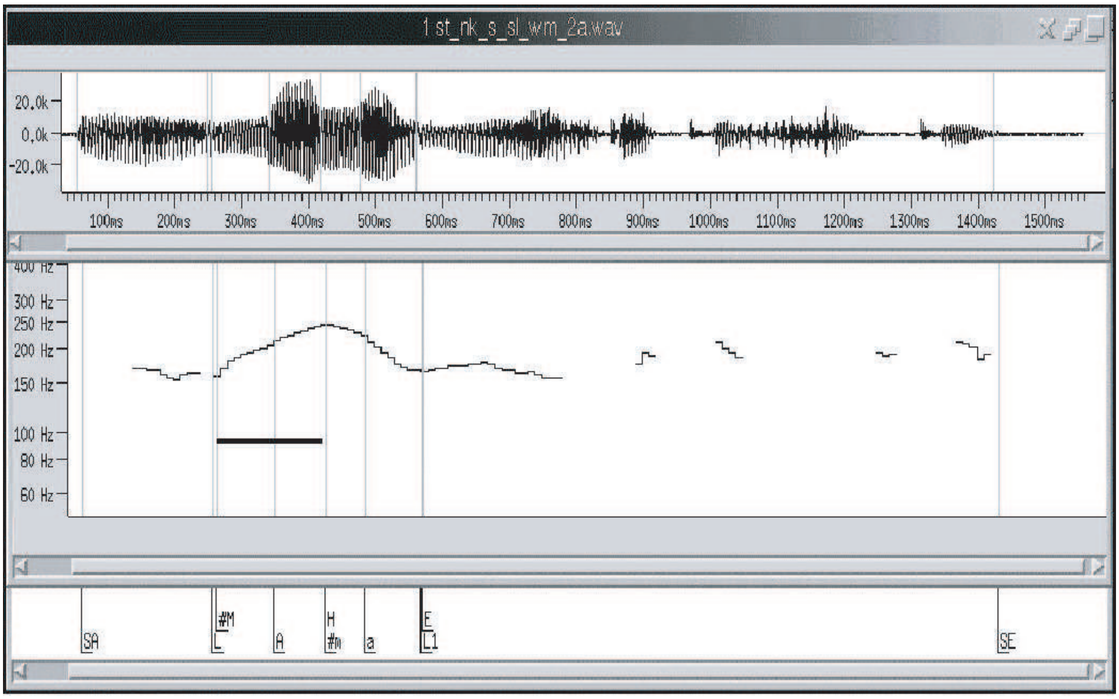


Figure 4.7.: *Realisation of L+H* by speaker WM*

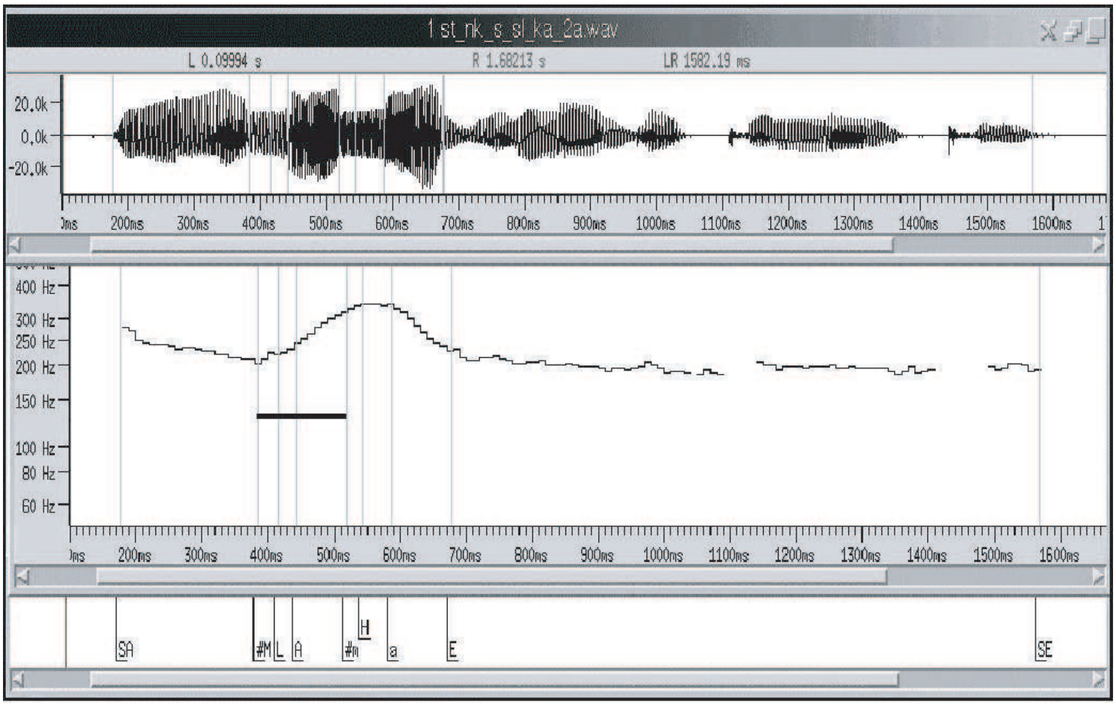


Figure 4.8.: *Realisation of L+H* by speaker KA*

the other hand disambiguated these cases in the frequency domain by using significantly higher F0 values in narrow contrastive focus.

Additional disambiguation was used by speaker WM in the case of focus on sentence-final items in narrow non-contrastive contrastive focus. This speaker used an H*+L in both cases and significantly shifted the F0 peak to later in the syllable in the contrastive condition. As reported in Andreeva & Oliver (2005) the following accent types are used in the examined focus conditions. Polish speakers employ (!)H+L*, H*+L and L+H* accent types. Different accent types were found in the same focus condition and the same accent types we found in different focus conditions.

To analyse the effects on peak alignment we carried out multivariate analyses of variance, with Scheffé post-hoc tests when appropriate. The statistical analysis of the data shows that at a 5% significance level, speech rate influences the absolute but not the relative peak alignment measure. Significant differences in the absolute peak alignment differences shed light on the nature of the anchoring points of the tonal targets in the segmental structure of the language. However, unlike in the studies above, the relative peak alignment is found to be significant both to the syllable onset and offset. As we no longer have one specific point but points suggests a different "anchoring" than that observed by Arvaniti et al. (1998), and Ladd et al. (2000). The statistical results of F0 alignment will be presented following the focus condition (broad, narrow, narrow-contrastive) and sentence position (initial, medial and final).

Broad focus

Statements with broad focus were mostly realised as having a falling nuclear pitch movement, i.e. a H(igh) accent followed by a L(ow). Mean peak delays from the onset of accented syllable were -19.1% ((!)H+L*) and 15.7% (H*+L) for speaker WM and 44% for the speaker KA measured as a proportion of the syllable length, cf. Figure 4.9.

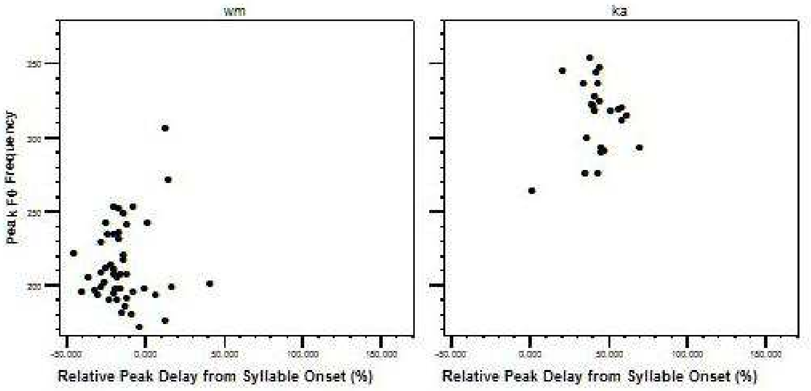


Figure 4.9.: *Speakers' peak height in broad focus condition - falling movement*

The difference in peak delay between the speakers reflects their choice of accent type: speaker KA uses H*+L exclusively whereas speaker WM prefers (!)H+L* (92% of the cases) and chooses H*+L accent in the remaining cases. Due to the different phonological peak alignment of the two accent types, (in the first case the high target occurs before the syllable onset and in the latter in the accented syllable), KA aligns her peaks much further into the syllable (3.5% into the rhyme onset) than speaker WM.

Narrow Non-contrastive Focus

Focus on subject (initial position) When the first content word (here subject) is in narrow non-contrastive focus, pitch accents across speakers are realised as falls as well as rises. Speaker WM uses both rises (L+H*) and falls (H*+L) in this condition. The observed tendency in fast speech for this speaker is to align falls 5.5% earlier (58.5% at normal speech rate) and rises 1.7% later (85.6% at normal speech rate)(cf. Fig. 4.10). For speaker KA, who uses only rising pitch accents (L+H*) in this condition, there is a main effect of tempo on peak alignment. The mean delay from syllable onset in normal speech rate for this speaker is 75.5% of the syllable duration, whereas in fast speech rate it occurs 26.5% later in the syllable (after syllable offset). Additionally we observe significantly higher F0 values in fast speech tempo for speaker KA (cf. Fig. 4.11).

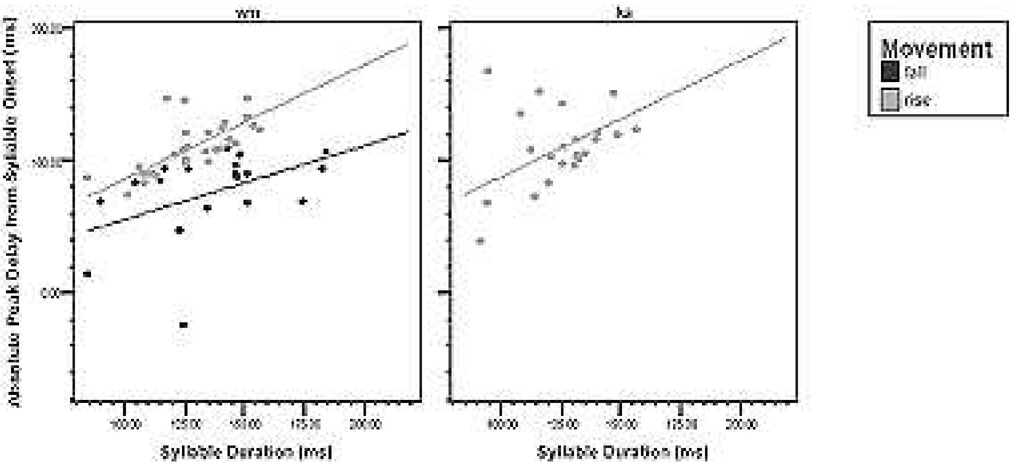


Figure 4.10.: *Speakers' peak positions in narrow non-contrastive focus condition - initial position*

Focus on object (final position) In this condition, both speakers produce falling pitch accents. Speaker WM uses H*+L as well as (!)H+L* whereas speaker KA uses exclusively H*+L. The analysis of peak alignment for the same accent type (H*+L) reveals significant speaker and tempo differences in that speaker WM puts her peak at 16% and 7.9% from the syllable onset in normal and fast speech tempo respectively (cf. Fig. 4.12), whereas speaker KA realises her peaks later (40.9%) without significant difference across tempo.

For both speakers we observe a significant effect of speech tempo on F0 peak values. However, they show contrasting behaviour: with increasing speech tempo, F0 values decrease for speaker WM and rise for speaker KA (cf. Fig. 4.13).

Narrow Contrastive Focus

Focus on subject (initial position) In narrow contrastive focus the accented syllable of the subject is realised with a rising pitch (L+H*) and the

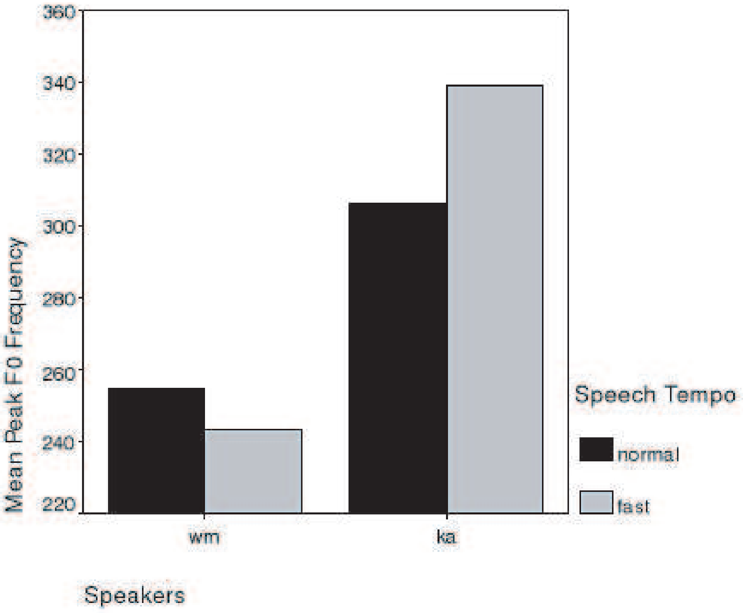


Figure 4.11.: *Speakers' peak height strategies in narrow non-contrastive focus condition - initial position*

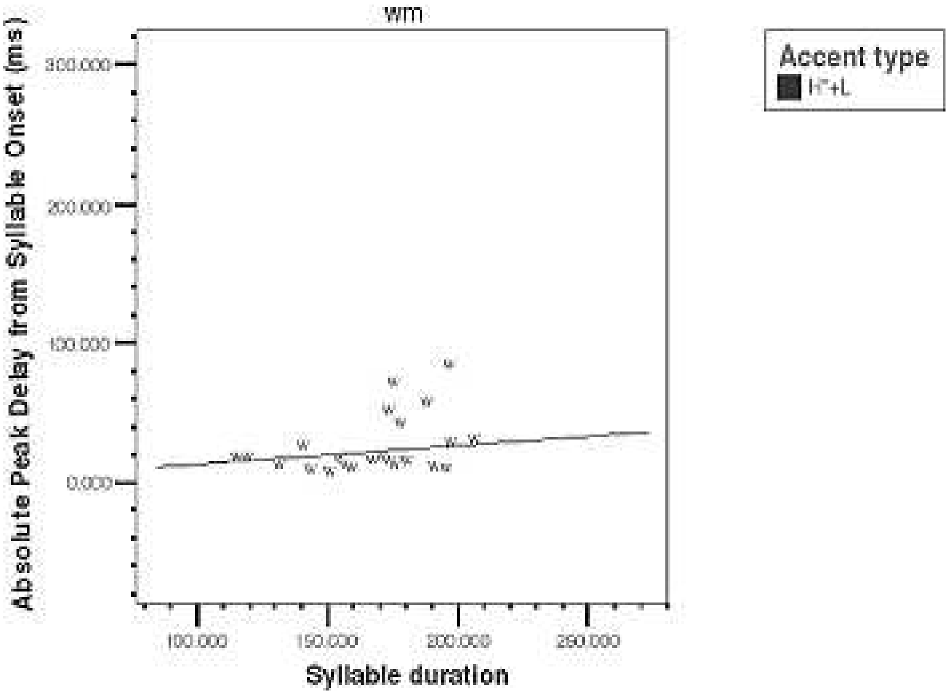


Figure 4.12.: *Peak alignment in narrow non-contrastive focus condition - speaker WM - final position*

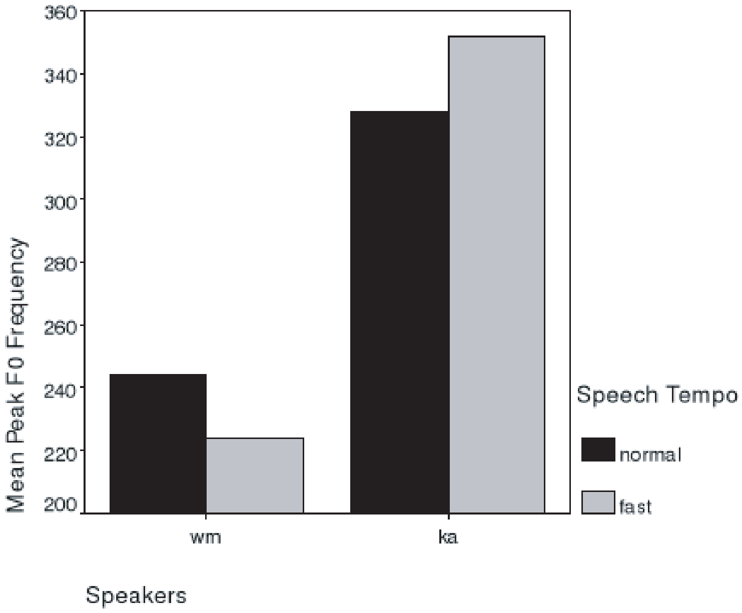


Figure 4.13.: *Speakers' peak height in narrow non-contrastive focus condition - final position*

peak delay is placed either towards the end of the syllable rhyme or after the syllable offset. In normal speech tempo, speakers WM and KA align the peak at 85.6% and 96.5% of the syllable duration respectively. In fast speech tempo, we observe peak alignment at 81.4% for speaker WM and 105.6% for speaker KA. Speaker WM aligns the peak earlier in fast speech as expected (cf. Fig. 4.14). The opposite behaviour found for speaker KA is not significant. Contrasting behaviour for the two speakers was also found with respect to the effect of speech tempo on peak height. The effect was significant for both speakers; WM lowered peak F0 with increasing tempo, whereas speaker KA showed higher peak F0 with increasing speech tempo (cf. Fig 4.15).

Focus on verb (medial position) In medial position, there were significant speaker differences. Speaker KA uses rises exclusively (L+H*), aligning them at 66.4% into the syllable. Speaker WM produces predominantly H*+L (there are only three cases of rising pitch accents). For this speaker the F0 peak is located at half the duration of the syllable for falls

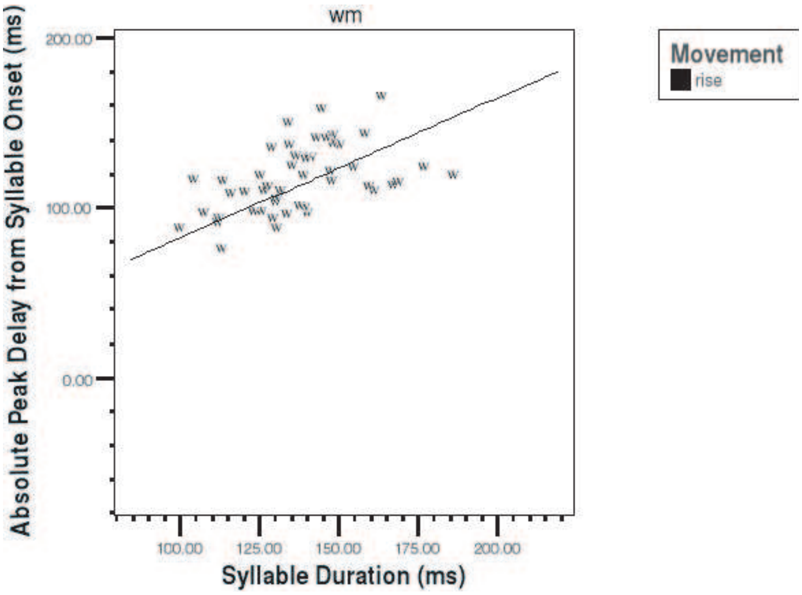


Figure 4.14.: *Speakers' peak alignment in narrow contrastive focus condition - initial position*

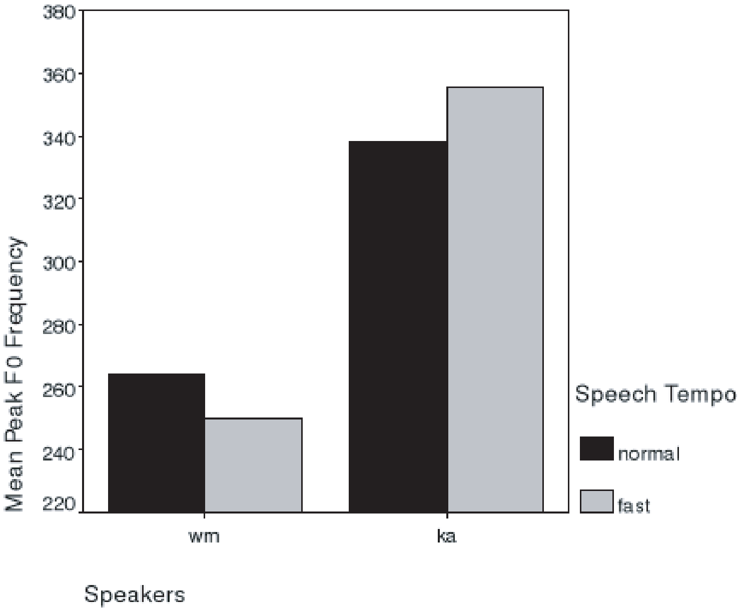


Figure 4.15.: *Speakers' peak height in narrow contrastive focus condition - initial position*

(51.3%) and at 86.7% for rises. The alignment analysis shows main effect of speech tempo for both speakers, whereby the peak occurs earlier in fast speech (cf. Fig. 4.16).

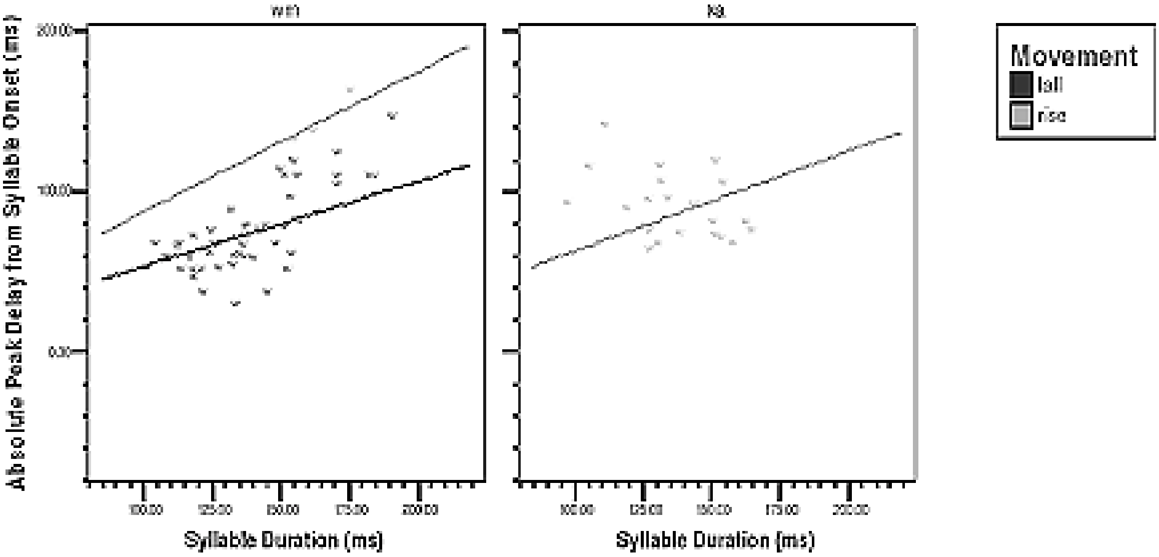


Figure 4.16.: *Speakers' peak alignment in narrow contrastive focus condition*

The falls produced by speaker WM are significantly lower in fast vs. normal speech rate. Again, the opposite is true for rising contours used by speaker KA (cf. Fig. 4.17).

Focus on object (final position) In terms of peak alignment this condition is realised exclusively by H*+L by both speaker KA and WM. For speaker WM the mean peak delay values are 16.6% from syllable onset, and the peak is placed before the rhyme onset. Speaker KA, on the other hand, aligns the peak later, at 44.1% of the syllable duration, and at 4.9% of the rhyme duration. Both speakers show significant peak alignment differences in two tempo conditions as noted for other conditions with falling pitch accents: we observe earlier peak alignment with increased speech rate (cf. Fig. 4.18).

Another main effect in the analysis is the relationship between the pitch alignment and the structure of the utterance. With respect to the

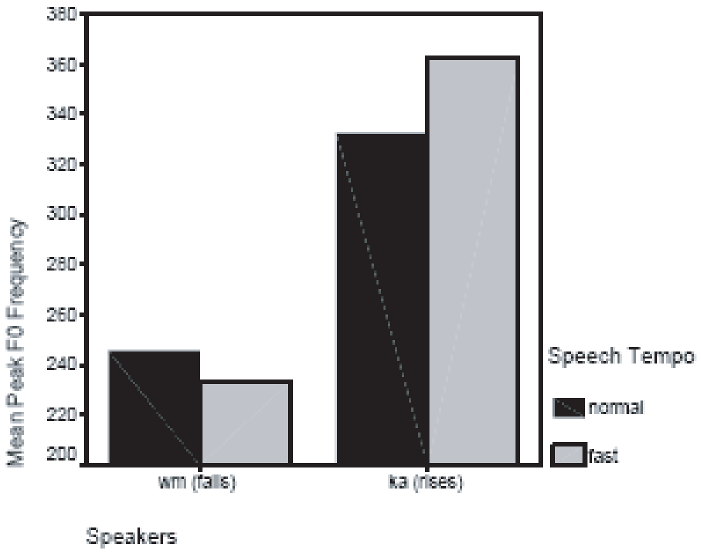


Figure 4.17.: *Speaker strategies in narrow contrastive focus condition*

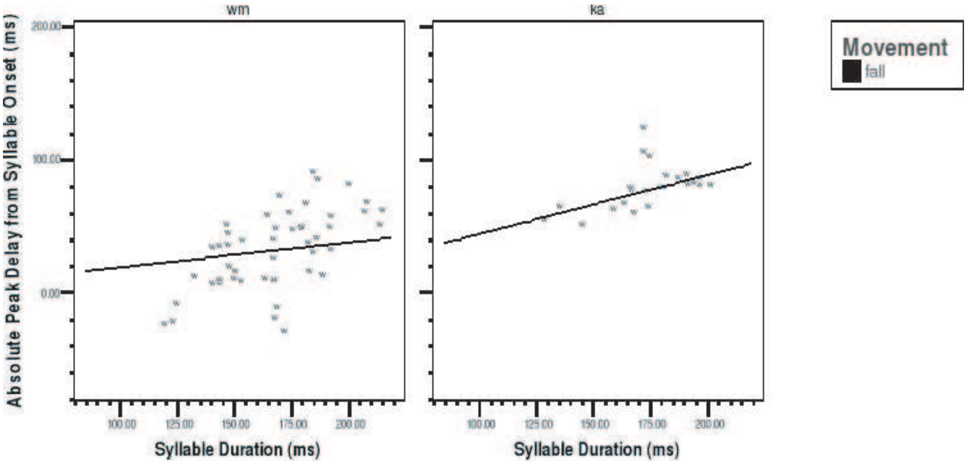


Figure 4.18.: *Speakers' peak alignment in narrow contrastive focus condition*

position of the focused item in the utterance we find the following tendency across focus conditions: the later the focused item in the utterance the earlier the peak alignment.

Non-contrastive focus

The analysis, summarised in Table 4.3, shows that in the non-contrastive focus condition within the same movement (falling or rising) the F0 peak is aligned later in sentence initial position than in sentence final. In the case of speaker WM, it is linked to different accent types used in the two positions, i.e. in the initial position we observe H*+L, whereas in the final (object) position this speaker additionally uses (!)H+L* and H+L*. The statistical analysis reveals, as expected, a significantly later peak alignment in the initial position (55.6%) than in the final (shown as a proportion of the syllable length). The alignment measures are not significantly different for the sentence final position between narrow non-contrastive and broad object for this speaker (-15.9% vs. -16.1%, 12.6% vs. 15.7% respectively), which reflects the fact that WM makes use of the same accent types (H*+L and H+L*) in both conditions. As speaker KA chooses to use different movement types in initial and final position (rise vs. fall, respectively), the comparison is limited to the final position in the broad and narrow non-contrastive condition. Here, speaker KA uses same accent type (H*+L) and aligns F0 peaks similarly for object in narrow non-contrastive and broad focus (43.3% vs. 43.9%).

Tempo	Normal		Fast	
Speaker	Rise	Fall	Rise	Fall
WM	x	subj > obj = broad	x	subj > obj = broad
KA	x	obj = broad	x	obj = broad

Table 4.3.: *The effect of position in a sentence on peak alignment across speakers in non-contrastive focus*

Contrastive focus

Within the contrastive focus condition, three sentence positions are compared: initial, medial and final. Speaker WM uses a rising accent type (L+H*) in initial and falling accent type (H*+L) in medial and final positions. With the falling movement, the F0 peak is aligned significantly later in medial than in the final sentence position (51.3% vs. 16.6% respectively). In the data from speaker KA, we observe a significantly later positioned F0 peak in initial versus medial position for the accent type L+H* (101% vs. 66.3%). The effect of position in a sentence is summarised in Table 4.4.

Tempo	Normal		Fast	
Speaker	Rise	Fall	Rise	Fall
WM	x	verb > obj	x	verb > obj
KA	subj > verb	x	subj > verb	x

Table 4.4.: *The effect of position in a sentence on peak alignment across focus and speakers in contrastive focus.*

With respect to the position of the focused item in the utterance, we found the following tendency: the later the focused item in the utterance, the earlier the peak alignment. A possible explanation is the phenomenon of "tonal repulsion". The proximity of the intonation phrase boundary tones leads to temporal readjustments of peak location (Silverman & Pierrehumbert, 1990).

4.2. Summary and conclusions

The goal of this study was to investigate how broad and narrow contrastive focus and non-contrastive focus are realised in terms of accent type and the temporal alignment of high tonal targets for different speech rates and positions within the utterance. The following accent types were found to be used by the speakers: (!)H+L*, H*+L, and L+H*. We found different accent types in the same focus condition and the same accent types in different focus conditions, which refutes our first hypothesis. Speakers employed both peak alignment and peak height to obtain a phonological contrast between the different focus conditions. With this evidence the prediction of our second hypothesis is confirmed.

The goal of this study was also to investigate the phonetic details of peak alignment in phonologically specified accent types across different focus conditions (broad and narrow contrastive vs. non-contrastive focus) in Polish. The relationship between peak location and segmental structure was explored for different speech rates (normal vs. fast) and positions within the utterance (final vs. non-final).

The statistical analysis of the data showed that apart from accent type itself, speaker, speech tempo, focus type and position in the sentence as well as the interaction between them influence peak alignment at the 5% significance level. Additionally, the peak F0 frequency was significantly influenced by the following factors: focus, speaker and accent type.

The results showed main speech tempo effects on temporal alignment of F0 targets and peak F0 height. The observed correlation between tempo and peak alignment is positive. With increasing speech tempo, speakers align their peaks significantly earlier. The F0 targets are conditioned by the accent type used and therefore we observe speaker specific strategies: either increasing or lowering F0 values with tempo changes. The fact that the absolute but not the relative measures for peak alignment differ significantly with changing speech rate clearly lends support to the claim that speakers control the peak alignment in a consistent way relative to the time-course of the syllable. This confirms the hypothesis about the existence of well-defined targets. However, in our data, they do not appear to be aligned relative to ”segmental anchor points” to which the tones would be aligned as defined by Arvaniti et al. (1998) and Atterer & Ladd (2004).

The results demonstrate that the target is sensitive to syllable onset and offset for Polish.

The alignment pattern found here emerged clearly due to the use of only one syllable type (/ma/) for all conditions. However, further investigations need to be carried out with the variety of syllable structures in order to assess the extent to which the target is aligned relative to onset and offset of the syllable or the beginning and end of the sonorant portion.

The following accent types were found to be used by the speakers: H+L*, !H+L*, H*+L and L+H*. Across the speakers, we found different accent types in the same focus condition and the same accent types in different focus conditions. This is reflected in the alignment measures where the following patterns are observed.

First, in non-contrastive focus, F0 peak targets are aligned significantly earlier than in contrastive focus. This is mainly due to the fact that, the speakers tend to use rising accents in contrastive and falling accents in non-contrastive focus conditions. Additionally, alignment of F0 events varies with respect to the position of the focused items in an utterance. Within each focus type utterance initial focused items (subject) are aligned significantly later than utterance medial (verb) and utterance final items in focus (object). A possible explanation is the phenomenon of ”tonal repulsion”. The proximity of the intonation phrase boundary tones leads to temporal readjustments of peak location (Silverman & Pierrehumbert, 1990).

Finally, the details of alignment for the H(igh) targets highlight the differences in the phonetic realisation of phonologically specified accent types. The experimental findings demonstrated a pattern, namely, that in all focus type conditions, speakers controlled the position of the F0 peak such that accented items positioned later in the utterance received an earlier F0 peak alignment than those in sentence medial and initial position. It is not clear how much this pattern is a result of what may be a universal tendency or whether it is partly a reflection of the choice of a different falling accent types used in different sentence positions.

The above study sheds light on the description of temporal alignment of F0 events in Polish across focus conditions. Nevertheless, as already stated above, further analysis is needed before the above claims and descriptions can be extended to a more general model of alignment in the language studied. This study by no means exhausts all the factors affecting the phonetics of tonal alignment in Polish. Further research is needed to determine what these factors are, which of them are language-specific, and which might be considered universal.

5. Speech Resources

Introduction

In order to accomplish the goal of the current study, a speech corpus is required. Over the last ten years, through various projects, a considerable number of linguistic resources has been created for the Polish language. As a major prerequisite for the current study is a speech corpus, the chapter describes the existing speech corpora for Polish, Section 5.1, and concentrates on the speech databases PoInt and Babel. Their structure and content will be described in Sections 5.2 and 5.3. The emphasis will be put on type of speech material used, speakers, annotation conventions and the preprocessing that is necessary to convert their contents to the needs of the current study. The acoustic analysis of speech samples will be presented in Sections 5.2.3 and 5.3.1. Finally the details of the application of the selected speech material in other parts of the thesis will be shown.

5.1. Speech Resources

As far as the speech resources for Polish are concerned there have been a few projects at a national and European level carried out in order to provide the scientific community with speech corpora which would meet the demand for fundamental research in speech technology, encompassing speech synthesis, recognition, dialogue analysis and spontaneous speech studies in general. In the following sections, these resources, including CORPORA, SpeechDat E, Speecon, Babel and PoInt, are described. As two of them, Babel and PoInt, have been used extensively in the current study, their description, additionally, details the necessary modifications which have been applied by the author in order to adapt them to the needs of the current study.

CORPORA Supported by the Polish National Research Committee was the creation of "CORPORA"; a speech database for Polish Diphones (Grocholewski, 1999), developed by a team from the Computer Science Department at the Poznań University of Technology, led by Stefan Grocholewski. Together, there are 16425 sentences (365 utterances per speaker) by 45 speakers (28 male, 11 female, 6 children) who were recorded producing the alphabet, numerals, first names, and 114 statements (Grocholewski, 2001). The utterances cover 966 Polish diphones and the length of the database is approximately 6 hours. Within the project, a dynamic time warping tool for automatic segmentation and labelling of continuous speech was also developed. As part of the project a tool for manual segmentation corrections of phonemic labels as well as numerous DSP tools for converting between different audio formats were included.

SpeechDat E Within the SpeechDat E project (Eastern European Speech Databases for Creation of Voice Driven Teleservices), a Polish telephone speech database was collected by the team led by Wojciech Majewski from ITA, Wrocław University of Technology (van den Heuvel et al., 1999). The database contains the recordings of 1,000 Polish speakers, from eight different regions (488 males, 512 females) recorded over the Polish fixed telephone network TPSA. There were 9 speakers under 16 years old, 428 speakers between 16 and 30, 291 speakers are between 31 and 45, 254 speakers between 46 and 60, and 18 speakers over the age of 60.

Each speaker uttered sequences of digits (telephone numbers, credit card numbers, PIN codes), dates, spelled words, currency amounts, surnames, city names, company names, as well as yes/no questions, phonetically rich sentences, time phrases and phonetically rich words. Automatically generated orthographic transcription were manually corrected and the database was phonemically annotated using SAMPA.

Speecon Another database of Polish speech built for speech recognition was created as part of SPEECON project (Siemund et al., 2000). SPEECON (Speech-Driven Interfaces for Consumer Devices) was a project which aimed to develop voice-driven interfaces for consumer applications. The projects goal was to collect speech data for at least 20 languages and 600 speakers per language. The Polish part contains two sets of data, which comprise 550 adults and 50 children's recording sessions, respectively. The adult speakers were recorded in various environments: offices, living rooms, cars and public places. Recordings contain free spontaneous speech passages, elicited spontaneous speech, "phonetically compact" words and sentences, general-purpose words and phrases, specific application words and utterances.

Babel Within the framework of European COPERNICUS programme, ("BABEL a multilingual-Language Database" project for Polish, Bulgarian, Estonian, Hungarian and Romanian), Ryszard Gubrynowicz from the Speech Acoustics Laboratory at the Institute of Fundamental Technology Research (Polish Academy of Science) created the BABEL database for Polish (Gubrynowicz, 1998). This will be described in detail in Section 5.3.

PoInt Finally, as part of a project run by Maciej Karpiński and coordinated by Wiktor Jassem, a first database for Polish intonation research, called PoInt (Karpiński & Kleśta, 2001), was created in 2001. Because of the goal of the project, namely, investigating intonation patterns in Polish semi-spontaneous and read speech, the current study heavily relies on the database and it formed the basis of the prosodic modelling discussed in Chapter 9.

5.2. Polish Intonation Database PoInt

The current study uses the PoInt speech database of Polish, which was recorded as part of the Polish Intonation Database Project (Karpiński & Kleśta, 2001). This database includes recordings from 47 speakers of mixed sex and contains a variety of discourse types; fragments of read literary texts and semi-spontaneous monologues as well as map task based dialogues. PoInt is the first database for Polish which has been annotated with prosodic information based on Jassem's revised description of intonation, marking intonation tone heights with labels xL, L, M, H, xH. The accented syllables contain one letter labels and if the accented word is followed by an IP boundary the post–accented syllable also has a tone height marker. Such annotation gives us quantitative information on the contour shapes through tone heights which have been impressionistically labelled by trained human labellers.

5.2.1. Contents

The PoInt Database is delivered with a graphical user interface, which allows quick access to its contents. The program contains 1144 signal files with intonograms, spectrograms, orthographic and phonological transcriptions (IPA), pragmalinguistic and other comments. The speech signals represent semi-spontaneous and read speech, up to four intonation phrases each, realised by male and female native speakers of Polish, cf. Table 5.1. In order to elicit the speech material, the PoInt authors developed monologue and dialogue tasks and also took texts from contemporary Polish literature, e.g. ”florentynka” and ”szafa” read speech parts.

The graphical user interface incorporates a search for a signal, using various criteria, the possibility of listening to wave files and comparing them with the graphical representation (oscillogram, intonogram, phonetic transcription).

Age	20 - 30	31 - 40	41 - 50	over 50
Female	25	0	0	0
Male	18	4	0	0

Table 5.1.: *Distribution of male and female voices in the PoInt Database*

Filename conventions

Here is an example file name which contains all information needed to identify the type of speech material, speaker's name, gender, and number of the recording: alpa_f_florentynka4 indicates:

- the name of the speaker – alpa (first two letter of first name and surname),
- the gender of the speaker – f or m,
- the genre of the text – here a read text from "florentynka",
- the number of the signal from a certain recording – 4.

Speech material genres

The four text genres with names indicating source are:

- florentynka, szafa – pieces of prose,
- entliczek – a poem for children
- przedmioty, muzyka, obraz, komiks, pochodzenie – semi-spontaneous monologues,
- dyskusja, mapa, quiz – task-oriented dialogues.

The parts which are of interest to the current study are the read speech parts "florentynka" and "szafa", as well as the semi-spontaneous monologues "przedmioty" (Eng. objects), "muzyka" (Eng. music), "obraz"

(Eng. painting), "komiks" (Eng. comic strip) and "pochodzenie" (Eng. childhood memories). The following sections will concentrate on these particular parts unless indicated otherwise.

5.2.2. Annotation

The program delivered with the PoInt database contains signal files with intonograms, spectrograms, phonological and orthographic transcriptions, as well as pragmalinguistic and other comments. For the current thesis the Praat TextGrid files with phonetic and intonation tier based annotation were kindly made available by Dr. Karpiński (Karpiński & Kleśta, 2001).

Each file may contain a few (usually 1–3) intonational phrases. Only the nuclear syllables are annotated. Five relative pitch heights were used: extra high (xH), high (H), middle (M), low (L) and extra low (xL). Extra high tones were very rare. A typical label consists of two tone symbols, with the first referring to the tone of the nuclear syllable (usually penultimate in Polish) and the second to the tone of the post-nuclear one (e.g. HL, LxL). In monosyllabic words, the labels consist of one tone (e.g. H, L, or M). In cases where monosyllabic words receive the nuclear stress, only one tone symbol was possible (e.g. H, L or M). When the nuclear syllable occurred earlier than on the penultimate syllable, a three tone symbol was used (e.g. HM–L). Additionally, in cases where the analysed material, especially ultimate syllables, were degraded either through devoicing [D], an extreme fall in energy [F], creak [C], or reduced syllable realisation [s], extra labels, not directly connected with the phonological level of analysis, were included.

Pitch Accents			Total
H	L	M	
287	613	495	1395

Table 5.2.: *Distribution of pitch accent labels in PoInt*

In addition to the above labels of the nuclear syllable, PoInt authors labelled the last syllable in the phrase when the nuclear syllable was in greater distance from the end of the phrase. Karpiński & Kleśta (2001) suggest that only a certain subset of all possible pitch movements is phonologically relevant. In the current study we distinguish between pitch accent labels and final syllable labels by marking the latter with a boundary marker '|'. Tables 5.2 and 5.3 show accent and boundary tone distribution excluding [D], [F], [C] and [s] elements.

Boundary Tones					Total
H|	L|	M|	xH|	xL|	
416	442	266	10	95	1229

Table 5.3.: *Distribution of boundary tones in PoInt*

We performed an additional check of the speech material to find out if any of the segments marked with these labels can be used i.e its quality is acceptable. In these cases the labels, mostly [C|] were included in the analysis and depending on their height relabelled to L|, M|, H|, xH| or xL|. In table 5.4 the pitch accent types are shown in relation to the boundary tone following them after the relabelling had been carried out.

		Boundary Tones				
		H|	L|	M|	xH|	xL|
Pitch Accents	H	41	104	65	9	0
	L	178	63	67	0	93
	M	125	154	87	0	0

Table 5.4.: *Distribution of pitch accent labels and boundaries in PoInt*

Accent types Due to the variety of utterance types, the melodic structure of the database is quite varied. By combining the five pitch levels, the intonation phrases contain three basic pitch movements: falling, rising and flat.

- falling: HL, HM, ML, LxL
- rising: LH, LM, MH, HxH
- flat: MM

It would be possible to extend the inventory movements by ’falling-rising’, for example, if the sequences of three tone labels were to be taken into account but due to their small number they will be treated as part of the three basic movements.

Table 5.5 shows the distribution of one, two and three-symbol labels for the whole of the database as reported by Francuzik et al. (2005).

Labels	Melody Type	Number of Occurrences
One-Symbol	Monosyllabic	104
	Creaky	9
	Other	21
Two-Symbol	Falling	592
	Rising	455
	Flat	240
	Creaky	319
	Other	57
Three-Symbol	Compound	132
	Creaky	35
	Other	11

Table 5.5.: *Intonation Symbols in PoInt Database*

Figure 5.1 shows an example of a typical falling contour.

Figure 5.2 shows an example of a rising contour.

Figure 5.3 shows an example of a typical flat contour.

To visualise the intonation labels applied in PoInt, plots showing 50 random examples of each label H, L and M were drawn, cf. Fig. 5.4, Fig. 5.5, and Fig. 5.6. Each plot shows three F0 values within an accented syllable, measured at the start, mid and end of it.

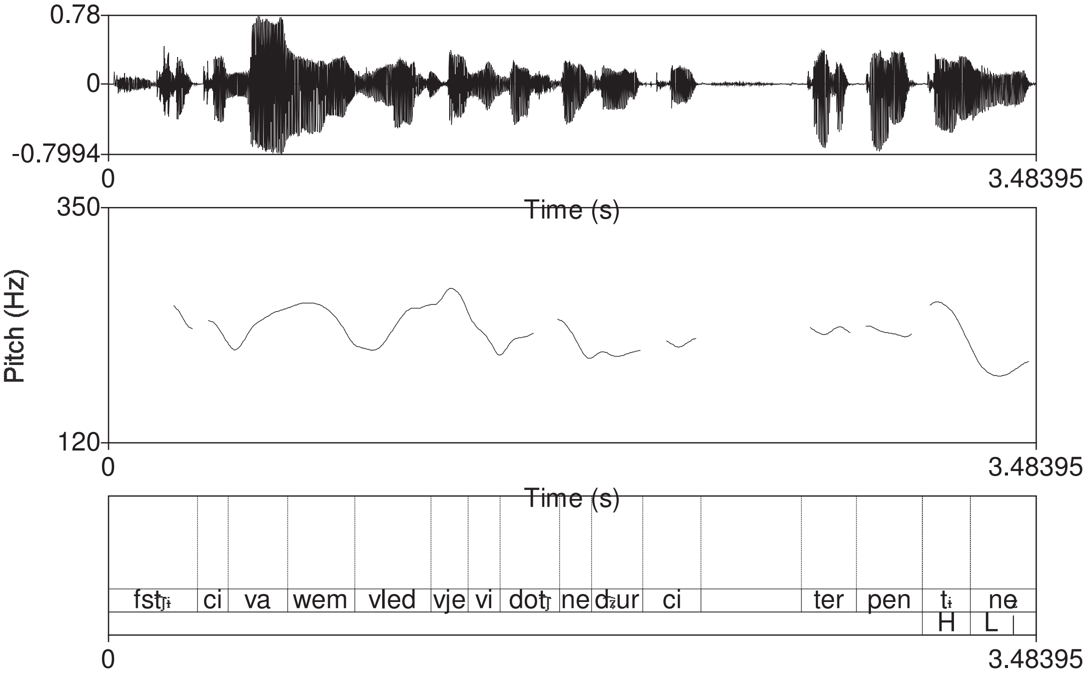


Figure 5.1.: *Accent HL : 'Wstrzykiwałem w ledwie widoczne dziurki terpentynę'*

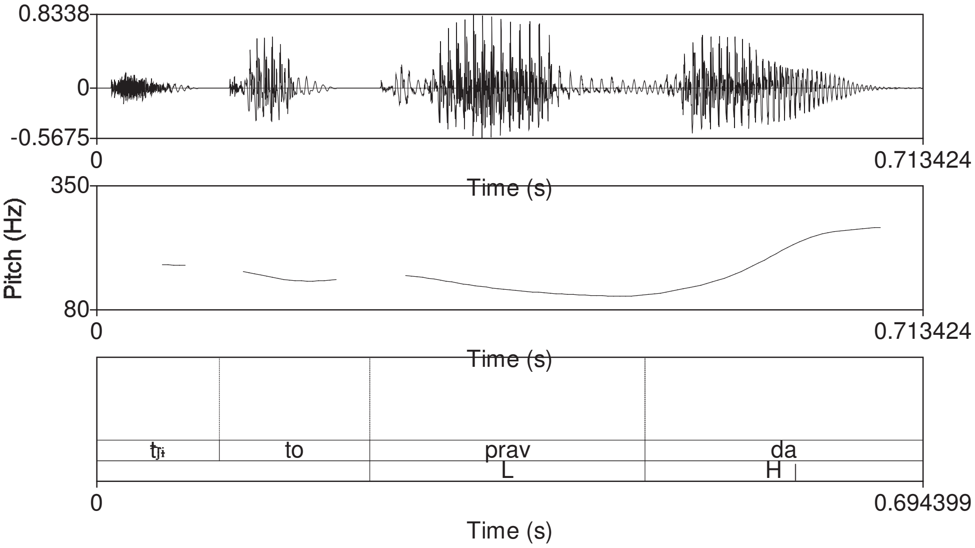


Figure 5.2.: *Accent LH : 'Czy to prawda ?'*

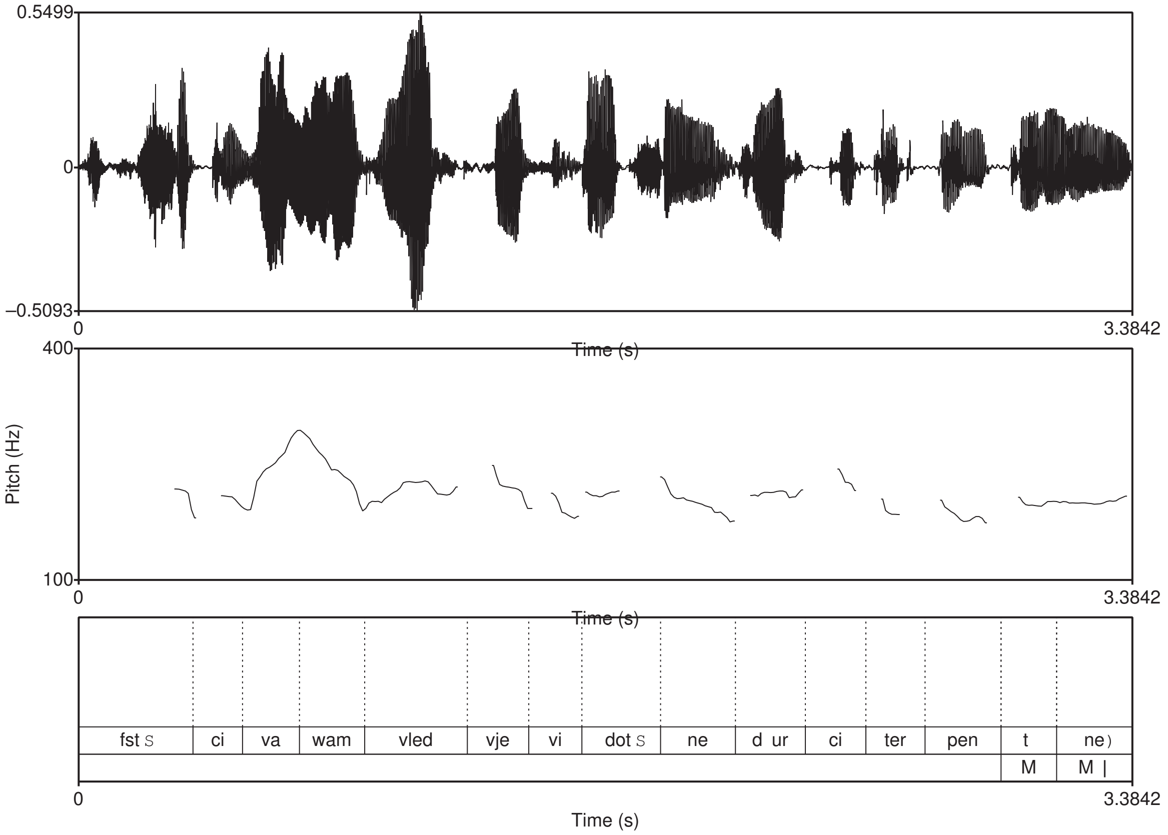


Figure 5.3.: *Accent MM : 'Wstrzykiwałem w ledwie widoczne dziurki terpentynę'*

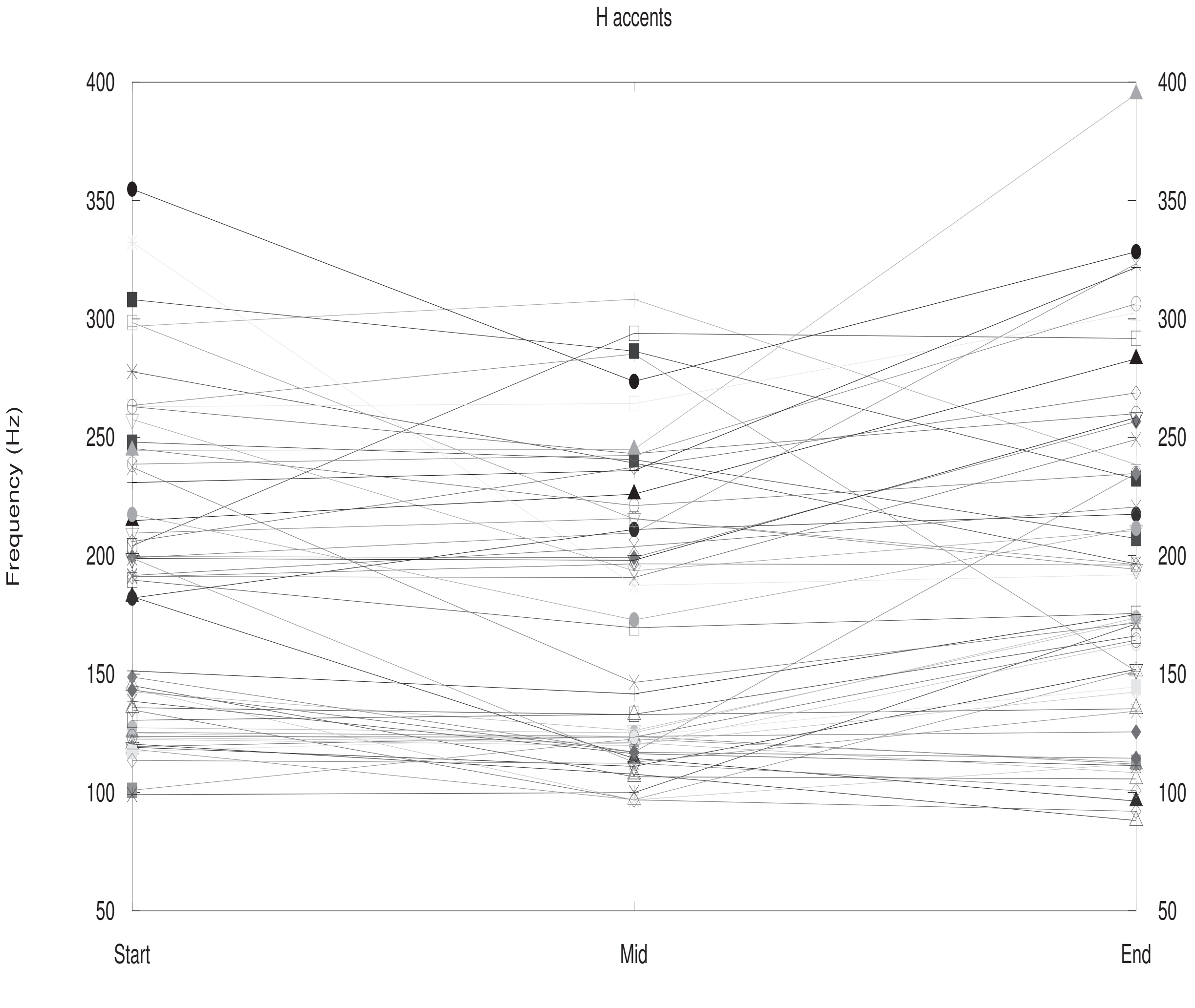


Figure 5.4.: *Accent H*

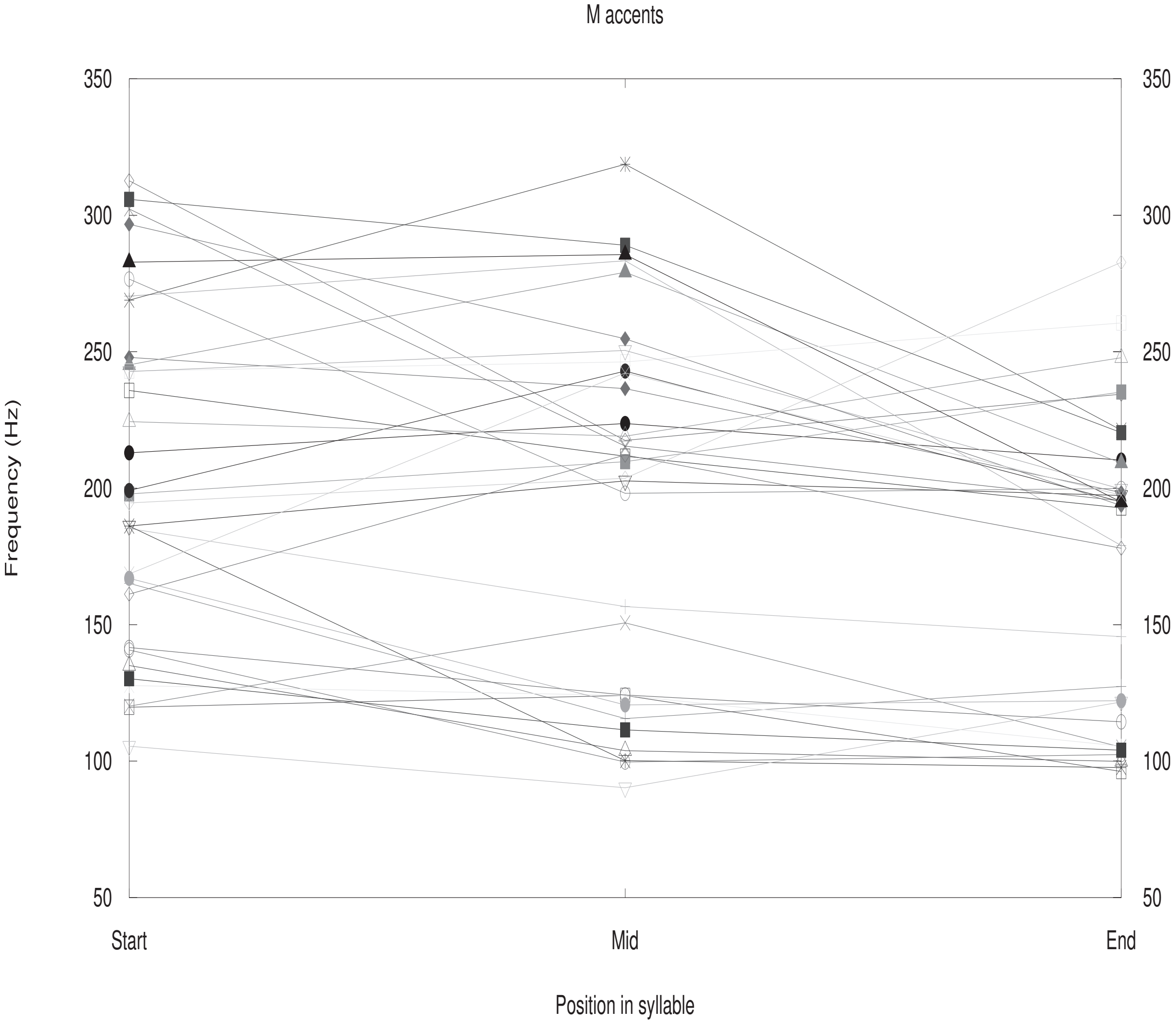


Figure 5.5.: *Accent M*

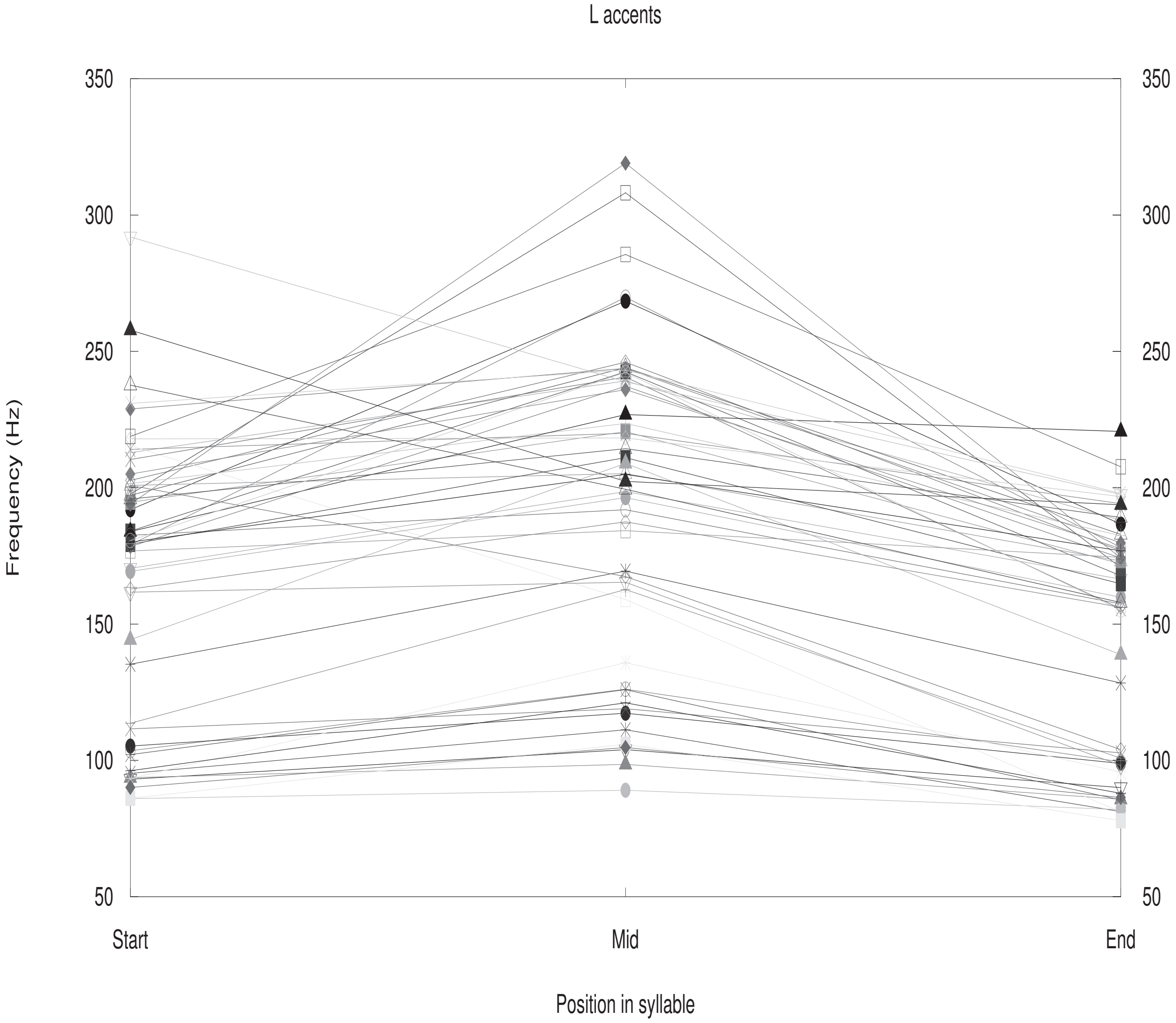


Figure 5.6.: *Accent L*

The values of interest were measured based on syllable boundaries markers present in the original annotation and no normalisation was carried out with respect to male and female voices. The scatter of values clearly reflects the sensitivity of auditory judgements to the context in which the pitch accent is heard. However the mean values for mid-syllable measurements are: H: 292.89 Hz (Female), 174.05 Hz (Male), M: 244.81 Hz (Female), 140.39 Hz (Male), L: 220.72 Hz (Female), 118.32 Hz (Male).

Some of the pitch accents were preferred depending on the gender. As reported in Francuzik et al. (2005), female speakers preferred the rising movements LH ($>$ 100 occurrences), whereas the most common movement for the male part of the database was a falling contour ML and HL. Nevertheless, four out of the five most frequent contours in the database (LH, MH, ML, LH) do not differ in frequency between the sexes.

5.2.3. PoInt Database parameters

The PoInt database has been parameterised using Praat software (Boersma, 2001) to gain insight into the prosodic structure and distributional properties of the database.

The parameters that are of interest in this study are those pertaining to the shape of the contours. Thus the F0 peak alignment, F0 peak height and a slope as represented by Tilt (Taylor, 1998) value were examined. The segmental points with relation to which the analysis was carried out were pitch syllable start and end point, and the analysis window spanned 3 syllables: pre-nuclear syllable, nuclear and post-nuclear.

F0 Peak Alignment

The F0 peak alignment turned out to be a significant measurement with respect to its anchoring to the points in the segmental structure of Polish (cf. Chapter 4). The location of the peak was closely related to the type of pitch accent realised on the nuclear syllable.

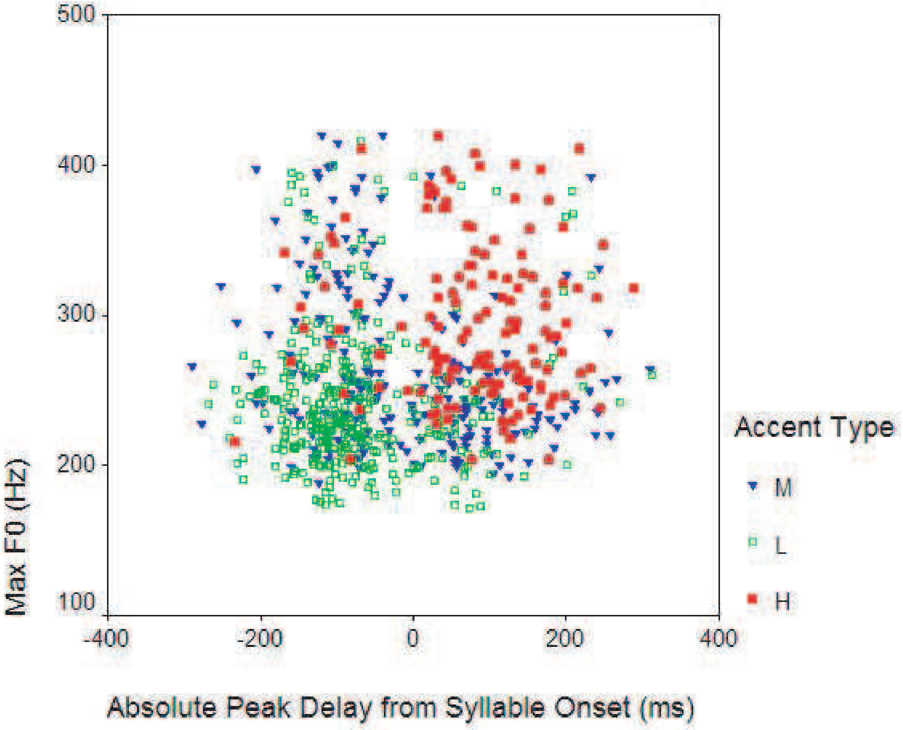


Figure 5.7.: *Absolute F0 peak delay from syllable onset - female data*

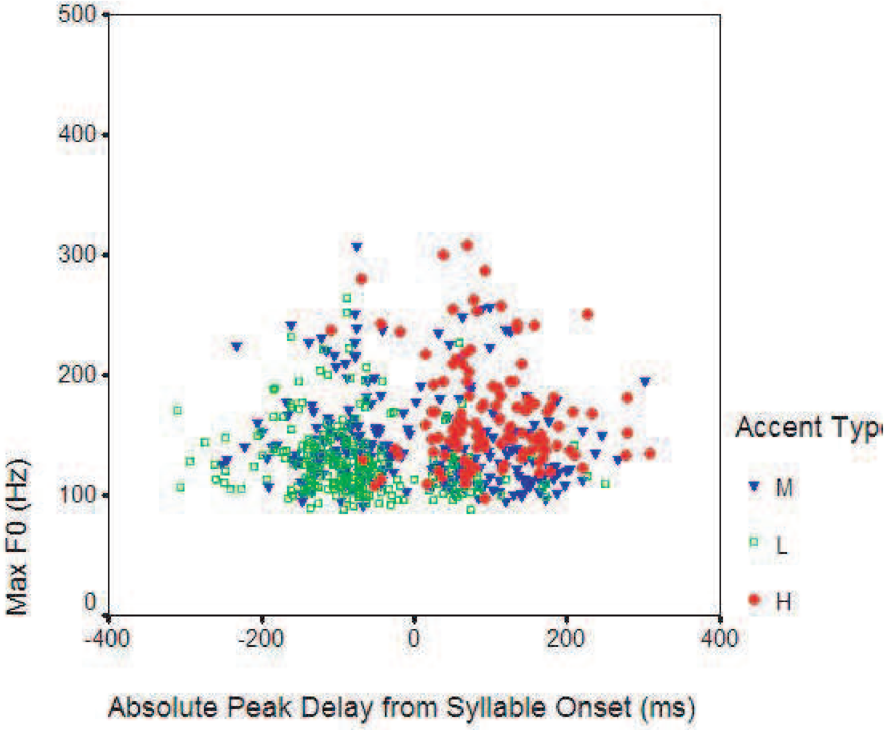


Figure 5.8.: *Absolute F0 peak delay from syllable onset - male data*

The PoInt data has, therefore, been analysed to find the delay of the F0 peak from the syllable onset for each label category that was used in the annotation. Figs. 5.7 and 5.8 show the location of the F0 peak for male and female data separately.

F0 Peak Height

As shown in 50 random examples of each pitch accent label H, L and M , cf. Fig. 5.4, Fig. 5.5, and Fig. 5.6, the distribution of F0 values vary within each accent type. Fig. 5.9 shows the average F0 values (male and female data) per pitch height at three points in the syllable, start, mid and end.

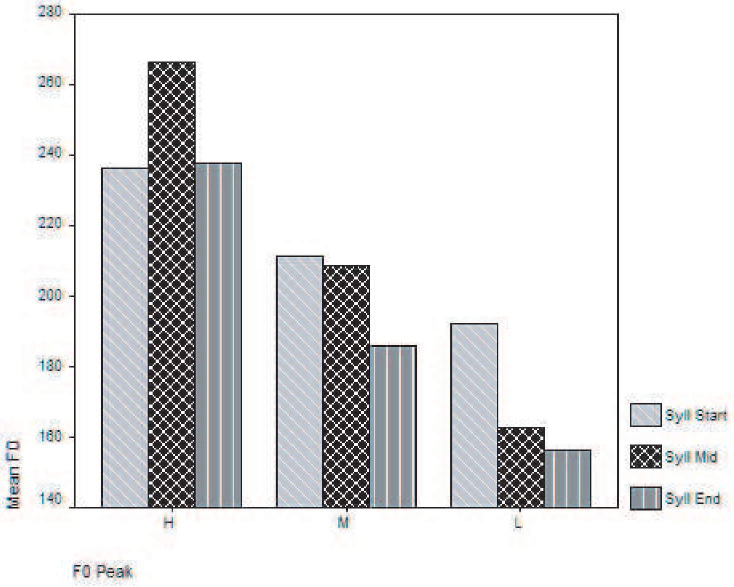


Figure 5.9.: *Tone heights values*

Tilt

The tilt value T1 cf. 5.1

$$T1 = \frac{F0[PostTonic] - F0[PreTonic]}{F0[max] - F0[min]} \tag{5.1}$$

tells us if we deal with a fall or a rise, like Taylor's Tilt parameter (Taylor, 2000), with values ranging between –1 (fall) and +1 (rise), cf. Section 5.10.

Taking a three syllable window, the F0 values for the pre-tonic, tonic, and post-tonic syllables are taken into account. F0[PreTonic], F0[Tonic], F0[PostTonic] refer to F0 values at mid points of these syllables, whereas F0[max] and F0[min] refer to maximum and minimum F0 values within the three syllable window analysed.

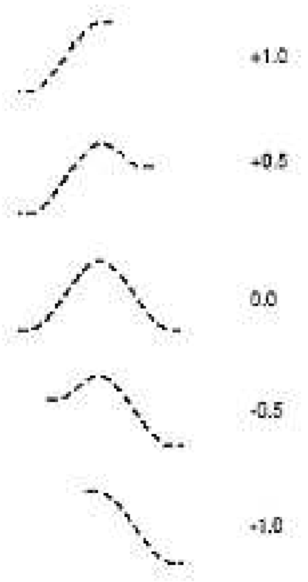


Figure 5.10.: *Values of Tilt depending on accent shape (Taylor 2000)*

5.3. BABEL Database

The BABEL database was constructed within the "BABEL a multilingual-Language Database" Project. It contains read speech from 100 speakers. Table 5.6 shows the distribution of male and female voices per age group. The sampled speech data for each speaker consists of 40 passages, 210 digits, 30 phrases and 144 syllables. The passages were labelled at the phonetic level. The database was recorded at two centres : Polish Academy of Science in Warsaw and Technical University of Wrocław. The part of

Age	20 - 30	31 - 40	41 - 50	over 50
Female	35	4	4	6
Male	28	7	10	6

Table 5.6.: *Distribution of male and female voices in Polish Babel*

the database obtained for the current study comes from the Warsaw part of the project. Prof. Gubrynowicz was kind enough to make the sound files together with their phonetic annotation available for the project.

5.3.1. Speech Material

Under the BABEL project all the languages had linguistically similar structure and consisted of similar elements. The Polish part of BABEL included four types of recordings.

1. blocks of five to seven situational sentences,
2. phonetically balanced individual sentences,
3. digits from 0 to 9999,
4. set of (C)CVC syllables in isolation or within a word.

In order to record the Polish version of BABEL, the English text was translated, keeping the same numbering of sentence blocks, adapted to the Polish language and compared with the French version. Both the French and Polish parts were on average 20% longer than the English version (Gubrynowicz, 1999).

The part which is of most interest to the current study is the situational utterances. They are based on real-life themes and required from the speakers a certain degree of role-play. They had to read or act out passages about customer complaints, telephone enquiries, ordering a service, or about someone relating a story.

Annotation A large part of the database was phonemically annotated following Polish SAMPA transcription conventions. It consisted of mapping IPA symbols onto ASCII codes. Additionally, the start and end of the utterances were marked with a double hash symbol ##. Any pauses made by the speakers within one set were marked with a hash symbol # (Gubrynowicz, 1999). In the Warsaw part of the database forty sets of sentences were annotated, belonging to one male and one female voice, in total thirty minutes long.

BABEL processing For the current study, a part of the BABEL database, namely, the recordings of speakers PJ (female) and SL (male), were used. PJ's speech material was 8.6 minute long, while SL's recordings amounted to 8.4 minutes. The recordings were mainly used as part of perception studies described in Chapter 7 and 10. The original format of the files was converted to Praat TextGrid in order to have uniformity within the two database sets, Babel and PoInt.

Figures 5.11 and 5.12 show examples of the two speakers' speech material.

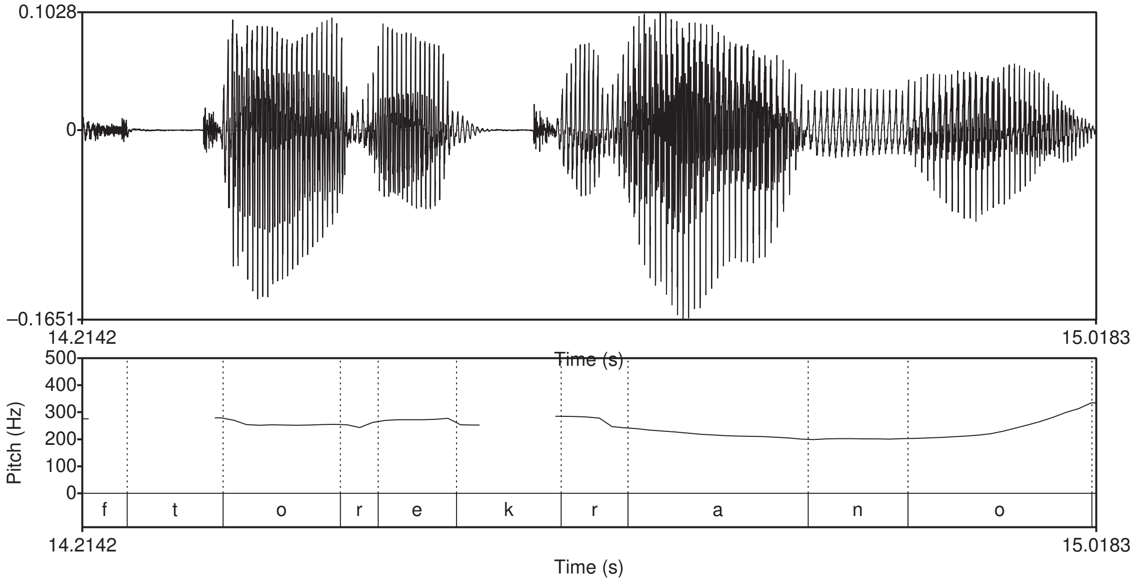


Figure 5.11.: *Babel sample - female data*

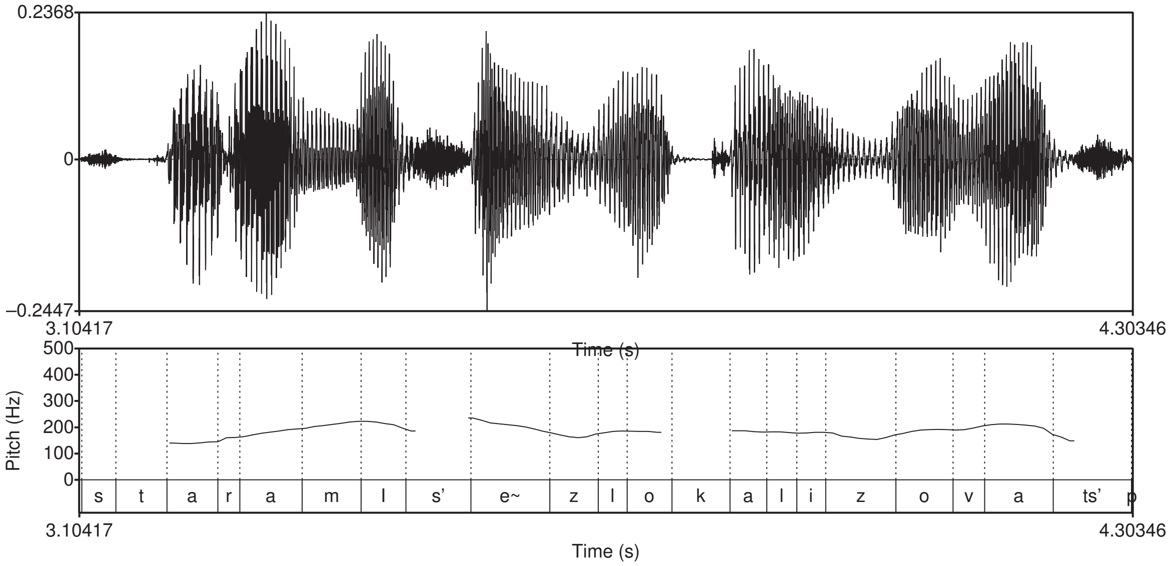


Figure 5.12.: *Babel sample - male data*

5.4. Conclusions

The speech data sources available at the time of the study provided ample prosodic material. After the necessary format post-processing and acoustic analysis of prosodic events, the parameters including information on F0 peak alignment, height, and slope as measured by Tilt, provided the entry point for the modelling of intonation events for speech synthesis and insight into the prosodic structure of Polish intonation contours. The statistical information obtained from these large speech data bases complements the findings from the peak alignment studies reported in Chapter 4 and is used in the rest of the thesis in the following way. In Chapter 7, it forms an input to the F0 stylisation algorithm, in Chapter 8 it serves to re-classify the prosodic events and in Chapter 9 to build an intonation model for pitch-accent type and place prediction in speech synthesis. Additionally, utterances from PoInt and BABEL serve as input to the perception studies described in Chapter 10.

The following chapter introduces the speech synthesis cycle and describes in detail the prosodic modelling process, introducing the necessary background information and related studies. An overview of existing Polish synthesis systems as well as the system within which the prosodic modelling will be carried out, namely Festival, are presented.

Part II.

Intonation in Speech Synthesis

6. Review of prosodic modelling in speech synthesis

Introduction

This chapter describes the development cycle of a speech synthesis system with emphasis on the prosodic modelling module. In section 6.1, we introduce the main building blocks of a text to speech system. Section 6.2 deals in detail with the description of the prosodic components. We describe in greater detail the components of a prosodic model, including duration modelling, phrase break location modelling, pitch accent prediction and assignment, as well as techniques of contour generation. Section 6.3 contains a survey of studies devoted to prosodic modelling for Polish. An overview of existing Polish synthesis systems is presented in Section 6.4. In Section 6.5 we introduce a research tool used for this study, namely Festival TTS, and in Section 6.5.1 we go on to describe the Polish TTS voice. Finally in Section 6.5.1 we present methods of intonation generation available in Festival.

6.1. Overview of speech synthesis cycle

A general text-to-speech synthesiser has two major components: a natural language processing module (NLP) and a digital signal processing module (DSP). The NLP module is responsible for producing a symbolic representation of the text to be read out by the synthesiser. It should tokenise and normalise the textual input by performing text analysis. Regardless of system architecture, it will include a) detecting end of sentence, b) resolving mark-up languages, c) identifying tokens in raw text, and d) mapping them into identifiable words. Text normalisation will follow and is a means of resolving abbreviations, acronyms, cardinal and ordinal numbers, dates, currency and time indications, addresses, telephone and bank account numbers, email and web addresses and emoticons, to name a few. The next component within the NLP module is the grapheme-to-phoneme conversion (G2P), also called letter-to-sound (LTS) conversion. This can be accomplished, depending on language and system through a set of rules either manually written or automatically trained, and/or a lexicon lookup. The rules need to accomplish a variety of tasks, accounting for various phonetic/phonological intra and inter-word processes present in languages. Being responsible for the automatic conversion to phonetic transcription they need to take into account vowel reduction, consonant cluster reduction, vowel lengthening, coarticulation, assimilation, liaison, pronunciation of heterophonic homographs, proper names as well as novel or foreign words.

Many systems perform a deep morphological decomposition, using stem and inflection lexicons together with grammars to assign part-of-speech category to words, and this information can later be used in homograph disambiguation, dealing with idioms, compounds and syntactic analysis. Apart from LTS conversion, words are syllabified and receive stress information which can later be modified (accent shift, deaccentuation) by phrasing and syntactic modules. At this stage it is important to group the input into clause- and phrase-like constituents. Apart from phonetic transcription, the final symbolic representation will include, phrase boundary placement, pitch accent place and type as well as duration specification. The duration/pitch accent specification is only required in pre-unit-selection systems, where there is only one instance of any diphone available. Whereas in unit-selection systems the goal is to find the most suitable unit in a huge speech database, in diphone systems the pitch and durations of units need to be manipulated.

After the NLP module has created a conversion from raw or marked-up text to a phonetic-prosodic representation, a digital processing module (DSP) needs to convert it into a speech waveform. This is the last step in a TTS cycle. Historically there are two main approaches to waveform creation: synthesis-by-rule and synthesis by concatenation. Rule synthesisers usually appear in the form of formant synthesisers, which aim to describe speech as composed of parameters related to formant frequencies and bandwidths. Due to the difficulty of writing rules estimating formant frequencies and bandwidths from speech data, they are error-prone and time consuming to develop. Another type of rule synthesisers are articulatory synthesisers, which model the movement of the articulators and the acoustics of the vocal tract (in terms of area functions or gestures which specify the target area functions). The acoustic signal is derived by means of an acoustic theory of speech production. The synthesis quality of rule synthesisers is suboptimal and suffers from unnatural sounding speech problem.

Concatenative synthesisers, are the present-day dominant approach to the DSP-stage of text-to-speech synthesis. As the name suggests, they concatenate pre-recorded units of speech. This is a completely different philosophy to the formant synthesisers. The actual task of the synthesiser is to produce an adequate sequence of concatenated segments, extracted from a speech database. The original intonation and the duration of segments needs to be adjusted to the parameters predicted by the NLP module. Various DSP techniques are employed to go from such an acoustic specification to an actual waveform, the two main ones being the hybrid Harmonic/Stochastic (H/S) model of Abrantes et al. (1991) and the Time-Domain Pitch-Synchronous-OveraLap-Add (TD-PSOLA) one (Moulines & Charpentier, 1990). PSOLA synthesisers, producing very high speech quality combined with a very low computational cost, are now widely used in the speech synthesis community.

6.2. Prosodic modelling techniques in speech synthesis

6.2.1. Component description

For synthesis of natural-sounding speech, it is essential to control prosody, to ensure appropriate rhythm, tempo, accent, intonation and stress. Segmental duration control is needed to model temporal characteristics just as fundamental frequency control is needed for tonal characteristics. For a model of prosody to be a description of a language, instead of just coding F0 contours, linguistically important features like phrasing, organization into prosodic phrase constituents which more closely relates to its expected prosodic realization, need to be taken into account.

There are thus four main components in a process of assigning prosody within a TTS system:

Prosodic phrase modelling

To generate suitable prosody, it is necessary to identify and break utterances into phrases and clauses and decide where intonational boundaries are. The placement of intonation phrase boundaries is important if the systems are to imitate natural speech, where longer utterances are cut into individual phrases characterised by their pitch accents and boundary tones.

There are statistical methods, e.g. CART (Breiman et al., 1984), for predicting prosodic phrases directly from POS information, position in a sentence, punctuation and other relevant factors. Some systems offer a full syntactic parse, e.g. MITalk (Allen et al., 1987), and others use a more superficial approach, identifying and using the position of function words to detect the phrase breaks, e.g. the simple phrase model in Festival. Some use punctuation information only, which can be useful but is not sufficient.

Duration modelling

Predicting the duration of each phone is dependent on many factors. What needs to be considered is the intrinsic duration of phonemes, their phonetic context, stress level, position within the word, the syntactic structure of the phrase and sentence, the type of word to name just a few. An example of a system, where duration rules proposed by Klatt, known as Klatt rules, were implemented is MITalk (Allen et al., 1987).

Klatt rules predict the segmental duration by multiplying the intrinsic duration of a given segment with a context dependent factor value. The result is then added to a segment specific minimal duration which can also be multiplied by a context-dependent factor. Klatt rules were modified and implemented for languages other than English, e.g. Swedish (Carlson & Granström, 1986), German (Kohler, 1988) or (Brinckmann & Trouvain, 2003) , French (Bartkova & Sorin, 1987) or Brazilian Portuguese (Simoes, 1990).

The most frequent methods of predicting duration in speech synthesis have been based on rule-driven models. However, statistical approaches have also been utilised where the development method is based on the statistical analysis of a prosodic database. Campbell (1989) has shown that a neural network can be trained to perform as well as the Klatt rules, and Riley (1992) used Classification and Regression Trees (CART) to automatically cluster segment durations according to their context.

The Bell Labs TTS System employs a so called 'sums-of-products' model (van Santen, 1994). Construction of the duration model is performed using an inferential-statistical analysis of the speech corpus and parameter fitting. The parameters of this quantitative duration model are fitted to a segmented speech database. Van Santen's method, shown to be superior to CART-based approaches, has been applied to American English (van Santen, 1994), Mandarin Chinese (Shih & Ao, 1997), and German (Möbius & van Santen, 1996).

Corpus-based techniques in TTS duration modelling have also been borrowing from advances in automatic speech recognition. For example, duration modelling using the sums-of-products model, has been extended to incorporate data transformations in the data-driven estimation process (Bellegarda et al., 2001).

As the quality of speech synthesis improves, a greater demand is put on the intonation system to produce more varied intonation tunes. Intonation prediction, the process of generating an intonation contour by means of a intonation model, can be split into two tasks:

1. Prediction of place and types of accents: (and/or boundary tones). This is done on a per syllable basis, identifying which syllables are to be accented as well as what type of accent is required (if appropriate for the theory).

2. Realisation of the F0 contour: given the accents/tones the F0 contour is generated.

Accent prediction and assignment

The task of predicting a prosodic contour consists of two tasks in most TTS systems: 1) the prediction of pitch accents and boundary tones coded as intonation labels and 2) the generation of contours from these labels. Decisions need to be reached as to which words should be accented. This is done on a per syllable basis, identifying which syllables are to be accented and then predicting the category of accent and/or boundary tones. Predicting the position and type of pitch accents and/ or boundary tones is still an on-going research topic.

Classification and regression trees (CART) (Breiman et al., 1984) are widely used in speech synthesis to model accent prediction (Syrdal et al., 2001; Hirschberg, 1993; Sun, 2002b). Decision trees are widespread machine learning algorithms that have also been extensively applied to F0 generation e.g. (Dusterhoff et al., 1999; Sun, 2002a), duration prediction (Ripley, 1993), intonation phrase prediction (Wang & Hirschberg, 1992; Sun & Applebaum, 2001), pitch accent prediction (Hirschberg, 1993; Sun, 2002b), and pitch contour generation (Dusterhoff & Black, 1997; Dusterhoff et al., 1999). In this thesis, the CART approach is used, and its application is described in detail in Chapter 9, section 9.1.

F0 contour prediction and generation

Once accents' place and type is specified, pitch accents/tones and boundaries need to be mapped to their acoustic correlates. Given the parametric representation in the model the final F0 contour can be generated.

How the F0 is generated depends on the type of intonation model chosen for its representation. A number of intonation theories have been utilised in various systems to try to undertake this task. From the three major types of intonation description formalisms (acoustic, linguistic and perceptual) two major classes of models of generating F0 have emerged: superpositional models and tone sequence models. They differ in how they define the relation between local movements and global trends in the intonation contour. Superpositional or hierarchical models define F0 contours as complex patterns resulting from the superposition of several components and thus generate an F0 contour by modelling factors separately (phone, syllable, word, phase, sentence) and then combining the partial models. Tone sequence models, on the other hand, claim that F0 contours are generated from a sequence of phonologically distinctive tones that are locally determined and do not interact with each other.

Superpositional models Superpositional model is a hierarchical model including several layered non-categorial components which concurrently determine the shape of F0 contour. This approach is used e.g. by Grønnum to model Danish intonation (Grønnum, 1992). In the model a text contour, a sentence contour, a phrase contour, and a component for stress group patterns are integrated. The stress group is viewed as an entity consisting of a stressed syllable and following unstressed ones. A final component models microprosodic effects. The basis for the intonation model is the source filter approach developed over two decades by Fujisaki & Hirose (1984) and Fujisaki (1992) for Japanese.

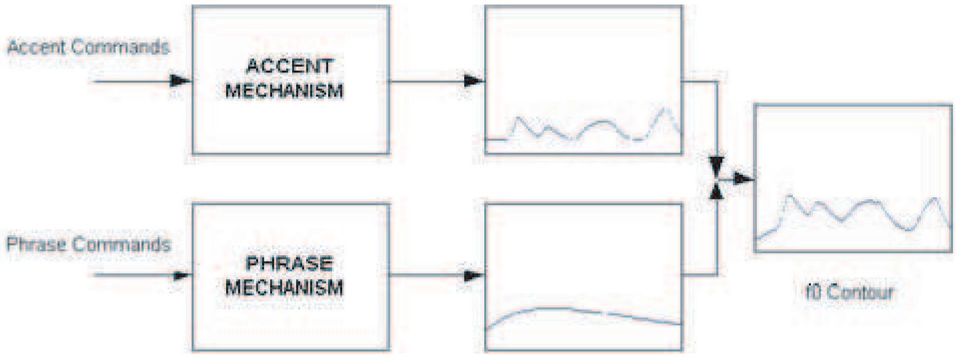


Figure 6.1.: *The Fujisaki Model*

In Fujisaki model, the F0 contour is treated as a linear superposition of accent and phrase commands. The phrase command acts over the domain of the intonation phrase, shaped as an initial rise followed by a long fall to an asymptote line. This is generated by a phrase control mechanism, activated by a pulse command with varying magnitude. The accent command is a local peak on an accented syllable, generated by the accent-control mechanism, see Figure 6.1. This is called by a binary step function with duration and amplitude parameters. The F0 contour is obtained by filtering each sequence of commands and combining the output of the two filters superimposed on a baseline F0 value. As a means of generating an F0 contour, the Fujisaki model has been used in various TTS systems and languages (Hirose et al., 1986; Allen et al., 1987; Bailly, 1989; Fujisaki & Ljungqvist, 1986; Möbius et al., 1993; Hirai et al., 1994).

Tone-sequence models Linear or Tone sequence models claim that F0 contours are generated from a sequence of phonologically distinctive tones, or categorially different pitch accents, that are locally determined and do not interact. Accordingly, they propose generating F0 contour from left to right as a sequence of discrete values or movements/tones.

Such a model is characterised by

- Linearity of tonal structure,
- Distinction between pitch accent and lexical stress,
- Analysis of pitch accents in terms of level tones,
- Local sources for global trends.

The most influential example is Pierrehumberts model for American English (Pierrehumbert, 1980). Pierrehumbert based her model on the autosegmental-metrical approach, with pitch movements being decomposed into pitch levels. Pierrehumbert defines three types of accents: pitch accents, boundary tones and phrase accents, to represent the pitch movements on stressed syllables, at phrase boundaries and between a stressed syllable and phrase boundary, respectively.

The original model has been extended by Beckman & Pierrehumbert (1986), and has evolved into a standard set of conventions for transcribing intonation of American English - Tone and Break Indices (ToBI) (Silverman et al., 1992).

Intonation phrases are modelled as sequences of (H) high and (L) low pitch levels. These level are represented with several symbols marked in the utterance to yield an orthographic representation, which is then converted to a pitch contour for that utterance, cf. Fig. 6.2. High (H) and Low (L) pitch targets are predicted from the text, and are aligned with the text by means of linguistic processing, preceding intonation prediction in the TTS process. The rules predicting the height of the targets as well as their temporal alignment need to take several levels of structural information into account:

- Tone specification,
- Segmental information,
- Downtrend effects (declination, downstep, sentence final F0 lowering),
- Prominence of each accent.

ToBI offers a well-defined and widely available labelling system. The ToBI labelling system itself does not define a mechanism to go from the labels to an F0 contour, or the reverse.

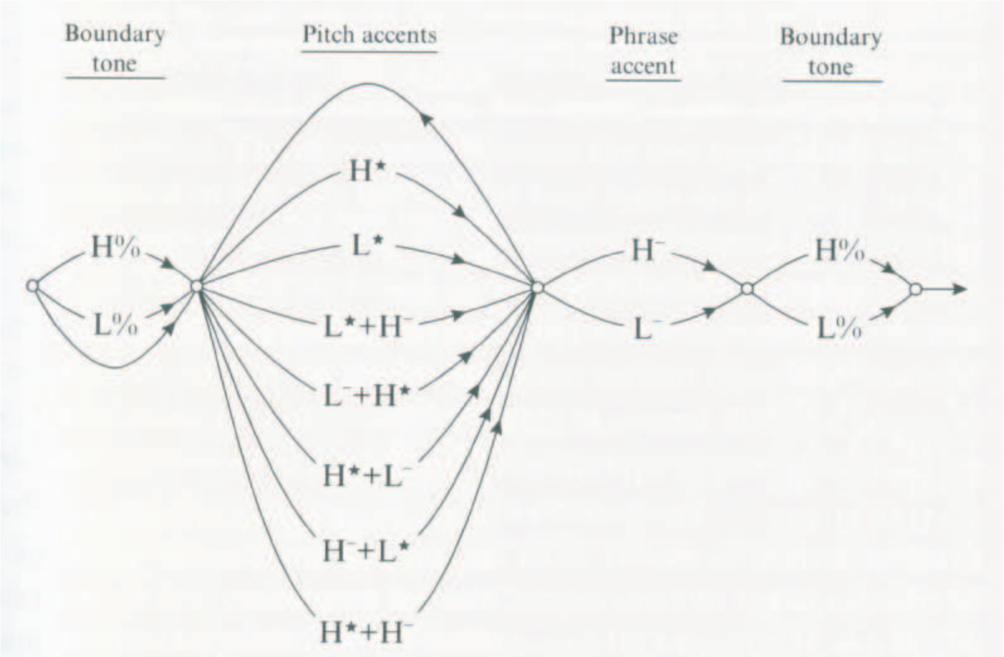


Figure 6.2.: *The ToBI Model*

However there are both hand written rule systems e.g. (Anderson et al., 1984) and statistically trained methods e.g. (Black & Hunt, 1996) which carry out this task for English. The popularity of ToBI can be measured by the number of languages which have been given a language specific ToBI version. Thus, transcriptions can be compared across dialects and languages: ToBI for English, GToBI for German (Grice & Baumann, 2002), SCToBI for Serbo-Croatian (Godjevac, 2001) or ToDI for Dutch (Gussenhoven, 2004).

Generating F0 contours from ToBI labels can be accomplished by a variety of methods, depending on the TTS system. One such method is what Black & Hunt (1996) call 'APL' see Anderson et al. (1984), which predicts by rule a number of pitch targets per syllable marked with a pitch or phrase accent, or a boundary tone (Black & Hunt, 1996). There are top, reference and base values, decreasing over time, set by hand through experimentation, which are smoothed to produce an F0 contour.

Linear regression is a method used to automatically predict the optimal F0 values for ToBI labels (Black & Hunt, 1996). This approach predicts three F0 values for every syllable (start, middle and end). Linear regression models assume that a predicted variable (p) can be modelled as the sum of a set of weighted real-valued factors 6.1.

$$p = w_0 + w_1 f_1 + w_2 f_2 + w_3 f_3 + ... + w_n f_n \tag{6.1}$$

The factors (f_i) represent parameterised properties of the data (accent type, positions in the phrase, etc.) and the weights (w_i) are trained using linear regression. Accents are represented by features representing the group an accent falls into (H*, !H*, L*, L+H*, etc.). The same strategy is used for Phrase accents and boundary tones. Break indices (1 to 4) can also be modelled in this approach using the same set of features. The features used for each syllable can comprise

- Accent type on current syllable,
- Boundary tone on current syllable,
- Break index type on current syllable,
- Lexical stress on current syllable,
- Number of syllables from start and to end of current phrase,

- Number of stressed syllables from start and to end of current phrase,
- Number of accented syllables from start and to end of current phrase,
- Number of syllables since last accented syllable.

Results for English and Japanese (Black & Hunt, 1996) suggest that a linear regression method offers a better modelling of the F0 contour than a rule-based model. One has to bear in mind that a fully labelled database is required to be able to build such a model, and lack of prosodically labelled data is the biggest hurdle to overcome before we can benefit from it.

Neural networks have also been applied extensively to prosody modelling e.g. (Buhmann et al., 2000; Chen et al., 1998; Sun, 2001; Traber, 1992). In this line of approach, a neural network is used to learn the mapping between the input linguistic/acoustic features and output targets, which are usually F0 values or some transformed parameters. With appropriate structures, a neural network can capture the complex relationship between linguistic features and intonation. In the current study, linear regression modelling is applied to the task of predicting F0. The model is described in Chapter 9, Section 9.2.

6.3. Modelling Polish Prosody for Speech Technology

In this section we are going to review studies on modelling prosody for the Polish language which have concentrated on some or all of the components described in the previous section.

One of the first attempts at automatic modelling of fundamental frequency contours and gaining control over F0 with view to applying it to speech synthesis was the study by Demenko et al. (1993). The authors applied the Fujisaki model cf. Section 6.2.1 to approximate the fundamental frequency contour. The basis of the analysis of Polish intonation were read newspaper passages with an acoustic analysis window of sentence size. On the sentence level, the aim was to establish values for declination and a way of controlling phrase and accent commands. At a lower, syllable level, typical pitch movements and their position in a phrase were determined. To accomplish the task, the F0 contours in all types of structures were parameterised using their F0 range (max and min F0 values) and the rate of F0 change (within word and clause). The clause was divided into three parts and F0 control rules were designed for each of them : the initial part (including the first stressed syllable), the medial and the final part, containing the penultimate and last stressed syllable. Overall, the study applied the Fujisaki concept, but on the basis of a perceptual study, the parameters and the way of controlling the phrase and accent command were modified.

Later attempts at modelling Polish intonation contours use neural networks to parameterise and model F0 contours. Demenko & Jassem (1999b,a) use a three layer MLP network to classify nine nuclear accents: HL, ML, LL, HM, LH, LM, MH, MM, LHL and two prenuclear accents High and Low, with L for low, M for mid and H for high, (for description of the labels see Section 2.1.2). This classification follows the British tradition of intonation analysis for English by O'Connor & Arnold (1973) and Jassem (1996). The material used in the study comprised of isolated constructed phrases as well as read texts. The constructed phrases were spoken by a phonetician and then imitated by students. Only the close imitations were later used in the classification tasks. In the case of the read text the nine nuclear tunes were combined into four classes of F0 contours : Rising, Falling, Rising-Falling and Level. Table 6.1 shows the membership of the classes. In addition, the study classified the prenuclear contours (H and L) and the anacrusis (P).

The parameters for describing the nuclear tunes are based on the shape and range of F0 excursions, and the speaker's range and F0 minimum. The features defined pertained to variations of F0 during vowels and inter-

Classes	Pitch Accents			
Rising	LH	MH	LM	
Falling	HL	HM	ML	LxL
Rising-Falling	LHL			
Level	MM			

Table 6.1.: *Classes of nuclear accents in read texts*

syllabic F0 pitch values. The average classification success rate was 82% (Demenko & Jassem, 1999b). In the case of initial unaccented syllables the classification rate was 60% and 67% in the training and verification set, respectively. The so-called 'primary accents', (LH, HL, LHL), were best predicted, in over 80% of the cases. Distinguishing between primary and secondary accents proved to be problematic due to the fact that most of the 11 acoustic features were related to the phrase final events. For prenuclear (secondary) accents, a classification rate of 79% - 83% was obtained.

Neural networks were also the preferred method in another study (Dymarski et al., 1995) where a diphone synthesiser uses it for its text-to-phoneme conversion. The pitch, energy and duration were controlled by time-domain interpolation of the excitation signal as well as pitch period insertion. The original prosody control in the system was revised and Kula et al. (2001) reported on a new, word- and phrase-based prosody model implemented in the original system. The sentence intonation model recognised three basic types of F0 contour:

- An affirmative sentence contour which descends towards the end of the phrase,
- A yes/no question contour which rises towards the end of the phrase,
- A wh-question contour which starts higher than the two above and also slightly rises at the end of the phrase.

Controlling the pitch on the syllable level distinguished between seven types of syllable, based on their location (penultimate or antepenultimate) and the presence of a punctuation mark (weak, strong or a question mark). Additionally, an algorithm was designed to compute lengthening factors used in duration control. It took into account type of syllable and phoneme, as well as final and pre-boundary lengthening.

The following section gives examples of some of the existing TTS systems for Polish.

6.4. Polish Speech Synthesis Systems

Following the interest in advances in text to speech synthesis for other languages, there has also been research on speech synthesis for the Polish language. The co–operation between Adam Mickiewicz University and the Institute of Fundamental Technology Research at the Polish Academy of Sciences brought about one of the first systems for Polish, a formant based TTS system (Imiołczyk et al., 1993). Input to the system were synthetic allophone parameters, declination line tables and accent movement tables, obtained by applying the Fujisaki method of F0 contour approximation.

As described in Imiołczyk et al. (1993), it was an example of a formant synthesiser based on a source-filter model, containing an algorithm for deriving phonemic transcription, segmental rules for allophonic variation as well as suprasegmental rules. The suprasegmental rules were divided into context dependent duration rules (300 phone lengthening and shortening instructions), and Fujisaki based rules for shaping intonation contours.

Adam Mickiewicz University and Poznań University of Technology designed a triphone synthesis system called ORATOR. It was based on a database of diphones and triphones (6000 triphones of type CVC). There are also hardware synthesisers for Polish, with Auto-Lektor, produced by Harpo, Apollo, Kubuś and others, which are, however, inferior in quality to the software systems described below.

The next generation of synthesisers, developed in the mid and late nineties, were diphone-based systems, examples of which are one developed at the Institute of Telecommunication, Technical University of Warsaw (Dymarski et al., 1995) or the author's own system, developed as part of Festival TTS at the University of Edinburgh (Oliver, 1998). A similar system based on the same linguistic module was created by Krzysztof Szklanny from the Polish–Japanese Institute of Information Technology who built a Polish voice for the MBROLA project. POLVOC, built as part of a co-operation between the Institute of Telecommunication at the Technical University of Warsaw and Elan Informatique is another example of a diphone based TTS system, and uses a TDPSOLA algorithm. Among the commercial systems are diphone voices "Speaker", "Ivan" and "Ivona" from IVO Software, a TTS system LEKTOR offered by a company called Drive, and Syntalk, developed by NEUROSOFT.

The most popular TTS systems at the moment are unit selection systems based on corpus synthesis and producing very natural sounding speech. The most widely known unit selection systems which exist for Polish are RealSpeak by Nuance (formerly Lernout and Hauspie, later Scansoft and currently Nuance), BOSS (Bonn Open Synthesis System) Polish voice developed jointly by Adam Mickiewicz University and the Institute for Communication Research and Phonetics (IKP), Bonn (Stöber et al., 2000), a Polish unit-selection voice developed in cooperation with the Polish–Japanese Institute of Information Technology (Oliver & Szklanny, 2006), as well as systems by Acapela and SVOX.

6.5. Festival Text-to-Speech Synthesis system

The base Polish synthesis was developed as a new voice in the Festival Speech Synthesis System (Black & Taylor, 1998), a diphone-based synthesiser which offers the standard method for producing speech from segments, durations and intonation targets.

Festival is a general speech synthesis system developed at the Centre for Speech Technology Research at the University of Edinburgh, UK. It offers a full text to speech system with various Application Programming Interfaces, as well an environment for the development and research of speech synthesis techniques. It is written in C++ with a Scheme-based command interpreter for general control. It offers residual-excited LPC based synthesis (Hunt et al., 1989), and PSOLA (Moulines & Charpentier, 1990). As an Open Source speech synthesis system it is designed for development and runtime use. It is being used in many commercial, e.g. AT&T , and academic systems. It is distributed with RedHat 8.x versions, supports many other linux/unix distributions and can also be used under Windows, e.g. Cygwin.

Festival is truly multilingual, has no built-in language and is designed to allow the addition of new languages. The process of building new synthetic voices has been made much easier thanks to the CMU Festvox Project, which enables the creation of tools, scripts, documentation, step by step walkthroughs for particular processes and examples for building voices. There are example voice databases for various languages ready to be used and there is support for external databases from the MBROLA project.

Researchers have a choice of waveform synthesis methods, as Festival supports diphone, unit selection and limited domain synthesis. It offers support in the form of text analysers (POS tagging, tokenisation, lexicon building), statistical and rule-based grapheme-to-phoneme conversion, lexicon building tools and a range of statistical learning tools designed to aid prosodic modelling techniques. Additional tools include autoaligner for labelling recorded speech, and the scripts for building duration, intonation and phrasing modules. Together with Edinburgh Speech Tools library, it can be used for signal processing; it performs F0 analysis and intonation analysis (Tilt, ToBI).

The system has a modular architecture and its modules are replaceable. All in all, Festival can be used as a generic text to speech system, FestVox project as a tool for building voices. Additionally, there is a smaller version of the system, a portable synthesis engine called Flite, designed for embedded systems, which can be used for enabling TTS on portable devices.

6.5.1. System overview

Polish module

The development of a Polish voice in Festival was carried out according to the following steps, standard for adding any voice to Festival.

1. Phone set
   For most new languages and often for new dialects, a new phone set is required. It is really the basic building block of a voice and most other parts are defined in terms of this set, see Appendix B.

2. Producing the diphone database
   The text of the database was designed in the form of pseudo-Polish nonsense words, including all possible phoneme pairs: V-C, C-V, C-C, V-V, V-#, C-#, #-C, #-V. The text was written in transcription to avoid coarticulation and orthographic ambiguities of real words. The nonsense words are three-syllable words with the medial syllable filled by all phoneme combinations. In addition, all phonemes appear initially and finally to derive silence-phoneme and phoneme-silence diphones.

3. Recording and processing the database
   A female native speaker of standard Polish was recorded. The recordings took place in an Edinburgh recording studio and were subsequently digitised.

4. Transcription conventions

   The broad phonetic transcription is based on SAMPA with some modifications, i.e., where SAMPA for Polish suggests a plus sign " + " for palatalisation, it was replaced by a single accent sign " ' " , to avoid Scheme language conflicts, see Appendix B.

5. Lexicon and Letter-to-Sound Rules

   A Lexicon in Festival is a subsystem that provides pronunciations for words. Polish is a language whose pronunciation can be predicted from its orthography so the lexicon is small and most of the work of generating phonetic transcription is accomplished with letter-to-sound rules designed by an expert.

6. Intonation

   For intonation there are a number of simple options which do not require training data. Within the MSc project (Oliver, 1998), the hat pattern was used on all stressed syllables in content words and on single syllable content words. The scope of the project did not allow any deep treatment of intonation, so the voice is neutral and at times sounds rather monotonous.

7. Duration

   Using the Classification and Regression tree method (CART) (Breiman et al., 1984), longer durations are predicted in stressed syllables and in clause initial and clause final syllables.

8. Waveform synthesis

   There are a number of choices for waveform synthesis currently supported. In this project LPC PSOLA and a diphone database for Polish were used.

9. Other linguistic processing

   This part enables the synthesiser to process numerals, homographs, and acronyms, and was adapted to cover Polish.

10. Evaluation

    The synthesiser is supposed to read any text. In this case any on-line publication, e.g. a newspaper. The assessment part followed SAM procedures for spoken language systems assessment (Gibbon et al., 1997)

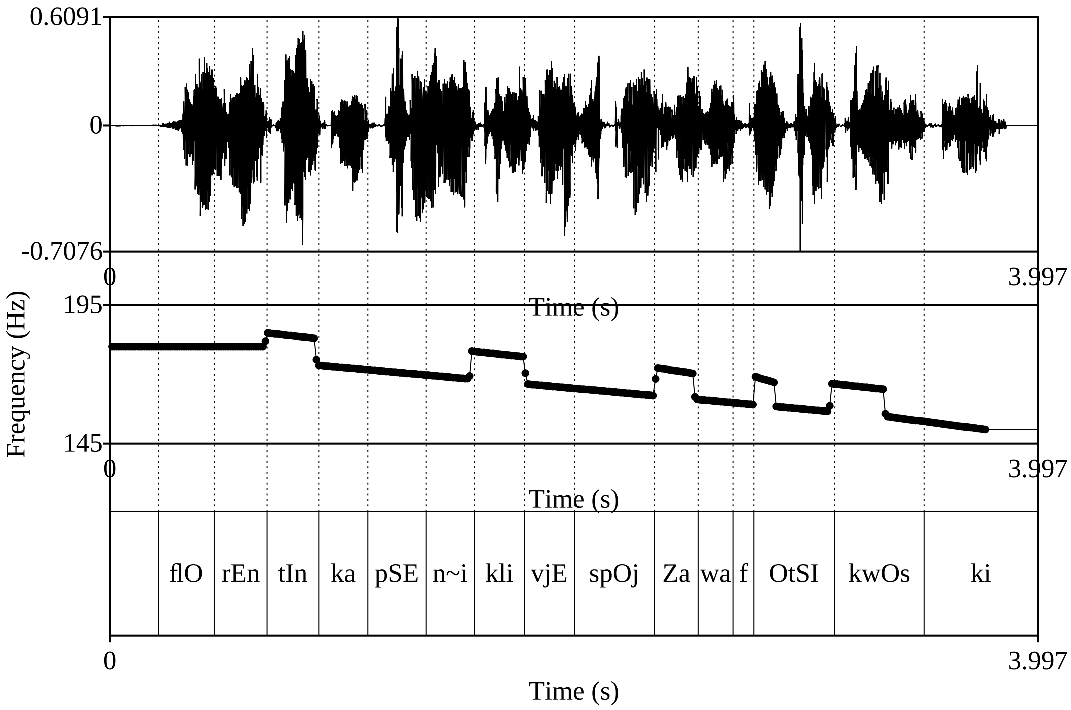


Figure 6.3.: *Hat pattern F0 contour for a sentence 'Florentynka przenikliwie spojrzała w oczy Kłoski'.*

Festival intonation generation

In general, intonation is generated in two steps:

- Prediction of accents (and/or boundary tones) on a per syllable basis,
- Prediction of F0 target values. This must be done after durations are predicted.

In Festival there is a choice of different intonation modules giving varying levels of control over pitch generation.

Default intonation This is the simplest form of intonation, which simply creates a target at the start of the utterance, and one at the end. The default values are 130 Hz and 110 Hz. Other values may be set through the parameter duffint_params. For example the following will generate a monotone at 150Hz.

```
(set! duffint_params '((start 150) (end 150)))
(Parameter.set 'Int_Method 'DuffInt)
(Parameter.set 'Int_Target_Method Int_Targets_Default)
(Black & Taylor, 1998)
```

Simple intonation This module uses a CART tree to predict whether each syllable is accented or not. A predicted value of NONE means no accent is generated by the function. Any other predicted value will causes 'hat' accent to be put on that syllable.

A default cart tree predicts accents on the stressed syllables in poly–syllabic content words, and on the only syllable in single syllable content words. There are two interesting parameters a) f0_mean, which gives the mean F0 for this speaker (default 110 Hz) and b) f0_std , which is the standard deviation of F0 for this speaker (default 25 Hz). This second value is used to determine the amount of random variation to be added to in the generated targets.

The F0, generated for each phrase, declines over the length of the phrase until the last syllable. An imaginary baseline is drawn from the start to the end (minus the final extra fall), For each accented syllable three targets are added, one at the start, one in mid-vowel, and one at the end. The start and end are at set to the baseline (as defined for that syllable) and the mid-vowel value is set to baseline+f0_std. This model is not supposed to be complex or comprehensive but it offers a very quick and easy way to generate something other than a straight-line F0.

Tree intonation This module is more flexible, as two different CART trees can be used to predict 'accents' and 'end-tones'. Although at present this module is used for an implementation of the ToBI intonation labelling system it could be used for many different types of intonation system. The target module for this method uses a Linear Regression model to predict start, mid-vowel and end targets for each syllable, using arbitrarily specified features. This follows the work described in Black & Hunt (1996).

The system within which the current work is carried out is a Polish voice module for Festival (Oliver, 1998).

The original intonation model uses a simple prosody prediction and generation module and does not require training data. The model uses hand written classification and regression tree to predict if a syllable is accented or not. It then predicts accents on the stressed syllables in content words and on single syllable content words, on which a 'hat' shaped pattern is applied (see Fig. 6.3). The module uses a simplified version of the above as the function words had been listed and put in a voice description file.

Based on the recorded material, speakers F0 mean and F0 standard deviation are added as parameters to the model. F0 standard deviation is used to determine how much F0 variation should be produced in the generated targets. For each phrase in the given utterance an F0 is generated which declines over the length of the phrase. A baseline is drawn from the start to the end excluding the final extra fall. For each accented syllable three targets are added, one at the start, one in mid vowel, and one at the end. The start and end have the baseline value as declined for that syllable and the mid vowel is set to the baseline value increased by the value of F0 standard deviation.

6.6. Conclusions

A general data-driven procedure for creating new prosodic modules for the FESTIVAL text-to-speech synthesiser comprises modules based on rules or on machine learning methods such as the Classification and Regression Trees (Breiman et al., 1984) and linear regression (Black & Hunt, 1996). The prosodic module of the text to speech system is aimed at computing the values of a set of prosodic variables. In a minimal configuration these variables are the phoneme duration and F0 values.

The work described in this thesis aims to improve upon the very simple intonation models that the original module relies upon. Modelling accent placement and type will be accomplished using classification and regression trees and the prediction of F0 values will utilise a linear prediction model. Prosodic modelling will be discussed in detail in Chapter 9. Prerequisites for the modelling are a good representation of the F0 signal in the database as well as a prosodically labelled database. In Chapter 7 methods of F0 contour stylisation will be presented and Chapter 8 will deal with the principles and procedures for re-labelling the existing speech database to obtain a set of pitch accent labels for pitch-accent modelling.

7. Analysis of accent types in Polish: F0 stylisation

Introduction

The analysis of intonation is a challenging process. Automating the analysis and interpretation of F0 data can be even more complicated. It can nevertheless provide tools for corpus analysis and can be used as a testbed for linguistic theories. In practical fields like speech synthesis F0 data provides a description from which synthetic contours can be generated in a way resembling the natural F0 contour. Efforts are made to reduce the amount of data required to generate such contours, isolate the functional part and define a valid representation of the underlying contour.

In this study we aim to derive prototypical pitch contour types found in Polish based on their acoustic characteristics derived from a stylised F0 curve which should be perceptually indistinguishable from the original.

We present a method for obtaining precise acoustic representation of underlying F0 contour within pitch accent classes in Polish by means of automatic F0 stylisation. The term is used here for a procedure that modifies the measured F0 contour of an utterance into a more simple but functionally equivalent form, preserving the perceptual identity of the original signal, i.e. 'a close copy' of it ('t Hart & Collier, 1975; Pijper, 1979) . This chapter is organised as follows: in Section 7.1.1 we give an overview of F0 stylisation methods, followed in Section 7.1.2 by the description of the speech resources used for the study. Sections 7.1.3 and 7.1.4 describe Prosogram and Momel F0 stylisation methods in detail and Section 7.2.1

introduces the modifications made to the Momel method. Finally, in Section 7.3 we present a perceptual study designed to evaluate the F0 stylisation methods.

7.1. Stylisation of fundamental frequency

7.1.1. F0 stylisation techniques

There are a number of methods used to stylise and model fundamental frequency curve; purely acoustic ones and those based on the knowledge about human perception of pitch. Notably, research at IPO, which developed into a general theory of intonation, first attempted to stylise F0 contours with the aim to develop an intonation model for Dutch speech synthesis (Cohen & 't Hart, 1965). The stylisation approach was based on the principle that the simplified F0 curve must be melodically identical to the original curve ('t Hart & Collier, 1975). Such stylisation, coined 'close copy stylisation' by Pijper (1979), involved piecewise linear approximations. Perceptual experiments showed it to be 'perceptually equivalent' to the original signal ('t Hart et al., 1990). The technique was subsequently applied to languages other than Dutch, notably English (Pijper, 1983; Willems et al., 1988) and German (Adriaens, 1991).

An approach using knowledge of perceptual models is also used in Prosogram, a method proposed by d'Alessandro & Mertens (1995), where stylisation is based on a model of tonal perception by humans taking the syllabic nucleus as a basic unit. The F0 contour is transformed into a sequence of tonal segments which are either static or dynamic as a function of a glissando threshold which varies with the duration of the syllable and the amplitude of the F0 change. The method takes two further factors into account, namely, segmentation into syllabic (or vocalic) nuclei and the differential glissando threshold (perceivable change in slope).

An example of a purely acoustic method is Taylor's 'Tilt' model (Taylor, 2000) analysing the F0 curve as a linear sequence of three events (Rise, Fall and Connection). Rise and Fall are interpreted as piecewise parabolas and the Connection element is interpreted as a linear transition. The representation of Rise and Fall elements is considered equivalent to the quadratic spline method, thus placing this approach mid-way between the stylisation of F0 using linear transitions and treating it as a sequence of parabolic segments. The latter is the approach used in the method discussed below.

The method applied in this thesis is is Momel (MOdélisation de MELodie) proposed by Hirst (1983) and automated by Hirst & Espesser (1993). This F0 stylisation method is based on the technique Hirst calls 'asymmetrical modal quadratic regression', for details see Section 7.1.4. The algorithm replaces the original F0 curve by a simpler numerical function and models the macroprosodic component of F0 as a quadratic spline function resulting in a continuous contour.

Unlike other methods, which use a sequence of straight line segments, MOMEL's quadratic spline function produces a continuous, smooth curve, without the angles which result from the concatenation of straight lines. By replacing the microvariations of the original F0 curve with a simpler numerical function it is assumed that his method preserves the original macroprosodic information carried by the gross F0 movements. The resulting curve maintains continuity because unvoiced segments are interpolated.

The argument for using quadratic spline is that

- quadratic spline produces a curve closer to original F0 curve than when using straight lines and thus introduces less noise into quantitative studies,
- stylisation by quadratic spline produces a macroprosodic contour which is practically identical to the F0 curve

> produced on utterances consisting solely of sonorant segments. (Campione et al., 2000)

In summary, we can divide the sylisation methods into two types: those that use linear interpolation and those that try to produce a smoother F0 curve using quadratic interpolation. Another division is to distinguish those methods which use a perceptual modelling as opposed to a purely acoustic approach. The following sections describe in detail the structure of a method using knowledge of perceptual models, namely the example of Prosogram, and an acoustic method, taking the example of MOMEL.

7.1.2. Resources

The current study uses the PoInt speech database of Polish (Karpiński & Kleśta, 2001), which has been recorded as part of the Polish Intonation Database Project. This database, as described in Chapter 5, contains a variety of discourse types: fragments of read literary texts, quasi-spontaneous monologues, as well as map-task-based dialogues. In order to avoid speech overlaps present in the map task material, we use a subset of the database, leaving out spontaneous speech utterances. The database contains recordings of male and female speakers and was phonetically and prosodically annotated on the syllabic level. Additionally, a Polish part of the Babel database (Gubrynowicz, 1998), described in Chapter 5, consisting of recordings of read passages was used.

7.1.3. Prosogram

The algorithm implemented by d'Alessandro & Mertens (1995) incorporates knowledge about speech prosody perception in the F0 stylisation process. This model takes into account the three factors:

- Segmentation into syllabic (or vocalic) nuclei,

- The glissando threshold (G),
- The differential glissando threshold (DG).

The model assumes that in speech perception syllabic decomposition takes place. The F0 contour of an utterance is seen as a sequence of short-duration tones, perceptually segmented on-line and highly dependent on their segmental context. This syllabic segmentation of the utterance should be instrumentally based on the phonetic segmentation, automatic or manual. The glissando threshold (G) relates to the auditory threshold for pitch variation, which depends on amplitude and duration of F0 variation. The effect, although investigated for years e.g. (Rossi, 1971, 1978), has not been studied on continuous speech. The optimal unit proposed for dealing with the glissando threshold is semitones per second (ST/s) ('h Hart, 1976).

The value of G can be calculated using:

$$G = \frac{0.16}{T^2} \tag{7.1}$$

where T is the duration of the pitch variation. Mertens (2004) compared different stylisations based on different thresholds to select G applicable to continuous speech. He proposes

$$G = \frac{0.32}{T^2} \tag{7.2}$$

twice the original threshold, as one closest to the annotation by expert human labellers.

The third factor, the differential glissando threshold (DG), refers to the minimum difference in slope necessary to distinguish between two successive tonal segments or glissandi. Its value set to DG=20 was suggested (the values reported in the literature are between 12 and 40) (d'Alessandro & Mertens, 1995).

The process of stylisation can be divided into three stages:

1. phonetic segmentation and syllabification

2. pitch determination and integration

3. stylisation of syllabic pitch contours.

phonetic segmentation and syllabification The segmentation of the speech material needs to be accurate, therefore both automatic or manual segmentation needs to be verified. The process takes the segmentation information, together with F0 intensity and voiced/unvoiced parameters, to organise input into phonetic syllables, each syllable around one local loudness peak with the neighbouring voiceless or less 'sonorous' material.

pitch determination and integration Next a pitch determination algorithm is applied, followed by a weighted time-average model (WTAM) derived from pitch perception research (d'Alessandro & Castellengo, 1994). The model performs a linear smoothing of the F0 obtained from the pitch tracker. The new contour obtained in this way is refered to as the Weighted Time Averaged Pitch (WTAP).

stylisation of syllabic pitch contours The stylisation proper consists of a segmentation of compound tones and an assignment of perceived pitch targets. The decomposition of tones is based on the differential glissando threshold (DG) and the glissando threshold (G). The content is segmented at points where important changes in the contour occur. The segmentation is recursive and stops when the overall glissando rate of the analysis window is below the glissando threshold. The segments, in turn, can be grouped together when the difference in slope is below the differential glissando threshold. The stylised contour is obtained by linear interpolation between targets from the voiced parts.

7.1.4. MOMEL algorithm

As mentioned in Section 7.1.1 MOMEL stylisation is based on an 'asymmetrical modal quadratic regression'. This means that during modal regression, parameters of the function are selected such that

> the values of the function are less than a given distance $\triangle$ from the largest possible number of items of a series. (Hirst et al., 2000)

The MOMEL algorithm consists of four stages:

1. preprocessing of F0
2. estimation of target-candidates $<< t, h >>$ (time and amplitude)
3. partition of candidates
4. reduction of candidates.

In stage (1), all F0 values more than a given ratio higher (typically 5%) than both their immediate neighbours are set to 0. Since unvoiced zones are coded to 0, this preprocessing eliminates one or two values (i.e. 10 to 20 ms) at the onset of voicing.

In stage (2) target candidates for each point(x) cf Fig. 7.1 are estimated in three steps

- within an analysis window of length A (typically 300 ms) centred on x, values of F0, including values for unvoiced segments, are neutralised if they are outside the range defined by two thresholds, hzmin and hzmax and are subsequently treated as missing values. Typical values for the thresholds are 50 Hz and 500 Hz respectively.
- a quadratic regression is applied within the window to all non-neutralised values. All predicted values which are more than a distance Delta (D) below the value of F0 are neutralised (typical D=5%). This step is reiterated until no new values are neutralised.

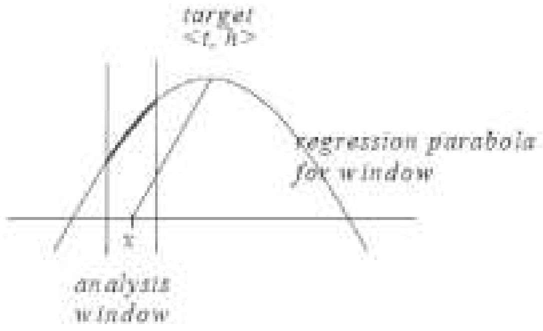


Figure 7.1.: *Calculation of a local target point*

- for each instance x, a target point $<< t, h >>$ (time and amplitude) is calculated from the regression coefficients. It corresponds to the extremum (min or max) of the corresponding parabola.

During stage (3) within a moving window of length R (typically 200 ms), the sequence of target candidates is partitioned. The moving window R is divided into two halves, left and right. The algorithm seeks values where there is a minimum difference between the targets in the left and right halves of the window. A partition boundary is inserted when the difference between the average weighted values of t and h in the left and right halves of the window corresponds to a local maximum which is greater than a threshold (set to the mean distance between left and right halves of all windows).

In stage (4) candidates are reduced. Within each section of the partition undertaken in stage (3), outlying candidates more than 1SD from the corresponding mean values for each section are eliminated. The mean value of the remaining targets in each segment is then calculated as the final estimate of t (time) and h (amplitude) for that segment.

7.2. Evaluation of Prosogram and Momel stylisation

In order to decide on the most suitable F0 stylisation method for further analysis, we present arguments for both acoustic and perceptual methods. Prosogram, as described in 7.1.3 is a perceptually motivated theory-independent stylisation method providing a transcription of tonal events without the bias of human transcribers. Its intrinsic limitations come from the fact that we still have limited knowledge about the way people perceive prosody. The method relies on the preliminary decomposition of the pitch contour into short-duration tones, which is assumed to take place in human perception. But it has not been widely investigated to what extent spectral and energy changes influence the perception of pitch. In this method it is therefore assumed that phonetic change results in intonation segmentation, which holds true for abrupt segments like stops, pauses but is not clear for more smooth transitions involving liquids or vowels. Additionally, the method requires phonetic segmentation since it bases its analysis on the voiced part of the syllable. It has only been tested for French and the model assumes a differential slope threshold for which there is very little data in the literature.

Momel, on the other hand, does not require any segmentation of the data or any pre-training. It is considered to be language independent, having been tested on French, English, Spanish, Italian, Arabic and Swedish. From the evaluation study on French (Campione & Véronis, 2000), there were 3.9% erroneous points and 6.09% silence (missing points). F0 curves stylised and interpolated with this method are considered to give a closer approximation of the original F0 curve, thus introducing less noise into quantitative studies. The method can be viewed as preserving the macro-prosodic information while eliminating microprosodic variations. However, our qualitative analysis of the method revealed systematic errors in the algorithm. In the preprocessing stage, it eliminates values (10 to 20 ms) at the onset of voicing. Target points were found in wrong places.

A second type of systematic error were redundant target points, and a third type were missing target points. Missed target points were often before a pause, which can result in a change from a rising to a falling contour, and vice versa. After a pause, it can result in an intonation event too high for the speaker's voice range.

In view of the fact that our data was not segmented, which precluded the use of the Prosogram model unless a separate segmentation was applied and manually verified, and since Momel errors were of a systematic nature, it was decided to apply and improve this acoustic method.

7.2.1. Momel modification

The modifications we undertook to improve the Momel algorithm were aimed at the third type of error, missing targets at the beginning and end of an utterance. In order to do this we needed to identify reasons why the algorithm is prone to error before and after a pause and what measures have been taken by the authors of the algorithm since its first version to correct it. Namely, when there is a concave tonal contour at the beginning or end of an utterance, the algorithm will only detect one target point, where intonation changes its direction.

Then, a target point at the same vertical and horizontal distance from the end or beginning of voicing as the end or beginning of voicing is from the previous (following) target is calculated (Hirst, personal communication) cf. Fig. 7.2. The assumption behind it is that, if a target point is put there, the stylised curve will capture the final rise or initial fall. Unfortunately, there is no way of knowing if this is the point that the speaker is actually aiming at.

The solution is not successful in all cases, see Figures 7.3 and 7.4 for encountered frequency errors if there are not enough points per signal.

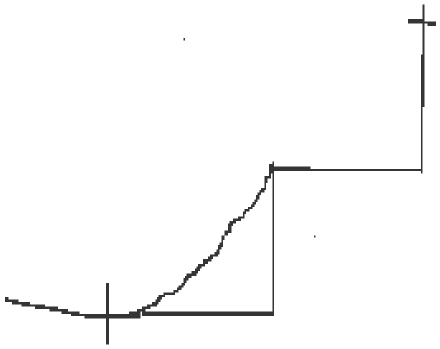


Figure 7.2.: *Calculating the last Momel point in an utterance*

We propose an alternative treatment of the problem. We do not calculate additional target points at the same vertical and horizontal distance from the end of voicing as the end of voicing is from the previous target, both at the beginning and end of the utterance. Instead, we take away the first and last Momel point computed according to the above strategy, as well as Momel points not in the speaker's range. We then check if the whole signal is covered by Momel points. In order to span the whole signal we add, if present, original F0 points which had previously been eliminated to the Momel target points at the boundary locations. In cases of unvoiced material at both ends of the signal, the values before the F0 onset and after the F0 offset are replaced with the first and last F0 value, respectively, found in the extracted contour, Fig. 7.5. This procedure is used by Mixdorff (2005) in the preprocessing stage for the Fujisaki-model parameters extraction.

In this way, elements of the stylised curve are enriched with the initial and final variations of F0, essential for capturing boundary prosodic events. The new continuous F0 curve is then derived by quadratic interpolation of the Momel target points and original F0 points added at the beginning and end of an utterance cf. Fig. 7.6.

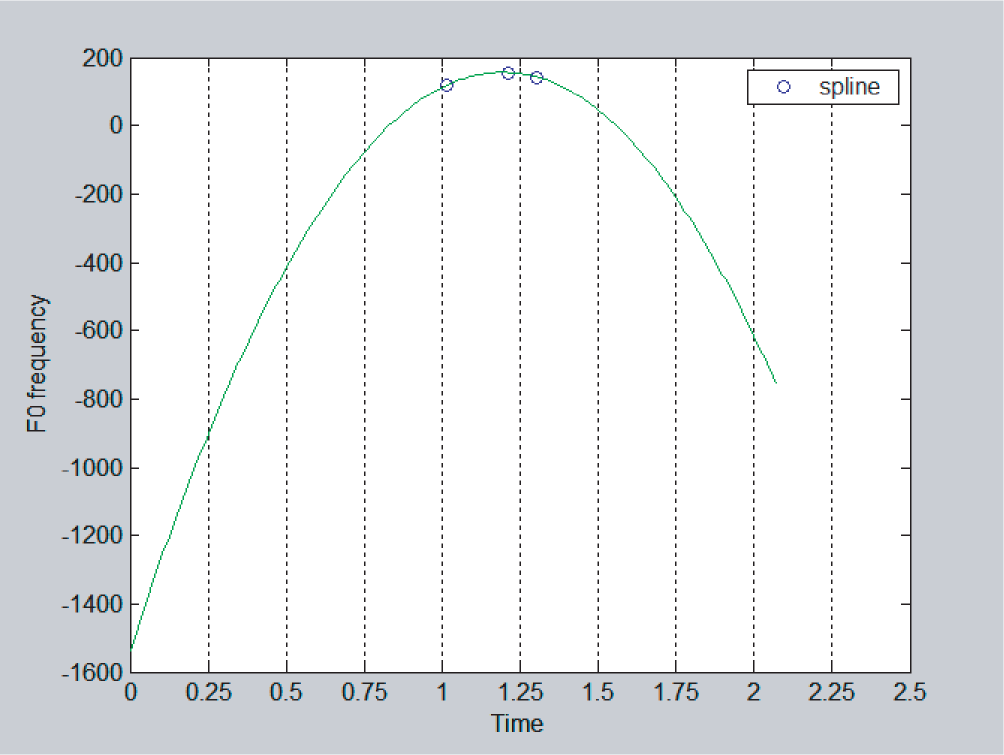


Figure 7.3.: *Momel frequency error; beginning of utterance*

7.3. Perceptual study

The main goal of the evaluation is to determine which F0 stylisation method offers the closest perceptually close representation of the F0 curve, so that it can then be used for deriving precise acoustic parameters of prosodic events. To judge the modified Momel algorithm's suitability for the task, a human perceptual study in form of a rating test was designed. In order to perceptually test the F0 stylisation methods, sentences stylised with the original Momel, the modified Momel and a perceptually based method Prosogram have to be directly compared with the original signal.

The hypothesis was that the modified Momel algorithm would be perceptually close to the original sentence, would offer a better representation of the original signal at the boundary of an utterance and would thus be rated higher than the original Momel. At the same time, the ratings of our augmented algorithm should not differ significantly from Prosogram.

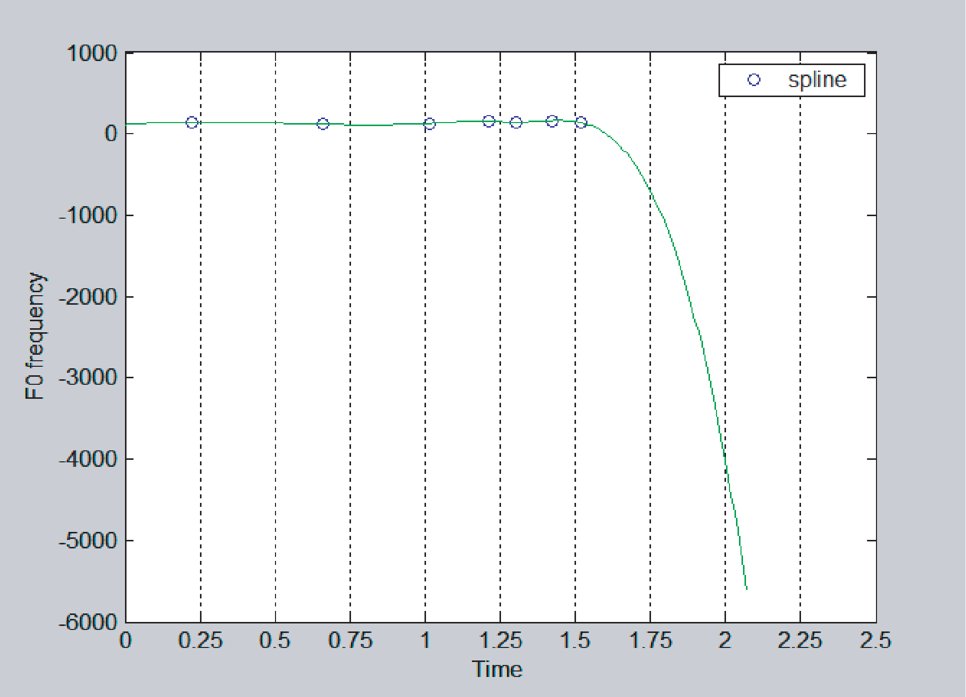


Figure 7.4.: *Momel frequency error; end of utterance*

7.3.1. Experimental design

Method It was decided that for the judgement of stimuli representing three F0 stylisation methods an enhanced Visual Analogue Scale (VAS), called the Visual Sort and Rate (VSR) method should be used (Granqvist, 2003). On a computer screen, subjects are presented with a window in which a number of icons are grouped together, each representing a stimulus which can be heard when the icon is clicked upon. The subjects' task is to sort the stimuli by moving the icons on the screen, adjusting their position according to perceived quality.

Granqvist's results show that the VSR method gives better reliability of subjects' ratings than VAS. The author reports significant improvements of the subjects' performance in a listening test situation compared to VAS. Subjects are reported to find it visually better and by introducing a sorting task it becomes more meaningful, rating is easier and gives more consistent results. The method allows multiple playback of the stimuli and the repositioning of icons corresponding to the stimuli. Thus, the rating can

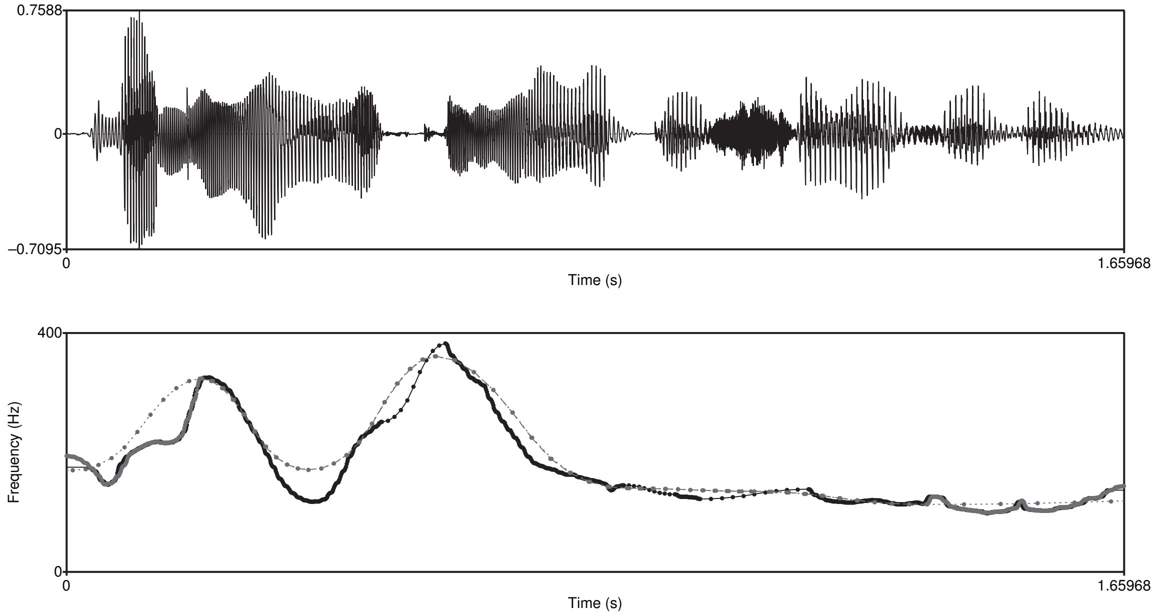


Figure 7.5.: *The waveform and its stylised pitch: original F0 track (black), original MOMEL (grey linked points), and the modified MOMEL (grey linked points and solid grey line at the beginning and end)*

be modified according to the listeners' impression until they are satisfied with the result.

Subjects The subjects for the experiment were Polish native speakers, selected from a group of students at Saarland University.They were mostly ERASMUS/SOCRATES exchange students in their first term abroad. A total of 15 participants, (5 male and 10 female) aged between 21 and 33 took part in the study.

Speech material The stimuli to be judged were taken from the BABEL and PoInt databases. In total, there were 30 sentences (ten in the practice stage and twenty for the main part of the experiment). The sentences varied in duration between 0.96 and 3.78 sec. We resynthesised three versions of the same sentence using PSOLA, based on the stylisation performed by i) the original MOMEL algorithm, ii) Prosogram and iii) the modified Momel algorithm. The stylisation was performed using Praat tools pro-

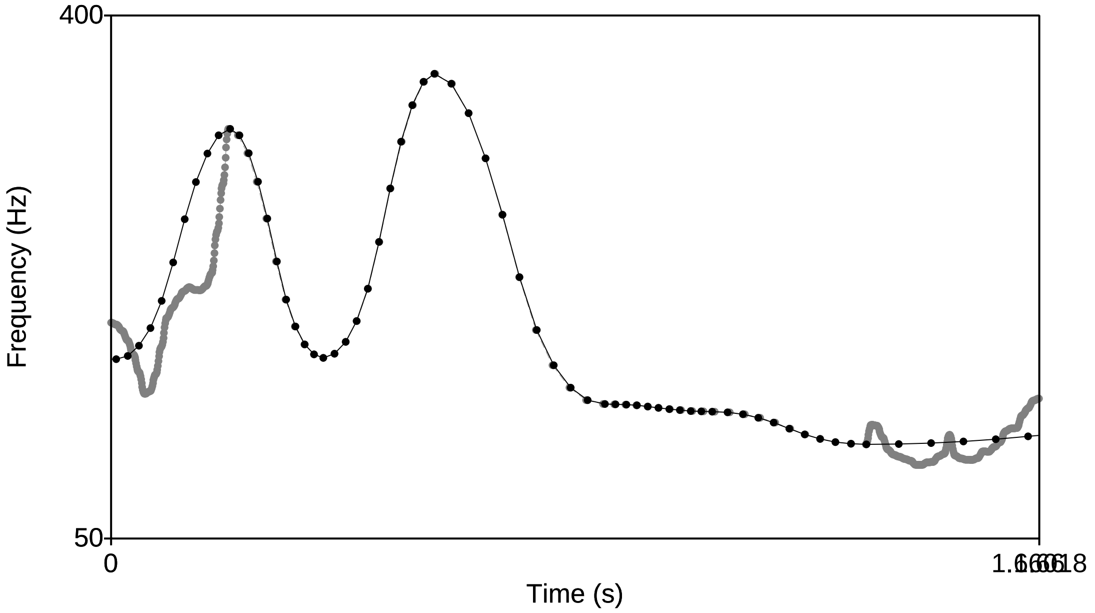
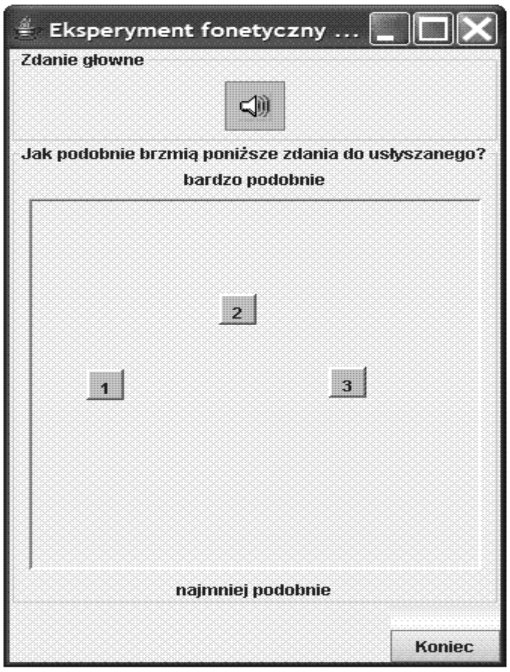


Figure 7.6.: *Stylised pitch: original MOMEL (linked points), and the modified MOMEL (points plus solid grey lines at the beginning and end)*

vided by the developers of the methods. In order to be able to compare Prosogram to both Momel methods, the target points were interpolated quadratically.

Procedure In this study, we used the Rating Test software by Schröder (2004), developed for rating of sentences in the context of emotional speech synthesis task. The graphical interface of the Rating Test, cf. Fig. 7.7, makes it possible to implement the VSR method.

The subjects could see a screen with icons representing three stylised versions of a stimulus placed in a window. At the top of the window there is an icon with a resynthesised version of the original stimulus with which the comparison is to be made. The experimental question: "How similar do the sentences shown below sound in relation to the one just heard?" is placed under the original sentence. The original sentence is referred to as 'the one just heard' as it is always played automatically at the beginning of each experimental sentence comparison. In addition, a text is placed at



Jak podobnie brzmią poniższe zdania do usłyszanego?

Figure 7.7.: *Perceptual Study Interface*

the top and bottom of the screen with the words 'very similar', and 'not at all similar', accordingly. By clicking on the icons, the stimulus files can be heard. Afterwards, the icons can be moved up and down or from side to side in the window, either closer or further from the main stimulus which they are being compared to.

Summary The subjects were always asked to rate the similarity of three target sentences to one main stimulus. The main stimulus was an original sentence from the database which had been resynthesised using the PSOLA method. The three target stimuli were the PSOLA-resynthesised versions of the main stimulus and differed from it only in the way their F0 contour was stylised. The participants were allowed to listen more than once to the stimuli and they could keep repositioning the icons corresponding to these stimuli until they were satisfied with their ratings. The stimuli were arranged randomly within each task.

7.3.2. Results

Figure 7.8 shows inter-subject rating scores for sentences synthesised using Prosogram, the original Momel method and the modified Momel stylisation method. Modified Momel target sentences were rated as significantly closer to the main sentence than those stylised using the original Momel algorithm ($p < 0.001$). The Prosogram was rated highest overall but did not differ significantly ($p = 0.204$) from the modified Momel ratings (receiving mean values 70 and 67 out of 100, respectively).

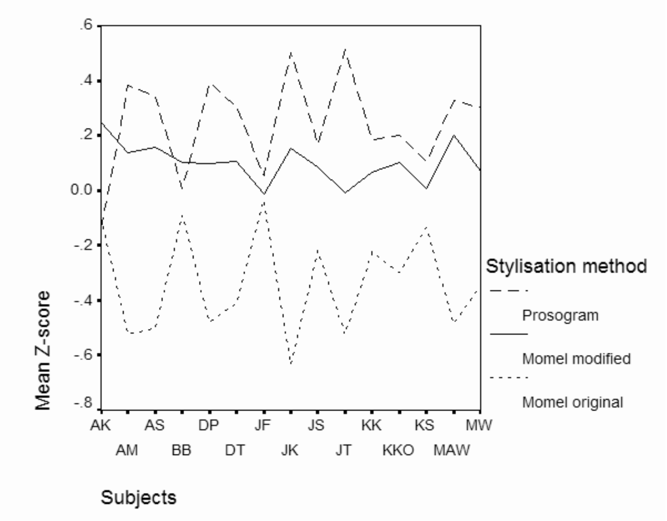


Figure 7.8.: *Inter-subject rating scores*

Both the results of the practice part and test proper are consistent and were not significantly different, thus showing no learning effect. Using a different test paradigm, d'Alessandro & Mertens (1995) claimed very close perceptual similarity between Prosogram versions and original utterances. The statistical equivalence found here between the modified Momel versions and the Prosogram versions may therefore be taken as an indication of the perceptual quality of the modified Momel approach.

There is a significant difference in the rating of tokens coming from the two databases, cf. Fig. 7.9.

Babel sentences, which were segmented into phonemes, received better ratings than PoInt which was only syllabically segmented ($p < 0.001$). The

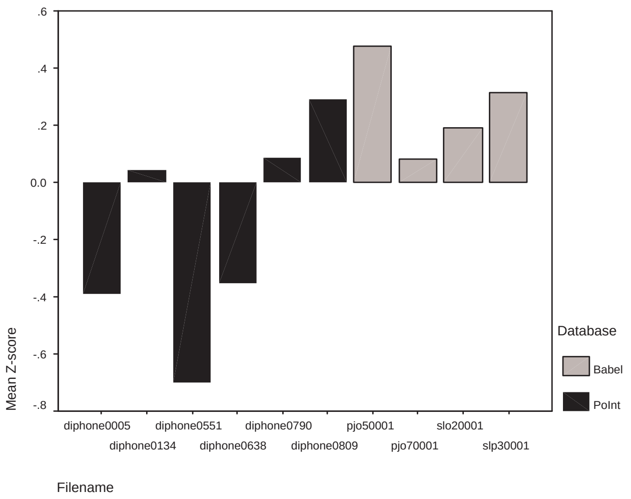


Figure 7.9.: *Z-score rating for tokens from the two databases*

difference in rating is visible in the scores for Prosogram stylised stimuli, cf. Fig. 7.10.

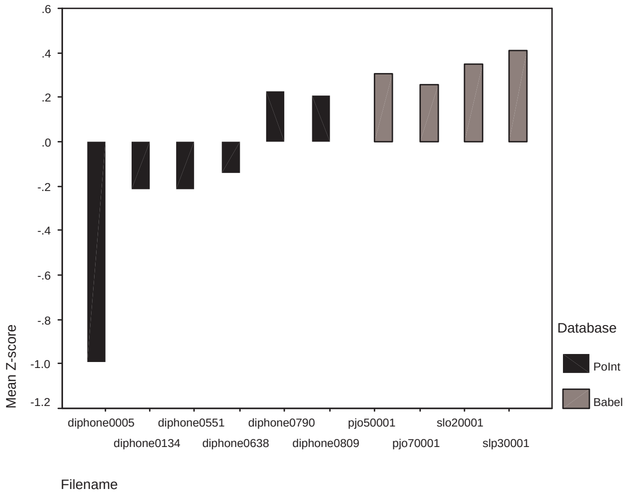


Figure 7.10.: *Z-score rating for tokens from the two databases using Prosogram stylisation method*

This is easily explained by the nature of Prosogram, which relies on syllabic nucleus identification. Polish syllables can contain complex consonantal material in the syllable onset, which is often voiceless, thus impeding the algorithm. The same material, namely four stimuli which were rated systematically worse than both the Momel versions when presented as input with a phonemic segmentation to the same program, yield significantly better resemblance to the original sentence than these when based on syllabic segmentation ($p < 0.001$). Prosogram finds, on average, six times more stylised target points in this condition.

7.4. Conclusions

We conclude that, when faced with speech material without segmental annotation, the stylisation using modified Momel can be considered a satisfactory method for representing F0 curve, yielding results approaching the ratings of the perceptually grounded method Prosogram. Moreover, the sentences stylised using both of these techniques are viewed as being perceptually close to the original speech signal.

To fulfill the goal of the study, acoustic parameters of tonal accent classes will be generated from an automatically stylised F0 curve, obtained using a Momel algorithm augmented for boundary locations. The revised algorithm for these locations has been tested in a perceptual study which shows that subjects judge the sentences stylised with the revised Momel to be better than the original algorithm. The parameters obtained from the stylised curve will be used as input to tonal accent classification described in Chapter 8.

8. Analysis of accent types in Polish: towards new division of accents

Introduction

This chapter presents a method for automatically deriving pitch accent classes for prosodic modelling in Polish by means of unsupervised clustering methods. The motivation behind choosing a stochastic method was a need for precise parameterisation for each pitch accent category and to reduce the bias and cost of database manual prosodic annotation. Moreover, the data exploratory techniques used here do not presuppose a number or type of accents desired and remains theory independent. Consistent specification of pitch accents through precise parameters of accent classes are generated from an automatically stylised F0 curve, obtained using a Momel algorithm augmented for boundary locations as described in Chapter 7. A revised algorithm for these locations has been tested in a perceptual study which shows that subjects judge the sentences stylised with the revised Momel to be perceptually close to the original signal. The parameters obtained from the stylised curve and pitch accent location information encoded in the speech database are used as input to Kohonen self-organising maps and a hierarchical clustering technique. The results identify three classes of pitch contours present. Manual inspection of the resulting clusters ensures that they are also linguistically motivated and not spuriously produced by the data mining techniques. The results give insight into the typology of accent types present in the language as represented by the speech database used for their classification. Utterances automatically labelled using this classification will be used as input to a prosody prediction and generation module of a speech synthesis system as described in Chapter 9.

This chapter is organised as follows: in Section 8.1 we describe machine learning-techniques used for the classification of accent types; we detail pitch accent clustering analysis in Section 8.2, where we also introduce the parameterisation and clustering techniques applied. Section 8.2.2 describes the results of the clustering and the characteristics of the derived pitch accent types, and Section 8.3 presents our conclusions.

8.1. Classification of accent types using machine learning techniques

In order to obtain a high quality speech synthesis, a good prosodic modelling needs to be developed. One of the steps in modelling prosody, apart from predicting duration and phrasing, is to predict and generate appropriate pitch accent contours. To train an accent type prediction system, it is necessary to have a good typology of accent types present in the language. Moreover, we need precise parameters corresponding to each category, which can be derived directly from the F0 curve if a prosodically pre-labelled database is available. However, the non-trivial task of manually labelling a speech database introduces human annotator bias, especially when the annotation scheme used is based on phonological categories (e.g. ToBI) (Braun, 2005). The resulting annotation is then partly phonological and phonetic, which makes the task difficult for a machine learner.

Automatic accent-classification studies has been carried out using the following techniques: i) hierarchical clustering (Klabbers & van Santen, 2004), ii) linear regression (Keller & Keller, 2003), iii) EM bagging and boosting (Sun, 2002b), iv) hidden Markov models (Kumpf & King, 1996;

Blackburn et al., 1993; Chan et al., 1994; Zervas et al., 2004; Batliner et al., 1999), v) self-organising maps (Werner, 2001) .

In this study, we aim to derive prototypical pitch contour types found in Polish, based on their acoustic characteristics (Oliver, 2005a). For our exploratory approach, a two stage procedure is used. First, we apply Self-Organising Maps (SOMs) (Kohonen, 1995), a vector quantisation method which performs clustering of instances, in the form of acoustic feature vectors, cf. Section 8.1.1. This method has been applied in prosodic event classification e.g. (Werner, 2001), and does not presuppose a number of clusters. This last feature is relevant in our case because we do not want to limit the pitch accent inventory to the set of labels used in the database we are using nor choose a particular inventory suggested in the literature cf. Chapter 2. Second, a hierarchical agglomerative clustering is applied to derive a final number of clusters from fuzzy sets obtained by SOMs.

The resulting classification would enable an automatic re-annotation of the database with intonation events grouped according to similarity of acoustic parameters. This way, any method responsible for prediction and generation of different pitch accents will be equipped with a consistent specification of pitch accents.

8.1.1. Self-organising maps

Self-Organising Map (SOM) is an example of an unsupervised artificial neural network approach. This approach has been around since the early 1980's and has widely been applied in engineering and many other fields. Through a process called self-organization, SOM configures the output nodes into a topological representation of the original data and organises similarity clusters that can be seen as soft edged classes or fuzzy sets emerging from statistical correlations. There is also a reduction of high-dimensional data so it can be projected into a 2D space. The SOM can thus serve as a clustering tool as well as a tool for visualizing high-dimensional data.

As mentioned above, self-organising maps are one application of neural networks. SOMs have been widely used for material classification and categorisation purposes especially within the cases of unknown classes. The basic Self-Organising Map can be visualised as a sheet-like neural-network array consisting of cells (or nodes). During the learning, these cells become specifically tuned to various input signal patterns or classes of patterns in an orderly fashion. The learning process is competitive and unsupervised, meaning that no teacher is needed to define the correct output for an input. In the basic version, only one map node (winner) at a time is activated corresponding to each input. The locations of the responses in the array tend to become ordered in the learning process as if some meaningful non-linear co-ordinate system for the different input features were being created over the network (Honkela, 1997). Some properties that distinguish the SOM from the other neural networks are that it is numerical instead of symbolic, non-parametric, and capable of learning without supervision. The numerical nature of the method enables it to treat numerical statistical data naturally, and to represent graded relationships. Because the method does not require supervision and is non-parametric, i.e. no assumptions about the distribution of the data need to be made beforehand, it may even find quite unexpected structures from the selected data (Kaski, 1997). The resulting clusters can be seen as soft edged classes or fuzzy sets emerging from statistical correlations. Each data sample is part of every cluster with membership proportional to the value of the neighbourhood function. The first benefit is that each class prototypes created are local averages of the data, therefore less sensitive to random vectors. This way SOM is an intermediate step that makes this clustering method a two-level approach: the data is first clustered using the SOM and then the SOM is clustered. To cluster the SOM, either partitive or hierarchical methods can be applied.

8.2. Pitch accent clustering analysis

Based on the results of the perceptual study described in Chapter 7, the F0 contours stylised using the modified Momel method are used for the extraction of parameters in accent categorisation. The goal of the study is to find prototypical phonetically distinct pitch accent classes, without the need for a priori knowledge about their phonological categories.

To derive the final number of clusters from the resulting topological representation of the original data (SOMs), hierarchical agglomerative clustering is applied. The idea is to minimise the distance within and maximise the distance between clusters. The HAC clustering method is a bottom-up greedy algorithm that starts with a separate cluster for each object. In each step, the two most similar clusters are determined and merged into a new cluster. The algorithm terminates when one large cluster containing all objects has been formed, which then is the only remaining cluster from the set of initial clusters (Manning & Schütze, 1999). The algorithm used here is a group-average clustering method where the criterion for merging clusters is average similarity. Hierarchical clustering results are usually represented by means of a dendrogram. It is a binary tree in which each data point corresponds to terminal nodes, and distance from the root to a subtree indicates the similarity of subtrees highly similar nodes or subtrees having joining points that are farther from the root.

By cutting the resulting dendrogram in the place where there is a large distance between two clusters, and merging the clusters based on their proximity, we determine the final number of clusters. The main motivation behind the proposed algorithms is not specifying but discovering how many pitch accent classes could describe the intonational variation present in the database used for training (i.e. read and semi-spontaneous speech) cf. Chapter 5, Section 5.2.

The tool used for the two-stage method was TANAGRA (Rakotomalala, 2005). It implements a variation of HAC, called hybrid clustering.

Based on linguistic knowledge about the acoustic data, only a small number of clusters is to be expected. Tanagra therefore uses a fast method for the step one of the clustering, namely, the SOM. HAC starts from the SOM created clusters and builds the dendrogram. A very useful feature of the program is that what is presented to the user is not the tree itself, as it would probably be too complicated to decipher, but the gap between the nodes, which is presented in a table. The best partitioning, corresponding to the biggest gap is highlighted. What's worth mentioning is that most of the time, partitioning into two clusters always shows the best gap but it is not interesting in this study. It is the user who in the end can decide on the optimal number of clusters, based on closer inspection of the clusters.

8.2.1. Parameterisation

Based on the prosodic annotation marking the positions of pitch accents within the utterance, the stylised contours were parameterised. For the description we used the alignment of F0 peak with respect to syllable onset, the form of movement comprising movement slope and the pitch amplitude as calculated by Tilt parameter (Taylor, 2000) cf. Chapter 5 Section 5.2.3. Since the corpus used consisted of both male and female speakers, we normalised the F0 values by first converting them to semitone values and z-scores calculated per sentence, and scaling to the mean of the database, for details see Section 9.2.

An ANOVA analysis determined that these parameters do not vary significantly between male and female speakers in the database ($p>0.5$). Here is a list of parameters that were input into the clustering algorithm:

- F0 peak delay from pitch accented syllable onset,
- F0 pitch event slope,
- Pitch event duration (three syllable window),

- Difference between F0 max and F0 min in pre-tonic, tonic and post-tonic syllables,
- Tilt measure.

8.2.2. Results

The clustering performed using SOMs and hierarchical clustering, based on the above parameters, results in three clusters. Figure 8.1 shows the prototypical contours found in each cluster associated with the accented syllable.

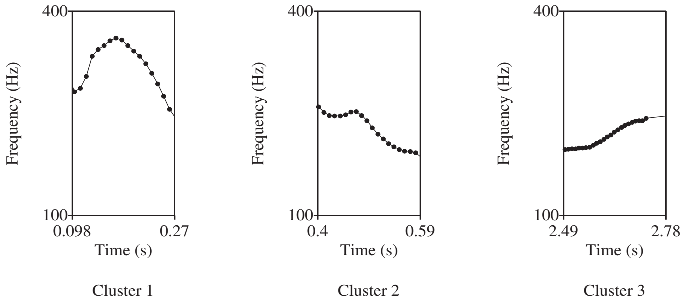


Figure 8.1.: *Prototypical contours in clusters 1-3*

An ANOVA analysis revealed that the acoustic features: the absolute F0 peak delay from pitch accented syllable onset, the slope (start and end) as well as pitch difference were significant between the three cluster groups (p<0.001), cf. Table 8.1.

The sex of the speakers was not found to be a significant factor. Table 8.2 shows the similar number of male and female speakers belong to all cluster groups.

Cluster 1 is the smallest cluster, containing 200 tokens (16.1% of the total number of pitch accents). The contours in this group exhibit the steepest rising and falling movement, with F0 peak positioned in the middle of the syllable. These contours reach the highest F0 maximum values and the lowest F0 minimum values of all clusters cf. Table 8.1 and thus biggest pitch differences within the accented syllable, cf. Figure 8.3. Fig. 8.6 shows

	Cluster 1	Cluster 2	Cluster 3
F0 Absolute Peak Delay (ms)	77.33	13.56	204.62
Slope	61.28	16.84	17.21
Slope (start)	22.4	0	9.89
Slope (end)	40.8	13.83	3.4
F0 difference (st)	9.39	4.18	1.85
F0 max (st)	13.22	9.37	10.49
F0 min (st)	5.35	6.47	7.69

Table 8.1.: *Clusters characteristics*

Sex	Cluster 1	Cluster 2	Cluster 3
Male	97	325	209
Female	103	317	189

Table 8.2.: *Male and female clusters membership*

an example of an utterance with an accent belonging to Cluster 1. The biggest number of original pitch accent labels from PoInt database falling into this class are HL (35.5%) and but it also has almost equal number of LL (12.5%), HM (12%), ML (11%), and LH (9.5%) and MH (9%), cf. Table 8.3.

20% of members in this group are phrase non-final pitch accents and mostly associated with emphatic focus or a wh-word in a question.

Cluster 2 is the largest cluster, containing 51.8% of all the pitch accents (642 tokens). This group is characterised by a very early F0 peak, cf. Figure 8.2 mostly starting at the syllable start or early in the onset forming a plateau, see Slope Start results for this cluster in Fig 8.4, which is followed by a gradual fall ending at the speaker's mid or bottom range. 90% of the pitch accents in this cluster are phrase final and are characteristic of a final fall. Fig. 8.7 shows an example of an utterance containing an accent belonging to this Cluster of contours. As seen in Table 8.3 the labels ML (25%), LH (14.6%), Lx (12%), and LL (11%) form a majority of contours in his class.

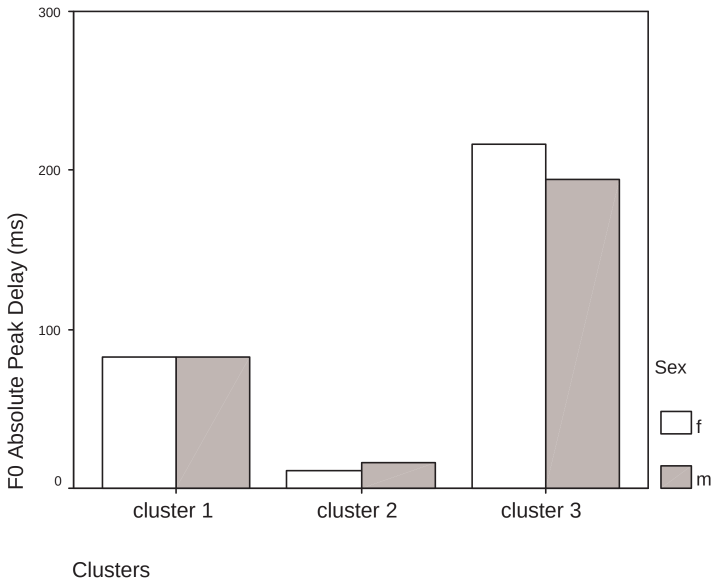


Figure 8.2.: *Absolute F0 Peak Delay in Clusters 1-3*

Cluster 3 contains 398 tokens (32.1% of the total pitch accents). Original pitch accent labels MH (22.6%), MM (16%), HM(16%) and LH (15.8%) are most common in this class, cf. Table 8.3. The pitch accents in this group exhibit a very late F0 peak, cf. Figure 8.2, either in the syllable offset or on the following syllable. This type of contour is found mostly in phrase final position and is typically associated with a continuation rise or a yes/no question. Fig. 8.8 shows an example of an utterance containing an accent belonging to Cluster 3.

Based on re-classification of pitch accent labels, the original PoInt annotation has been changed to reflect new pitch accent class membership. Table 8.3 shows the change in membership from original contour labels to new clusters. Based on the linguistic analysis of clusters, the new labels will reflect their movement, which will also make the comparison with the original labelling convention feasible. Cluster 1 will be referred to from now on as ”RF” (rise-fall), cluster 2 will be labelled ”F” (fall) and cluster 3 will be relabelled as ”R” (rise).

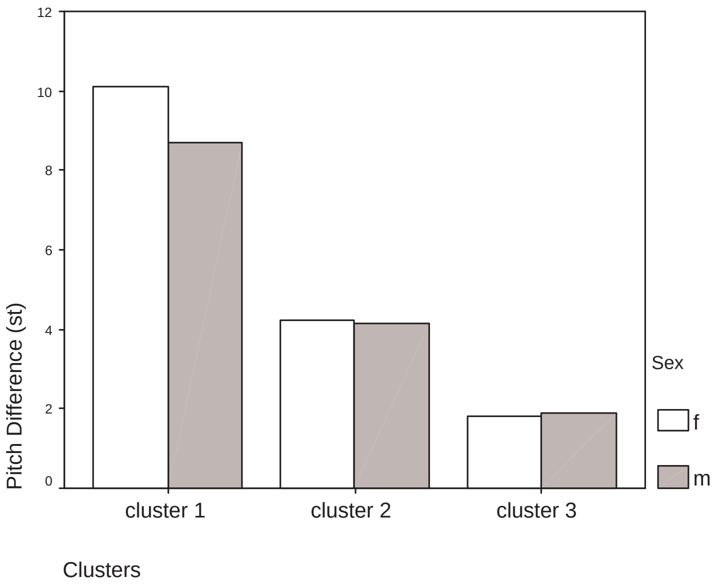


Figure 8.3.: *F0 Pitch Difference in Clusters 1-3*

8.3. Conclusions

As the speech material used in this study does not cover all aspects of speech variations (e.g. expressive speech), it probably does not represent a complete inventory of the tonal accents of Polish. Nevertheless, it is considered sufficient to propose collapsing the number of accent types to be used for intonation generation to three types cf. Table 8.3. The analysis shows that automatic classification can be successfully applied to the task of deriving pitch classes if an acceptable stylisation of F0 is carried out. From an automatically stylised F0 curve, precise parameters of accent classes can be generated. Furthermore, the viability of using unsupervised clustering techniques which do not require a pre-determined number of categories is confirmed.

Self-organising maps are thus considered a reasonable method for accent categorisation. Their results are linguistically meaningful and consistent. Features like F0 peak delay and slope seem to provide sufficient prosodic information for such a classification. The derived pitch accent classes, parameterised in this way are used for the pitch accent modelling

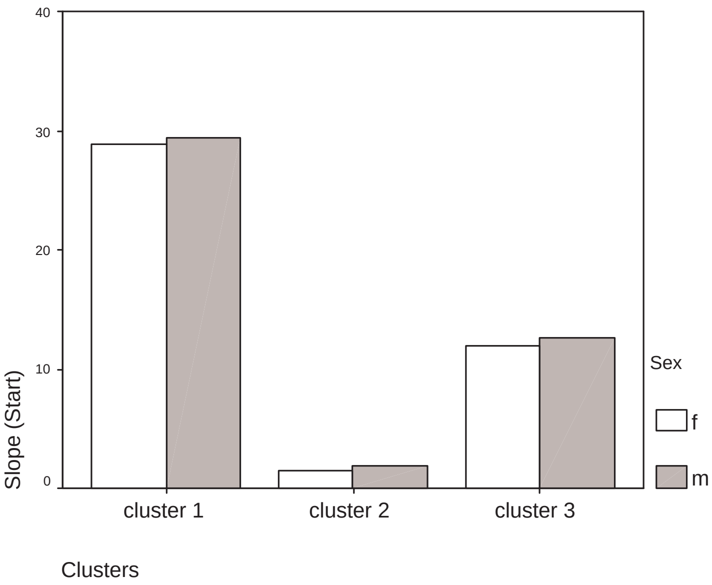


Figure 8.4.: *Start part of Slope in Clusters 1-3*

presented in Chapter 9, where they are used as input to the prosody prediction and generation module of a speech synthesis system. As these results are based on statistical analysis, they had to be further tested in a human perceptual study, details of which are included as part of the evaluation process described in Chapter 10.

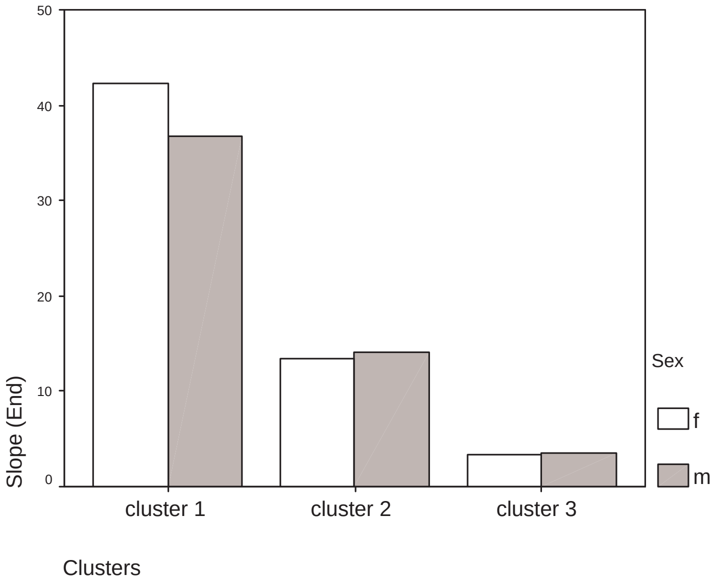


Figure 8.5.: *End part of Slope in Clusters 1-3*

Pitch Accents	Cluster 1	Cluster 2	Cluster 3
HH	6	5	36
HL	71	46	22
HM	24	37	64
Hx	1	0	7
LH	19	94	63
LL	25	70	10
LM	3	51	17
Lx	7	77	3
MH	18	41	90
ML	22	160	21
MM	4	61	65
Total	200	642	398

Table 8.3.: *Original Pitch Accent labels membership in clusters 1-3*

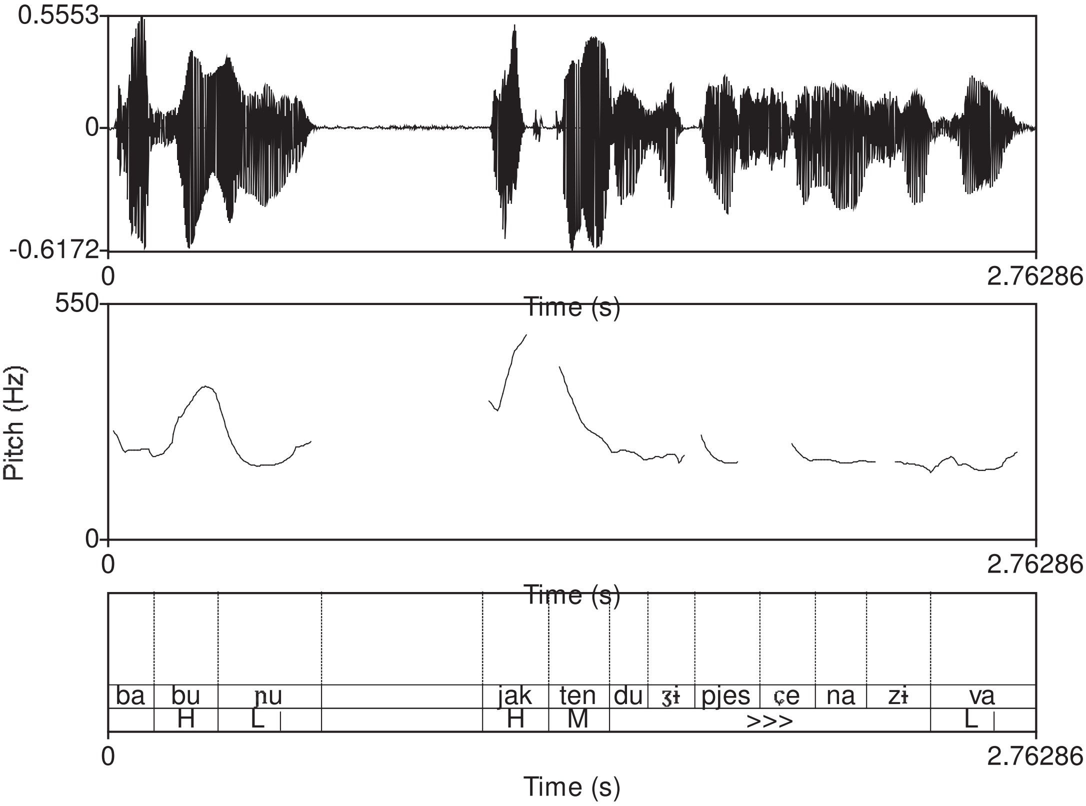


Figure 8.6.: *PoInt utterance with a pitch accent (original F0 and label) classified as Cluster 1*

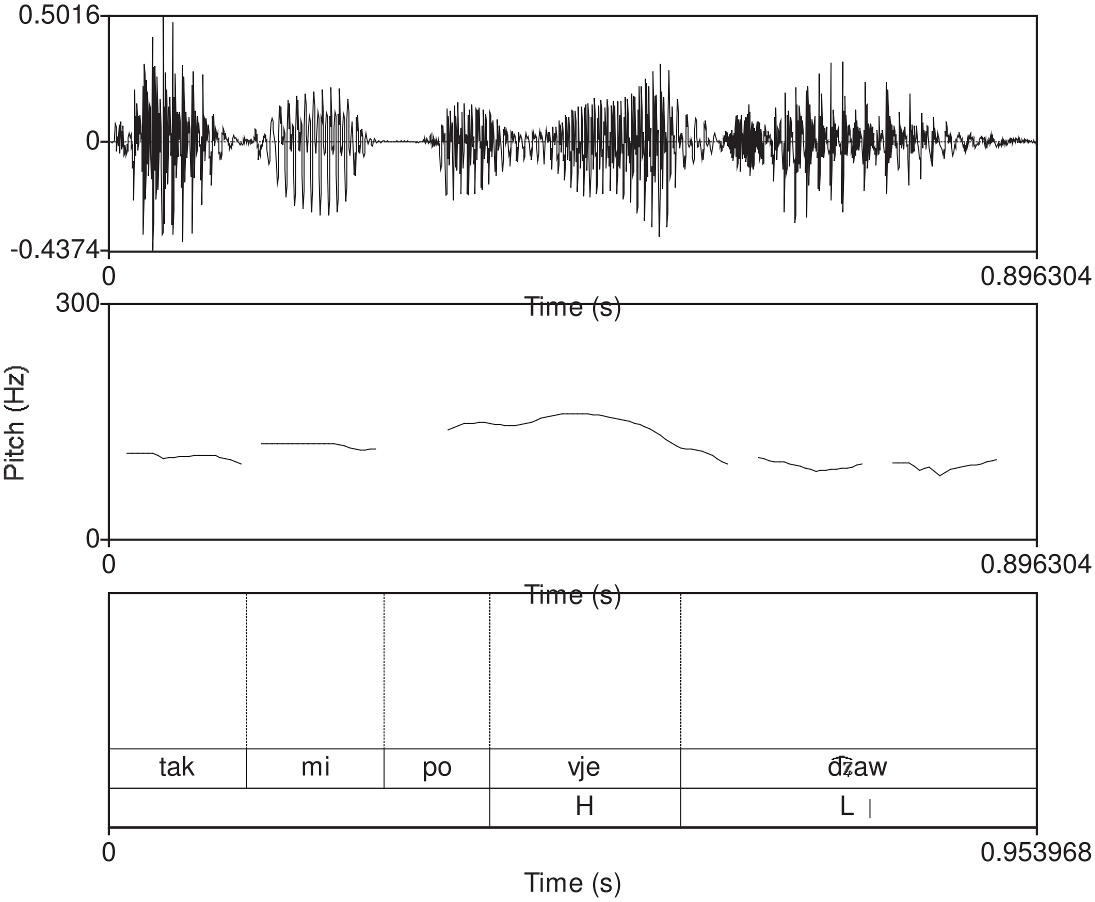


Figure 8.7.: *PoInt utterance with a pitch accent (original F0 and label) classified as Cluster 2*

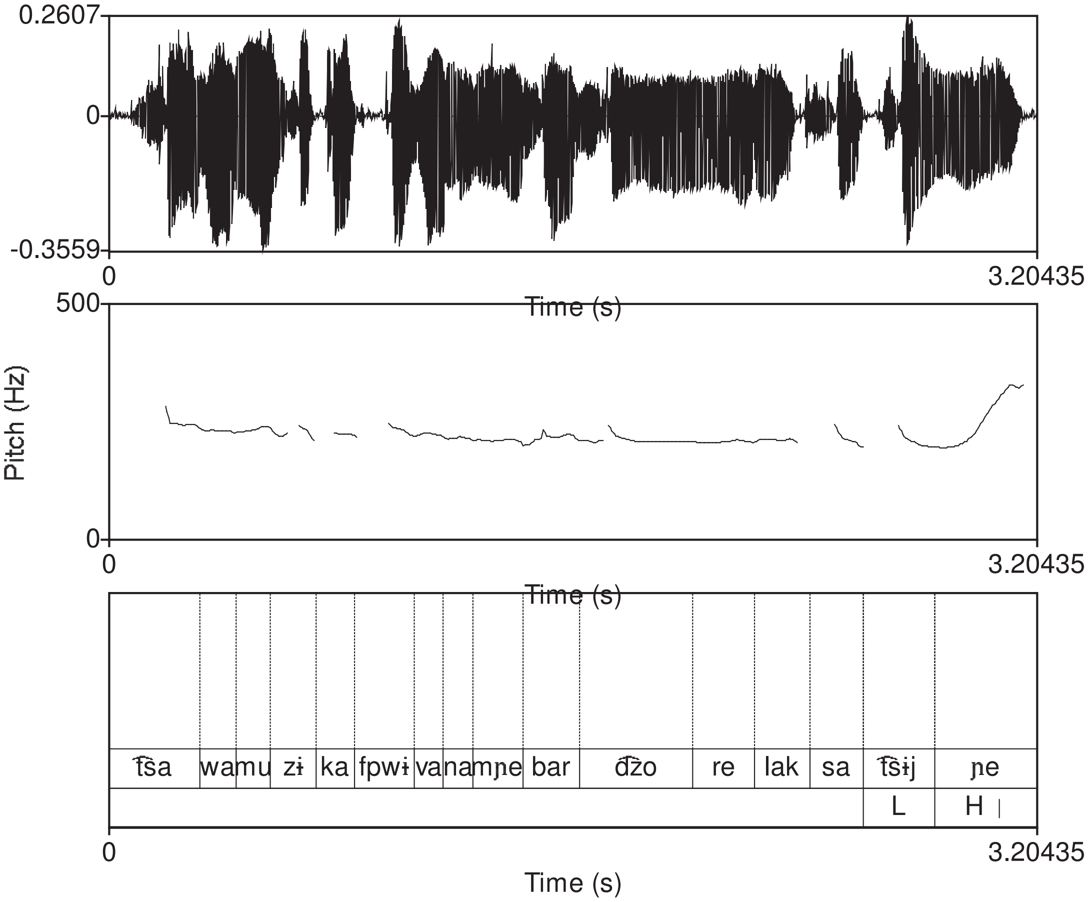


Figure 8.8.: *PoInt utterance with a pitch accent (original F0 and label) classified as Cluster 3*

9. Statistical modelling of Polish intonation

Introduction

In order to get high quality speech synthesis, a good prosodic model needs to be developed. Modelling prosody for diphone-based concatenative speech synthesis involves generating pitch contours based on a linguistic description of a language. The training of a model requires a viable linguistic description of the intonation of the language and a suitably annotated speech database.

To successfully develop a prosodic model, we need three components: a language specific description of intonation, an annotated speech corpus on which the model can be trained and a system in which the model can be implemented and tested. This chapter describes methodology and results of the modelling of previously established accent types described in Chapter 8. The process is composed of three parts: i) Accent place and ii) type prediction and iii) the prediction of F0 values for curve generation.

We describe a Polish prosody modelling module for the Festival speech synthesis system. The module uses classification and regression trees for accent place and type prediction, and a linear regression technique for F0 contour generation for these accents. The techniques used to address problems with the limited amount data are shown. We demonstrate how improvements were achieved by the using a modified F0 stylisation, and language specific features. Results of a formal perception study show a significant preference for the new intonation model over the original one cf. Section 9.2.3.

We then discuss two models we have built for use with the Polish voice within Festival (Oliver, 1998), a CART (Breiman et al., 1984) model for accent prediction and a Linear Regression (Black & Hunt, 1996) model for contour generation. We pay particular attention to the language specific details of these models and the techniques used to overcome problems with the resources that are available in Polish.

9.1. Accent place and type prediction

9.1.1. Training of models

First, an accent prediction model will be built by using original data. Then based on new accent types the improvement will be shown.

As described in Section 6.2 Classification and regression trees (CARTs) are often used in speech synthesis for duration modelling, but they can also be applied to accent prediction (Dusterhoff et al., 1999). Trees are constructed by a data driven training process. A tree is derived by a set of yes/no or if/then questions relating to the data in order to predict the dependent variable. As a classifier, the tree partitions the training data into classes so that the class with the highest probability is chosen to classify an unknown case. As a regression tree, the tree is designed to minimise the error in the predicted variable.

The main advantage of this method is that it can deal with non-linear data in a reasonable way. However, the training procedure can get stuck at local maxima. This can be overcome to an extent by providing the model with features that are thought to most likely to influence the placement of pitch events and avoiding correlating fatures.

In the original implementation for Polish accent prediction there was no support for different accent types. In contrast, for languages like English, models built from databases with labels based on ToBI (Pierrehumbert, 1980) or similar descriptions have been implemented. In our case, this paradigm has been modified to use the PoInt annotation after relabelling described in Chapter 8 based on a set comprising of R-F(rise-fall), R(rising), and F(falling) accents.

Feature set The following features are suggested for place and type of accent prediction:

- position in sentence,
- syllable structure of the accented syllable,
- number of syllables in the accent unit.

For accent type and place prediction we initially used the following features in the CART:

- position of syllable in phrase and word,
- strength of break,
- whether the syllable is lexically stressed or not,
- number of (stressed, unstressed, accented) syllables since last and until next phrase break,
- number of syllables since last and until next accented syllable,
- number of minor phrase breaks since last major phrase break.

We then introduced language specific features to improve accent type prediction, namely, additional information regarding the position of a syllable within a word. This is motivated by the fact that in Polish the last content word usually receives the phrase accent and stress generally occurs on the penultimate syllable of the word. To accommodate this the following features were added:

- the word is final in a phrase,
- the syllable is penultimate in a word,
- a simplified part of speech tag (content/function word).

9.1.2. Results

In Appendix E, Tables E.1 and E show the confusion matrix for accent type assignment depending on the method and data set used (training and test). The preliminary results were based on modelling using the original PoInt labels.

Table 9.1 shows the confusion matrix for accent type assignment in the final model including category NONE corresponding to unaccented tokens. A 7% improvement was noted as a result of including the language specific features.

Accents	R-F	F	R	NONE	Accuracy
R-F	673	59	91	2	81.6%
F	92	725	146	2	75.1%
R	22	129	774	0	83.7%
NONE	16	24	14	797	93.7%
	Correct 2969/3566			Total Accuracy 83.3%	

Table 9.1.: *Accent prediction confusion matrix*

The features that contributed most to the model were: i) the number of syllables since (0.9045) and ii) till a phrase break (0.9034), iii) position in word (0.9065), iv) number of accented syllables before next phrase break (0.9086), v) strength of the break after next syllable (0.9107) and vi) stress on the syllable (0.9117).

9.2. F0 contour prediction

Contour generation in Festival is traditionally carried out using three Linear regression (LR) models (Black & Hunt, 1996). Linear regression methods are also found elsewhere in literature, cf. Section 6.2.

Linear regression models assume that a predicted variable (p) can be modelled as the sum of a set of weighted real-valued factors 9.1.

$$p = w_0 + w_1 f_1 + w_2 f_2 + w_3 f_3 + ... + w_n f_n \tag{9.1}$$

The factors (f_i) represent parameterised properties of the data, and the weights (w_i) are trained, usually using a stepwise least squares technique. Each of the three models predicts the F0 at a different point in a syllable (start, middle and end, respectively). The factors incorporate information like the type of accent present, the position of phrase breaks, syllable stress and syllable position within the text. Each model considers this information for a five syllable window centred on the current syllable. This allows the pitch on syllables around an accented syllable to be affected by the presence of the accent, so that pitch movement is not restricted to occurring on the syllable that is marked with the pitch event. For example the peak of an L+H* could occur in the syllable following the one the accent is assigned to.

9.2.1. Training of models

Based on the accents predicted in Section 9.1 we built linear regression models to predict F0 at the start, mid point and end of syllables. These replaced the original rule based model which imposed the hat-shaped accents onto a declining baseline.

The advantage of using Linear Regression is that relatively little knowledge about the intonation system in question needs to be known (although arguably this is needed to complete the accent prediction stage).

However, we show that the inclusion of language specific features can lead to improvements in the resulting contour.

The training of these models generally requires at least an hour of good quality speech from a single speaker. Our database did not contain enough data per single speaker from which to build a model, so we decided to investigate the use of multi-speaker data. To be able to use PoInt multi-speaker data, an F0 normalisation procedure was necessary. We adapted the procedure used by Clark (2003) to normalise across Tone Groups (here referred to as intonation phrases) to normalise across speakers.

Fundamental frequency normalisation The normalisation was carried out using:

$$F0_n = \frac{F0 - \mu_i}{\sigma_i} \tag{9.2}$$

where μ_i is F0 mean and σ_i is the F0 standard deviation of the utterance in question. Next, rescaling is performed. The rescalling uses mean and standard deviation of the database:

$$F0 = F0_n \sigma_D + \mu_D \tag{9.3}$$

where μ_D is F0 mean of the database and σ_D the F0 standard deviation of the database.

Feature set The Linear Regression (LR) Model 1 uses the Festival English default set of features except that the features relating to pitch accents are changed to accommodate the relabelled PoInt prosodic annotation scheme (F, R, R-F). Fig. 9.1 shows an example waveform and F0 contour synthesised using the LR Model 1.

Preliminary results are presented in Appendix F in Table F.1. In all methods F0 was stylised, normalised, tokens were randomised, stop value set to 10. The number of features was constant and set to 41, and the differences were also in treatment of low F0 values. In method 'min' all

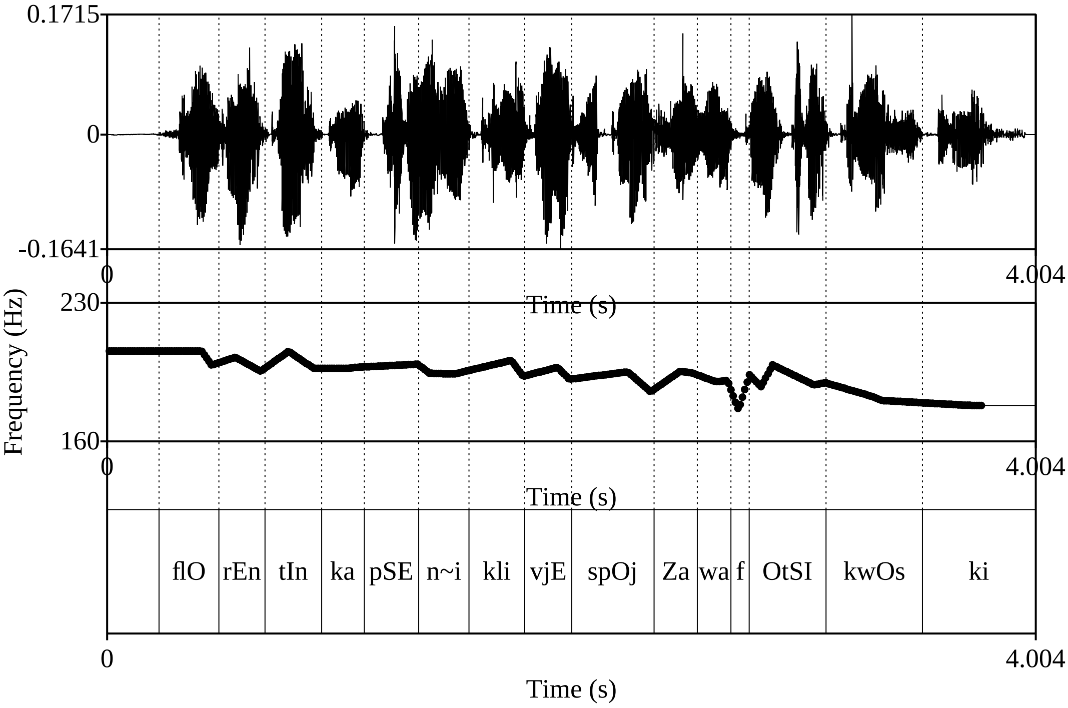


Figure 9.1.: *F0 contour synthesised using the Linear Regression Model 1 for a sentence: 'Florentynka przenikliwie spojrzała w oczy Kłoski'*

values below 40 Hz were set to 0. In method 'max' all values below 50 were set to 50.

9.2.2. Results

Table 9.2 shows the RMSE and correlation values for the LR model 1 for the training and test data, computed for the whole signal, first reported in Oliver & Clark (2005).

	LR train		LR test	
Position	RMSE	Correlation	RMSE	Correlation
Start	42.51	0.72	41.85	0.75
Mid	43.5	0.66	43.63	0.68
End	40.3	0.67	40.19	0.7

Table 9.2.: *RMSE (Hz) and correlation for LR Model 1*

The results are based on the normalised F0 data from the female and male speakers, which consisted of 865 utterances. The fact that RMSE and correlation figures are not as good as those obtained from a single speaker database reported in literature e.g. in Italian (Vegnaduzzo, 2003), is due to the great variability between the speakers and the fact that the normalisation method (see Section 9.2.1) is not that sophisticated.

9.2.3. Pilot perception study

To obtain an auditory evaluation of the resulting model for F0 contour generation and to measure its relation both to the original Festival TTS model a perception study was carried out. To perform the experiment, 15 sentences were selected from the database used to train the LR model. The stimuli were created using Festival TTS system (creation of F0 track) and Praat (pitch replacement). In order to imitate the intonation contour of the original database sentences, APML (Carolis et al., 2004), an XML-based mark-up language, was used to generate the same accent placement and type as marked in the automatically relabelled annotation. Both sets of stimuli varied only in the intonation model applied (original vs. new) keeping other acoustic characteristics intact. Ten native Polish speakers who took part in the study were presented with pairs of utterances and asked to choose the realisation of the sentence they preferred, paying attention to its melody. The experiment was conducted via the Internet using web forms and the participants had the possibility to listen to stimuli more than once.

The analysis of the results showed that the new intonation model was significantly preferred with a mean 78% preference over the original model ($p < 0.001$). Following the pilot, an extended, formal study was carried out and its results are presented in detail in Chapter 10.

9.3. Conclusions

We have outlined the characteristics of an original rule based model producing a hat pattern F0 contour over a declining baseline and compared it to the corpus based model using statistical techniques. In contrast to the baseline model, the modified intonation model incorporates CART based accent type prediction model and a linear regression model to predict F0 target values for a given sentence.

The new prosodic models are much better than the original baseline model, given the nature and size of the corpus. Multiple speakers in particular made it difficult to produce a consistent dataset from which to train models. Nevertheless, it proves both that Festival can be successfully used as a tool for prosody research for languages other than English or German and that within the limits of the resources, reasonable improvements can be achieved.

Automatic relabelling of the database based on F0 stylised contour resulted in consistent and linguistically meaningful accent type labels which in turn improved CART results. Adding language specific features improved accent prediction, demonstrating the importance of prior linguistic knowledge of the prosody of the language being modelled.

The normalisation of the pitch contours of the different speakers to the same pitch range resulted in limited success. Further work is being carried out to investigate other transforms that are more appropriate than the normalisation technique applied here, for example, the piecewise linear transform used by Patterson (2000).

In this work both objective and subjective evaluation has been used to assess the new prosody model on single sentences. The extended procedure and tests to evaluate the acceptance of the new intonation model are described in Chapter 10.

10. Evaluation

Introduction

This chapter describes the evaluation of pitch accents modelled for Polish text–to–speech synthesis (TTS). We introduce methods of speech synthesis assessment with an emphasis on prosody. We, then, describe a subjective assessment test carried out here. It consists of three-part perception study involving Polish native speakers who judge sentences using a default prosodic model against sentences produced with the new prosodic model discussed in Chapter 9.

The chapter is organised as follows: Section 10.1 describes methods of TTS assessment, in Section 10.2 we present experimental design in the form of three perceptual studies, followed by results in Section 10.3. In Section 10.4 we present the main conclusions and future directions of this work.

10.1. Assessment of prosody in TTS

In the development of modern text to speech systems, emphasis needs to be put on standardisation of tests capable of testing the true quality of the generated speech and which would be able to compare directly different systems or research techniques. Not only should they answer the question of preference for a particular system but should also provide feedback on the shortcomings of one particular system.

The fact that there are such subtle issues to cope with in the area of intonation perception makes it difficult to introduce an objective evaluation metric for naturalness of synthetic output. Thus, while intelligibility tests are abundant, naturalness appears to be more difficult to measure automatically. Currently RMSE (Root Mean Square Error) and Correlation with the natural contour are used as objective metrics to evaluate synthetic intonation, but there appears to be much room for further research and improvement in this area (Clark & Dusterhoff, 1999).

The recommended subjective evaluation method designed to test the quality of components within one system is Mean Opinion Scale (MOS) measure, in its original form as proposed by ITU-T (ITU–T, 1996) or in its recently modified form (Viswanathan & Viswanathan, 2005). The two main properties being measured by the tests are intelligibility and naturalness. Having achieved segment intelligibility most efforts in modern synthesisers are directed towards naturalness. The naturalness of synthetic speech depends greatly on its prosodic quality. In this work we described two prosodic models developed within one system cf. Chapter 9. The main goal is to assess the relative naturalness and acceptability of the corpus-based intonation model and to compare it with the original rule-based model. At the same time, the overall acceptability of the models is measured by comparing them with the natural sentences. By selecting stimuli from the database the corpus-based model was trained on, we assess the suitability of it as a training data.

10.2. Perception study

Three perception tests were designed in order to address the main goals of the study, results of which were first reported in Oliver (2005b). In the first test, we measure the similarity of segmentally naturally produced sentences with synthetic F0 curves produced by the two intonation models to the natural original sentences and evaluate how much they differ. Apart from comparing two models, this test, above all, would directly indicate the usefulness of using a particular corpus for building an intonation model.

Synthetic intonation is expected to be different from the natural realisation but this does not necessarily mean that it is inappropriate.

The method selected for the first study is based on the Visual Sort and Rate method, introduced in Chapter 7, which has been proved to provide consistent results in an audio listening task, increasing inter and intra-subject agreement (Granqvist, 2003). A second test was designed to test the overall acceptability and quality of the new synthesis model.

In this study, fully synthetic signals were generated by a TTS system. In the third study, two intonation models were compared with each other but unlike the first study, they were generated fully by a TTS system, using the same procedure as in the second study. The sentences were designed in such a way that they only differed in the intonation model used. The methods used in the second and third study follow ITU-T test recommendations and evaluation procedures (Vazquez-Alvarez & Huckvale, 2002). All three perceptual tests were carried out in a laboratory environment through high-quality Sennheiser headphones and supervised by the experimenter.

10.2.1. Subjects

20 native speakers of Polish (7 males and 13 females) aged 22-32 took part in the study. They were selected from university students at the Saarland University who at the time of the experiment were on an ERASMUS/SOCRATES exchange program. None had speech or hearing problems, most were inexperienced with speech synthesis systems and only one had received any teaching in phonetics and phonology.

10.2.2. Speech Material and Procedure

The material used in all the parts of the perception study was taken from the Babel and the PoInt database both read, semi-spontaneous and the map task part.

In order to prepare the stimuli three versions of each sentence needs to be created

1. original sentences resynthesised with PSOLA,
2. sentences synthesised with the default hat pattern model,
3. sentences synthesised with the linear regression model.

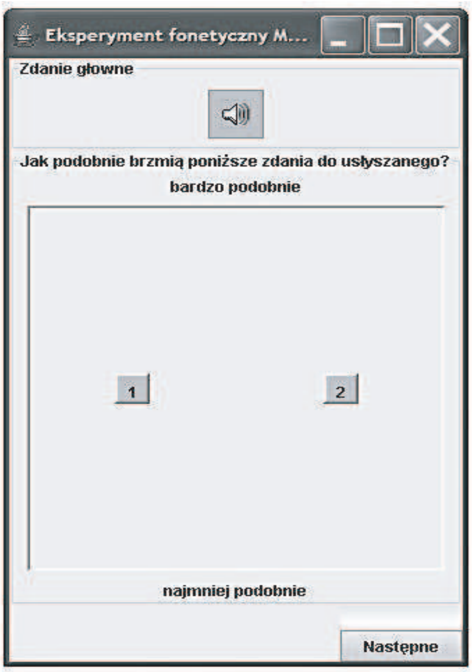
By switching the prosody module to be used, the Festival TTS system could produce two versions of all the sentences corresponding to the default hat pattern model and the new linear regression model. The speech analysis program Praat was used to PSOLA resynthesise the original sentences from the database.

Before the perception experiment started, the subjects were given a short introduction to its nature. They were told that there would be three parts and that they would hear synthetic speech, in particular pairs of sentences featuring two versions of the same utterance. It was also explained that the differences between the two versions could be quite subtle and the decision that two sentences sound the same is not wrong (Part I).

It was always pointed out that if they considered the differences to be too subtle to decide which version of the sentence sound better or worse, they should trust their intuition and still choose one that they would prefer to hear from a machine talking to them (Part III).

The subjects were also encouraged to ask any questions or make any comments, and were asked to report any strange program behaviour, e.g. a crash, both in the practice session and in the main part of the study, if necessary.

Part I The first experiment measured how similar the intonation curves produced by two prosody models are to the original sentences taken from the PoInt database (Karpiński & Kleśta, 2001).



Jak podobnie brzmią poniższe zdania do usłyszanego?

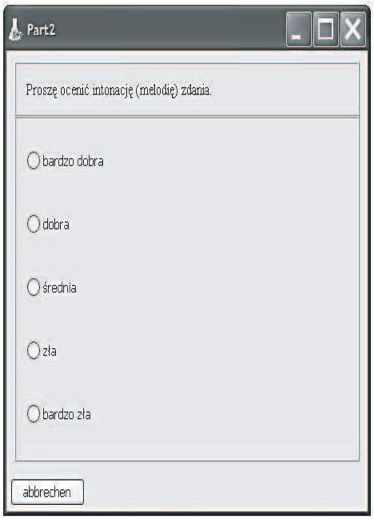
Figure 10.1.: *Perception study part I interface*

In total, 25 sentences (five in the practice stage and twenty for the experiment's main part) were randomly selected for the first experiment. The method used in the experiment follows the same procedure as that used in the previous experiment, see Chapter 7. Figure 10.1 shows the interface used for the first part using Rating Test software (Schröder, 2004). The subjects were always asked to rate the relative similarity of the two target sentences to one main stimulus.

Using the Visual Sort and Rate method (Granqvist, 2003), the main stimulus was an original sentence from the database, which had also been resynthesised. The two target stimuli were versions of the main stimulus and differed from it only in the way their F0 contour was generated. The stimuli were arranged randomly within each task. The subjects saw a screen with icons representing two versions of the original sentence placed in a window. At the top of the window there was an icon with the original version to which the comparison was to be made. The written question "how similar do the sentences placed below sound in relation to the one just heard" was placed under the original version. The original sentence was referred to as 'the one just heard' as it was always played at the beginning of each comparison.

In addition, a text was placed at the top and bottom of the screen with the words 'very similar', and 'very different', accordingly. For the calculation of the similarity scores, the label 'very similar' corresponded to a value of one hundred, and the label 'very different' corresponded to a value of zero.

By clicking on the stimulus button, the versions could be heard. The icons could be moved up and down the window, either closer or further from the original version. At the same time, the stimuli were being rated according to their similarity to each another.


Part2

Proszę ocenić intonację (melodię) zdania.

○ bardzo dobra

○ dobra

○ średnia

○ zła

○ bardzo zła

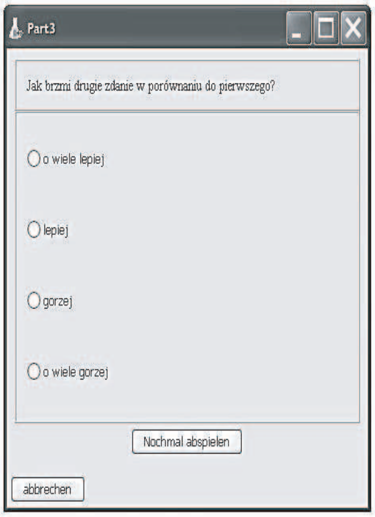
Figure 10.2.: *Perception study part II interface*

Part II In the second experiment, sentences were presented individually for judgement in an Absolute Category Rating test to measure overall acceptability of sentences produced with the linear regression F0 model. This part was presented to the listeners after a short break during which feedback from Part I was collected and the procedure for Part II was explained. The break was motivated by the fact that a different interface was used for the rest of the study (experiment two and three).

The task involved direct judging of intonation, a concept which had to be explained to listeners. Also, the stimuli used in Part II were fully synthesised by a TTS system as opposed to the PSOLA resynthesis used in Part I.

The scale used for the study was based on the five-point category-judgement scale recommended by ITU-T (ITU–T, 1996). Using the Scape experimental program (Grabowski & Bauer, 2004) 25 randomly selected sentences, different to those used in Part I, were synthesised with Festival TTS using the linear regression model. They were presented in a random order and were ranked on the scale from perceived as having a very bad intonation (1/5) to perceived as having a very good intonation (5/5).

The listeners were allowed to play the stimuli repeatedly, but were asked to make a decision as early as possible, i.e., ideally after first or second hearing. The visual interface used is shown in Fig. 10.2.



Jak brzmi drugie zdanie w porównaniu do pierwszego?

Figure 10.3.: *Perception study part III interface*

Part III The third part was designed to determine the preference ranking by direct comparison of the two intonation models in a pairwise comparison task. In this method, each model was matched head-to-head with the other using the Comparison Category Rating method recommended by ITU-T.

Listeners were always presented with the same sentence produced by two different systems and were asked to judge how the second sentence compared to the first, giving labels ’better’, ’much better’, ’worse’ or ’much worse’. The subjects were forced to give rank orders in all cases.

In order to eliminate order effects we designed 2x20 pairs with each model as second in the pair and additionally added fillers consisting of same model pairs. As in experiment two, we again used Scape (Grabowski & Bauer, 2004) as an interface for the evaluation, see Figure 10.3.

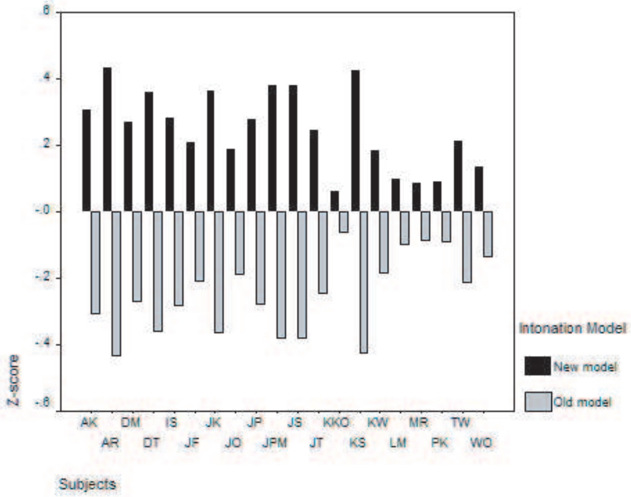


Figure 10.4.: *Across-subjects Z-score distribution*

10.3. Results

Part I Part I was designed to measure the overall acceptability of the two intonation models by comparing them to the natural sentences and evaluating how much they differ. This can then to be used as an indication of the usefulness of using a particular corpus for building an intonation model. With an alpha level of .05, ANOVA revealed a significant effect of intonation model across listeners, F(1,798)=54.346, p<0.001. The new linear regression model always received a score above average, see Figure 10.4.

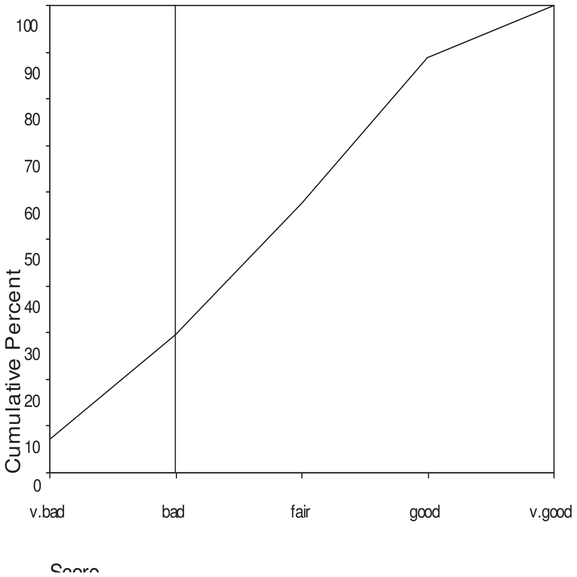


Figure 10.5.: *Cumulative frequency for the linear regression model ratings*

Part II Part II assessed the naturalness and acceptability of the corpus based intonation model described in Chapter 9 using sentences generated fully by the TTS system. The mean score for the new model was 3.17, with a variance 1.14 across the subjects. On average, subjects rated the synthetic intonation produced by the new modified model in 70.4% of the cases in categories 'fair' to 'very good'. Figure 10.5 shows the cumulative score in this part.

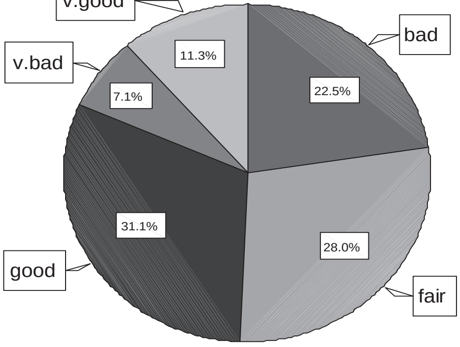


Figure 10.6.: *Intonation score distribution in part II*

In Figure 10.6 we detail the distribution of all the ratings as a percentage of all the stimuli.

Part III In part III, the two intonation models were compared directly with each other. As in part two, the sentences were generated fully by a TTS system. Pairwise comparisons revealed statistically significant differences in ratings for the two models, F(1,878)=106.494, p<0.001. The new intonation model was preferred to the original model in all cases by on average a 66% to 34% ratio across orders of presentation. In Table 10.1 we summarise the preference scores for the two orders of presentation: pairs composed of a) old model - new model and b) new model - old model.

Order of presentation	New model preference	Old model preference
a) old model - new model	70%	30%
b) new model - old model	62%	38%
Average	66%	34%

Table 10.1.: *Preference scores for the two intonation models in two orders of presentation*

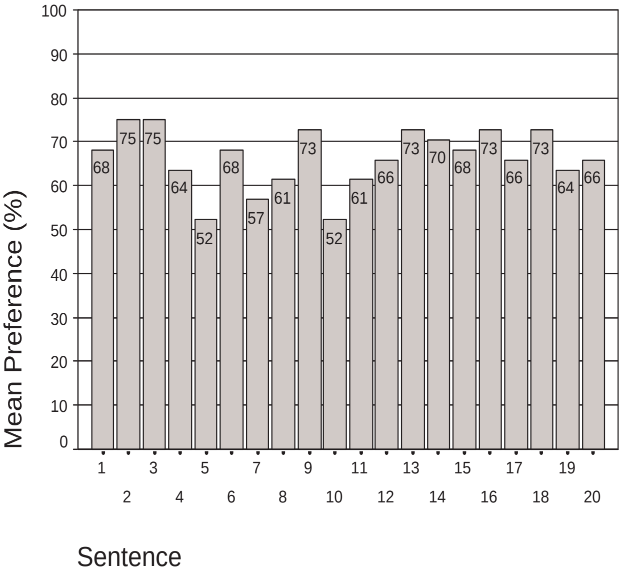


Figure 10.7.: *Per sentence preference for the linear regression intonation model*

Figure 10.7 summarises the test results. For each sentence we show the preference for the new model measured in per cent values of judgements 'better' and 'much better'. Apart from sentences 5 and 10, listeners show significant preference for the linear regression model. Sentences 5 and 10 have almost equal preference for either intonation model (52%). Examination of these sentences revealed that they both have a negative semantic load and are also the two shortest sentences.

10.4. Conclusions

The current study offered a description and an evaluation of intonation generation in speech synthesis based on the example of the Polish module implementation in Festival TTS system. We compared an original rule based model producing a hat pattern F0 contour over a declining baseline to the corpus based model using statistical techniques. A subjective evaluation composed of three perceptual tests was designed to measure:

a) the similarity of sentences with synthetic F0 curves produced by the two intonation models to the natural sentences in the database they were trained on, thus also to test the appropriateness of the database used,

b) to measure an overall acceptability and quality of the intonation curve produced by the new model, and

c) to examine listeners' preference for a particular model in a direct comparison task.

In the evaluation experiment listeners confirmed that for isolated sentences the corpus based model produces F0 contours that are more similar to the natural sentences than the original simple rule-based model, thus validating the use of PoInt database as a basis for constructing prosodic models. The assessment of overall appropriateness and quality of synthetic intonation produced with this method showed that majority of tokens was judged positively by the respondents (fair to very good). The pair comparison test ranking the two models revealed that the new intonation model was significantly and consistently preferred to the original model and represents a promising approach for implementing natural prosody in the context of a text-to-speech synthesis system. Further work is needed to improve the corpus based model to produce appropriate intonation patterns that would reflect different prosodic contexts required in e.g. dialogue domain, and be capable of capturing individual characteristics of speakers in specific domains.

11. Conclusions

Introduction

The main body of the research in this thesis concerns intonation modelling for speech synthesis. The research undertaken in the analysis and synthesis of Polish intonation has resulted in several findings which are summarised in the rest of this Chapter, which points out the thesis contributions as well as its limitations and possible directions for future research.

11.1. Main Findings

We presented a study of Polish intonation modelling, the results of its implementation in the Festival speech synthesis system and a formal evaluation. Based on the description of Polish intonation, the PoInt speech database and an existing TTS system, we developed a model capable of generating a suitable F0 contour. The development consisted of three stages:

- Automatic prosodic annotation using parameters derived from stylised F0 contours and pitch accent classes obtained through a clustering process,
- Prediction of pitch accent types using classification and regression trees, and
- A linear regression model to predict F0 contour values.

The research undertaken in this thesis raised questions concerning the application of stylisation, clustering and statistical modelling techniques to languages other than English and resulted in innovative approaches necessitated by the nature of the Polish language and the speech resources available. At the same time the study highlighted language specific theoretical and practical issues which resulted in their closer investigation.

First, we reviewed the existing intonation descriptions available for Polish, and decided that because Demenko's acoustically motivated findings were consistent with computer assisted classification of basic Polish intonation contours by Jassem (1987) and earlier impressionistic findings (Steffen-Batog, 1996; Dłuska, 1964), and decided that it would to be a reliable starting point for the validation of the classes of nuclear patterns and for the prosodic modelling.

Second, two aspects of the Polish prosodic system were chosen for closer investigation: lexical stress placement due to its unclear nature on a subset of past tense verb forms, and F0 peak alignment, as this was seen as crucial for any further acoustic parameterisation of F0 contours required in the modelling stage.

Polish has a strong tendency towards fixed stress on the penultimate syllable. Based on the database recordings of 40 speakers of Polish reading a narrative text, contrary to stress exception claims in the literature, accent on the penultimate rather than the antepenultimate syllable seems to be prevalent in modern literary texts read aloud (produced by 31 out of 40 speakers).

We investigated how reliably stress can be perceived and, with a view to implementation in speech synthesis, the nature of the acoustic correlates of perceived stress in these verbs. A perceptual experiment revealed that these accented syllable can be reliably perceived by native listeners (95% inter-listener agreement).

A second experiment was carried out to shed light on the potential acceptability of using penultimate stress in these verb forms. In a perception task naive listeners evaluated the synthesised texts for acceptability of the alternative stress placement. Accent on the penultimate syllable was preferred for synthesised texts of different registers; literary and colloquial texts. The overall preference for penultimate stress indicates the new status of the person-number marker as an affix, rather than a clitic. As for the implications for speech synthesis systems, a penultimate stress on verbs can be assigned across the board.

The lexical stress placement study was followed by a tonal accent peak alignment study. The purpose of this study was to attempt to produce a descriptive model of peak placement and to analyse the relationship between peak location and its segmental anchors. Questions posed in this study further pertain to the phonetic realisation of the phonologically specified tonal accents in Polish as a function of the following factors: speech tempo, information structure, and position within utterance. We investigated whether the information structure affects the choice and realisation of the nuclear pitch accents in Polish with respect to peak alignment and whether the phenomenon of segmental anchoring can be observed.

Different accent types were found in the same focus condition and the same accent types in different focus conditions. Speakers employ both peak alignment and peak height to obtain a phonological contrast between the different focus conditions. The fact that the relative measures for peak alignment did not differ significantly with changing speech rate supports the claim that speakers control the peak alignment in an extremely consistent way. The results show that the segmental anchoring for Polish is defined by the syllable onset and offset. With respect to the position of the focused item in the utterance we find the following tendency: the later the focused item in the utterance the earlier the peak alignment.

Focus is seen as an influencing factor in the phonetic realisation of phonological accent. In Polish, broad and narrow focus utterances are produced with falling contour with early, often downstepped H+L* and late H*+L peak realisation, whereas contrastive focus is characterised by a rise L+H*.

Next, we presented a method for obtaining precise acoustic representation of underlying F0 contour within pitch accent classes in Polish by means of automatic F0 stylisation. After preliminary tests, the chosen method, MOMEL (Hirst & Espesser, 1993), was found to produce systematic errors (targets redundant, missing or put in wrong places), which had already been reported in literature. As the missing target points could result in the change from rising to falling contour, and vice versa, or in an intonation event too high for the range of the speaker, a study was conducted to eliminate these errors. The MOMEL algorithm was augmented for boundary locations. Elements of the stylised curve were enriched with the initial and final variations of F0 essential for capturing boundary prosodic events. Additional target points based on existing F0 were used at boundaries when stylisation on MOMEL target points computed there did not correspond to the original F0 points in the speech signal or exceeded the speaker F0 range. A perceptual study revealed a significant preference for modified MOMEL F0 stylisation ($p < 0.001$) over the original MOMEL method. In a comparison with another stylisation technique, Prosogram, both the modified MOMEL and Prosogram were judged as perceptually much closer copies of the original speech signal.

As a next step, the acoustic parameters characterising the identified accent types were re-examined. The speech database was enriched with segmental and textual annotation layers, and F0 normalisation was performed across female and male speakers. Based on this pre-processing, automatic accent feature extraction was carried out. The development consisted of creating an automatic prosodic annotation. The pitch accent classes used in this annotation were obtained by means of a clustering process based on parameters derived from a stylised pitch contours.

Accent type study involved classification of contour types using machine learning techniques. Two step method was used :

- Self-Organising Maps (Kaski, 1997), chosen due to it having no preconceptions about the number of resulting clusters (intonation categories), and its ability to represent high-dimensional acoustic feature vectors in a low-dimensional form,

- Hierarchical agglomerative clustering (HAC), which starts with the low-level clusters built by SOMs and builds the hierarchy by progressively merging clusters using average similarity.

A close inspection of the resulting clusters made it possible to derive linguistically motivated classes by aligning their members with the original prosodic annotation and sentence type information. In this way three accent types were selected (Falling, Rising, Rising-Falling), which served as input to a prosody prediction module.

Following the analysis of the prosodic system, the intonation prediction and generation implementation was carried out in two stages:

- Prediction of accent placement and accent type using classification and regression trees (Breiman et al., 1984; Dusterhoff et al., 1999),
- Prediction/generation of F0 contour using linear regression (Black & Hunt, 1996).

The system within which the current work is carried out is a Polish voice module (Oliver, 1998) for Festival (Black et al., 1998) developed as part of an MSc. degree at Edinburgh University. The Polish module uses concatenative diphone-based synthesis. The work described here improved upon the simple intonation models that this module uses.

The modified intonation model is statistical, relying on a prosodically labelled corpus. First, CART trees are used to predict pitch accents and boundary tones. In their default form, these modules are designed to be used for an implementation of the ToBI intonation labelling system. Only after the accent type and place have been predicted, can an F0 contour be generated in this framework. In the original implementation for Polish accent prediction there was no support for different accent types. In contrast, for a language like English, models built from databases with labels based on ToBI or similar descriptions have been implemented. In our case, this paradigm was modified to use the PoInt annotation after re-labelling described in Chapter 8 based on a set of R, F, and R-F pitch accents. For accent type and place prediction we initially used the following features: position of syllable in phrase/word, strength of break, stress, number of (stressed, unstressed, accented) syllables since/till last phrase break, number of syllables since/till last accented syllable, number of minor phrase breaks since last major phrase break. We then introduced language specific features to improve accent type prediction, namely, additional information regarding the position of a syllable within a word. It is motivated by the fact that in Polish the last content word usually receives the phrase accent and stress occurs generally on the penultimate syllable of a word. To accommodate this the following features were provided: the word is final in a phrase, the syllable is penultimate in a word, a simplified part of speech (content/function word). We obtained a satisfactory 83.3% accuracy across the pitch accent types.

Contour generation in the modified model was carried out using three linear regression (LR) models. Each of the three models predicts the F0 at a different point of a syllable (start, middle and end respectively). The factors incorporate information like the type of accent present, the position of phrase breaks, syllable stress and syllable position within the text. Each model considers this information for a five syllable window centered on the current syllable. This allows the pitch on syllables around an accented syllable to be affected by the presence of the accent, so that pitch movement is not restricted to occur on the syllable that is marked with the pitch event. The symbolic results for F0 generation, as measured by RMSE and correlation with the original signal reached an average 0.71 correlation and 41.98 RMSE values.

The symbolic results were evaluated by means of perception tests using 20 subjects. The main goal was to assess the naturalness and acceptability of the corpus-based linear regression intonation model and to compare it to the original rule-based 'hat accent' model. The new linear regression intonation model was significantly preferred to the original model by on average a 66% to 34% ratio across conditions. Apart from comparing different implementations of intonation models to each other, these perceptual studies measured the relation of synthetic stimuli to the original signal, thus evaluating the overall acceptability of synthetic intonation. Subjects rated the synthetic intonation produced by the new modified model in 70.4% of the cases as 'fair' to 'very good'. Thus, the results of the perceptual experiments show that corpus based intonation model is perceived closer to the original signal than the rule-based model.

The new intonation model's perceived success can be attributed to a number of causes. One of them is that based on the description of Polish intonation, an evaluation of the typology of accent types present in the language was carried out using the speech database. Evaluating theoretical points by means of experimental studies, as in the case of stress assignment and F0 peak alignment was very informative.

Very close attention was paid to all the building blocks of the intonation modelling process. Both theoretical and computational considerations resulted in the decisions to remain theory independent and to use a test platform which enables different theories to be tested, is reliable, and which, furthermore, already contains a language specific TTS module onto which a new intonation model can be added.

In addition, manual and automatic improvement of the data annotation is perceived as one of the main improvement boosters. Eliminating outliers, annotation misalignments, correcting segmental and prosodic labels made it possible to derive more reliable acoustic parameters and identify linguistic features.

Furthermore, encouraging results are thought to stem from the ability to build a precise parameterisation for each pitch accent category by using an augmented interpolation through unvoiced segments and thus having a close copy of original signal preserving perceptually relevant information.

Finally, we feel that the quality of the intonation model was greatly improved by the automatic pitch accent type classification which reduced the bias and cost of prosodic database annotation and resulted in a consistent specification of pitch accents.

11.2. Future directions

This thesis describes a Polish intonation modelling module intended for use in a speech synthesis system. The main question arising is whether the model captures all relevant intonation events present in the language. It should be kept in mind that the corpora used in the study covers a particular subset of intonation patterns, and the model is capable of producing only those events that were present in the speech base. Moreover, due to the mixture of read and semi-spontaneous monologue, it is not suggested that the same contours will be completely appropriate in a dialogue system. Extending the model with more varied speech data would be desirable.

The model is statistically trainable and adaptable to deal with new speech data; the only limitation is the symbolic representation created by the linguistic part of the synthesis. We might have a model covering an adequate number of accent types reflecting different discourse situations, but if the textual analysis fails to predict the right accents from the prosodic part, the end result will not be as expected.

In today's multi–media world, with talking gadgets and speaking agents, there is need for more discourse specific intonation as well as more expressive intonation. Whether this should be achieved with diphone systems or unit selection systems is debatable. While a diphone-based system gives us more control over the F0 contour (but at the expense of quality), a unit-selection system's overall quality is better but the lack of full control over F0 leaves it prone to errors if a wrong database target is chosen. If the desired contour is underrepresented or absent in the speech base, the desired intonation pattern will not be produced. Although this and many other issues are left open, it is hoped that the research undertaken here brings us a step closer to providing more speech technology resources for minority languages.

A. Stress placement acceptability test

Stimuli(30)

1. Na ostatnim spotkaniu musieliśmy nawet wystąpić bez jednego zawodnika.
2. Latem byliśmy bliżsi pozyskania zawodnika niż obecnie.
3. Choć na pewno nie graliśmy jeszcze tak jakbyśmy o tym marzyli.
4. W drugiej połowie zagraliśmy jednak bardzo dobrze w obronie.
5. Po rannym treningu poszliśmy cała ekipą do klubu na małe piwko.
6. Pogodziliśmy się z tym że spadliśmy na czwartą pozycję.
7. Już dawno nie zdobyliśmy tyle punktów.
8. W drugiej połowie zwolniliśmy nieco.
9. Dzisiejszy mecz przegraliśmy w obronie.
10. Obawialiśmy się tego spotkania.
11. A kiedy to mówiłem usłyszeliśmy niecierpliwy poświst.
12. I płyneliśmy dzień jeden i dzień wtóry.
13. My członkowie szeroko rozpostartej rodziny szlacheckiej stanowiliśmy odrębną społeczność.
14. Schodziliśmy się nad wieczorem pijać kawę i herbatę.

15. Jadaliśmy ciastka i lody i naturalnie rozmawialiśmy o wypadkach bieżących.

16. Przyjeliśmy pana nie na to żebyś tutaj produkował się ze swoją filantropią.

17. Toż zeszłego roku w samą Wilię wzieliśmy z nimi dzika prawie żywcem.

18. Przyszliśmy w nocy na pozycję i nie mogliśmy obwarować miejsc mocną straż ą.

19. Masz przecie serce i nie popsujesz tej świę tej zgody w której dotąd żyliśmy.

20. Zaraz za Kiejdanami wykręciliśmy do Dałnowa i Kroków.

21. Kogoście widzieli - widzieliśmy Ewę.

22. Czestośmy go widzieli zanim wyjechaliśmy z miasta.

23. Kiedyśmy go zobaczyli zaświeciliśmy lampę.

24. Kiedyśmy wreszcie wrócili przebraliśmy go w nowe rzeczy.

25. Dlaczegoście to zrobili? Ponieważ nie znaleźliśmy innego rozwiązania.

26. Potem gdyśmy schodzili z ostatniego już mostku zwróciliśmy uwagę na bogactwo inwencji kapitelów.

27. Więc z tą panną ktorąśmy jadąc spotkali, mieszkaliśmy razem u kuzynki państwa.

28. A gdyśmy zakończyli sprawę zapiliśmy ją dobrym winem żydowskim.

29. Któż mi uwierzy żeśmy biegli dwa dni bez ustanku i deptaliśmy im po pietach.

30. A przecie kiedyśmy się pobrali nie dziwowaliśmy się niczemu.

Fillers (20)

1. Czy to prawda że baranki zjadają krzaki?
2. Wiadomości te gromadziły się z wolna i przypadkowo.
3. Bardzo lubię zachody słońca.
4. Proszę cię narysuj mi baranka.
5. Co to jest za przedmiot?
6. Więc ty przybyłeś z innej planety?
7. Wiele pracy kosztował mnie ten rysunek.
8. Byłem zdenerwowany stanem mojej maszyny.
9. Urodziłam się równocześnie ze słońcem.
10. To się bardzo przydaje do podgrzewania śniadań.
11. Stał całkowicie zbity z tropu trzymając klosz w powietrzu.
12. W jaki więc sposób mogłeś mnie rozpoznać?
13. Najjaśniejszy panie proszę mi wybaczyć moje pytania.
14. Nie odchodź mianuję cię ministrem.
15. Wobec tego będziesz sam siebie sądzić.
16. Jeśli potrafisz dobrze siebie osądzić będziesz naprawdę mądry.
17. Dorośli są bardzo dziwni.
18. A co trzeba zrobić aby kapelusz spadł?

19. Dorośli są zdecydowanie śmieszni.

20. Nie mam czasu na włóczęgę.

B. Sampa Transcription for Polish

Vowels

The vowel system comprises 8 phonemes, as follows. Those symbolized with a tilde are nasalised.

SAMPA	Orthography	Example
i	pit	pit
I	typ	tIp
e	test	test
a	pat	pat
o	pot	pot
u	puk	puk
e~	gęś	ge~s’
o~	wąs	vo~s

Consonants

The consonant system comprises 29 phonemes, as follows. The symbol ’ indicates palatalization.

p	pik	pik
b	bit	bit
t	test	test
d	dym	dIm
k	kit	kit
g	gen	gen
f	fan	fan
v	wilk	vilk
s	syk	sIk
z	zbir	zbir

S	szyk	SIk
Z	żyto	ZIto
s’	świt	s’vit
z’	źle	z’le
x	hymn	xImn
ts	cyk	tsIk
dz	dzwon	dzvon
tS	czyn	tSIn
dZ	dżem	dZem
ts’	ćma	ts’ma
dz’	dźwig	dz’vik
m	mysz	mIS
n	nasz	naS
n’	koń	kon’
N	pęk	peNk
l	luk	luk
r	ryk	rIk
w	łyk	wIk
j	jak	jak

C. Features Used in CART modelling

Features Used in CART modelling

lisp_sylaccent_accent Syllable is accented

lisp_syl_final_syl_in_phrase Syllable is final in phrase

lisp_syl_penult_syl_in_word Syllable is penultimate in word

R:SylStructure.parent.R:Word.p.gpos Part of speech of previous word

R:SylStructure.parent.gpos Part of speech of current word

R:SylStructure.parent.R:Word.n.gpos Part of speech of next word

syl_in Number of syllables since last phrase break

syl_out Number of syllables before next phrase break

ssyl_in Number of stressed syllables since last phrase break

ssyl_out Number of stressed syllables before next phrase break

asyl_in Number of accented syllables since last phrase break

asyl_out Number of accented syllables before next phrase break

pp.stress Syllable before previous is stressed

p.stress Previous syllable is stressed

stress Syllable is stressed

n.stress Next syllable is stressed

nn.stress Sylllabe after next is stressed

pp.syl_break Break level after the syllable before previous syllable

p.syl_break Break level after the previous syllable

syl_break Break level after the syllable

n.syl_break Break level after the next syllable

nn.syl_break Break level after the syllable following next syllable

syl_numphones Number of phones in the syllable

syl_onset_type Syllable onset type

pos_in_word Syllable position in word

position_type The type of syllable with respect to the word it is related to

last_accent Number of syllables since last accented syllable

next_accent Number of syllables to the next accented syllable

sub_phrases Number of non major phrase breaks since last major phrase break

old_syl_break Strength of break after the syllable

D. Features Used in Linear Regression modelling

Feature descriptions:

syl_startpitch F0 at syllable start

syl_midpitch F0 at syllable mid

syl_endpitch F0 at syllable end

lisp_sylaccent_accent Accented syllable

p.lisp_sylaccent_accent Accent on previous syllable

pp.lisp_sylaccent_accent Accent on syllable before previous syllable

n.lisp_sylaccent_accent Accent on next syllable

nn.lisp_sylaccent_accent Accent on syllable after next

lisp_sylaccent_hstar R accent on syllable

lisp_sylaccent_mstar R-F accent on syllable

lisp_sylaccent_lstar F accent on syllable

p.lisp_sylaccent_hstar R accent on previous syllable

p.lisp_sylaccent_mstar R-F accent on previous syllable

p.lisp_sylaccent_lstar F accent on previous syllable

pp.lisp_sylaccent_hstar R accent on syllable before previous syllable

pp.lisp_sylaccent_mstar R -Faccent on syllable before previous syllable

pp.lisp_sylaccent_lstar F accent on syllable before previous syllable

n.lisp_sylaccent_hstar R accent on next syllable

n.lisp_sylaccent_mstar R-F accent on next syllable

n.lisp_sylaccent_lstar F accent on next syllable

nn.lisp_sylaccent_hstar R accent on syllable after next syllable

nn.lisp_sylaccent_mstar R-F accent on syllable after next syllable

nn.lisp_sylaccent_lstar F accent on syllable after next syllable

pos_in_word Syllable position in word

pp.syl_break Break level after the syllable before previous syllable

p.syl_break Break level after the previous syllable

syl_break Break level after the syllable

n.syl_break Break level after the next syllable

nn.syl_break Break level after the syllable following next syllable

pp.stress Syllable before previous is stressed

p.stress Previous syllable is stressed

stress Syllable is stressed

n.stress Next syllable is stressed

nn.stress Syllable after next is stressed

syl_in Number of syllables since last phrase break

syl_out Number of syllables before next phrase break

ssyl_in Number of stressed syllables since last phrase break

ssyl_out Number of stressed syllables before next phrase break

asyl_in Number of accented syllables since last phrase break

asyl_out Number of accented syllables before next phrase break

last_accent Number of syllables since last accented syllable

next_accent Number of syllables before next accented syllable

syl_numphones Number of phones in the syllable

E. Accent Type Prediction Preliminary Results

Method	wagon7lim_a	wagon7lim	wagon7unlim	wagon7unlim_a
Total	3566	3566	3566	3566
Total H	849	828	825	841
Total M	901	934	925	867
Total L	922	923	965	972
Total NONE	894	881	851	886
Correct (%)	84.55	83.65	82.26	85.17
H (%)	83.27	82.97	81.58	83.35
M (%)	82.8	85.76	83.68	85.01
L (%)	72.45	74.87	75.13	73.35
NONE (%)	100	91.2	93.6	100

Table E.1.: *Accent type prediction confusion matrix Training Data.*

Method	wagon7lim_a	wagon7lim	wagon7unlim	wagon7unlim_a
Total	3962	3962	3962	3962
Correct	2491	2737	2648	2390
Correct (%)	62.87	69.08	66.84	60.32
H (%)	72.47	75.92	73.62	76.35
M (%)	31.36	53.53	47.36	16.2
L (%)	50.54	60.57	60.7	52.9
NONE (%)	100	88.03	87.41	100
Entropy	4.33	2.91	3.29	5.10
Perplexity	20.13	7.52	9.75	34.32

Table E.2.: *Accent type prediction confusion matrix Test Data.*

F. F0 Prediction Preliminary Results

Training Data		
Method	ols_min	ols_max
Set size	4491	4491
RMSE Start	41.31	42.21
Correlation Start	0.49	0.48
RMSE Mid	53.02	53.15
Correlation Mid	0.52	0.53
RMSE End	51.36	51.63
Correlation End	0.55	0.55
Test Data		
Method	ols_min	ols_max
Set size	499	499
RMSE Start	41.71	41.64
Correlation Start	0.45	0.45
RMSE Mid	57.33	57.36
Correlation Mid	0.44	0.44
RMSE End	54.40	54.52
Correlation. End	0.51	0.51

Table F.1.: *Linear Regression F0 Prediction Models.*

Bibliography

Abrantes, A.J., Marques, J.S. & Trancoso, I.M. (1991). Hybrid sinusoidal modeling of speech without voicing decision. In: *Proceedings of Eurospeech 1991.* Genova, 231–234.

Adriaens, L.M.H. (1991). *Ein Modell deutscher Intonation: eine experimentell– phonetische Untersuchung nach den perzeptiv relevanten Grunfrequenzünderrungen in vorgelesenem Text.* Ph.D. thesis, Eindhoven University of Technology.

Allen, J., Hunnicutt, M., Klatt, D.H., Armstrong, R.C. & Pisoni, D.B. (1987). *From Text to Speech: The MITalk System.* Cambridge Studies in Speech Science and Communication. New York, NY: Cambridge University Press.

Anderson, M., Pierrehumbert, J. & Liberman, M. (1984). Synthesis by rule of english intonation patterns. In: *Proceedings of ICASSP'84.* 2.8.1–2.8.4.

Andreeva, B. & Oliver, D. (2005). Information structure in polish and bulgarian: Accent types and peak alignment in broad and narrow focus. In: S. Franks, F.Y. Gladney & M. Tasseva-Kurktchieva (eds.) *In Formal Approaches to Slavic Linguistics 13: The South Carolina Meeting.* Ann Arbor, MI: Michigan Slavic Publications, 1–12.

Arvaniti, A., Ladd, D.R. & Mennen, I. (1998). Stability of tonal alignment: the case of Greek prenuclear accents. *Journal of Phonetics* **26(1)**, 3–25.

Atterer, M. & Ladd, D.R. (2004). On the phonetics and phonology of segmental anchoring of F0: evidence from German. *Journal of Phonetics* **32**, 177–197.

Bailly, G. (1989). Integration of rhythmic and syntactic constraints in a model of generation of french prosody. *Speech Communication* **8**, 137–146.

Bąk, P. (1995). *Gramatyka Języka Polskiego. Zarys populary.* Warszawa: Wiedza Powszechna.

Bard, E.G., Robertson, D. & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language* **72(1)**, 32–68.

Bartkova, K. & Sorin, C. (1987). A model of segmental duration for speech synthesis in french. *Speech Communication* **6**, 245–260.

Batliner, A., Nutt, M., Warnke, V., Nöth, E., Buckow, J., Huber, R. & Niemann, H. (1999). Automatic annotation and classification of phrase accents in spontaneous speech. In: *Proceedings of Eurospeech 1999*. 519–522.

Beckman, M.E. & Ayers, G.M. (1994). *Guidelines for ToBI Labeling*. Tech. rep., Department of Linguistics Ohio State University.

Beckman, M.E. & Pierrehumbert, J.B. (1986). Intonational structure in japanese and english. *Phonology Yearbook* **3**, 255–309.

Bellegarda, J., Silverman, K., Lenzo, K. & Anderson, V. (2001). Statistical prosodic modeling: From corpus design to parameter estimation. In: *Proceedings of IEEE Trans. Speech and Audio Processing*, vol. 9(1). 52–66.

Black, A. & Taylor, P. (1998). *Festival Speech Synthesis System: system documentation*. Tech. Rep. HCRC/TR–83, University of Edinburgh, Human Communication Research Centre.

Black, A., Taylor, P. & Caley, R. (1998). *The Festival speech synthesis system*. Retrieved from <http://festvox.org/festival/>.

Black, A.W. & Hunt, A.J. (1996). Generating f0 contours from ToBI labels using linear regression. In: *Proceedings of ICSLP 96*. Philadelphia, 1385–1388.

Blackburn, C., Vonwiller, J. & King, R. (1993). Automatic accent classification using artificial neural networks. In: *Proceedings of Eurospeech 1993*. 1241–1244.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International* **5**(9/10), 341–345.

Booij, G. & Rubach, J. (1987). Postcyclic versus postlexical rules in lexical phonology. *Linguistic Inquiry* **18.1**, 1–44.

Braun, B. (2005). *Production and Perception of Thematic Contrast in German*, *Contemporary Series in Descriptive Linguistics*, vol. 9. Oxford: Peter Lang Publishing.

Breiman, L., Friedman, J., Olhsen, J. & Stone, C.J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.

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

Bruce, G., Dogil, G., Filipsson, M., Jilka, M., Lastow, B., Mayer, J. & Mohler, G. (1996). Testing intonational models by computer simulation. In: *Proceedings of 5th Conference on Laboratory Phonology*. Northwestern University, Evanston/Illinois.

Buhmann, J., Vereecken, H., Fackrell, J., Martens, J.P. & van Coile, B. (2000). Data driven intonation modelling of 6 languages. In: *Proceedings of ICSLP-2000*. 179–182.

Campbell, W.N. (1989). Syllable–level duration determination. In: *Proceedings of Eurospeech 1989*. Paris, 2698–2701.

Campione, E., Hirst, D. & Véronis, J. (2000). Automatic stylisation and symbolic coding of f0: Implementations of the intsint model. In: A. Botinis (ed.) *Intonation: Analysis, Modelling and Technology*, *Text, Speech and Language Technology*, vol. 15. Dortrecht: Kluwer Academic Publishers, 185–208.

Campione, E. & Véronis, J. (2000). Une évaluation de l'algorithme de stylisation mélodique MOMEL. *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence* **19**, 27–44.

Carlson, R. & Granström, B. (1986). A search for durational rules in a real–speech data base. *Phonetica* **43**, 140–154.

Carolis, B.D., Pelachaud, C., Poggi, I. & Steedman, M. (2004). Apml, a mark-up language. In: H. Prendinger (ed.) *Life-like Characters, Tools, Affective Functions and Applications*. Berlin: Springer, 65–85.

Chan, M.V., Feng, X., Heinen, J.A. & Niederjohn, R.J. (1994). Classification of speech accents with neural networks. In: *Proceedings of the 1994 IEEE International Conference on Neural Networks-IEEE World Congress on Computational Intelligence*, vol. 7. 4483–4486.

Chen, S.H., Hwang, S. & Wang, Y.R. (1998). An rnn-based prosodic information synthesizer for mandarin text-to-speech. *IEEE transactions on speech and audio processing* **6**(3), 226–239.

Clark, R. (2003). *Generating Synthetic Pitch Contours Using Prosodic Structure.* Ph.D. thesis, University of Edinburgh.

Clark, R. & Dusterhoff, K. (1999). Objective Methods for Evaluation Synthetic Intonation. In: *Proceedings of Eurospeech 1999*, vol. 4. 1623–1626.

Cohen, A. & 't Hart, J. (1965). Perceptual analysis of intonation pattern. In: *Actes du 5ème Congrès International dAcoustique.* Liège, 1–4.

Cowart, W. (1997). *Experimental Syntax: Applying Objective Methods to Sentence Judgments.* Thousand Oaks/London/New Delhi: Sage Publications.

Cruttenden, A. (1997). *Intonation.* New York: Cambridge University Press, 2 edn.

Crystal, D. (1969). *Prosodic systems and intonation in English.* Cambridge studies in linguistics, 1. Cambridge University Press.

d'Alessandro, C. & Castellengo, M. (1994). The pitch of short-duration vibrato tones. *Journal of the Acoustical Society of America* **95(3)**, 1617–1630.

d'Alessandro, C. & Mertens, P. (1995). Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language* **9**(3), 257–288.

Demenko, G. (1999). *Analiza Cech Suprasegmentalnych Języka Polskiego na Potrzeby Technologii Mowy.* Poznań: Wydawnictwo Naukowe UAM.

Demenko, G. & Jassem, W. (1999a). A classification of Polish pitch accents using neural networks. In: *Proceedings of ICPhS 1999.* San Francisco, 1525–1528.

Demenko, G. & Jassem, W. (1999b). Modelling intonational phrase structure with artificial neural networks. In: *Proceedings of Eurospeech 1999.* San Francisco, 711–714.

Demenko, G., Nowak, I. & Imiołczyk, J. (1993). Analysis and Synthesis of Pitch Movements in a read Polish Text. In: *Proceedings of Eurospeech 1993*. Berlin, 797–800.

Dłuska, M. (1964). *Prozodia języka polskiego.* Warsaw: PWN.

Dogil, G. (1987). Lexical phonology and floating inflection in Polish. In: W.U. Dressler, H.C. Luschützky, O.E. Pfeiffer & J.R. Rennison (eds.) *Proceedings of the 5th International Phonology Meeting (Phonologica 1984)*. Cambridge: Cambridge University Press, 39–47.

Dogil, G. (1995a). Articulatory correlates of secondary stress in Polish and Spanish. In: *Phonetic AIMS*. Universität Stuttgart, 241–264.

Dogil, G. (1995b). Phonetic correlates of word stress. In: *Phonetic AIMS*. Universität Stuttgart, 1–60.

Durand, P., Durand-Deska, A., Gubrynowicz, R. & Marek, B. (2002). Polish: Prosodic aspects of czy questions. In: *Proceedings of Prosody 2002*. Aix-en-Provence, 255–258.

Dusterhoff, K. & Black, A. (1997). Generating f0 contours for speech synthesis using the tilt intonation theory. In: *Proceedings of ESCA Workshop of Intonation 1997*. Athens, Greece, 107–110.

Dusterhoff, K.E., Black, A. & Taylor, P. (1999). Using decision trees within the Tilt intonation model to predict f0 contours. In: *Proceedings of Eurospeech 1999*. 1627–1630.

Dymarski, P., Kuklinski, S. & Kula, S. (1995). A text-to-speech synethesizer for the Polish language. In: *Proceedings of Eurospeech 1995*. Madrid, 1101–1104.

Francuzik, K., Karpiński, M., Kleśta, J. & Szalkowska, E. (2005). Nuclear melody in polish semi-spontaneous and read speech: Evidence from the polish intonational database point. *Studia Phonetica Posnaniensia* **7**, 97–128.

Franks, S. & King, T.H. (2000). *A Handbook of Slavic Clitics.* Oxford: Oxford University Press.

Franks, S. (1985). Extrametricality and Stress in Polish. *Linguistic Inquiry* **1**, 144–150.

Frota, S. (2000). *Prosody and Focus in European Portuguese. Phonological Phrasing and Intonation.* Ph.D. thesis, University of Lisbon, New York: Garland.

Fujisaki, H. (1992). Modelling the process of fundamental frequency contour generation. In: Y. Tohkura, E. Vatikiotis-Bateson & Y. Sagisasaka (eds.) *Speech Perception, Production and Linguistic Structure.* IOS Press, 31–328.

Fujisaki, H. & Hirose, K. (1984). Analysis of voice fundamental frequency contours for declarative sentences of japanese. *Journal of Acoustic Society of Japan* **5(4)**, 233–242.

Fujisaki, H. & Ljungqvist, M. (1986). Proposal and evaluation of models for the glottal source waveform. In: *Proceedings of ICASSP-86.* 1605–1608.

Gibbon, D., Moore, R. & Winski, R. (1997). *Handbook of Standards and Resources for Spoken Language Systems.* Berlin, New York: Mouton de Gruyter.

Godjevac, S. (2001). *Serbo–Croatian ToBI* SC_ToBI. Retrieved from <http://www.ling.ohiostate.edu/t̃obi/>.

Grabowski, R. & Bauer, D. (2004). *System for Computer–Aided Perception Experiments SCAPE.* Retrieved April 20, 2005, from <http://www.coli.uni-saarland.de/˜doba/scape/>.

Granqvist, S. (2003). The visual sort and rate method for perceptual evaluation in listening tests. *Logopedics, phoniatrics, vocology* **28**(3), 109–116.

Grice, M. & Baumann, S. (2002). Deutsche Intonation und GToBI. *Linguistische Berichte* **191**, 267–298.

Grocholewski, S. (1999). CORPORA Speech Database for Polish Diphones. In: *Proceedings of Eurospeech 1997.* Rhodes, Greece, 1735–1738.

Grocholewski, S. (2001). Hidden Markov Models for Polish. In: *Proceedings of Prosody 2000: speech recognition and synthesis.* Poznań: Adam Mickiewicz University, 69–74.

Grønnum, N. (1992). *The Groundworks of Danish Intonation: An Introduction.* University of Copenhagen: Museum Tusculanum Press.

Gubrynowicz, R. (1998). The Polish Database of Spoken Language. In: *Proceedings of LREC 1998*, vol. 1. Granada, 1031–1037.

Gubrynowicz, R. (1999). Design and Implementation of Polish Speech Database under the BABEL Project. In: W. Jassem, C. Basztura, G. Demenko & K. Jassem (eds.) *Speech and Language Technology: Papers Reports and Technical Notes*, vol. 3. Poznań: UAM, 257–275.

Gussenhoven, C. (2004). Transcription of dutch intonation. In: J. Sun-Ah (ed.) *Prosodic Typology: The Phonology of Intonation and Phrasing.* Oxford: Oxford University Press, 118–145.

'h Hart, J. (1976). Psychoacoustic backgrounds of pitch contour stylisation. *I.P.O Annual Progress Report* **11**, 11–19.

Hirai, T., Iwahashi, N., Higuchi, N. & Sagisaka, Y. (1994). Automatic extraction of fo control parameters using statistical analysis. In: *Proceedings of Second ESCA/IEEE Workshop on Speech Synthesis SSW2-1994.* Mohonk Mountain House, New Paltz, NY, 57–60.

Hirose, K., Fujisaki, H. & Kawai, H. (1986). Generation of prosodic symbols for rule-synthesis of connected speech of japanese. In: *Proceedings of ICASSP86.* Tokyo, 2415–2418.

Hirschberg, J. (1993). Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence* **63**, 305–340.

Hirst, D.J. & Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix* **15**, 75–85.

Hirst, D. (1983). Structures and categories in prosodic representations. In: A. Cutler & R. Ladd (eds.) *Prosody: Models and Measurements.* Berlin: Springer, 93–109.

Hirst, D., Cristo, A.D. & Espesser, R. (2000). Levels of representation and levels of analysis for the description of intonation systems. In: M. Horne (ed.) *Prosody: Theory and Experiment. Studies presented to Gösta Bruce*, vol. 14. Dordrecht: Kluwer Academic, 51–87.

Honkela, T. (1997). *Self-organizing Maps in Natural Language Processing.* Ph.D. thesis, Helsinki University of Technology, Neural Networks Research Centre, Finland.

Hunt, M., Zwierynski, D. & Carr, R. (1989). Issues in high quality LPC analysis and synthesis. In: *Proceedings of Eurospeech 1989*, vol. 2. Paris, France, 348–351.

Imiołczyk, J., Nowak, I. & Demenko, G. (1993). A text-to-Speech System for Polish. In: *Proceedings of Eurospeech 1993*. Berlin, 889–892.

IPDS (1997). xassp (Advanced Speech Signal Processor under the X Window System) – User's Manual. In: *xassp (Advanced Speech Signal Processor under the X Window System) – User's Manual*, vol. 32. 31–115. Version 1.2.15.

ITU–T (1996). Methods for subjective determination of transmission quality. ITU–T Recommendation p.800.

Jassem, W. (1952). *Intonation of Conversational English (Educated Southern British)*. Wrocław: Wrocławskie Towarzystwo Naukowe.

Jassem, W. (1954). *Fonetyka języka angielskiego*. Warszawa: PWN.

Jassem, W. (1961). *Akcent Języka Polskiego*, *Prace Językoznawcze*, vol. 31. Kraków: PAN.

Jassem, W. (1987). Computer-based classification of basic polish intonations. In: *Proceedings of the 11th International Congress of Phonetic Sciences*. Tallinn, 253–256.

Jassem, W. (1996). A quantitative analysis of standard british-english nuclear tones. *Journal of Quantitative Linguistics* **3(3)**, 229–243.

Karpiński, M. & Kleśta, J. (2001). The project of an intonational database for the Polish language. In: *Proceedings of Prosody 2000: speech recognition and synthesis*. Poznań: Adam Mickiewicz University, 113–118.

Kaski, S. (1997). *Data Exploration Using Self-Organizing Maps*. Ph.D. thesis, Helsinki University of Technology. Department of Computer Science and Engineering.

Keller, E. & Keller, B.Z. (2003). How much prosody can you learn from twenty utterances? *Linguistik online* **17**, 57–79.

Keller, F., Corley, M., Corley, S., Konieczny, L. & Todirascu, A. (1998). *WebExp: A Java Toolbox for web-based psychological experiments*. Tech. Rep. HCRC/TR–99, University of Edinburgh, Human Communication Research Centre.

Keller, F. (1998). *Grammaticality Judgements and Linguistic Methodology.* Tech. Rep. EUCCS–RP–1998–3, University of Edinburgh, Human Communication Research Centre.

Keller, F. (1999). Book review: The empirical base of linguistics: Grammaticality judgments and linguistic methodology, Carson T. Schütze. *Journal of Logic, Language and Information* **8**(1), 114–121.

Kenstowicz, M. & Kisseberth, C. (1977). *Topics in Phonological Theory.* New York: Academic Press.

Kingdon, R. (1958). *The groundwork of English stress.* London: Longmans, Green and Co.

Klabbers, E. & van Santen, J.P.H. (2004). Clustering of foot-based pitch contours in expressive speech. In: *Fifth ISCA ITRW on Speech Synthesis (SSW5).* Pittsburgh, PA, USA, 73–78.

Kohler, K. (1987). Categorical pitch perception. In: *Proceedings of the XIth International Congress of Phonetic Sciences*, vol. 5. Tallinn, 331–333.

Kohler, K. (1988). Zeitstrukturierung in der Sprachsynthese. In: *ITG-Fachbericht*, vol. 105. 165–170.

Kohonen, T. (1995). *Self-Organizing Maps*, *Springer Series in Information Sciences*, vol. 30. Berlin, Heidelberg: Springer-Verlag.

Kula, S., Dymarski, P., Janicki, A., Jobin, C. & de Mareuil, P. (2001). Prosody control in diphone-based speech synthesis system for Polish. In: *Proceedings of Prosody 2000: speech recognition and synthesis.* Poznań: Adam Mickiewicz University, 135–142.

Kumpf, K. & King, R.W. (1996). Automatic accent classification of foreign accented australian english speech. In: *Proceedings of ICSLP 1996.* Philadelphia, 1740–1743.

Ladd, D.R. (1996). *Intonation Phonology.* Cambridge Studies in Linguistics 79. Cambridge: Cambridge University Press.

Ladd, D.R., Faulkner, D., Faulkner, H. & Schepman, A. (1999). Constant ”segmental anchoring” of F0 movements under changes in speech rate. *Journal of Acoustical Society of America* **106**, 1543–1554.

Ladd, D.R., Mennen, I. & Schepman, A. (2000). Phonological conditioning of peak alignment of rising pitch accents in Dutch. *Journal of Acoustical Society of America* **107**, 2685–2696.

Ladd, D.R. & Schepman, A. (2003). Sagging transitions between high accent peaks in English: experimental evidence. *Journal of Phonetics* **31**, 81–112.

Lodge, M. (1981). *Magnitude Scaling: Quantitative Measurement of Opinions.* Beverly Hills, CA: Sage Publications.

Manning, C.D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge, MA: MIT.

Mertens, P. (2004). The prosogram: Semi-automatic transcription of prosody based on a tonal perception model. In: *Proceedings of the 2nd International Conference on Speech Prosody.* Nara,Japan, 23–26.

Misheva, A. (1991). *Intonacionna sistema na bâlgarskija ezik.* Sofia: BAN.

Mixdorff, H. (2005). *Fujisaki Parameter Extraction Environment.* Retrieved January 20, 2005, from <http://www.tfh-berlin.de/˜mixdorff>.

Möbius, B. & van Santen, J.P.H. (1996). Modeling segmental duration in german text-to-speech synthesis. In: *Proceedings of ICSLP 96.* 2395–2398.

Möbius, B., Pätzold, M. & Hess, W. (1993). Analysis and synthesis of german f0 contours by means of fujisaki's model. *Speech Communication* **13**, 53–61.

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

O'Connor, J. & Arnold, G. (1973). *Intonation of Colloquial English.* London: Longman.

Oliver, D. (1998). *Polish text-to-speech Synthesis.* Master's thesis, University of Edinburgh.

Oliver, D. (2005a). Deriving pitch accent classes using automatic F0 stylisation and unsupervised clustering techniques. In: *Proceedings of Second Baltic Conference on Human Language Technologies.* Tallinn, Estonia, 161–166.

Oliver, D. (2005b). Perception of Intonation in Speech Synthesis: Subjective Evaluation of Polish Intonation model. In: *Archives of Acoustics*, vol. 30. Warsaw: Polish Academy of Sciences, 443–444.

Oliver, D. & Andreeva, B. (2008). Peak Alignment in Broad and Narrow Focus. A Cross–language Study. In: G. Zybatow, L. Szucsich, U. Junghanns & R. Meyer (eds.) *Formal Description of Slavic Languages: The Fifth Conference, Leipzig 2003*. Frankfurt am Main: Peter Lang, 134–148.

Oliver, D. & Clark, R. (2005). Modelling Pitch Accent Types for Polish Speech Synthesis. In: *Proceedings of Interspeech 2005*. Lisbon, 1965–1968.

Oliver, D. & Grice, M. (2003). The Phonetics and Phonology of Lexical Stress in Polish Verbs. In: *Proceedings of the 15th International Congress of Phonetic Science*. Barcelona, 2027–2030.

Oliver, D. & Szklanny, K. (2006). Creation and analysis of a polish speech database for use in unit selection synthesis. In: *Proceedings of The fifth International Conference on Language Resources and Evaluation LREC 2006*. Genoa, Italy.

Palmer, H.E. (1922). *English intonation with systematic exercises*. Cambridge: Heffer.

Patterson, D. (2000). *A Linguistic Approach to Pitch Range Modelling*. Ph.D. thesis, University of Edinburgh.

Pierrehumbert, J. & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In: P. Cohen, J. Morgan & M. Pollack (eds.) *Intentions in Communication*. Cambridge, MA,: MIT Press, 271–311.

Pierrehumbert, J.B. (1980). *The Phonology and Phonetics of English intonation*. Ph.D. thesis, MIT, Cambridge.

Pijper, J.D. (1979). Close-copy stylisation of British English intonation contour. *IPO Annual Progress Report* **14**, 66–71.

Pijper, J.D. (1983). *Modelling British English Intonation: An Analysis by Resynthesis of British English Intonation*. Netherlands Phonetic Archives 3. Dordrecht: Foris.

Prieto, P.J., van Santen, J.P.H. & Hirschberg, J. (1995). Tonal alignment patterns in Spanish. *Journal of Phonetics* **23**, 429–451.

Prieto, P. (2004). The search for phonological targets in the tonal space: Evidence from five sentence-types in peninsular spanish. In: T.L. Face (ed.) *Laboratory Approaches to Spanish Phonology*, no. 7 in Phonology and Phonetics. The Hague: Mouton de Gruyter, 29–59.

Rakotomalala, R. (2005). TANAGRA: a free software for research and academic purposes. *RNTI-E-3* **2**, 697–702.

Rietveld, T. & Gussenhoven, C. (1995). Aligning pitch targets in speech synthesis: effects of syllable structure. *Journal of Phonetics* **23**, 375–385.

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

Ripley, B. (1993). Statistical aspects of neural networks. In: O.E. Barndorff-Nielsen, J.L. Jensen & W.S. Kendall (eds.) *Networks and Chaos – Statistical and Probabilistic Aspects*, chap. Statistical aspects of neural networks. Department of Statistics, University of Oxford: Chapman and Hall., 40–123.

Rossi, M. (1971). Le seuil de glissando ou seuil de perception des variations tonales pour la parole. *Phonetica* **23**, 1–33.

Rossi, M. (1978). La perception des glissandos descendants dans les contours prosodiques. *Phonetica* **35**, 11–40.

Rubach, J. & Booij, G. (1990). Syllable structure assignment in Polish. *Phonology* **7(1)**, 121–158.

Schepman, A., Lickley, R. & Ladd, D. (2006). Effects of vowel length and right context on the alignment of Dutch nuclear accents. *Journal of Phonetics* **34**, 1–28.

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

Shih, C. & Ao, B. (1997). Duration study for the bell laboratories mandarin text-to-speech system. In: J. van Santen, R.W. Sproat, J. Olive &

J. Hirschberg (eds.) *Progress in Speech Synthesis*. New York: Springer-Verlag, 382–399.

Siemund, R., Höge, H., Kunzmann, S. & Marasek, K. (2000). SPEECON - Speech Data for Consumer Devices. In: *Proceedings of LREC 2000*, vol. II. Athens, 883–886.

Silverman, K., Beckman, M.E., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. & Hirschberg, J. (1992). ToBI: a standard for labelling English prosody. In: *Proceedings of ICSLP 1992*, vol. 2.2. 867–870.

Silverman, K. & Pierrehumbert, J. (1990). The timing of prenuclear high accents in english. In: J. Kingston & M. Beckman (eds.) *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*. Cambridge: Cambridge University Press, 72–106.

Simoes, A.R. (1990). Predicting sound segment duration in connected speech: An acoustical study of Brazilian Portuguese. In: *In ESCA Workshop on Speech Synthesis*. Autrans, 173–176.

Spencer, A. (ed.) (1991). *Morphological Theory*. Oxford, Cambridge: Blackwell.

Steele, J. (1986). Nuclear accent F0 peak location: effect of rate, vowel, and number of syllables. *Journal of the Acoustical Society of America* **80**(1), S51.

Steele, S.A. & Altom, M.J. (1986). *Nuclear Stress : Changes in Vowel Duration and F0 Peak Location*. Technical Memorandum 11227-860625-09, AT and T Bell Laboratories.

Steffen-Batog, M. (1996). *Struktura przebiegu melodii polskiego języka ogólnego*. Poznań: Sorus.

Stöber, K., Breuer, S., Wagner, P. & Abresch, J. (2000). Dokumentation zum Bonn Open Synthesis System (BOSS) II. Unveröffentlichtes Dokument.

Sun, X. (2001). Predicting underlying pitch targets for intonation modeling. In: *Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis*. Perthshire, Scotland, 143–148.

Sun, X. (2002a). F0 generation for speech synthesis using a multi–tier approach. In: *Proceedings of International Conference on Spoken Language Processing ICSLP 2002*. Denver, Colorado, 2077–2080.

Sun, X. (2002b). Pitch accent prediction using ensemble machine learning. In: *Proceedings of International Conference on Spoken Language Processing ICSLP 2002*. Denver, Colorado, 561—564.

Sun, X. & Applebaum, T.H. (2001). Intonational phrase break prediction using decision tree and n-gram model. In: *Proceedings of Eurospeech 2001*. 537–540.

Syrdal, A.K., Hirschberg, J., McGory, J.T. & Beckman, M.E. (2001). Automatic tobi prediction and alignment to speed manual labeling of prosody. *Speech Communication* **33**, 135–151.

Syrdal, A.K., Möhler, G., Dusterhoff, K., Conkie, A. & Black, A.W. (1998). Three methods of intonation modeling. In: *Proceedings of 3rd ESCA Workshop on Speech Synthesis*. 305–310.

’t Hart, J. & Collier, R. (1975). Integrating different levels of intonation analysis. *Journal of Phonetics* **3**, 235–255.

’t Hart, J., Collier, R. & Cohen, A. (1990). *A perceptual study of intonation: An experimental-phonetic approach to speech melody*. Cambridge studies in speech science and communication. Cambridge: Cambridge University Press.

Taylor, P. (1998). The Tilt Intonation Model. In: *Proceedings of ICSLP 1998*. Sydney, 1383–1386.

Taylor, P. (2000). Analysis and synthesis of intonation using the Tilt model. *Journal of the Acoustical Society of America* **107**(3), 1697–1714.

Traber, C. (1992). F0 generation with a data base of natural f0 patterns and with a neural. In: G. Bailly, C. Benoît & T. Sawallis (eds.) *Talking Machines: Theories, Models, and Designs*. North-Holland, 287–304.

van den Heuvel, H., Boudy, J., Bakcsi, Z., Cernocky, J., Galunov, V., Kochanina, J., Majewski, W., Pollak, P., Rusko, M., Sadowski, J., Staroniewicz, P. & Tropf, H.S. (1999). SpeechDat-E: Five Eastern European Speech Databases for Voice-Operated Teleservices Completed. In: *Proceedings of Eurospeech 2001*. 2059–2062.

van Santen, J.P.H. (1994). Assignment of segmental duration in text–to–speech synthesis. *Computer, Speech and Language* **8**(2), 95–128.

Vazquez-Alvarez, Y. & Huckvale, M. (2002). The Reliability of the ITU–P.85 Standard for the Evaluation of Text-to-Speech Systems. In: *Proceedings of ICSLP 2002*. 329–332.

Vegnaduzzo, M. (2003). *Modeling Intonation for the Italian Festival TTS using Linear Regression*. Master's thesis, University of Edinburgh.

Viswanathan, M. & Viswanathan, M. (2005). Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale. *Computer Speech and Language* **19**, 55–83.

Wang, M.Q. & Hirschberg, J. (1992). Automatic classification of intonational phrasing boundaries. *Computer Speech and Language* **6(2)**, 175–196.

Wells, J.C. (1997). Sampa computer readable phonetic alphabet. In: D. Gibbon, R. Moore & R. Winski (eds.) *Handbook of Standards and Resources for Spoken Language Systems*, chap. Part IV, section B. Berlin and New York: Mouton de Gruyter.

Werner, S. (2001). Acoustic Classification of Intonational Events. In: *Proceedings of Prosody in Speech Recognition and Understanding*. Red Bank, NJ, USA: ISCA.

Willems, N., Collier, R. & Pijper, J.D. (1988). A synthesis scheme for british english intonation. *Journal of the Acoustical Society of America* **1984**, 1250–1260.

Zervas, P., Fakotakis, N., Kokkinakis, G., Kouroupetroglou, G. & Xydas, G. (2004). Evaluation of corpus based tone prediction in mismatched environments for greek tts synthesis. In: *Proceedings of Interspeech 2004*. 761–764.

Issues of PHONUS published between 1995 and 2008

(electronic versions as pdf files are available online under
http://www.coli.uni-saarland.de => Phonetics => Phonus)

PHONUS 1, 1995, edited by W.J. Barry & J. Koreman
Sammelband

PHONUS 2, 1996, edited by W.J. Barry & A. Addison
Proceedings Workshop on Pronunciation Training at German Universities, Colleges of Education and Polytechnics

PHONUS 3, 1997, edited by W.J. Barry & J. Koreman
Sammelband/Volume

PHONUS 4, 1999, edited by W.J. Barry & J. Koreman
Sammelband/Volume

PHONUS 5, 2000, edited by W.J. Barry & J. Koreman, with K. Kirchhoff
Proceedings Workshop on Phonetics and Phonology in ASR

PHONUS 6, 2002, edited by W.J. Barry & M. Pützer
Festschrift für Max Mangold zum 80. Geburtstag

PHONUS 7, 2004, edited by W.J. Barry
PhD dissertation
Marc Schröder: *Speech and Emotion Research: An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis*.

PHONUS 8, 2004, edited by W.J. Barry
PhD dissertation
Jürgen Trouvain: *Tempo Variation in Speech Production. Implications for Speech Synthesis.*

PHONUS 9, 2005, edited by W.J. Barry & J. Trouvain
Magisterarbeiten
Köser, Stephanie: *['zOl?n] oder 'sollten'? Zur Glottalisierung alveolarer Plosive im Deutschen.*

Jarmut, Silke: *Instrumentelle Analyse von Sprechern mit Rekurrensparese - Vier longitudinale Fallstudien.*
Klein, Cordula: *Acoustic and Perceptual Gender Characteristics in the Voices of Pre-adolescent Children.*

PHONUS 10, 2006, edited by W.J. Barry & J. Trouvain
Masters thesis/Diploma thesis
Caren Brinckmann: *Improving Prosody Prediction for Speech Synthesis - With and Without Symbolic Prosody Features*.

PHONUS 11, 2007, edited by W.J. Barry & J. Trouvain
Monographie
Roland Marti: *ó w dolnoserbšćinje* (ó in Lower Sorbian / ó im Niedersorbischen).

PHONUS 12, 2008, edited by W.J. Barry
Dissertation
Bistra Andreeva: *Zur Phonetik und Phonologie der Intonation der Sofioter Varietät des Bulgarischen*.

PHONUS 13, 2008, edited by W.J. Barry
Habilitationsschrift
Manfred Pützer: *Stimmqualität und Artikulation bei Dysarthrophonien.*

PHONUS 14, 2008, edited by W.J. Barry & J. Trouvain
PhD dissertation
Dominika Oliver: *Modelling Polish Intonation for Speech Synthesis.*