University of West Bohemia in Pilsen
Faculty of Applied Sciences

AUTOMATIC SEGMENTATION OF SPEECH INTO SENTENCE-LIKE UNITS

Ing. Jáchym Kolář

A dissertation submitted for the degree of

Doctor of Philosophy

in *Cybernetics*

Major Advisor:	Prof. Ing. Josef Psutka, CSc.
Department:	Department of Cybernetics

Pilsen, 2008

Západočeská univerzita v Plzni
Fakulta aplikovaných věd

AUTOMATICKÁ SEGMENTACE MLUVENÉ ŘEČI DO VĚTNÝCH JEDNOTEK

Ing. Jáchym Kolář

disertační práce
k získání akademického titulu doktor
v oboru *Kybernetika*

Školitel: Prof. Ing. Josef Psutka, CSc.
Katedra: Katedra kybernetiky

Plzeň, 2008

Declaration of Originality

I hereby declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Pilsen, August 7, 2008

Ing. Jáchym Kolář

Dedicated to my parents and all my true friends.

Acknowledgments

First of all, I would like to thank my advisor, Prof. Josef Psutka, for providing me with the opportunity to do this interesting research and for academic guidance throughout my postgraduate study at the Department of Cybernetics, University of West Bohemia in Pilsen.

I also gratefully acknowledge Elizabeth Shriberg for giving me excellent research guidance during my internship at the International Computer Science Institute in Berkeley. She has been a great source of scientific inspiration. In particular, my work on speaker-specific modeling for sentence segmentation of speech presented in Chapter 9 was largely inspired by her ideas. Furthermore, I thank her for useful comments on drafts of Chapters 2 and 3.

I am much obliged to Yang Liu at University of Texas at Dallas for always being willing to answer all the questions in my numerous emails. She has been my counselor in many areas ranging from scientific writing to machine learning and natural language processing techniques. She also provided me with her models for the TnT tagger, which I used for part-of-speech tagging of the ICSI meeting corpus. In addition, I thank her for valuable comments on a draft of this thesis.

Several people deserve acknowledgment for helping me create the Czech MDE corpora: Stephanie Strassel and Christopher Walker from the Linguistic Data Consortium helped me understand all the subtle details of the MDE annotation, Dagmar Kozlíková gave me valuable advice about some detailed features of Czech syntax, and Jan Švec implemented the annotation tool.

Jan Švec also helped me with the implementation of prosodic feature extraction. Furthermore, Aleš Pražák and Josef Psutka Jr. converted my ASR models into the UWB speech recognizer, and Drahomíra "johanka" Spoustová at Charles University in Prague helped me with automatic morphological tagging of Czech speech transcripts.

In addition, I acknowledge all my colleagues at University of West Bohemia and ICSI who have somehow contributed to my research – by giving me useful advice, showing me how to use various software toolkits, or just by being kind and helpful colleagues and neighbors. Last but not least, I thank my family and friends for their support (not only) during my work on this thesis.

The research presented in this thesis was financially supported by the Ministry of Education of the Czech Republic (projects No. 2C06020, LC536, and ME909), and by the European Union 6th FWP ISR Integrated Project AMI (FP6-506811). The access to the METACentrum computing facilities provided under the research program MSM6383917201 was also appreciated. The views expressed in the thesis are those of the author, and not the funding agencies.

Contents

Abstract

Current automatic speech recognition (ASR) systems are able to transcribe large volumes of speech data with reasonable accuracy. However, automatic transcripts produced by these systems often do not have a form convenient for subsequent processing. The problem is that standard speech recognizers output only a raw stream of words, leaving out important structural information such as locations of sentence or dialog act boundaries. Natural language processing techniques used in downstream processing (e.g., text summarization, information extraction, machine translation) are typically trained on well-formatted text, and fail on unstructured streams of words. This thesis deals with the problem of automatic segmentation of speech recognition output into sentence-like units. The work is focused on two languages – English and Czech.

Since no Czech speech corpora with appropriate annotation of sentence-like units had been available, they had to be prepared as part of this thesis. I describe creation of two Czech speech corpora with structural metadata annotation in two different domains: broadcast news (mostly read-aloud speech) and broadcast conversations (spontaneous speech). The employed annotation scheme creates boundaries between natural breakpoints in the flow of speech, flags non-content words for optional removal, and identifies sections of disfluent speech. Then I present a detailed analysis of the annotated corpora in terms of structural metadata statistics.

The main goal of this work is to develop automatic systems for dialog act segmentation of English multiparty meetings and sentence unit segmentation of the two new Czech corpora. I use and compare three modeling approaches – hidden Markov models, maximum entropy, and a boosting-based algorithm called BoosTexter. All of these approaches rely on two information sources – recognized words (what was said) and prosody (how it was said).

Features extracted from the recognized words describe lexical patterns associated with sentence-external and sentence-internal interword boundaries. I explore features capturing word identities, parts of speech, and automatically induced word classes. Prosodic features are used to reflect breaks in temporal, intonational, and loudness contours in an utterance. In the approaches I employ in this thesis, prosodic features for automatic classification are extracted directly from the speech signal based on ASR time alignments, without any need for hand-labeling of prosodic events.

All methods are evaluated on two types of speech transcripts – manual transcripts (reference conditions) and automatically-generated transcripts (ASR conditions). The results indicate that superior performance is achieved when the three statistical models are combined via posterior probability interpolation. Furthermore, feature analysis reveals that English and Czech slightly differ in overall feature usage patterns. In addition to experiments in a standard speaker-independent fashion, I also explore speaker-dependent modeling for the English multiparty meeting domain. The experimental results show that speaker adaptation of both prosodic and language models yields modest yet significant improvement for sentence-like unit segmentation.

Abstrakt

Systémy automatického rozpoznávání řeči (automatic speech recognition, ASR) jsou již dnes schopny přepisovat velké objemy řečových dat s uspokojivou přesností. Problémem ale je, že tyto automatické přepisy nemají formu vhodnou pro následné zpracování. Standardní ASR systémy generují pouze nestrukturovaný proud slov, který neobsahuje žádnou interpunkci. Tento fakt nejenom snižuje čitelnost těchto přepisů, ale také způsobuje problémy při jejich následném automatickém zpracování. Systémy zpracování přirozeného jazyka (např. automatická sumarizace textu, automatická extrakce informací, automatický překlad) jsou obvykle natrénovány na formátovaném textu, a proto na automatických přepisech nezřídka selhávají. Tato disertační práce se zabývá problémem automatické segmentace těchto přepisů do větných jednotek. Zaměřuje se na dva jazyky – angličtinu a češtinu.

Jelikož nebyly k dispozici žádné české řečové korpusy s vhodnou anotací hranic větných jednotek, musely být připraveny v rámci této práce. Byly vytvořeny dva české korpusy s anotací tzv. strukturálních metadat. Tyto korpusy obsahují řeč ze dvou rozdílných oblastí – rozhlasových a televizních zpráv (převážně čtená řeč) a diskusních pořadů (převážně spontánní řeč). Použité anotační schéma v řeči definuje hranice syntakticko-sémantických jednotek, označuje výplňová slova a vymezuje oblasti neplynulé řeči. V práci dále porovnávám četnosti výskytu jednotlivých druhů strukturálních událostí v obou korpusech.

Hlavním cílem práce je vytvořit systémy pro automatickou segmentaci řeči do větných jednotek pro tři různé korpusy – anglický korpus pracovních schůzek a dva výše zmíněné české korpusy. V práci používám a srovnávám tři statistické modely – skrytý Markovův model, model maximální entropie a boostingový model BoosTexter. Všechny tři modely využívají dva základní zdroje informací – rozpoznaná slova (co bylo řečeno) a prozódii (jak to bylo řečeno).

Klasifikační příznaky získané z rozpoznaných slov popisují mimo slovních tvarů také jejich morfologické značky. Dále využívám i metody automatického shlukování slov do tříd. Prozodické příznaky zachycují nespojitosti v melodických, hlasitostních a temporálních konturách řeči. Všechny prozodické příznaky jsou získány přímo z řečového signálu na základě časových značek získaných ze systému ASR. Pro trénování systému tedy nejsou třeba data s ruční anotací prozodických jevů.

Všechny použité metody jsou vyhodnoceny na dvou druzích řečových přepisů – ručních a automatických. Výsledky ukazují, že největší přesnosti segmentace je dosaženo, když jsou všechny tři zkoumané statistické modely zkombinovány pomocí interpolace aposteriorních pravděpodobností. Analýza použitých příznaků dále odhaluje, že angličtina a čeština se mírně liší v tom, jaké příznaky jsou pro jejich segmentaci nejdůležitější. Mimo standardních experimentů nezávislých na řečníkovi zkoumám také možnosti adaptace modelů na konkrétního řečníka, která je zajímavá pro oblast pracovních schůzek, ve kterých se obvykle řečníci opakují. Experimentální výsledky signalizují, že adaptace jazykových i prozodických modelů přináší malé, ale statisticky významné zlepšení celkové přesnosti segmentace do větných jednotek.

List of Figures

List of Tables

Glossary of Abbreviations and Acronyms

A/P	......	Aside/Parenthetical
AIC	......	Automatically Induced Classes
ASR	......	Automatic Speech Recognition
AuxWords	......	Auxiliary Words
BER	......	Boundary Error Rate
BL	......	Baseline
BN	......	Broadcast News
CART	......	Classification and Regression Tree
CRF	......	Conditional Random Field
CTS	......	Conversational Telephone Speech
DA	......	Dialog Act
DelReg	......	Deletable Region
DM	......	Discourse Marker
DR	......	Discourse Response
EET	......	Explicit Editing Term
EM	......	Expectation-Maximization
EARS	......	Effective, Affordable, Reusable Speech-to-Text
Eq.	......	Equation
F	......	F-measure
F_0	......	Fundamental Frequency
Fig.	......	Figure
FP	......	Filled Pause
GALE	......	Global Autonomous Language Exploitation
GMM	......	Gaussian Mixture Model
HELM	......	Hidden Event Language Model
HMM	......	Hidden Markov Model
IAA	......	Inter-Annotator Agreement
ICSI	......	International Computer Science Institute
IP	......	Interruption Point

L-BFGS	Limited Memory Broyden-Fletcher-Goldfarb-Shanno Method
LDC	Linguistic Data Consortium
LM	Language Model
LTM	Lognormal Tied Mixture
LVCSR	Large Vocabulary Continuous Speech Recognition
MaxEnt	Maximum Entropy
MDE	Metadata Extraction
MLP	Multi-Layer Perceptron
MRDA	Meeting Recorder Dialog Act
MT	Machine Translation
NCCF	Normalized Cross Correlation Function
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
P	Precision
PC	Personal Computer
PDT	Prague Dependency Treebank
PLP	Perceptual Linear Prediction
POS	Part of Speech
POSmix	Parts of Speech mixed with words
PWL	Piece-Wise Linear
R	Recall
RAPT	Robust Algorithm for Pitch Tracking
REF	Reference Conditions
RF	Radioforum Corpus
RMS	Root Mean Square
RT	Rich Transcription
SD	Speaker-Dependent
SDA	Sequential Dependency Analysis
SI	Speaker-Independent
STT	Speech To Text
SU	Sentence/Syntactic/Semantic/Slash Unit
SVM	Support Vector Machine
ToBI	Tone and Break Indices
TTS	Text To Speech
TV	Television
UWB	University of West Bohemia
WER	Word Error Rate
WHG	Word Hypotheses Graph

Chapter 1

Introduction

> *For me, the big chore is always the same:*
> *how to begin a sentence, how to continue it,*
> *how to complete it.*
>
> Claude Simon

1.1 Motivation

Recent years have witnessed significant progress in the area of automatic speech recognition (ASR). Nowadays, large volumes of audio data can be transcribed automatically with reasonable accuracy. However, automatic transcripts often do not have a form convenient for subsequent processing. The problem is that standard ASR systems output only a raw stream of words, leaving out important structural information such as locations of sentence or dialog act boundaries. Such locations are overt in standard text via punctuation and capitalization, but "hidden" in speech.

As shown by a number of studies, the absence of sentence boundaries is confusing both for humans and computers. For example, Jones et al. [1] demonstrated that sentence breaks are critical for legibility of speech transcripts. Moreover, missing sentence segmentation makes meaning of some utterances ambiguous. If an automatic speech recognizer outputs the stream of words *"no jobs are running"*, it is not clear what was said – whether it is *"No jobs are running."* or *"No. Jobs are running."* The two possible interpretations have completely opposite meaning.

Likewise, missing sentence boundaries cause significant problems to automatic downstream processes. Many natural language processing (NLP) techniques (e.g., parsing, automatic summarization, information extraction and retrieval, machine translation) are typically trained on well-formatted input, such as text, and fail when dealing with unstructured streams of words. For instance, Kahn et al. [2] achieved a significant error reduction in parsing performance by using an automatic sentence boundary detection system, Furui et al. [3] reported that speech summarization improved when sentence boundaries were provided, and Matusov et al. [4] showed that the use of sentence boundaries is beneficial for machine translation. Thus, automatic segmentation of speech into sentence-like units is, without a doubt, a very important task that is essential for linking automatic speech recognition and downstream NLP processes.

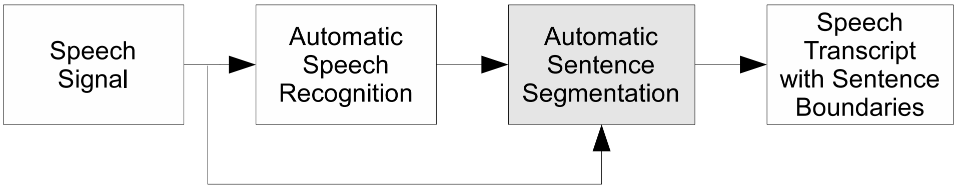

Speech Signal
Automatic Speech Recognition
Automatic Sentence Segmentation
Speech Transcript with Sentence Boundaries

Figure 1.1: Diagram of the sentence segmentation task

1.2 General Task and Approach

The previous section outlined the motivation for this thesis. In this section, I take a closer look at the general task of this work. First of all, note that the sentence segmentation task should not be confused with cepstrum-based signal segmentation tasks such as automatic speech/non-speech detection. Automatic speech/non-speech detectors are only able to split the speech signal into segments based on a non-speech region duration threshold. This kind of speech segmentation is quite easy to obtain, however, it does not meet the requirements on syntactic and semantic completeness imposed by the downstream NLP processes. Thus, the goal of this work is to segment speech into *linguistic* (defined as syntactically and semantically coherent) units, instead of acoustic units (defined as bounded by silence on either end).

When addressing the sentence segmentation task, it must be taken into account that speech (especially spontaneous) is not as clearly structured as written text. In conversational speech, "sentence" is not a straightforward notion because spontaneous utterances do not consist of sentences as we know them from prose. The linguistic units into which I aim to segment speech can generally be referred to as *sentence-like units*. Hence, the official title of this thesis is *Automatic Segmentation of Speech into Sentence-like Units*. In the following text, I sometimes also use a shorter term "sentence segmentation" in place of "segmentation into sentence-like units" for the sake of brevity. Exact definitions of sentence-like units slightly differ for individual data sets used in this thesis. They are always precisely defined in corresponding chapters.

A diagram of the automatic sentence segmentation task is shown in Fig. 1.1. It is assumed that a speech signal and an ASR output containing recognized words with corresponding timestamps are available to be used to solve the task. The goal is to create an enriched speech transcript containing automatically inserted sentence unit boundaries. Therefore, the problem is addressed as a post-processing step that generates the structural information after the recognition results are produced. The segmentation task can be viewed as a two-way classification problem, in which each inter-word boundary has to be labeled as either a within-unit boundary, or a boundary between two units.

There are two basic sources of information that can be used to solve this task — *recognized words* (what was said) and *prosody* (how it was said). By prosody we mean information about temporal, pitch, and energy characteristics of utterances. This information is basically independent of word identities. To detect possible sentence boundaries in the recognized word stream, we utilize prosodic features extracted from the speech signal, and combine them with textual cues obtained from the word string. The question is how to model and combine the available knowledge sources to find the most accurate hypotheses.

The first issue is how to exploit the information contained in the word sequence. It is obvious that, in some word contexts, sentence boundaries are more probable than in others. For example, the English pronoun *I* is a strong indicator of the beginning of a new sentence. This subtask is referred to as *language* or *lexical modeling*. On the other hand, *prosodic modeling* aims to find acoustic-prosodic contexts indicating sentence boundaries. Such examples include long pauses or marked pitch falls. The final issue is how to effectively combine prosodic and language models. It is not only possible to combine two independent models, but we can also consider joint lexical and prosodic modeling.

1.3 Scope of the Thesis

The thesis deals with sentence segmentation of speech in two languages – Czech and English. For English, there has been quite a lot of material to build upon, while the starting point for Czech was completely different. At the time I started my research, no related previous work in Czech had been published and no corpora with appropriate annotation were available. Hence, I had to start everything from scratch. My work on Czech not only included building an automatic sentence segmentation system, but also a design and creation of Czech speech corpora with appropriate annotation.

I decided to create two corpora corresponding to two distinct speaking styles: a read speech (broadcast news) corpus and a spontaneous speech (broadcast conversation) corpus. The first corpus was created by enriching an existing Czech broadcast news corpus, whereas the preparation of the latter started from scratch. The annotation scheme I used went far beyond labeling of sentence-like unit boundaries – speech disfluencies, filler words, and some other spontaneous speech phenomena were also annotated. Although automatic detection of the additionally annotated phenomena exceeds the planned scope of this work, I wanted to prepare corpora useful for studying a broad spectrum of spontaneous speech phenomena. After the corpora had been prepared, I used them to create a sentence segmentation system for Czech and performed a number of various experiments.

For English, the scope of the work was a bit different. I focused on the meeting domain, which is an area of growing interest in the spoken language technology community. For all experiments, I used the publicly available ICSI meeting corpus. In addition to experiments in a standard speaker-independent fashion, I also explored speaker-dependent modeling for the multiparty meeting domain.[1]

All sentence segmentation models used in this thesis are statistical, i.e. trained from data using machine learning methods. The segmentation systems for Czech and English are trained on different data, however, they share a common modeling groundwork. For example, they employ the same prosodic feature extraction methods and combine the same statistical modeling techniques. Three different statistical models have been explored – a hidden Markov model, a maximum entropy model, and a boosting-based model called Boostexter.

1.4 Thesis Organization

The remainder of this thesis is organized as follows. The text consists of eleven chapters and three appendices. Chapter 2 summarizes important linguistic topics – differences between structure of spoken and written language, prosody, and spontaneous speech. Chapter 3 surveys literature about perception of prosodic boundaries and automatic sentence segmentation of speech. Chapter 4 explicitly lists goals of this thesis.

After four introductory chapters, Chapter 5 presents design, creation, and analysis of Czech corpora with structural metadata annotation. Chapter 6 describes implementation of prosodic features used for automatic classification. Chapter 7 overviews three statistical

[1] Note that an explicit list of this thesis objectives is presented in a separate chapter (Chapter 4), as required by the dissertation guidelines of our department.

models used in this work – hidden Markov model, maximum entropy, and BoosTexter. Chapter 8 reports automatic dialog act segmentation of English multiparty meetings. Chapter 9 investigates speaker-specific modeling for dialog act segmentation of meetings. Chapter 10 describes experiments with sentence unit segmentation of Czech data. Finally, Chapter 11 draws conclusions.

The main part of the thesis is followed by an appendix part. Appendix A introduces the software tool used for metadata annotation of Czech corpora. Appendix B shows a complete list of all implemented prosodic features. Appendix C presents examples of automatically sentence segmented speech transcripts. At the end of the text, there is a complete list of references and a list of my publications.

Chapter 2

Linguistic Background

Thought is the fountain of speech.
Chrysippus

The pen is the tongue of the mind.
Miguel de Cervantes

This chapter briefly presents linguistic background important for this thesis. The chapter describes differences between structure of spoken and written language, suprasegmental (prosodic) features of speech, and particularities of spontaneous speech. Although these topics may seem to be disparate, all pertain to the sentence segmentation task. Since this work aims to automatically segment speech into units that have primarily been defined for text, it is important to note differences between linguistic units into which we can decompose spoken utterances and text. Speech prosody is overviewed here since I use prosodic features as important cues for automatic sentence boundary detection. Spontaneous speech is described because two of the three corpora used in this thesis contain spontaneous conversations and it is important to know which phenomena we should expect to observe.

The chapter is organized as follows. Section 2.1 compares linguistic structural units in spoken and written language. Section 2.2 describes prosody and prosodic terminology. Note that this section focuses mainly on a qualitative description of prosody. A description of prosodic features as used for automatic classification is presented in Chapter 6. Section 2.3 is devoted to the particularities of spontaneous speech. Finally, Section 2.4 gives a summary of the whole chapter.

2.1 Structural Units in Spoken and Written Language

Language has two basic forms: *spoken* (speech) and *written* (text). In this section, I briefly compare linguistic unit hierarchies of the two forms of language. Because the main goal of this thesis is to automatically segment speech into units primarily defined for written language, it is important to be aware of the differences between spoken and written language units.

In terms of a formal description, both forms of natural language can be viewed as hierarchical systems of symbols. It is possible to represent them as a set of mutually separated units and a set of their relations. Elements of higher orders are composed of lower order elements and boundaries of the higher order segments are typically consistent with the lower order boundaries. Given the different natures of speech and text, these hierarchies are not identical. The written form of language is a system of discrete symbols. In most languages, strings of these symbols are interrupted at word boundaries. On the other hand, speech is a continuous acoustic signal. For the recipient, it is much more difficult to divide it into meaningful linguistic segments since the majority of their boundaries is not explicitly marked in the speech signal [5].

Spoken language is directly dependent on speech production physiology. Thus, its hierarchy depends on the way an utterance is produced. The smallest segmental unit of speech is a *phoneme*, however, the smallest stretch of speech into which a speaker is able to divide his/her utterance is a *syllable*. Syllables are often considered the phonological "building blocks" of words. The general structure of a syllable consists of *onset*, *nucleus*, and *coda*. The nucleus (sometimes also called peak) is the central part of the syllable and typically consists of a vowel. The onset is the syllable part preceding the nucleus, whereas the coda comprises all consonant sounds that follow the nucleus. The final part of each syllable (nucleus plus coda) is called a *rhyme*.

The smallest lingual unit that carries a semantic interpretation is a *morpheme*. Examples of morphemes are prefixes, suffixes, or roots. They are defined in both spoken and written language. It is not possible to specify whether morphemes are on the higher or lower level than syllables in the spoken language unit hierarchy. Their boundaries may overlap since morphemes can be smaller than a syllable as well as span over several syllables; phonologic and semantic boundaries apparently do not have to coincide.

In continuous speech, syllables tend to form rhythmical groups of similar lengths. Such groups are typically indicated by *stress*, the relative emphasis given to certain syllables. In the linguistic literature, these groups are referred to as *feet*, *stress-groups*, *prosodic words*, or *phonemic words*. Several feet fluently pronounced one after another form a larger intonational unit – *intonational phrase*. Speakers indicate relationships of different intensities between the intonational phrases to express their mutual relation.

A compact group of intonational phrases compose an *utterance unit* (or *sentence-like unit*). In spoken language as well as in text, it is also possible to define units larger than a sentence. For instance, a portion of speech in a dialog uttered by a single speaker while he/she holds the floor is defined as a *turn*. Likewise, a stretch of speech relating to a single topic may in a longer monolog may, in analogy with text, be marked as a *paragraph*. It has been shown that boundaries between topics are signalized in speech [6, 7, 8, 9]. An *utterance* is the unit on the highest level of the hierarchy.

Written language shows a different hierarchy. The smallest segmental unit is a *grapheme* (character), however, analogous to spoken language, the smallest unit that carries semantic meaning is a *morpheme*. While in spoken language, morphemes are composed of phones, in written language, they are composed of graphemes. By joining morphemes we get *words*, words constitute *clauses* and *sentences*, sentences form *paragraphs*, and a sequence of paragraphs is *text*.

The alignment between the two language unit hierarchies may be outlined as follows. On the segment level, there are phonemes vs. graphemes. The degree of agreement between them differs language to language. For example, in Czech, there is a tight agreement given the phonetically-based spelling rules, whereas for English, which does not have exact letter-to-sound rules, the correspondence is not that good. On further levels, there is a relationship between feet and words. For instance in Czech, feet are often identical with written words, but also may comprise several words (e.g., a preposition plus a noun). Intonational phrases typically have an equal or shorter length than clauses, however, in some special cases they may span over two clauses. The average length of an intonational phrase in Czech is between two and three feet. The relationship of written and spoken "paragraphs" has already been mentioned above, text and utterance are well-corresponding units on the highest level of the hierarchy.

For this work, a critical part of the comparison is the alignment of written sentences and spoken sentence-like units. While a sentence is a clearly defined notion for text, the notion “sentence-like unit” does not have a standardized definition in the literature. In particular, it is often difficult to distinguish between boundaries of compound clauses within a sentence and boundaries between two (compound) “sentences” in speech. I use two distinct definitions of sentence-like units in this thesis. For English meeting data, these units are defined as dialog acts (DAs), while for Czech broadcast corpora, they are defined as syntactic-semantic units (SUs) according to the MDE standard (detailed definition is presented in Chapter 5). Although the two definitions differ slightly (in particular, some types of grammatically subordinate clauses may form a complete DA, but not a complete SU), both of them rely on the idea of dividing speech into meaningful units having a minimal possible length. Thus, we can say that these units are typically shorter (or at most equally long) than sentences in standard text.

2.2 Prosody

The description of prosody is useful for this work because prosodic cues are important indicators of sentence-like unit boundaries in speech, and may be employed in automatic sentence segmentation systems. In this section, I describe the most important prosodic phenomena and also overview standard prosodic terminology.

2.2.1 Suprasegmental Information in Speech

A speaker’s utterance contain more information than just identities of the uttered words. It is not only important to know *what* was said, but also *how* it was said. This additional information which may not be encoded by grammar is referred to as *prosody.* A typical feature of prosody is that it is linked to stretches of speech larger than a segment (phoneme), i.e. syllables, words, phrases, or entire utterances. That is why we also refer to prosodic features as *suprasegmental* features of speech. Among others, prosody may convey to the listener the following information:

- Whether an utterance is a question, a statement, or an imperative;
- Speaker’s state of mind;
- Whether the speaker is being sarcastic or ironic;
- Which word(s) in the message carry the most important information;
- Syntactic structure of an ambiguous utterance;
- Locations of word, phrase, and sentence boundaries.

The term *prosody* is derived from the Greek word $\pi\rho o\sigma\omega\delta\iota\alpha$, which is a musical term meaning something like “song sung to music” or “sung accompaniment”. In more recent usage, the term has come to be used to refer to rhythmical structure of verse in poetry. Linguistics has taken over the term prosody to refer to “musical” features of actual language utterances [10].

In the area of automatic speech processing, prosody has largely been used in Text-to-Speech (TTS) synthesis. Since monotonic speech sounds highly unnatural, generating prosodically rich speech is critical for success of TTS systems. In recent years, speech scientists have also started to more often use prosody in speech recognition and understanding applications. The tasks for which use of prosodic features has been studied include linguistic segmentation of speech (focus of this thesis), fast endpointing in spoken dialog systems, emotion recognition, or speech recognition itself.

2.2.2 Levels of Prosody Description

Suprasegmental features perceived by the listener include *length*, *pitch*, *loudness*, *rhythm*, and *speaking rate*. It is not possible to measure these human-perceived features from a speech signal directly; we can only measure their correlates. For instance, loudness is correlated with short term energy of the speech signal and pitch is correlated with fundamental frequency (F_0) of the signal [11]. Thus, we should distinguish the *acoustic* (measurable) and *psychoacoustic* (perceived) level of prosody.

Prosody also has a *linguistic* description. In this description, we only take into account those features that are "linguistically significant". These are only those that represent speaker's linguistic intention and can be recorded as a sequence of symbols [12]. Such features are typically expressed as a combination of several acoustic features. These combinations are not exactly determined since the same linguistic intention may be expressed using different means – less expressive use of one acoustic feature may be compensated by stronger use of another.

Prosody is also influenced by *paralinguistic* events. These phenomena are defined as all nonlinguistic and nonverbal elements of communication, regardless whether they are perceptible from voice (voice quality, voice color) or not (gestures, facial expressions). They express an emotional and visceral state of the speaker by affecting realization of the verbal message.

2.2.3 Prosodic Terminology

The following subsections overview the most important terms of prosodic terminology. The terminology is not standardized in the literature. A number of terms have varying definitions and some terms (such as *intonation* or *melody*) are used to refer to more than one phenomenon. Thus, I find important to clarify here in which meaning the prosodic terms are used in this thesis.

2.2.3.1 Stress, Accent, Rhythm

In phonology, stress is a relative emphasis given to certain syllables in a word. We often distinguish two terms: *stress* and *accent* (or *pitch-accent*). The term stress usually denotes a "potential feature", while the term accent stands for a realized emphasis [5, 13]. However, some other authors use these two terms as synonyms.

The accent is a complex event comprising intensity, pitch, and duration. An accented syllable is typically not recognized based on absolute values of acoustic-prosodic quantities, but rather based on contrast with syllables in its vicinity. In other words, the accented syllable may show, for instance, either a lower or higher pitch than the surrounding syllables. We can also observe that phonemes in accented syllables are usually more carefully articulated [14].

A placement of stress within a foot may either be fixed to a given syllable (Czech, Finnish – stress always on the first syllable, Polish – always on the penult syllable, etc.), or comes as part of the word and must be memorized (so called *lexical* stress – English or Italian). Beside the primary stress, longer feet may also show secondary (less prominent) stress. Overall distribution of stressed syllables within an utterance sets the *rhythm* (or *timing*) of speech.

The timing is determined by segmentation of the continuous stream of syllables into groups having similar length and acoustic characteristics. There are two basic ways in which languages distribute syllables across time: *stress timing* and *syllable timing*. In syllable-timed languages, every syllable is perceived as taking up roughly the same amount of time (e.g., Czech or French). By contrast, syllables in stress-timed languages may have different duration, but there is a roughly constant time period between each pair of adjacent stressed syllables (e.g., English or German).

2.2.3.2 Focus

Besides word accent, we also recognize semantic accent – *focus* (also *focal accent*). This accent is not linked with words but with sentences or clauses. It does not indicate prominence of a syllable but prominence of a word. The emphasized word usually shows a stronger stressed syllable with higher or lower pitch, or eventually it is longer or shorter. In Czech, it is usual that focus is on the last foot of the sentence. By changing its placement, the speaker may modify meaning of the utterance as it is outlined in the following example.

A better result was achieved by Czech ***women.***	(in contrast to Czech *men*)
A better result was achieved by ***Czech*** *women.*	(in contrast to *Slovak* women)

2.2.3.3 Melody, Intonation, Intonation Patterns

The term *melody* of speech refers to a pitch curve within an utterance. The term *intonation* is also often used as a synonym of melody. One way how to formally describe a melody of an utterance is to disarticulate its pitch curve into abstract melodic patterns – *intonation patterns* (cadenza). Linguists particularly focus on intonation patterns of sentence-final intonation phrases. The following patterns are usually distinguished: falling, rising, flat, falling-rising, and rising-falling. This classification system is phonemic – it is meant to express contrasts in a succinct manner rather than specify the realization. More information on the concept of intonation patterns is given, among others, in [5].

2.2.3.4 Pitch Declination, Pitch Reset

When analyzing the pitch contour within a single utterance unit, in general we observe a gradual pitch lowering. In other words, running averages of F_0 are higher on the beginning of an utterance unit than on the end. This feature of speech is usually denoted as a *pitch declination.* There are many ways how to quantitatively describe the declination. Most frequently it is displayed as a line connecting local pitch peaks (*topline*) or local pitch minima (baseline). Alternatively it can be captured using linear regression from all F_0 values [15]. The declination trend does not necessarily have to be linear, rarely we also observe an exponential declination [16].

So-called *pitch reset* is often observed at utterance unit (sentence, paragraph) boundaries. The speaker returns to higher F_0 values and the declination slope is usually changed. It was shown that stronger pitch resets usually correspond with more significant boundaries. For instance, there typically occur stronger pitch resets at paragraph boundaries than at sentence boundaries [8, 17, 18]. An important finding is that a breath is not necessary for pitch resets, and vice versa, a breath does not imply a pitch reset [19].

An issue is whether the pitch declination is phonetically-motivated (caused by a decrease of air pressure in the lungs and a consequent decrease of sub-glottal pressure) or phonologically-motivated (intentionally controlled by the speaker). Since several experiments showed that shorter sentences have a steeper slope of declination, it is considered that the declination is rather purposeful although physiological causes may not be neglected [20].

The form of the declination depends on speaking style. The differences are rather quantitative than qualitative. A steeper declination and larger pitch reset is usually observed in read-aloud speech, however, even in spontaneous speech, we may observe correlations between resets and linguistic unit boundaries.

2.2.4 Prosodic Boundaries in Speech

Speakers typically join words (feet) into phrases separated by prosodic means. These *intonational phrases* are defined as groups of feet perceived as compact intonational units. Their compactness is most typically based on prosodic marking of their boundaries [5, 21]. If an utterance is short, the intonational phrase may correspond to the entire utterance. On the other hand, longer utterance units cannot be uttered as a single intonation unit. Even if it was possible, it would only make understanding harder for the listener.

As was mentioned above, intonational phrases are primarily marked by their boundaries. In general, we distinguish several types of boundaries in speech. In the literature about Czech prosody, the only recognized type of a prosodic phrase is an intonation phrase (frequently referred to as an "*utterance stretch*"). By contrast, English literature also recognizes smaller phrasal units – *intermediate phrases*. In this theory, intonation phrases consist of one or several intermediate phrases and a boundary tone.

A most frequently used standard for annotation of English intonation is called ToBI (Tones and Break Indices) [22]. When using the ToBI annotation scheme, an index number reflecting strength of prosodic juncture with the following word is assigned to each word in an utterance. These so-called break indices range from 0 (the strongest juncture – boundary within a clitic group) to 4 (the weakest juncture – boundary of an intonational phrase). More information on the ToBI labeling system is given, for example, in [13].

The placement of prosodic boundaries is influenced by syntactic and semantic structure of the utterance. Speakers generally tend to realize their utterances as a sequence of intonation units that are linguistically meaningful. In most of languages, prosodic phrasing is more related to syntax than to semantics. Besides these two factors, the phrasing is also influenced by rhythmic constraints. If all these three factors are in harmony and lead the speaker to the same prosodic structure, then this structure is more markedly prosodically expressed in the produced utterance. On the other hand, if they are not in harmony, the speaker has to deal with their discrepancy and decide which of them to prefer. While the prosodic structure following semantic and syntactic structure of the message is more convenient for listeners, for the speaker, it is easier to follow the rhythmic constraints.

2.3 Spontaneous Speech

The quality and fluency of speech is highly dependent on whether the utterance is prepared beforehand or not. The most fluent speech is typically produced when it is read-aloud from a paper or screen, while the least fluent utterances are usually produced when the speaker is completely surprised by an unexpected question or unforeseen situation and has to respond spontaneously. According to the level of spontaneity, we distinguish *planned* and *spontaneous* speech.

There is no general agreement on the spontaneous speech definition in the ASR literature. Most typically, all speech that is not read-aloud is considered to be spontaneous. When using this definition, we must be aware of the fact that it includes a number of distinct speaking styles with different levels of spontaneity [23]. On the one hand, this definition includes just repeating some prompted words, while, on the other hand, it also includes completely

$$\texttt{I believe } \underbrace{\texttt{it's not}}_{\textit{Reparandum}} \; \texttt{*} \; \underbrace{\texttt{uh pardon me}}_{\textit{Editing Phase}} \; \underbrace{\texttt{it is}}_{\textit{Correction}} \; \texttt{a significant difference.}$$

Figure 2.1: Structure of an edit disfluency. The asterisk (*) denotes the interruption point.

unprepared spontaneous conversations. For the remainder of this thesis, the term "spontaneous speech" will stand for such utterances which were not prepared in detail in advance.

Spontaneous speech is described here because two of the three corpora used in this thesis contain spontaneous conversations and it is good to know which phenomena we can expect to observe. The description herein primarily focus on the most prominent spontaneous speech phenomenon – speech disfluencies. I also mention differences between prosody in planned and spontaneous speech.

2.3.1 Disfluencies in Spontaneous Speech

Involuntary speaker's lapses disturbing a fluent flow of speech are referred to as *speech disfluencies*. These failures may show themselves as filled or unfilled pauses, word repetitions, corrections, or false starts. According to the terminology introduced by Shriberg in [24], disfluencies consist of the following four parts:

- Reparandum – stretch of speech that was later revised (i.e., repeated, corrected, or abandoned), and may be deleted without losing an "important" piece of information.
- Interruption Point (IP) – interword location at which point fluent speech becomes disfluent.
- Editing phase – temporal region spanning from the end of the reparandum to the onset of the correction phase which may contain filled or silent pauses and/or explicit editing term.
- Corrrection – stretch of speech that "repairs" the material in the reparandum.

The structure of disfluency is displayed in Fig. 2.1. This structure may be used to describe complex disfluencies such as repairs or false starts as well simpler disfluencies such as filled pauses. In the latter case, the reparandum as well as the correction are empty [25].

Shriberg also showed that the probability of observing a disfluency grows exponentially with the length of the uttered sentence [26]. It also has been shown that disfluencies occur more frequently in the initial part of the sentence than in the final part. The number of disfluencies also differ between monologs and dialogs. The disfluencies occur more often in dialogs since they are more difficult to be managed by the speakers. An interesting observation is that speakers produce less disfluencies when speaking with computers than with humans. However, this is expected to change when automatic spoken dialog systems become more natural.

Despite all the difficulties in producing spontaneous speech, we cannot say that it represent an inferior form of communication than read-aloud speech. For example, professional TV and radio anchors try to speak spontaneously when interviewing their guests in studio. Thus, spontaneous speech is not a synonym for sloppy speech [27]. The following subsections describe the most frequent types of disfluencies.

2.3.1.1 Filled Pauses

If a speaker gets in trouble and does not know which word should he/she use at the moment, an inarticulate sound filling a pause in the utterance is usually produced. It signals that the speaker is not done, and wants to keep holding the floor. In English, we recognize two typical instances of filled pauses (FPs): *"um"* (longer with a nasal component), and *"uh"* (shorter without a nasal component). Speakers often pronounce filled pauses with a lower pitch and their duration is longer than standard duration of the spectrally most similar phonemes [7]. Since it has turned out that filled pauses discomfort machines more than people, spoken language processing researchers have increased their attention to them in recent years. An example of an utterance with an FP follows.

$$\texttt{This has already been reported by * } \underbrace{\texttt{uh}}_{\textit{Editing Phase}} \texttt{ Ivanov.}$$

FPs can occur separately as well as in editing phases of various complex disfluencies. Freestanding FPs are frequently produced when the speaker searches for a proper word. In such cases, the filled pause is typically followed by a highly informative word or phrase. Filled pauses also frequently appear at the very beginning of an utterance when the speaker holds the floor but does not know how to start. Another typical occurrence locations are boundaries between syntactic constituents since speakers tend to produce their utterances in consistent units [28]. Filled pauses may also mark start of a new topic in a longer stretch of speech [27]. For spoken English, it has also been reported that *um* is more frequently used at utterance unit boundaries, while *uh* is typically used inside a unit [24]. A detailed survey on using *uh* and *um* in spontaneous English is given in [29].

A frequently discussed problem is how listeners use filled pauses for processing and understanding of an utterance. The view that listeners "ignore" them was prevailed by the estimation that they perceive their communication function. This finding is in correspondence with the theory of the full-value of spontaneous speech. FP locations may be important features for automatic linguistic segmentation of speech, however, the problem is that their use is strongly dependent on an individual speaking style. Fox Tree [30] argued that *um*s seem to aid comprehension by directing listeners' attention to the upcoming phase, while, by contrast, *uh*s have no effect on human word recognition, possibly because the effects were masked by pausing.

Beside by FPs, a disfluency may also be signaled by a silent pause. Such disfluencies are especially overt at places where there is no syntactic or semantic motivation for a silent pause or where the pause is inadequately long.

2.3.1.2 Repetitions

Repetition disfluencies occur when a speaker repeats a word or phrase. By repeating, the talker gains some additional time to plan the rest of the utterance. In our disfluency notation, the reparandum of the repetition disfluency contains all but the last repetition and the editing phase contains a filled pause or is empty. If there is more than just one repetition, we can describe its structure using nested disfluencies. The number of editing phases and IPs is then equal to the number of times the phrase is repeated. Thus, the ultimate repair only contains the last repetition.

Repetitions in spontaneous speech were analyzed by Clark and Wasow [31]. They divided repeats into four stages: *initial commitment*, *suspension of speech*, *hiatus*, and *restart of the* *constituent*. In the notation used in this thesis, these stages correspond to reparandum, interruption point, editing phase, and correction. In a so-called *commit-and-restore model* of repeated words, Clark and Wasow hypothesized that the more complex is a constituent, the more likely a talker is to suspend it after an initial commitment to it. Moreover, they claimed that speakers prefer to produce constituents with a *continuous delivery*, and that speakers make a preliminary commitment to constituents, expecting to suspend them later. Their hypotheses are supported by the finding that in spontaneous English, function words are repeated more frequently than content words. An example of a repetition disfluency follows.

$$\texttt{So}\ \underbrace{\texttt{let's}}_{\textit{Reparandum}}\ \texttt{*}\ \underbrace{\texttt{uh}}_{\textit{Editing Phase}}\ \underbrace{\texttt{let's}}_{\textit{Correction}}\ \texttt{do it the other way round.}$$

2.3.1.3 Corrections

Corrections are the most typical examples of a disfluency. A typical example of a correction has already been displayed in Fig. 2.1.

2.3.1.4 False starts

False starts occur when a talker decides not to repair an inaccurate part of the utterance, but rather abandons it completely and starts to reformulate the message from the beginning. The false starts apparently aggravate speech understanding [30]. An example of a false start disfluency follows.

$$\underbrace{\texttt{This is really * really}}_{\textit{Reparandum}}\ \texttt{*}\ \underbrace{\texttt{uh}}_{\textit{Editing Phase}}\ \underbrace{\texttt{you can see that it works}}_{\textit{Correction}}\texttt{.}$$

2.3.1.5 Disfluencies with Explicit Editing Terms

Talkers sometimes utter words or phrases not relating to the semantic content of the utterance but having a different metalinguistic function. Explicit editing terms are examples of such phrases. These terms may appear in disfluency editing phases to explicitly indicate that the speaker made a mistake in his/her utterance. An example of a disfluency with an explicit editing term follows.

$$\texttt{I know}\ \underbrace{\texttt{the kids}}_{\textit{Reparandum}}\ *\ \underbrace{\texttt{uh or rather}}_{\textit{Editing Phase}}\ \underbrace{\texttt{some of the kids}}_{\textit{Correction}}\ \texttt{will like this idea.}$$

2.3.2 Prosody in Spontaneous Speech

When reading aloud, speakers know how long sentences they are going to read and thus may plan their prosody in advance. On the other hand, when speaking spontaneously, the planning of prosody is much more difficult. Moreover, prosody of spontaneous speech is largely affected by disfluencies disrupting continuity of prosodic trends. Consequently, we can conclude that spontaneous prosody is much less regular than prosody of planned speech. Various linguistic literature [11, 15, 32, 33, 34, 35] reports the following particular differences between prosody of planned and spontaneous speech:

- Differences in *pausing* – pauses are more frequent in spontaneous speaking and they also occur at locations where they do not have syntactic or semantic motivation.

- Differences in *pitch declination* – the declination is less steep (or completely missing) in spontaneous speech and pitch resets are less perceptible.
- Differences in *mean pitch* – average of pitch values is usually lower in spontaneous speech.
- Differences in *duration* – preboundary lengthening is not that regular and expressive in spontaneous speaking as well as its relation with the strength of a boundary may differ; regularity of segment duration may be disrupted by disfluencies, duration of segments in corrections is usually shorter.

2.4 Chapter Summary

In this chapter, I have presented linguistic background to aid understanding of my thesis, particularly for non-linguists. First, I described differences between linguistic structural units in spoken and written language. These differences must be taken into account since the goal of this thesis is to automatically segment speech into units that have primarily been defined for text.

Second, I overviewed suprasegmental features of speech – prosody. The description of prosody is useful herein because prosody is an important indicator of sentence-like unit boundaries in speech. I described the most important prosodic phenomena and also overviewed prosodic terminology. This description is referred to in Chapter 6, which is devoted to extraction of prosodic features useful for automatic classification.

Third, I summarized phenomena typical for spontaneous speech. Spontaneous speech is described here because two of the three speech corpora used in this thesis contain spontaneous conversations and it is important to know which phenomena we can expect to observe. The presented description was primarily focused on the most prominent spontaneous speech phenomenon – speech disfluencies. Differences between prosody in planned and spontaneous speech were also mentioned.

Chapter 3

Related Work

If I have seen further, it is by standing on the shoulders of giants.

Isaac Newton

In recent years, a substantial amount of research has been conducted in the areas relating to automatic segmentation of speech into linguistic units. In this chapter, I survey the literature presenting that research. Because I have tried to make the survey as exhaustive as possible, I have also included some very recent papers that have been published parallel with my ongoing work. Some specific research performed in the domains that are investigated in this thesis (corpora with structural metadata, dialog act segmentation of meetings, and sentence segmentation of spoken Czech) is not presented here but in the respective chapters (Chapters 5, 8, and 10), in order to enhance their readability. The survey in this chapter is categorized based on what knowledge sources and modeling approaches have been used.

The chapter is structured as follows. Section 3.1 surveys important psycholinguistic studies on perception of intonational boundaries in speech. In that section, separate subsections overview studies about preboundary lengthening, pausing, local coding of prosodic boundary information, prosodic boundaries in Czech, and relations between syntactic and prosodic structures. Section 3.2 refers to technology-motivated research. Before presenting results of individual research groups, Subsection 3.2.1 overviews scoring metrics used for evaluation of the tasks related to sentence segmentation. Subsection 3.2.2 summarizes signal-based approaches to sentence segmentation that do not rely on speech recognition output. In contrast, Subsection 3.2.3 describes approaches only relying on recognized words. Section 3.2.4 presents approaches combining textual and prosodic knowledge. Finally, Section 3.3 gives a summary of the whole chapter.

3.1 Survey of Psycholinguistic Studies about Realization and Perception of Prosodic Boundaries

Studies about realization and perception of prosodic boundaries are important for this thesis since they provide insights that may be useful for design of prosodic features for automatic sentence segmentation systems. A number of papers has been published about perception of prosodic boundaries in various languages – in particular English, French, Swedish, and Dutch. Although this thesis focuses only on Czech and English, I report interesting results for other languages as well. Part of these results is directly applicable since some prosodic features are language-independent. The inter-language similarity is caused by physiologic properties of vocal organs that are common for all speakers. Prosodic differences among individual languages are mainly displayed in the rhythm, the relation among prosodic cues, and the way of expressing of overlapping prosodic events [16, 36].

3.1.1 Studies about Preboundary Lengthening

A lot of effort has been put into analyses of segmental duration before the boundaries. These phenomena are referred to as *final* or *preboundary lengthening*. It is considered that the region in which this lengthening is most prominently displayed is the final rhyme of the last word before a boundary. The final lengthening is considered not to be inborn, but the speaker has to learn it. This view is based on the finding that children do not use it. Thus, it is necessary to study it for each language individually. An interesting fact is that the final lengthening do not only appear in human speech, but is also present in bird singing, cicada chirp, or music [16].

Wightman et al. [37] reported that the lengthening in spoken English is progressive. It means that the last consonant in a word (if present) is relatively more lengthened than the preceding vowel. For spoken Swedish (radio stock market reports), it was shown that the lengthening is affected by the presence or absence of the semantic accent on the sentence-final word since the stressed words are lengthened more [38]. The authors also reported a negative correlation of the final lengthening and pause duration. In such cases, a *compensation rule* is applied – prosodic events may be marked using different means and less intensive use of one mean may be compensated by stronger use of another. Another finding of that work was that a longer lengthening corresponds to a stronger boundary. In contrast, Heldner and Megyesi [39] got contradictory results for spontaneous Swedish – the lengthening was more prominent before weaker boundaries, while stronger boundaries were usually marked by longer pauses.

Strangert analyzed features of intonational phrases in spontaneous Swedish [40]. She found that 80% of intonational phrase boundaries were syntactically motivated. Most of other boundaries occurred after function words on onsets of syntactic constituents. Moreover, she also studied preboundary lengthening. The lengthening of final rhymes was more prominent preceding weaker boundaring. On the other hand, pauses were longer after stronger boundaries. It was also found that the lengthening was more prominent in shorter intonational phrases.

Yang [41] focused on an analysis of occurrences of pauses and final lengthening in various forms of English broadcast news. Regarding preboundary lengthening, it was reported that the syllables before boundaries are the longest and syllable duration is decreasing toward a phrase onset. The lengthening usually began on the fifth syllable before the boundary.

3.1.2 Studies about Pausing

A substantial amount of research has been conducted to analyze *pausing*. Pauses are without a doubt the most expressive instruments for marking of strong boundaries. Their expressive power is strong enough to override all other prosodic means [9]. Megyesi [35] reported that, as expected, pauses were more frequent in spontaneous speech. In the already mentioned paper [41], Yang reported that the pauses marking boundaries were longer than other pauses. The percentage of phrases separated by a pause was ranging from 35% in a spontaneous interview to 75% in a radio story read by a single speaker. Moreover, it was found that longer pauses correspond to stronger boundaries. Van Donzel and Koopmans reported that clause, sentence and paragraph boundaries are realized using silent pauses and high boundary tones in spontaneous Dutch [42, 43].

3.1.3 Studies about Local Coding of Prosodic Boundaries

Interesting experiments were performed by Grosjean [44, 45]. He let a single speaker read four English sentences of different lengths. These sentences were special in that a shorter sentence was always identical with the beginning of a longer sentence. The shared sentence-initial part was in individual sentences followed by zero, three, six, and nine words, respectively. In a perception experiment, the author replayed the common part of the sentences to a group of subjects. He found that the listeners were able to guess whether the sentence ends after the last replayed word, and even to predict how many words come after in a particular recording. In a consecutive experiment, he found that the subjects were also able to do the same prediction when they only heard last words of particular recordings. In that case, the predictions were only slightly less accurate. Later, he repeated the same experiment with using French sentences instead of English to study language-dependency. Then, the listeners were also able to tell whether the replayed word is the last one in a sentence, but were not able to recognize how many words would follow.

In a similar study, Carlson, Swerts, and Hirschberg focused on the issue whether speakers prosodically mark boundaries “in advance”, that is whether the listener has to hear prosody after the last word before the boundary to recognize the boundary [46, 47]. To analyze it, they used a radio interview in Swedish. The subjects had to recognize the boundaries from stimuli comprising either only a single word or a two second stretch of speech. Thus, the listeners could not use pause duration or pitch resets as cues. Despite this fact the results showed that the listeners were able to successfully predict the boundaries. An interesting finding was that using the longer, two-second stimuli did not yield a prediction improvement. It motivates a question whether the listeners use lexico-grammatical information at all. Hence, they repeated the same experiments with American subjects who did not have any knowledge of Swedish. The result was that the American group of listeners was in predicting the boundaries almost as good as the Swedish group. These results as well as findings of Grosjean support the hypothesis that a significant amount of prosodic information is contained in the last word before the boundary. It implies that it is possible to achieve good results in automatic boundary detection only using local prosodic features.

3.1.4 Studies about Prosodic Boundaries in Czech

There is not a lot of published work on perception of prosodic boundaries in spoken Czech. An interesting study was done by Palková [5] who evaluated listening tests and reported the following findings regarding boundaries of intonational phrases. The subjects marked as strongest the boundaries that were marked both by a pause and a characteristic intonation pattern (falling or rising). If the pause was not present, the perceptual recognition of boundary become more difficult. In that case, the intonation tone had to be much stronger. Other prosodic means were not that heavily utilized and usually were only used as an accompaniment of the basic two means. For instance, slowing down toward the boundary was frequently used together with a less emphatic melodic change. Another finding was that structuring the utterance into intonational phrases facilitates its understanding. Moreover, non-professional talkers tended to structure their speech into stretches having a roughly constant length and only rarely used intonational phrases containing just one foot. The listeners were also able to recognize intonational phrases in spontaneous speech – the inter-subject agreement was 80 %.

3.1.5 Studies about Relation between Syntactic and Prosodic Boundaries

Finally, I also cite here work analyzing relations between syntactic structure and prosody. This relationship was explored by Fach [48]. He used a large broadcast speech corpus which was annotated by an automatic syntactic parser and measured collocations of syntactic and prosodic boundaries. The comparison of syntactic and prosodic structure is not easy to perform because while the syntactic structure is hierarchical, the prosodic structure is considered to be linear. Fach solved the problem by leaving the problematic segments in the syntactic tree unattached. After adjusting the data for disagreements caused by individual speaking style, he observed that 84 % of syntactic boundaries were in correspondence with prosodic boundaries. However, although these results are interesting, the alignment of prosody and syntax remains a debated area requiring further research.

3.2 Survey of Related Work on Sentence Segmentation and Automatic Punctuation of Speech

3.2.1 Evaluation Metrics for Sentence Segmentation

Several different evaluation metrics have been used for performance scoring of automatic sentence segmentation systems. In general, there is no measure considered as a standard. Instead, various metrics have been used in various projects. The most straightforward metric is a "*boundary error rate*"[1] (BER) [49] defined as the number of incorrectly classified samples divided by the total number of samples

$$BER = \frac{I+M}{N_W} \qquad [\%] \tag{3.1}$$

where I denotes the number of false DA boundary insertions, M the number of misses, and N_W the number of words (thus also interword boundaries) in the test set. A complement of BER defined as $Acc = 100\% - BER$, called *Classification Accuracy*, is often used as well.

When using BER for performance measuring, one must be aware of the fact that a strong majority of words (approximately 80–95% depending on the speaking style and genre) in speech transcripts are not followed by sentence unit boundaries. Hence, relatively low absolute error rates may be achieved by simply classifying all test samples as "non-sentence" boundaries. Thus, we always should report a "chance error rate". This error rate coresponds to an imaginary model that classifies all test samples as belonging to the class having the highest prior (i.e. "no-boundary" in all our tasks). An advantage of BER is that it has a clear interpretation. The differences in performance between two systems may easily be tested for statistical significance, e.g. using the Sign test [50].

Another frequently used measure is the *NIST* error rate, which was used for evaluation in the DARPA EARS project. For sentence boundary detection, the NIST error rate is defined as the number of misclassified boundaries divided by the total number of sentence boundaries in the reference

$$NIST = \frac{I+M}{N_{SU}} \qquad [\%] \tag{3.2}$$

where N_{SU} denotes the number of sentence unit boundaries in reference. The difference between BER and NIST is only in the denominator. While BER has the total number of word boundaries in the denominator, the NIST metric only calculates the number of sentence boundaries. An unnatural feature of the NIST error rate is that its values may exceed 100%. Another

[1] Sometimes also referred to as Classification Error Rate.

drawback of this scoring method is that we cannot directly use any standard statistical test for testing of differences in NIST for statistical significance because this metric is not based on consistent segments [51].

Yet another popular scoring approach evaluates performance using a pair of complementary measures – *precision* (P) and *recall* (R). These measures are well-known from information retrieval systems where it is also necessary to deal with a highly skewed prior distribution. Let TP denote the number of true positives, TN true negatives, FP false positives, and FN false negatives. Then, precision, defined as a measure of the proportion of the detected event labels that the system got right, may be expressed as

$$P = \frac{TP}{TP + FP} \tag{3.3}$$

Recall, defined as the proportion of the detected event labels that the system found, may be expressed as

$$R = \frac{TP}{TP + FN} \tag{3.4}$$

Precision reflects substitution and insertion errors (purity of retrieval), while recall reflects substitution and deletion errors (completeness of retrieval). The higher precision and recall scores are, the better we consider the evaluated system. There is typically a trade-off relation between P and R; for example, many empirical studies of information retrieval performance have shown a tendency for precision to decline as recall increases.

It is often favorable to express system performance by a single number. Then, we can use a metric called F-measure, which is the harmonic mean of P and R

$$F = \frac{2PR}{P + R} \qquad [\%] \tag{3.5}$$

F-measure is a very popular metric in the NLP community. However, it also has some drawbacks. If we detect more than one event (for example in automatic punctuation), we must be aware of the fact that F-measure deweights deletion and insertion errors in comparison with substitution errors by a factor of two, as shown by Makhoul et al. [52]. Nevertheless, this problem is not relevant to sentence segmentation where there are no substitution errors since we only detect sentence boundaries. Thus, using F-measure for scoring sentence segmentation systems is well-founded.

3.2.2 Signal-Based Approaches Not Using Textual Information

Some research has been conducted on segmentation of speech into sentence units without using an ASR system. These methods are based on analyzing pitch, energy, and spectral features of speech signal. In certain conditions, these techniques may be useful since they are not affected by speech recognition errors.

Haase et al. [53] developed a method for determining utterance unit boundaries in German broadcast news using F_0 contours and energy envelopes. They employed an interesting approach in which classified units were voiceless regions of the speech signal. For each such region, a set of features relating to normalized F_0 and RMS energy values was extracted. These features were then used in a decision tree classifier. In the first classification step, it was decided whether there was a boundary or not. If yes, it was determined whether the boundary corresponded to sentence, paragraph, or "article". The authors reported a boundary detection precision of 85 % at a recall of 84 %. For determining type of the boundary, they achieved a precision 93 % and recall 88 %.

A sentence boundary detection method based on prosodic features aligned with automatically recognized vowels, consonants, and pauses (V/C/P) was introduced by Wang et al. [54]. They proposed the following technique. First, the V/C/P classification based on energy and spectral features and tracked F_0 was performed. Then, sentence boundary candidates were found in the segmented speech signal according to pause durations. Subsequently, a set of features corresponding to each boundary candidate was extracted. The feature set included speaking rate (syllables per second), pause duration, and pitch-related features (F_0 range, F_0 maximum and minimum, F_0 onset and offset). These features were used in an AdaBoost classifier. The achieved classification accuracy was 77.8%.

A multipass linear fold algorithm for sentence segmentation based on using discontinuities in the F_0 contour was proposed by Wang[2] and Narayanan [55]. At first, F_0 was measured from the speech signal. Regions in which F_0 was undefined (so-called pitch breaks) were then sorted in the ascending order with respect to their durations. In the sorted pitch break map, it was possible to recognize two clusters: the first one corresponded to most of unvoiced regions of speech and interword boundaries; the second corresponded to sentence boundaries and disfluencies. The two clusters were identified using a single approach smoothing the data points with a two-piece linear function. All pitch breaks preceding the line break were then excluded and the same algorithm was applied on the remaining data points. This procedure was repeated until only a small group of sentence boundary candidates remained. Then, sentence boundaries were identified in this candidate group by applying a set of heuristic rules reflecting typical features of sentence boundaries and disfluencies. For example, disfluencies more frequently occur close to sentence beginnings, sentences have approximately the same length, sentence boundaries are typically not close to each other, pitch resets occur at sentence boundaries, and so on. The method was tested on a subset of the Switchboard corpus and a classification accuracy of 75.0 % was achieved. This accuracy corresponds to a precision of $P = 83.8$ % at a recall of $R = 92.8$ %. The results are interesting because only pitch-based features were employed; adding other prosodic features and textual information from an ASR system should yield further improvement.

3.2.3 Text-Based Approaches

Several techniques only based on using "recognized words"[3] have been proposed in the literature. Although I refer to these approaches as *text-based*, the term *text* is considered in terms of unstructured ASR output. Hence, no features relating to text formatting (such as capitalization or punctuation) are used in any of the methods overviewed in this section.

Gavalda et al. [56] was inspired by Palmer and Hearst's method for sentence boundary detection in standard text [57]. Their goal was to find clause boundaries in the speech transcripts from the Switchboard corpus. Their features were based on "trigger" words (30 most frequent words occurring at clause and sentence boundaries) and parts-of-speech. A multilayer perceptron with two neurons in the hidden layer was employed for classification. The best results in terms of precision and recall ($P = 84.5$ %, $R = 86.0$ % and $F = 85.2$ %) were achieved when the features were extracted from a window containing six surrounding words (three words in each direction).

Beeferman et al. introduced an automatic punctuation system called Cyberpunc [58]. Their system was only focused on automatic insertion of commas, assuming predetermined sentence boundaries. Thus, their setup largely differed from my sentence segmentation task.

[2] It is not the same person as Wang mentioned in the previous paragraph.

[3] However, note that although these methods have primarily been proposed for use in speech processing systems, many of them have only been tested on true words, ignoring word errors in speech recognition.

The system searched for the optimal locations of commas maximizing the overall probability in the HMM framework. The authors achieved a precision $P = 78.4\,\%$, recall $R = 65.6\,\%$, and F-measure $F = 70.2\,\%$, using the Wall Street Journal corpus.

Stevenson and Gaizauskas [59] reported some experiments with sentence segmentation of transcripts from the Wall Street Journal corpus. They used a memory-based learning algorithm. They used a bunch of text-related features including a probability that a current word starts or ends a sentence. The achieved performance was only $P = 36$ and $R = 35$ when case information was removed from testing transcripts. In contrast, the results significantly improved ($P = 78$ and $R = 75$) when capitalization information was accessible for their system. It implies that their approach is not very suitable for ASR conditions.

In order to aid automatic speech understanding, Gupta et al. [60] developed an automatic system for splitting strings of recognized words into clauses. Moreover, their system was also identifying edit disfluencies. In successive steps, the recognized text was divided into sentences, cleaned of edit disfluencies, and split into clauses. The segmentation was viewed as a tagging problem and a boosting based approach was employed. Their features included three words and POS tags in both directions from the boundary of interest, and numbers of identical words and tags in the analyzed window. The method was tested on reference transcripts from the Switchboard corpus. For the clause segmentation subtask, they reported $P = 63.8$ %, $R = 58.5\,\%$, and $F = 61.1\,\%$. They also compared their method with a baseline trigram model and reported improved F-measure by 21.4 % absolute.

Same as Beeferman et al., Shieber and Tao [61] focused on automatic comma restoration for English speech transcripts. However, in contrast with Cyberpunc, their method employed parsing. They found that the presence of comma is correlated with a number of syntactic subtrees having their left boundary at the same place. By adding this information into their language model, the number of correctly annotated sentences raised from 47 % to 58 %.

Mrozinski et al. [62] employed a combination of word- and class-based N-gram LMs for automatic sentence segmentation of speech. Instead of using standard Viterbi search algorithm, they employed a procedure in which local probabilities were combined with matching recursive paths by keeping track of N^2 most probable paths leading from w_{i-N} to w_i. However, the two search methods were not compared so that no conclusions about their efficiency could have been drawn. The method was tested on broadcast news data as well as on spontaneous conference lectures. They also showed that proper automatic sentence segmentation was essential to achieve good results with automatic speech summarization systems.

Lee at al. [63] used text-based automatic punctuation restoration as part of their automatic speech-to-speech translation system. In their approach, sentence boundaries were already known, so only commas were predicted. They used a very simple method in which a comma was recognized if its N-gram probability given the word context exceeded a fixed threshold.

A number of papers have been published on text-based segmentation of spoken Japanese. However, given a specific syntactic structure of Japanese language (clause boundaries are marked by conjugated forms of verb phrases or conjunctive particles, so various types of boundaries may be identified quite precisely by referring to POS tags), it seems not to be possible to easily port these methods to other languages. For example, Takanashi et al. [64] developed a semi-automatic method for identification of clause boundaries in spontaneous Japanese. Their approach was to detect boundaries using a set of conversion rules referring to POS tags.

A sentence boundary detector based on the Structured Language Model (SLM) [65] was developed by Mori et al. [66]. SLMs are LMs based on a combination of parsing and N-gram models using a probabilistic parametrization of a shift-reduced parser. The idea behind the SLMs is that the next word can be more accurately predicted from words on which it potentially depends rather than from simple N-grams. In the first step of Mori's sentence boundary detection method, sentence boundary candidates were found according to pause durations (threshold 300 ms). These sentence boundary candidates were then classified using SLMs. The authors reported a slight performance improvement compared to an N-gram model testing on a Japanese radio lecture corpus.

A text-based sentence boundary detector for Japanese relying on using sequential dependency analysis (SDA) combined with automatic chunking was presented by Oba et al. [67]. The proposed SDA method extracts a dependency structure of unsegmented word sequences using a subsidiary mechanism of sentence boundary detection. To reflect local properties of word sequences as well as their appropriateness to form a sentence, the SDA method was combined with automatic chunking in the conditional random field framework. The method was only tested on human transcriptions from the Corpus of Spontaneous Japanese, so it is not clear how sensitive it is to word errors.

3.2.4 Approaches Combining Textual and Prosodic Cues

3.2.4.1 SRI-ICSI Approach Based on HMMs

In the past decade, a substantial amount of research on automatic sentence segmentation of speech has been conducted by Shriberg, Stolcke, and their colleagues at SRI and later at the International Computer Science Institute (ICSI). Their methods are based on the idea of *direct modeling of prosody* [68]. In this approach, prosodic features relating to detected events are extracted *directly* from the speech signal and ASR output, and the estimation of prosodic model parameters is done via a machine learning algorithm. The prosodic model directly outputs detected event posteriors. The direct modeling approach is in contrast with the approaches utilizing prosodic information via some abstract prosodic labels (such as ToBI). In those "indirect" approaches, prosodic classifiers are used to automatically recognize the prosodic labels [69, 70], that are subsequently used as features in a downstream classifier [71, 72, 73].

The direct modeling approach has several advantages. First, building such models is less time consuming since it does not involve any hand-labeling of prosodic events in training data. In addition, it simply bypass potential problems with subjective perception of prosody by human labelers. Moreover, this approach is not only labor saving but also yields good results. It benefit from the fact that the prosodic model is optimized for detection of predicted events rather than for automatic assignment of prosodic labels. Besides sentence segmentation of speech, the direct modeling approach has also been applied to classification of dialog acts, emotion recognition, disfluency detection, and speaker verification. A diagram of a system based on the direct modeling approach is displayed in Fig. 3.1.

Shriberg, Stolcke et al. originally proposed to combine lexical and prosodic features in the HMM framework [49]. The HMM-based approach is one of the approaches I employ in my work. Since the combination approach itself will in detail be described in Chapter 7, along with other used techniques, I only summarize the results achieved by this approach in this section.

The SRI-ICSI group tested the HMM approach both in read-aloud and spontaneous speech corpora [74, 49]. They utilized prosodic features relating to pause, phone, and rhyme durations, and pitch and energy. Their results demonstrated that the combined model employing both lexical and prosodic features generally outperforms models utilizing just one information source. They also showed that for testing on real ASR transcripts, word recognition errors generally cause more degradation for lexical than for prosodic features. For broadcast news data, the overall BER was 3.3% in reference conditions (forced alignment of human transcripts) and

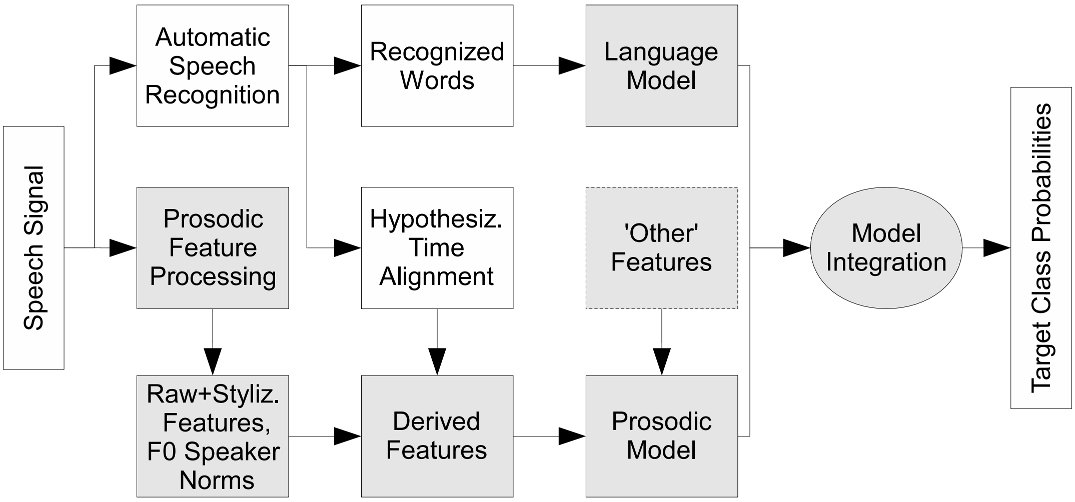

Speech Signal
Automatic Speech Recognition
Prosodic Feature Processing
Recognized Words
Hypothesiz. Time Alignment
Language Model
'Other' Features
Raw+Styliz. Features, F0 Speaker Norms
Derived Features
Prosodic Model
Model Integration
Target Class Probabilities

Figure 3.1: Schematic diagram of the direct prosody modeling approach. White boxes indicate processing in standard ASR systems, gray boxes indicate processing added in the prosody modeling approach [68].

10.8% in ASR conditions (fully automatic transcripts). For the telephone speech corpus, they reported $BER = 4.0\%$ in reference and $BER = 22.2\%$ in ASR conditions.

It was also found that usage of individual prosodic features is corpus-dependent. Though pause duration was always the most used feature, other top features differed across corpora. While F_0 features prevailed for read-aloud speech, duration features dominated in the conversational speech corpus.

More recent results of the same team using the HMM approach were published by Liu et al. in [75]. However, the results are difficult to compare with the original experiments because: (1) they only reported results in the NIST metric; (2) the experimental setup including a sentence boundary definition differed from the original experiments. The reported NIST error rates were 48.7% and 55.4% for Broadcast News, and 31.5% and 43.0% for telephone conversations, respectively.

Hillard et al. [76] extended the HMM approach by leveraging multiple ASR hypotheses in order to provide more robustness to ASR errors. Posterior probabilities for each hypothesized word sequence were estimated via HMMs, and subsequently, the hypotheses were combined using confusion networks to determine the overall most likely sequence of events. Moderate but statistically significant improvement was reported for conversational speech. On the other hand, for broadcast news no significant improvement was achieved.

The HMM approach with CART-style decision trees for prosody modeling was also adopted by Kim and Woodland [77]. They experimented with automatic punctuation detection in broadcast news speech. The authors measured performance of their system using modified P,R, and F. In their measure, a half score was given when a punctuation mark was located correctly but recognized as a different type of punctuation. Two series of experiments were performed. The first one was focused on automatic punctuation annotation in reference manual transcripts. For this setup, they reported $P = 76\%$, $R = 80\%$, and $F = 78\%$. In the second series of experiments, punctuation generation was directly integrated into the ASR system by adding punctuation marks as pseudo-words into the recognizer's vocabulary (with acoustic baseforms corresponding to silence). In this setup, prosodic information was also used for ASR lattice rescoring. Automatic punctuation performance within this integrated system was reported as $P = 58\%$, $R = 35\%$, and $F = 44\%$. In addition, a small reduction in WER was achieved by using the prosodic rescoring.

3.2.4.2 VERBMOBIL Approach

VERBMOBIL was a research project funded by the German government in years 1996–2000. The goal of this program was to develop a mobile automatic translator able to translate spontaneous speech from a limited domain (appointment scheduling) in real-time. The target languages were German, English, and Japanese. An essential subtask of the system was to automatically detect various linguistic boundaries. I particularly describe the results of VERBMOBIL herein in more detail because it was the first end-to-end application in which prosody modeling was successfully employed.

From the viewpoint of this thesis, the most interesting part of the system is the *prosody module*. The task of the module was to automatically annotate accents, sentence mood (question vs. non-question), and acoustic-prosodic and syntactic-prosodic boundaries. For reference annotation of syntactic and prosodic boundaries, a very complex annotation scheme distinguishing acoustic-prosodic, syntactic, syntactic-prosodic, and dialog act boundaries has been developed [78]. Input to the prosodic module was a speech signal and a Word Hypotheses Graph (WHG),[4], while module output was a WHG annotated with the above mentioned "prosodic information". This prosodically annotated WHG was consequently used in other VERBMOBIL modules (such as the syntactic analysis module or the semantic module). As opposed to off-line segmentation experiments conducted by the SRI-ICSI group, VERBMOBIL operated in real-time. On the other hand, it only worked with a much smaller vocabulary from a limited domain.

The VERBMOBIL approach to dialog act segmentation (and classification) was based on using *multi-layer perceptrons* for prosodic classification (MLPs), *polygram* language models [79], and an A^*-based search algorithm [80, 81]. The MLP classifier had two neurons in the output layer, one for DA boundaries and one for other word boundaries. Output scores from these two neurons were normalized in order to obtain posterior probabilities. The polygram language model was a set of interpolated N-grams with varying N. The polygrams for boundary detection modeled the probability $P(w_{i-2}, w_{i-1}, w_i, e_i, w_{i+1}, w_{i+2})$ where e_i corresponds to the classified boundary and $w_{i-2}, \ldots, w_{i+2}$ to surrounding words. The optimal boundary sequence was then found using the A^*-based search algorithm combining the scores from the MLP and polygrams. This approach directly enabled them to join the DA segmentation with DA classification. For German data, the authors reported $R = 80\%$ for DA boundaries, and $R = 96\%$ for non-DA boundaries.

Gallwitz et al. from the same team proposed a method for simultaneous generation of word hypothesis and linguistic boundaries [82, 83]. In this method, the language model of the speech recognizer treated linguistic boundaries as pseudo-words, and the acoustic model was augmented by special HMMs representing acoustic realizations of recognized boundaries (e.g., loud breath, filled pause, silence). In addition, a special HMM with one emitting state without a loop corresponded to boundaries that were not acoustically marked. To combine spectral and prosodic features, they employed a hybrid architecture that joined the MLP with semi-continuous HMMs. It involved using a soft vector quantization based on a Gaussian codebook. The output of the integrated system was an annotated WHG. The system was tested on a spontaneous speech database and the results showed modest improvement in WER. Moreover, the authors evaluated detection of syntactic-prosodic boundaries. In comparison with a prosody-only boundary detection model, recall improved from 75.1% to 88.2%, and precision

[4] In other words, it was an ASR lattice.

improved from 75.3% to 78.5%. Unfortunately, a comparison with a system combining prosody and language model was not reported.

3.2.4.3 Other Work Combining Textual and Prosodic Cues

Earlier than Gallwitz et al., Chen [84] used a combined approach of synchronous speech recognition and punctuation generation. His target set of punctuation was quite broad, including comma, full-stop, question mark, exclamation mark, colon, and semicolon. He found that pauses in speech were closely related to punctuation, and thus decided to treat punctuation marks as pseudo-words in the vocabulary. In his most successful experiment on a business letter corpus, Chen achieved punctuation detection accuracy of 57%, and correct punctuation placement (independent of punctuation type) accuracy of 83%.

Gotoh and Renals [85] presented a sentence boundary detector for broadcast news speech. They used an N-gram language model in which a label c_i indicating the presence of a sentence boundary was assigned to each word w_i. The only used feature was pause duration s_i. The combination of the language and prosodic model was performed via an approximation in the form $P(s_1^m, w_1^m, c_1^m) = \prod_{i=1}^{m} P_{PM}(w_i, c_i|s_i)^{\lambda} \cdot P(w_i, c_i|w_{i-N+1}^{i-1}, c_{i-N+1}^{i-1})$. This score was maximized using the Viterbi search. On ASR data with $WER = 26\%$, their model performed sentence boundary detection at $F = 70\%$.

The work of Gotoh and Renals was extended by Christensen et al. [86]. The followers focused on automatic punctuation annotation in broadcast news speech. They used a richer prosodic feature set including phone duration and pitch features. The finite state automaton approach combining the prosody and language model was adopted. The results indicated that pause duration had significant impact on full-stop detection, but only a little influence on detection of other punctuation marks. Overall, $F = 40\%$ was the best result on an automatically recognized broadcast news corpus with $WER = 23\%$.

Huang and Zweig [13] investigated a maximum entropy (MaxEnt) based approach for automatic punctuation from speech. Their motivation for using MaxEnt was that it allows a natural combination of lexical and prosodic features within a single model. The punctuation detection task was viewed as a tagging problem in which one of the allowed punctuation symbols (comma, period, question mark, and empty symbol) was assigned to each word. Their lexical features were extracted from the context including $w_i, w_{i+1}, w_{i+2}, w_{i-1}, w_{i-2}, t_{i-1}, t_{i-2}$ (where w_x correspond to words, and t_x correspond to punctuation marks) using special unigram and bigram feature templates. The lexical features were extended by pause duration features measured with a precision of 10 ms. The third group of features combined word identities and pause durations directly on the feature level. The method was tested on the Switchboard database, both in reference and ASR conditions ($WER = 20\%$). To achieve accurate evaluation, automatic transcripts were manually punctuated. The results indicated that commas could not be reliably detected using just pause information since they were heavily dependent on lexical information. It was also shown that question marks were often confused with periods. Using ASR output instead of human transcripts decreased overall F-measure from 80% to 73%.

Srivastava and Kubala [87] focused on sentence segmentation of Arabic broadcast news. They experimented with an MLP having input consisting of 47 prosodic features extracted from a one-second window around the boundary of interest. For language modeling, a trigram model combined with prosodic scores in a way similar to [86] was employed. The authors reported “detection error rate” (sum of false alarms and false rejections) to be 50.38%.

In [88], Liu et al. compared generative and posterior probability models (namely an HMM and MaxEnt) for sentence boundary detection in speech. Both models combined lexical, syntactic, and prosodic information. Features for the MaxEnt model involved word N-grams, POS N-grams, class N-grams, chunking tags, speaker change flags. In addition, in order to combine prosodic and lexical features in a single model, prosodic decision tree posteriors were discretized in a cumulative fashion to form binary features. The methods were compared on telephone coversations and broadcast news speech in both reference and ASR conditions. The results indicated that the MaxEnt model slightly outperformed HMM in STT conditions, while the HMM was better in STT conditions. The combination of the two approaches achieved the best performance for all reported tasks. The MaxEnt model showed much better accuracy than the HMM in dealing with lexical information. On the other hand, the HMM made more effective use of prosodic features, which were more robust to word errors.

In [89], Liu et al. extended their work by using Conditional Random Fields (CRFs), which combine the benefits of HMM and MaxEnt models. The features used for CRF training were identical with the features used in the MaxEnt model. For conversational speech, the CRF model was slightly superior to both the HMM and the MaxEnt model. The gain from using CRFs was highest when only using the N-gram features were used. The differences among individual models diminished when also other features were added. In contrast, the CRF showed less gain on broadcast news data. When all features were used, the CRF performance was identical with performance of the HMM. Across all conditions, the overall best results were achieved by three-way voting among the classifiers.

Roark et al. [72] presented a sentence segmentation method based on reranking of N-best lists. First, the N-best lists were generated based using a baseline system. Then, the lists were reranked using a MaxEnt reranker. The features were mainly based on outputs from a number of parsers, but the authors also used prosody in terms of automatically predicted ToBI labels, N-gram scores, and some other features. The empirical results on conversational data showed 2.6 % gain in NIST error rate for reference conditions, and a modest, yet statistically significant improvement for ASR conditions.

For sentence boundary detection and classification in conversational telephone speech, Tomalin and Woodland [90] exploited prosodic classifiers based on discriminatively trained Gaussian mixture models. In their approach, fourgram LMs were combined with the prosody model in a lattice-based 1-best Viterbi decoding framework using empirically determined grammar scaling factors. The probabilities obtained from the prosody model were divided by event priors, and the resulting likelihoods were placed on the arcs of initial lattices, which were then expanded using the LMs in the HTK lattice tools [91]. The prosodic posteriors were generated using two different approaches. The first approach was based on using CART-style decision trees, the latter on the dicriminatively trained GMMs. For each SU subtype (statement, question, backchannel, incomplete, no-boundary), GMMs were built using maximum likelihood training. Then, the GMMs were reestimated using Maximum Mutual Information training. The results indicated that the GMMs performed as well as decision trees, and, when both models were interpolated, the NIST error rate dropped by 0.8% absolute over the decision tree baseline.

Zimmermann et al. [92] presented a multilingual system for sentence segmentation of English and Mandarin broadcast programs. They tested several different models, including hidden event LM, MaxEnt, decision trees, and BoosTexter[5] [93], however, their comparison was only complete for lexical and pause features because a richer prosodic feature set was only tested with the BoosTexter model. The comparison showed that the best results were achieved when the BoosTexter model was combined with the hidden event LM. The reported results were $NIST = 62.4\%$ and $F = 67.3\%$ for English and $NIST = 58.7\%$ and $F = 70.8\%$ for Mandarin.

Matusov et al. [94] presented a sentence segmentation and punctuation prediction system

[5] BoosTexter model will be discussed in more detail in Section 7.3.

tailored for spoken language translation. The method was based on a log-linear combination of several independent models. The language model probability was factored into a product of a segment start, segment internal, and segment end probability. The pause model reflected normalized pause duration at the hypothesized word boundary. Other prosodic features were not utilized, but the authors claimed these could be added with a separate scaling factor, assuming they would provide a single posterior of a segment boundary. The authors also included an explicit sentence length probability feature having a log-normal distribution with parameters determined via a maximum likelihood estimate. Finally, a recursive search algorithm was employed to determine the globally optimal sentence segmentation of the document. The obtained results were slightly better than results achieved by the HMM approach trained on the same data.

In a later paper [4], in order to couple the segmentation with the predictive power of the phrase translation model, the same team introduced a novel feature – phrase coverage. They also performed comma prediction to produce commas as “soft boundaries” constraining reordering in the MT search. They concluded that the best translation results were achieved when segmentation algorithms were directly optimized for translation quality.

Batista et al. [95] developed an automatic punctuation restoration module as part of a Portuguese broadcast news transcription system. They focused on full stops and commas and employed a MaxEnt approach. Their features captured word and POS information, pause duration, and speaker changes. For sentence boundary detection, they reported $P = 76\%$, $R = 69\%$, and $F = 73\%$ for reference, and $P = 69\%$, $R = 48\%$, and $F = 56\%$ for ASR transcripts

Once again, there is also some specific work on Japanese. Shitaoka and his colleagues [96] presented two methods for sentence boundary detection in spontaneous Japanese. The first method was based on dependency information, the latter was based on SVMs viewing sentence boundary detection as a text chunking problem. The latter method, which was found to be superior, used the following features: morphological information of three preceding and subsequent words (character strings, pronunciation, POS, inflection type, inflection form), normalized pause duration, clause boundary information, and dependency probabilities of the target bunsetsu.[6] In a more recent paper, Akita et al. [97] adopted very similar approaches, and tested them on real ASR output. Again, the SVM-based method performed better than the statistical language model alone. As expected, the SVM-based method was also more robust to ASR errors.

3.3 Chapter Summary

In this chapter, I have reviewed studies about perception of prosodic boundaries as well as past work on automatic sentence segmentation of speech. First, I should point out the work of Grosjean followed by the joint work by Carlson, Swerts, and Hirschberg. Their listening tests showed that a significant amount of prosodic boundary markup is contained in the last word before the boundary. This finding has important implications for automatic boundary prediction systems since it suggests that good results may be achieved even if only local prosodic features are used. Noticeable results were also presented by Fach, who showed that most syntactic boundaries were in correspondence with prosodic boundaries in broadcast news speech. However, although his results were interesting, the alignment of prosody and syntax remains a debated area requiring further research.

A number of papers studied the phenomenon of preboundary lengthening. Some results

[6] *Bunsetsu* is a Japanese phrasal unit.

of perception tests were contradictory. For example, while a positive correlation of strength of boundary and lengthening was reported for broadcast speech, a negative correlation was observed in spontaneous speech. Regarding pausing, it was reported that pauses are more frequent in spontaneous conversations, in which they also often appear without a syntactic motivation.

In the second section, I have presented a survey of automatic sentence segmentation systems categorized based on what knowledge sources have been used. Among the signal-based approaches not relying on textual information, the multipass linear fold algorithm by Wang and Narayanan deserves most attention. They achieved encouraging results only using pitch-related features. The text-only based systems are apparently suboptimal since at least pause information is essential to achieve good results. However, some of them, such as Cyberpunc, are also worth mentioning since they introduced some ideas upon which more recent research has been build.

Past work on automatic sentence boundary detection have shown that both lexical and prosodic cues are important and should be combined. In this thesis, I largely build upon the framework proposed by Shriberg and Stolcke. They introduced a "direct modeling" approach in which prosodic features are extracted directly from the speech signal automatically aligned with speech recognition output. They had originally proposed to combine lexical and prosodic features in HMMs but later also used other combination approaches.

A mobile speech-to-speech translator VERBMOBIL was the first end-to-end application in which prosody and language modeling was successfully combined. From the viewpoint of this thesis, the most interesting part of the system is the prosody module. The task of the module was to automatically annotate accents, sentence modalities, and acoustic-prosodic and syntactic-prosodic boundaries. The VERBMOBIL approach to dialog act segmentation and classification was based on using MLPs for prosodic classification, polygram language models, and an A^*-based search algorithm.

I also got inspired by the MaxEnt approach as used by Huang and Zweig, and later, in a different fashion, by Liu et al. Furthermore, same as Zimmermann et al., I also employ the BoosTexter algorithm in my work. I also should highlight the work by Roark et al. Their complex approach is particularly interesting by using parsing features for hypotheses reranking. The parsing features should generally be helpful for sentence segmentation of speech, but their drawback is that parsing performance is largely affected by ASR errors.

Finally, note that a lot of the work summarized in this chapter has only been tested on manual transcripts. Even though such studies may provide interesting insights, the goal of this thesis is to develop robust methods also working well with real ASR output that contains word errors. Thus, my experiments in this thesis are evaluated using both human and automatically generated speech transcripts.

Chapter 4

Thesis Objectives

From now on, ending a sentence with a preposition
is something up with which I will not put.

Winston Churchill

The three preceding chapters, outlining my motivation, important linguistic background, and the state of the art in automatic sentence segmentation of speech, represented an introductory part of the thesis. Before describing my own work, this chapter explicitly lists particular objectives of this thesis. The objectives can be divided into two groups – *Creation and analysis of data resources* and *Development of automatic sentence-like unit segmentation systems.*

Since a suitable English corpus has already been available, the data creation part of this work only focuses on Czech. The objective of this part is the following:

1. *Prepare and analyze Czech speech corpora with appropriate annotation of sentence-like unit boundaries.* In order to analyze difference between planned and spontaneous speech, it was decided to create two distinct corpora – one in the domain of broadcast news and the other in the domain of broadcast conversations. The employed annotation scheme goes far beyond just marking sentence-like unit boundaries since it also includes annotation of fillers and edit disfluencies. This work is described in Chapter 5.

In the experimental part, I deal with both English and Czech. Since my work in both languages builds upon similar modeling approaches, before presenting the experiments themselves, I first describe the methods for prosodic feature extraction (Chapter 6), and the statistical models used for automatic sentence segmentation (Chapter 7). The experiments focusing on English were performed on the publicly available ICSI meeting corpus, whereas my experiments with spoken Czech used the new corpora created as part of this thesis. The explicit objectives in the experimental part are the following:

2. *Explore dialog act segmentation of multiparty meetings in English.* The automatic processing of multiparty meetings is an area of growing interest. Since prior to this work, this domain has not been well explored for sentence or dialog act segmentation, this has been done as part of this thesis. Besides standard speaker-independent experiments (Chapter 8), I have also focused on investigating speaker-specific modeling in this domain (Chapter 9).

3. *Develop and evaluate a baseline sentence unit segmentation system for Czech.* In this subtask, the goal is to design a sentence-like unit segmentation system and evaluate it using the corpora created in Objective 1. This work is described in Chapter 10.

Chapter 5

Design, Creation, and Analysis of Czech Speech Corpora with Structural Metadata Annotation

Words are the coins making up the currency of sentences, and there are always too many small coins.

Jules Renard

One of the main goals of this thesis was to create a sentence boundary detection system for spoken Czech. However, at the time when this work started, no speech corpora with annotation of sentence-like units necessary for training and testing Czech sentence segmentation systems were available. Hence, such corpora had to be prepared as a very important part of this work. I decided to create two corpora in two different domains: broadcast news (mostly read-aloud speech) and broadcast conversations (mostly spontaneous speech). The first corpus was created "just" by enriching an existing broadcast news corpus with structural metadata annotation, whereas the second had to be created from scratch.

The annotation scheme I use is based on the LDC's "Simple Metadata Annotation Specification". This structural annotation goes far beyond just labeling of sentence-like unit boundaries. Speech disfluencies, filler words, and some other phenomena were also annotated since I wanted to create corpora useful for studying a broad spectrum of spoken language phenomena. In this chapter, I also present a detailed analysis of the annotated corpora in terms of structural metadata statistics.

This chapter is organized as follows. Section 5.1 describes the used audio data and briefly outlines how the data were transcribed. Section 5.2 overviews approaches to annotating structure of spontaneous utterances and states the reasons why I chose to adopt the structural metadata annotation approach. Section 5.3 presents the structural metadata annotation guidelines for Czech. Section 5.4 analyzes a number of structural metadata statistics relating to the two annotated corpora. Section 5.5 summarizes the whole chapter.

5.1 Speech Data

5.1.1 Czech Broadcast News Corpus

The broadcast news (BN) speech data I used for this work were taken from the Czech Broadcast News Corpus. This corpus was recorded at UWB and is publicly available from the Linguistic

Table 5.1: Basic numbers about Czech Broadcast News corpus (BN) and Radioforum corpus (RF)

	BN	**RF**
Number of shows	342	52
Number of word tokens	234.2k	207.8k
Number of unique words	31.9k	25.3k
Duration of transcr. speech	26.7h	24.0h
Total number of speakers	284	94
— male speakers	188	77
— female speakers	96	17

Data Consortium (LDC) [98]. The corpus is spanning the period February 1, 2000 through April 22, 2000. During this time, news broadcasts on 3 TV channels and 4 radio stations were recorded. The whole corpus contains over 60 hours of audio stored on 342 waveform files, which yield more than 26 hours of pure transcribed speech.[1] The recordings do not contain weather forecasts, sport news, and traffic announcements. The signal is single channel. It was originally sampled at 44.10 kHz with 16-bit resolution, but for the official release, the waveforms were downsampled to 22.05 kHz. Basic numbers about the BN corpus are listed in Table 5.1. Details about corpus orthographic transcriptions were given in [99].

5.1.2 Radioforum – Czech Broadcast Conversation Corpus

The UWB speech research group has mainly gained its Czech spontaneous speech processing experience within the MALACH project [100]. However, at the time this work started, the testimonies of Holocaust survivors that comprise this corpus could not be freely distributed. Hence, a new corpus was recorded to support broader research on the problem of spontaneous Czech.

5.1.3 Audio Data

The newly recorded spontaneous speech database consists of 52 recordings of a radio discussion program called *Radioforum* (RF), which is broadcast by Czech Radio 1 every weekday evening. Radioforum is a live talk show where invited guests (most often politicians but also journalists, economists, doctors, teachers, soldiers, crime victims, and so on) spontaneously answer topical questions asked by one or two interviewers. The number of interviewees in a single program ranges from one to three. Most frequently, one interviewer and two interviewees appear in one show. The material includes passages of interactive dialog, but longer stretches of monolog-like speech slightly prevail.

Although the corpus was recorded from public radio where standard (literary) Czech would be expected, many speakers, especially those not used to talking on the radio, use colloquial language as well. Literary and colloquial word forms are often mixed in a single sentence. The usage of colloquial language, however, is not as frequent as in unconstrained informal conversations.

[1] Because of copyright issues, only 286 of the 342 recorded shows yielding 22.8 hours of transcribed speech could have been published at the LDC. However, I used all 342 recordings for all my experiments described in this thesis.

The recordings were acquired during the period from February 12, 2003 through June 6, 2003. The signal is single channel, sampled at 44 kHz with 16-bit resolution. Typical duration of a single discussion is 33–35 minutes (shortened to 26–29 minutes after removing compact segments of telephonic questions asked by radio listeners, which were not transcribed). Some basic numbers about the corpus are presented in the third column of Table 5.1.

5.1.4 Speech Transcription

The recorded shows were manually transcribed based on detailed annotation guidelines. The goal of the transcription was to produce precise time-aligned verbatim transcripts of the audio recordings. The transcription guidelines for this Czech corpus were based on the guidelines published in [99, 12]. However, the original guidelines were adjusted to better accommodate specifics of the recorded spontaneous speech corpus. Some of the modifications were inspired by the LDC's "Guidelines for RT-04 Transcription" [101]. For example, in order to increase inter-labeler consistency, the number of tags for labeling speaker and background noises was significantly reduced. For the same reason, I also changed the rules for transcription of filled pauses because the original rules had been too vague. The filled pause issue is discussed in more detail below in Section 5.3.1.1.

Among others, the transcription guidelines instructed annotators how to deal with the following phenomena:

- *Speaker turns* – a corresponding time stamp and speaker ID are inserted every time there is a speaker change in the audio.

- *Turn-internal breakpoints* – to break up long turns, breakpoints roughly corresponding to "sentence" boundaries within a speaker turn are inserted.

- *Overlapping speech* – an overlapping speech region is recognized when more than one speaker talks simultaneously; within this region, each speaker's speech is transcribed separately (if intelligible).

- *Background noises* – *[NOISE]* tags are used to mark noticeable *background* noises.

- *Speaker noises* – speaker-produced noises are identified with one of the following tags: *[LOUD_BREATH]*, *[COUGH]*, *[LAUGHTER]*, *[CLICK]*.

- *Filled pauses* – filled pauses produced by a speaker to indicate hesitation or to maintain control of a conversation are transcribed either as *[EE-HESITATION]* or as *[MM-HESITATION]*, based on their pronunciation.

- *Interjections* – certain interjections typically used as backchannels or to express speaker's agreement or disagreement are transcribed using the *[MHM]* (disagreement) and *[HM]* (agreement or backchannel) tags.

- *Unintelligible speech* – regions of unintelligible speech are marked with a special symbol.

- *Numbers* – all numerals are transcribed as complete words.

- *Foreign names* – foreign names in the transcript are marked using special symbols; if pronunciation of a foreign language name differs from that expected by Czech spelling rules, it is added to the transcript as a comment.

- *Mispronounced words* – mispronounced words (reading errors, slips of the tongue) are transcribed in the spelling corresponding to their pronunciation in the audio (i.e., the incorrect pronunciation is represented[2]) and marked with a special symbol.
- *Word fragments* – the pronounced part of the word is transcribed and a single dash is used to indicate point at which word was broken off.
- *Punctuation* – standard punctuation (limited to commas, periods, and question marks) is used to enhance transcript readability.

The verbatim transcripts of this corpus were created by a large number of annotators. To keep them maximally correct and consistent, all submitted annotations were manually revised.

5.2 Annotation of Spontaneous Speech Structure

In the previous section, the creation of time-aligned *verbatim* transcripts was described. This annotation is usually sufficient for training and testing standard ASR systems. However, raw streams of words do not convey complete information because the structural information beyond the words (metadata) is equally important as the words themselves. As mentioned in the Introduction, structural information is critical to both increasing human readability of the transcripts and allowing application of downstream NLP methods, which typically require a fluent and formatted input.

This thesis is focused on automatic generation of *rich* speech transcripts[3] which not only contain words but also boundaries of sentence-like units. Thus, the key problem is how to annotate sentence boundaries in speech. In principle, there are two basic options. The first option is to use standard punctuation, whereas the other is to employ a special annotation scheme tailored for spoken language. Although the former approach is quite convenient for read speech, its convenience for spontaneous speech is at least questionable.

Because spontaneous utterances are not as well-structured as read speech and written text, there exist a number of reasons why annotating structure by simply making reference to standard punctuation is inadequate for many applications. First, there are no agreed-upon rules for punctuating faulty syntactic structures, which are quite frequent in spontaneous speech. Second, punctuation marks are ambiguous; commas may indicate several different structural/syntactic events (e.g., clausal break, apposition, parenthesis, etc.). Third, even for written text, the rules for applying punctuation are quite variable; for instance commas are optional in many cases. Fourth, standard punctuation does not convey all structural information contained in spontaneous speech. Spontaneous utterances are often incomplete or disfluent. Since dealing with the specific spontaneous speech phenomena is crucial to spontaneous speech understanding, more precise annotation of disfluencies and other structural phenomena is required. On the other hand, we must take into account that special annotation of spontaneous speech phenomena is extremely labor-intensive and thus expensive.

5.2.1 Related Work on Spontaneous Speech Annotation

Several different annotation schemes has been presented for similar annotation tasks. Earliest efforts include the Meteer manual for disfluency tagging of the Switchboard corpus [102]. A detailed annotation scheme for the Trains dialog corpus was proposed by Heeman [103]. His

[2] Unlike English, this is possible in Czech since spelling rules are phonetically based.

[3] Besides sentence boundaries, such rich transcripts may also include speaker diarization (who speaks when), disfluency annotation, or other structural information.

annotation scheme included labeling of intonational boundaries within the ToBI framework, identification of discourse markers, and a very detailed annotation of speech repairs.

Within the VERBMOBIL project, Batliner et al. [78] presented a syntactic-prosodic labeling system for large spontaneous speech databases called "M". This annotation was only based on word transcripts; the annotators did not have access to audio recordings. Using a rough syntactic analysis, each word in a turn was assigned to one of 25 "M" classes. Depending on the target task, these 25 classes were grouped either into 3 main M classes (M3 - clause boundary, M0 - clause internal, MU - boundaries that cannot be determined without listening to audio or knowing particular pragmatic context), or into 5 syntactic classes (S0 - no boundary, S1 - boundary after a particle, S2 - phrase boundary, S3 - clause boundary, S4 - sentence boundary). An apparent drawback of this approach is that the annotation scheme is very complex and requires experienced linguist annotators.

At the time I started my work on this task, there was no similar annotation system for Czech. In addition, there was almost no published work focusing on syntax of conversational Czech. The exception is a treatise by Müllerová [104]. This monograph describes syntactic phenomena specific for conversational Czech, surveys various types of speech repairs, and briefly discusses Czech discourse markers. Although Müllerová's work is interesting, it does not offer any clear clues to explicit annotation of sentence-like units in spoken Czech. The definitions provided in this work are too vague to be applicable to NLP tasks; the author only studies spoken Czech in terms of a qualitative linguistic description.

Müllerová argues that it is often impossible to decide whether neighboring syntactic constituents correspond to a compound sentence or to a pair of independent syntactic structures. She proposes to distinguish four types of syntactic boundaries: (1) overt sentence boundaries without any valence ambiguity; (2) subordinate boundaries delimited by subordinate conjunctions; (3) boundaries with ambiguous syntactic constituents (not clear whether linked to the preceding or the following predicate); and (4) boundaries in regions with ill-formed syntactic structure.

In addition to the above mentioned syntactic boundaries, Müllerová also defines *content-pragmatic* units. She defines these units as syntactically and semantically coherent segments, which correspond to elementary illocutionary acts. Boundaries between these units are recognized based on semantic and pragmatic features. Although this general definition of the content-pragmatic units is in good agreement with our definition of SUs (presented below in Section 5.3.3), the author herself admits that boundaries of these units are often ambiguous, and, again, does not give any clear clues how to identify them with satisfactory consistency.

5.2.2 Structural Metadata Annotation Approach

For my work, I have decided to adopt the "Simple Metadata Annotation" approach [105], which was introduced by the LDC as part of the DARPA EARS (Efficient, Affordable, Reusable Speech-to-Text) program [106]. This annotation was defined for the EARS Metadata Extraction (MDE) subtask [107]. The goal of MDE is to create automatic transcripts that are maximally readable. This readability may be achieved in a number of ways: creating boundaries between natural breakpoints in the flow of speech; flagging non-content words like filled pauses and discourse markers for optional removal; and identifying sections of disfluent speech.

The word "simple" in the name of the approach only emphasizes a contrast with an early MDE definition known as "Full MDE". Since a pilot annotation study had revealed a number of problems with consistency of the "full" annotation, LDC developed the "simple" definition that eliminated some annotation tasks entirely and simplified others. As a result, the current MDE annotation can be performed by non-linguist annotators with reasonable consistency.

I have chosen to adopt the simple MDE approach for the following reasons:

- Because of the EARS project, the MDE annotation has become widely accepted as an annotation scheme for spontaneous speech.
- In comparison with other annotation schemes, the MDE system is relatively simple, which makes annotator training easier.
- It is not domain- or style-dependent, so that it can directly be used for both broadcast news and conversations.
- Speech transcripts segmented according to the MDE standard have already been successfully tested in downstream NLP applications (speech summarization, information retrieval, machine translation, etc.).

Although this thesis primarily focus on segmentation of speech into sentence-like units in this work, we decided to create corpora with complete MDE annotation in order to support future research of other spontaneous speech phenomena. Besides its importance to MDE research, the MDE-annotated corpora may also be useful for linguistic analysis of spontaneous Czech. In spite of the fact that the MDE annotation guidelines were primarily designed for spontaneous speech, the same guidelines may also be used for annotation of broadcast news corpora. Planned speech only makes the annotation easier since utterances that are difficult to annotate are much less frequent. Thus, the same guidelines were used for both the RF and BN corpus.

5.3 Structural Metadata Annotation for Czech

Originally, the structural MDE annotation standard was defined for English. When developing structural metadata annotation guidelines for Czech, I tried to follow the LDC guidelines for English as much as possible. However, it would not be correct to simply translate and copy all conventions from one language to another. Individual rules must be adjusted to accommodate specific phenomena of the target language. The language-dependent modifications are mainly based on the description of syntax of Czech compound and complex sentences as given by [108]. I also used two other Czech syntax handbooks [109, 110].

In the following text, all illustrative examples are presented in Czech and then in their English translations. Note that, because of significant differences between Czech and English, it is often impossible to present a good verbatim translation. I tried to use English translations that best illustrate the linguistic phenomena of interest. Also note that SU symbols in all English translations are not displayed as based on the English MDE standard, but rather illustrate their placement with respect to the Czech guidelines. Furthermore, the examples do not contain standard punctuation but only SU symbols. All examples are typed in a typewriter font. If an example represents a conversation, the speakers are distinguished using capital letter IDs (A:, B:).

In all examples in this section, I use a notation that is very similar to the notation used in the original English guidelines [105]. The notation is the following:[4]

Fillers:

`word`	–	"word" is a filler (discourse marker or explicit editing term)
`{ word }`	–	"word" is an Aside/Parenthetical

[4] In this introductory section, I just list the employed MDE symbols. Their meaning is explained below in corresponding sections.

Edit Disfluencies:

`[ word ] *`	–	"word" is a Deletable Region; * denotes an interruption point
<u>`word`</u>	–	"word" is the corrected portion of a disfluency

SUs:

`/.`	–	statement SU break
`/?`	–	question SU break
`/-`	–	incomplete SU break – arbitrarily abandoned
`/∼`	–	incomplete SU break – interrupted
`/&`	–	coordination break
`/,`	–	clause break
∅	–	no break at a place where one might be expected

The remainder of this section is organized as follows. Section 5.3.1 describes annotation of fillers, Section 5.3.2 presents annotation of edit disfluencies, and Section 5.3.3 is devoted to annotation of syntactic-semantic units (SUs).

5.3.1 Fillers

Fillers are words, short phrases, or non-verbal hesitation sounds that do not alter the propositional content of the utterance in which they are inserted. Their characteristic feature is that they do not depend on identities of surrounding words. In general, fillers are those parts of the utterance which could be removed from its transcript without losing any "important" piece of information. Four types of fillers are distinguished within the MDE system:

- Filled Pauses (FP),
- Discourse Markers (DM),
- Explicit Editing Terms (EET),
- Asides/Parentheticals (A/P).

Annotating fillers consists of identifying the filler words and assigning them an appropriate label.

5.3.1.1 Filled Pauses

FP is a non-verbal hesitation sound produced by speakers (either intentionally or not) to indicate uncertainty or to keep control of a conversation while thinking what to say next. In general, FPs can appear anywhere in the flow of speech. By their definition, they make no contribution to the semantic proposition of the utterance. Thus, FPs should not be confused with certain interjections that function to express agreement or disagreement, or as backchannels (such as English *uh-huh*). FP as a linguistic phenomenon was mentioned in Section 2.3.1.1.

An important (and also very interesting) fact about FPs is that they vary across languages. For instance, FPs in American English are known as *uh* and *um*, while Japanese speakers use *ahh*, *ano*, or *eto*, and French talkers most frequently vocalize a sound similar to *euh* [111].

No thorough linguistic study on FPs has been conducted for Czech. Consequently, there is no general agreement on how to transcribe them in text – their transcription differs corpus to corpus. For instance, Kaderka and Svobodová [112] propose to distinguish ten non-verbal sounds, six of which correspond to FPs[5] (*e*, *ee*, *eee*, *eh*, *ehm*, *em*). The first three FPs from this list differ only in length, the other three differ in phonetic makeup. My opinion is that to distinguish six different FPs is too many.

When developing transcription rules for FPs, one must be aware of the fact that there is always a trade off relation between transcription accuracy and consistency. If we choose too many categories, annotators will not be consistent in their recognition. On the other hand, too broad categories might cluster FPs that are wholly different both phonetically and functionally.

In order to be able to design annotation guidelines for Czech FPs, I spent a lot of time listening to Czech spontaneous speech recordings. Based on this experience, I decided to distinguish the following two FP categories:

- *EE* – this FP category is most typically represented by sounds similar to Czech *é*, but in my annotation guidelines, it also includes all hesitation sounds that are phonetically closer to vowels than consonants – for example, sounds similar to Czech long vowel *á* also may function as FPs. Also note that EEs are sometimes accompanied with a creaky voice quality.

- *MM* – this FP category contains all hesitation sounds that are phonetically more similar to consonants or mumble-like sounds. The most frequent hesitation sound from this group is similar to *mmm*. Another not infrequent example of an *MM* is an FP resembling a lengthened Czech consonant *v*. *MMs* typically pronounced with a closed (or almost closed) mouth – openness of mouth is also a good feature distinguishing *MMs* from *EEs*.

Overall, *EEs* are significantly more frequent than *MMs*. Experience with annotation of the two Czech corpora presented herein indicates that these two categories very well cover a vast majority of all FPs occurring in spontaneous Czech. Moreover, our annotators felt comfortable with using these two FP labels. Besides the positive experience, the number of recognized FP categories is also in line with the number of FP categories in American English.

The only problematic instances of FPs in terms of this transcription approach are those similar to *emm*. In such FPs, vowel-like and consonant-like components immediately follow each other. Since such FPs are really rare in spontaneous Czech, I decided not to introduce a special tag for them. However, I had to prepare instructions specifying their transcription. Annotators were instructed as follows. Only the *MM* symbol is used when the vowel-like component is much shorter than a dominant consonant-like component. By analogy, only the *EE* tag is used when the vowel-like component is strongly dominant. When both components are strong, the FP is transcribed using both symbols as *EE MM*. This notation is also used when instances of *EE* and *MM* appear separated by a pause. An example of an FP-annotated utterance follows.

`To je` *`EE`* `jenom` *`EE MM`* `jeho sen /.`

`This is` *`EE`* `just` *`EE MM`* `his dream /.`

Since I did not allow annotators to transcribe FPs using other words or symbols than *EE* and *MM*, the MDE annotation of FPs was in principle performed during the verbatim transcription stage. However, the annotators in the MDE annotation stage had the right to

[5] They do not discriminate between interjections altering content and hesitations in their guidelines.

insert or change FP symbols. I find this two-pass annotation setup useful because FPs are quite often missed by human transcribers.

Note that the presented novel approach to annotating FPs was only used for the RF corpus. The verbatim transcripts of the BN corpus had been created earlier using an FP annotation based on different guidelines. These guidelines paid only little attention to description of FP types. The vast majority of FPs in this corpus were transcribed as *ERs*.[6] Annotators could also use another English FP transcription, *UM*, but this symbol occurs only several times in the transcripts – apparently because FPs of this type are very rare in Czech.

5.3.1.2 Discourse Markers

DMs are words or phrases, such as the well-known "*you know*", that function primarily as structuring elements of spoken language. They do not carry separate meaning but signal such activities as a change of speaker, taking or holding control of the floor, giving up the floor, or beginning of a new topic. There exists a number of diverse definitions of DMs in the linguistic literature. Within MDE, we are only interested in such DMs whose presence in the utterance is unnecessary and whose cleanup do not lead to loss of "important" structuring information. Thus, structuring units such as *"Za prvé, . . . " ("First, . . . ")* do not receive DM labels. If multiple DMs occur in succession, each DM is tagged separately, rather than labeling one long DM spanning over all successive DM instances. An example of DM annotation follows.

Tak já *jako* nevím /.

So I *like* don't know /.

For any language, it is not possible to create a closed list of possible DMs. The of use DMs is dependent on a dialectal variation and rhetorical style of a particular talker. The list of popular DMs in Czech includes: *dobře (well), jako (like), jaksi (sort of), no (well), podívejte se (you see), prostě (simply), tak (so), takže (thus), tedy (then), víte (you know), víte co (you know what), vlastně (actually), v podstatě (basically)*, among others.

Some of the frequent DM words and phrases also have other literal meanings, which sometimes makes identification of DMs more difficult. For example, it is often difficult to decide whether the word *takže* serves as a DM or not. When annotating instances of this word, one must analyze whether the speaker intented to express relation to his/her preceding proposition, or to mark a discourse boundary. Another ambiguous word is *jako*, as illustrated in the following example. In the first sentence, it expresses a comparison, while in the second, it functions as a DM.

Je rychlý jako blesk /.
vs.
To *jako* není nic neobvyklého /.

He is fast like lightning /.
vs.
It is *like* nothing unusual /.

Besides general DMs, the MDE annotation system also recognizes its special case – Discourse Response (DR). DRs are DMs that are employed to express an active response to what another speaker said, in addition to mark the discourse structure. For instance, a speaker may also initiate his/her attempt to take the floor. DRs typically occur turn-initially. Importantly,

[6] *Er* is a British variant of American *uh.*

DRs should not be confused with direct answers to questions. Distinction between DRs and direct responses to questions is discussed below in Section 5.3.3.14. An example of a DR follows.

A: Já bych to tak udělal /.

B: *Hele* já si tím nejsem tak jistej /.

A: I'd do it that way /.

B: *See* I'm not that sure about it /.

5.3.1.3 Asides/Parentheticals

Asides and parentheticals occur when the speaker utters a short side comment and then returns to the original sentence pattern. Asides are comments on a new topic, while parentheticals are on the same topic as the main utterance. For annotation purposes, asides and parentheticals are not distinguished but treated as a single filler type. A/Ps are often prosodically marked. Speakers usually pause or shift their intonation. Strictly speaking, A/Ps are not fillers, but because as with other filler types, annotators must identify the full span of text functioning as an A/P, they are included with fillers in the guidelines. An example of an A/P follows.

Potom k němu přišel { moment musím si vypnout telefon } s tím velkým psem /.

Then he came to him { moment I must switch off my cell phone } with the big dog /.

Some very common Czech words or short phrases that can be denoted as "lexicalized parentheticals" (e.g., *řekněme (say)*,*myslím (I think)*) are not annotated as A/Ps. They usually lack the prosodic features that typically accompany A/Ps. In order to ensure a high IAA, a preliminary illustrative list of those "lexicalized parentheticals" was prepared. In addition, the maximal allowed length of a lexicalized parenthetical was limited to two words.

An important restriction of A/Ps within our MDE guidelines is that they cannot occur as SU-initial or SU-final. Such grammatical parentheticals occurring not in the middle of an SU but on their onset or end should rather be separated by an SU symbol (clausal break or an SU-external break).

Je to sto hlasů i s tím ministrem /, jak jsme dneska četli /.

It's one hudred votes including the minister /, as we read today /.

5.3.1.4 Explicit Editing Terms

Another type of fillers, EET, may only occur accompanying an edit disfluency. EETs are explicit expressions by which speakers signal that they are aware of the existence of a disfluency on their part. Basically, they can appear anywhere within the disfluency, but most frequently occur right after the end of the reparandum. EETs are rather rare in actual conversational language. Typical Czech EETs are e.g., *nebo (or), či (or), spíše (rather), vlastně (actually)*, or *chtěl jsem říct (I wanted to say)*.

Tohle je naše [koherentní]* *EE spíše* konzistentní stanovisko /.

This is our [coherent]* *EE rather* consistent statement /.

5.3.2 Edit Disfluencies

Edit disfluencies are portions of speech in which a speaker corrects or alters his/her utterance, or abandons it entirely. Annotation of edit disfluencies within the MDE scheme respects their structure described in Section 2.3.1, but individual phases of a disfluency are denoted using a slightly different naming convention. Herein, the phases of an edit disfluency are referred to as Deletable Region (DelReg, speaker's initial attempt to formulate an utterance that later gets corrected), interruption point (IP, the point at which the speaker breaks off the DelReg with an EET, repetition, revision or restart), optional explicit editing terms (an overt statement from the speaker recognizing the existence of a disfluency), and correction (portion of speech in which speaker corrects or alters the DelReg). Whereas corrections were not explicitly tagged within the English MDE project, I decided to label them in order to obtain relevant data for further research of spontaneous Czech. Their labeling is not very time consuming and the obtained data may be very useful – some typical correction patterns may be learned. An example of an edit disfluency follows:

Naše děti milují [kočku]* _EE vlastně_ psa pana Krause /.

Our children love [the cat]* _EE actually_ the dog of Mr Kraus /.

Moreover, it often happens that a speaker produces several disluencies in succession, either as serial or nested. In case of serial disfluencies, we simply mark the maximal extent of the disfluency as a single DelReg with multiple IPs that are explicitly tagged.

Ale [ta * myšl- * ten * ten zlej]* ten podivnej pocit to se nedá dobře popsat /.

But [the * ide- * the * the bad]* the strange feeling it can't be described well /.

Nested disfluencies (some component of the disfluency is disfluent itself) are more difficult to annotate. To keep annotation as simple as possible, the MDE standard does not allow using nested disfluency labels, so that all such disfluencies must be annotated using just simple, non-nested DelRegs. The following example shows a correction that contains an additional disfluency.

Přijel jsem [do Brna]* do [Plz-]* Plzně dnes ráno /.

I arrived [to Brno]* to [Pil-]* Pilsen today morning /.

Since Czech disfluencies have the same pattern as English, the rules about complex disfluencies from [105] can basically be directly applied to Czech. Thus, I do not survey all particular rules for annotating complex disfluencies herein because interested readers may consult the original English guidelines.

5.3.3 SUs

Dividing the stream of words into sentence-like units is a crucial component of the MDE annotation. The goal of this part of annotation is to improve transcript readability and processability by presenting it in small coherent chunks rather than long unstructured turns. Because speakers often tend to use long continuous compound sentences in spontaneous speech, it is nearly impossible to identify the end-of-sentence boundary with consistency using only a vague notion of a "conversational equivalent" of a written sentence – strict segmentation rules are necessary. Past experience with similar annotation problems indicates that acceptable inter-annotator agreement (IAA) can only be achieved in the context of rules grounded in "surface features", i.e. mainly syntax and prosody. Semantic features have not proved to be reliable.

One possible solution to the "conversational sentence" definition problem is to divide the flow of speech into "minimal meaningful units" functioning to express one complete idea on the speaker's part. It means that we divide the stream of words wherever it is grammatically possible and meaningful. The resulting units are either shorter or equally long as sentences in standard writing. These smaller units also seem to be convenient for downstream automatic applications. For example, speech translation applications usually prefer to process shorter segments [113].

The target utterance units are called SUs within the MDE task. In the MDE definition, the abbreviation SU may stand for one of the following possibilities: Sentential/Syntactic/Semantic/Slash Unit. Every word within the discourse is assigned to an SU (each word contained between two SU boundaries is considered part of the same SU), and all SUs are classified according to their function within the discourse. The following list shows the employed SU symbols (breaks) along with brief descriptions of their meaning:

- /. – Statement break – end of a complete SU functioning as a declarative statement
  (`Kate loves roses /.`)
- /? – Question break – end of an interrogative
  (`Do you like roses /?`)
- /, – Clausal break – identifies non-sentence clauses joined by subordination
  (`If it happens again /, I'll try a new cable /.`)
- /& – Coordination break – identifies coordination of either two dependent clauses or two main clauses that cannot stand alone
  (`Not only she is beautiful /& but also she is kind /.`)
- /- – Incomplete (arbitrary abandoned) SU
  (`Because my mother was born in Russia /, I know a lot about the /-`
  `They must fight the crime /.`)
- /∼ – Incomplete SU interrupted by another speaker
  (`A: Tell me about /∼      B: Just a moment /.`)

In contrast to the English MDE, we do not use an SU symbol for backchannels because both Czech corpora are single-channel. Therefore, backchannels that do not overlap with words uttered by the dominant speaker, and thus can be captured in a single-channel transcript[7], are treated as a special type of a filler.[8] In the illustrative examples below, I also use a special symbol "∅" which denotes "no break" at places where one might be expected. This symbol is only used for illustration purposes herein, and does not occur in real MDE annotations.

The SU symbols may be divided into two categories: sentence-internal (/& and /,) and sentence-external (others). Sentence-external breaks are fundamental and directly support the SU research task. They are used to indicate the presence of a main (independent) clause. These independent main clauses can stand alone as a sentence and do not depend directly on the surrounding clauses for their meaning. Sentence-level breaks may also appear after a short

[7] Overlapping backchannels are treated as noises since they cannot be explicitly transcribed within the dominant speaker turn.

[8] Since these non-overlapping backchannels are extremely rare in the corpus, I did not present their annotation in a separate section.

phrase that nonetheless functions as a "complete" sentence. In many cases, these breaks would be represented in standard writing with end-of-sentence punctuation. Sentence-internal breaks are secondary and have mainly been introduced to support IAA. However, it should be noted that it is important to have these symbols in the MDE annotations since some future task may require them to be automatically detected. Sentence-internal breaks delimit units that are smaller than a main clause and cannot stand alone as a complete sentence. In standard writing, these breaks often correspond to commas.

In SU annotation, the fundamental problem is to determine when to insert a new SU boundary and when to place two segments within the same SU. External breaks are inserted between SU boundaries, internal breaks (if exist) may further refine each SU. Besides a few exceptions[9], candidate locations for both sentence-internal and sentence-external SU labels are usually boundaries between two adjacent clauses. Thus, the key problem is to recognize the type of each clause boundary.

The above presented set of SU symbols corresponds to the original MDE standard. However, I did not use it as it was originally defined but introduced two significant modifications. Both modifications are language-independent. First, the original set contains only one symbol for incomplete SUs, but I propose to distinguish two types of incomplete SUs: /- — indicating that the speaker abandoned the SU arbitrary; and /∼ — indicating that the speaker was interrupted by another speaker. This distinction of incompletes is very useful since their patterns differ significantly in prosody, semantics, and syntax.

Second, in order to identify some "core boundaries" that could be both easier to detect automatically based on prosodic cues, and also relevant for spontaneous discourse analysis, I introduced two new symbols: //. and //? — the double slashes indicate a strong prosodic marking on the SU boundary, i.e. pause, final lengthening, and/or strong pitch fall/rise. The additional annotation refinement does not seem to cause a corresponding growth in annotation complexity. A rule of thumb instructs annotators to use the double-slash SU symbols when in doubt. IAA for this additional subtask is evaluated in Section 5.4.1. Note that, in contrast to ToBI-like systems, our system only involves labeling prosodic boundaries on SU boundaries, rather than on all word boundaries, which is much less time-consuming.

The proposed guideline modifications did not only include changes in the SU symbol set. Another modification pertains to the pause threshold. In the English SimpleMDE V6.2 standard, in order to support IAA, the pause longer than 0.5 sec automatically induces the end of a speaker turn and thereby requires a corresponding SU-external break. But the 0.5 sec threshold is problematic because some speakers produce long pauses in places where other speakers might produce filled pauses. Hence, I decided to drop the threshold rule and to rely solely on syntax. Likewise, I do not require the presence of a noticeable pause after incomplete (abandoned) SU breaks (/-) when syntax provides an overt evidence of incompleteness.

The following subsections provide descriptions of particular rules for SU annotation. To keep the description reasonably long, I only present the most important examples – especially those that emphasize differences between Czech and English. Full annotation guidelines may be found at `http://www.mde.zcu.cz`.[10]

5.3.3.1 Short Stand-alone Phrases Not Containing Verbs

SUs do not necessarily have to contain a verb. Even though some phrases do not constitute grammatically complete sentences, they may function as a complete utterance. To identify

[9] These are mentioned in the following sections.

[10] Unfortunately, the full guidelines are available only in Czech.

them correctly, annotators must be sure that the phrases are not syntactically connected with the previous SU.

```
Vítejte u Radiofóra /. Hosté Jan Novák poslanec a Pavel Kučera stínový
ministr obrany /.
```

```
Welcome to Radioforum /. Guests Jan Novák a deputy and Pavel Kučera a
shadow minister of defense /.
```

These stand-alone phrases often occur following a question – talkers sometimes repeat the question's topic to establish common ground before answering.

```
A: Jsou pro vás Glasgow Rangers těžkým soupeřem /?
B: No Rangers /. My jsme s losem spokojeni /.
```

```
A: Are Glasgow Rangers a tough opponent for you /?
B: Well Rangers /. We are satisfied with the draw /.
```

Another examples are headline news in broadcast news data which should also be annotated as individual SUs even if they do not form a complete clause.

```
Mušaraf neoficiálním vítězem pákistánských prezidentských voleb /.
Německo ochromeno stávkou železničářů /. Madelaine Albrightová
v exkluzivním interview pro Českou televizi /.
```

```
Musharraf the unofficial winner of Pakistan's presidential vote /.
Germany paralyzed by rail strike /. Madelaine Albright in an exclusive
interview for the Czech TV /.
```

All these rules are identical to the corresponding English rules.

5.3.3.2 Juxtaposition of Clauses

In general, juxtaposition means an absence of linking elements in a group of words that are listed together. As juxtaposition of an "introductory clause", we understand a connection of two main clauses that cannot be classified using any of the standard semantic relations defined by normative Czech grammar (copulative, disjunctive, etc.). The second clause is syntactically and semantically determined by the preceding clause, but there is no formal syntactic relationship. Thus, it is a kind of parenthetical clause in terms of grammar. In most of the cases, such clauses could be connected using the Czech conjunction *"že (lit. that)"* without any change of meaning. In English, this phenomenon does not have a separate rule since, unlike Czech, dropping of the conjunction *that* is standard. Czech guidelines instruct annotators to separate the clauses in juxtaposition using a clausal break.

```
Já vím /, vy to nemáte rád /.
```

```
I know /, you don't like it /.
```

5.3.3.3 Quotations

Since no quotation marks are used within the MDE annotation, direct or indirect quotations impose clausal breaks. The quote and its attribution typically form a single SU. This rule is identical to the corresponding English rule.

`Půjdu tam /, řekl David /.`
`Martin řekl /, že tam nepůjde /.`

`I'll go there /, David said /.`
`Martin said /, that he wouldn't go there /.`

If the quote is long and the quoted portion of the utterance contains several sentences, additional SUs are recognized, as shown in the following example:

`Ale premiér řekl /, nikdy jsem ho neviděl /.  Já toho člověka vůbec neznám /.  Tahle aféra je směšná /.`

`But the prime minister said /, I have never seen him /.  I don't know that man at all /.  This affair is ridiculous /.`

5.3.3.4 Idiomatic Expressions

Similarly to English, frozen idiomatic expressions are not separated by any SU symbols even if they contain multiple finite verbs.

`To je` ⊘ `prašť` ⊘ `jako uhoď /.`

`It is` ⊘ `hit` ⊘ `or punch /.`

5.3.3.5 Independent Subordinate-like Clauses

A complete SU may also be composed of stand-alone independent clauses starting with subordinate conjunctions. In these cases, the subordinate conjunctions basically functions as particles rather than conjunctions.

`Protože toto je opravdu jednoduché /.`

`Because this is really easy /.`

5.3.3.6 Parcelation

In spontaneous speaking, speakers often do not precisely plan the structure of their utterances in advance. As a result, we sometimes observe discontinuous appending of additional utterance constituents. The talker composes several successive elliptic utterance units, which typically have separate focal accents. This phenomenon is referred to as parcelation in Czech literature. If this parcelation is strong (which is typically recognized from short pauses between constituents), the utterance is segmented into multiple SUs.

`Chceš s ním mluvit /?  Sama /?  Beze svědků /?`

`Do you want to speak with him /?  Alone /?  Without witnesses /?`

5.3.3.7 Appositions

Apposition is a grammatical construction in which two elements are placed side by side, with one element serving to define or refine the other. In standard Czech writing, these constituents are typically separated by a comma, however, in the MDE annotation, we do not separate them by any SU breaks.

`Daniel Mach ⊘ ředitel místní školy ⊘ je můj přítel /.`

`Daniel Mach ⊘ the local school principal ⊘ is my friend /.`

In most cases, noun phrases form appositions. However, based on the broad definition of appositions (as defined by Vladimír Šmilauer), verbs, adverbs, or even clauses may form appositions, too. If a clause embedded in another clause appears in a apposition, we separate it by a clausal break.

`Řeka se kroutí /, tedy tvoří meandry /, blízko u lesa /.`

`The river twirls /, thus it forms meanders /, close to the wood /.`

I should also mention special introductory phrases such as *“to jest (*lit. *that is)”* or *“to znamená (*lit. *it means)”* (i.e., frozen phrases containing a finite verb), which frequently accompany clausal appositions. In terms of MDE, these phrases should not be understood as clauses but rather as introductory particles. As a result, they do not motivate any SU breaks. An illustrative example follows.

`Řeka se kroutí /, to znamená ⊘ tvoří meandry /, blízko u lesa /.`

`The river twirls /, it means ⊘ forms meanders /, close to the wood /.`

In contrast to the previous examples, if a clause that seems to be appositional is not embedded in another clause and may stand alone, it should be annotated as an independent SU. This is in agreement with the rule of thumb for problematic decisions – “segment wherever it is possible!”.

`Důkazy byly takové /, že soudy je osvobodily /. To znamená zbavily je toho obvinění /.`

`There were such evidence /, that the court set them free /. It means they found them not guilty /.`

5.3.3.8 Anacolutha

An anacoluthon in spoken language can be defined as an abrupt change of syntax within an utterance. In other words, an utterance begins in a way that implies a certain logical resolution, but concludes differently from the form grammar leads us to expect. An example of an anacoluthon in Czech is a disagreement between subject and predicate within a clause. Although anacolutha can also be used as a purposeful stylistic virtue, they more frequently occur as a consequence of an unintentional grammatical fault in conversational language. In the Czech corpora, anacolutha often occur in the vicinity of parentheticals and asides, as shown in the following example.

Pokud se skupina států {a nejsou to jenom Spojené státy je to také Velká Británie a další} **rozhodnou** použít sílu /, tak ...

If the group of states {and it's not just the United States it's also Great Britain and others} **decide** to use power /, then ...

In the example above, to match the singular subject *"skupina (group)"*, Czech grammar strictly requires to use the singular *"rozhodne (decides)"* instead of the plural *"rozhodnou (decide)"*. However, the speaker got confused by using a plural form in the parenthetical, and continued to use the plural form in the completion of the original message.

Annotators were instructed not to use any special annotation for these "small" anacolutha, pretending that grammar in the disturbed utterances is correct. However, note that anacolutha should not be confused with edit disfluencies.

5.3.3.9 Tag Questions

Tag questions are short phrases added to the end of a statement in order to appeal to the listener to give feedback. In terms of MDE, the statement plus the added phrase form a single SU that should be labeled as interrogative. The added phrase is separated from the preceding statement using a clausal break. This rule is identical to the corresponding English rule.

To si děláte legraci /, že jo /?

You must be joking /, aren't you /?

However, if intonation gives a clear clue that the added phrase does not function as a question, the whole SU is labeled as a statement and the added phrase is labeled as a DM.

Přišli tam včera *jo* /.

They came there yesterday *yeah* /.

5.3.3.10 Embedded Questions

When a question is embedded in a larger carrier clause, SU type is assigned according to the function of the whole utterance, and not according to the embedded question. Embedded questions most frequently occur in quoted direct speech. This rule is also identical to the corresponding English rule.

Zeptala se /, přidáš se k nám (?) /.

She asked /, will you join us (?) /.

5.3.3.11 Incomplete SUs

When a speaker's utterance does not express a complete thought, an incomplete SU is recognized. Boundaries of the incompletes are labeled with either "/–" or "/∼". If the utterance is interrupted and cut short by another speaker, then the "/∼" symbol is used. On the other hand, if the speaker abandons his/her utterance arbitrarily, the SU is annotated as "/–". It implies that "/∼" may only occur at a turn boundary, whereas "/–" may also occur as turn-internal. In standard text, "/–" may correspond to ellipses (. . .). The first example illustrates the use of "/∼":

```
A: Pokud vložíte dostatek peněz do /∼
B: Ale to není jen otázka peněz /.

A: If you put enough money into /∼
B: But it is not just a matter of money /.
```

The second example illustrates the use of "/–":

```
Jeho boty vypadaly jako /- On je divnej kluk /.

His shoes looked like /- He is a weird guy /.
```

5.3.3.12 Distinguishing Incomplete SUs and Restart Disfluencies

Incomplete SUs are sometimes difficult to distinguish from restart disfluencies, which do not receive incomplete SU labels, but are annotated as DelRegs. The distinction between these two phenomena within the MDE standard is based on the following rules. In comparison with the original MDE standard for English, the rules for Czech are slightly more complex.[11] Our experience indicates it does not lead to a significant decrease of IAA, and the accuracy of annotation is increased.

1. Restart disfluency may never appear as turn-final. In such cases, the incomplete utterance is always identified as an incomplete SU.
2. If the speaker immediately restructures the interrupted utterance and continues speaking on the same topic, restart disfluency is recognized. On the other hand, if he/she does not return to the incomplete message, an incomplete SU is recognized.
3. Incomplete SUs of the "/–" type must always contain either one or more SU-internal breaks (/& or /,), or "useful information" that is not repeated in the same turn. However, this does not mean that the occurrence of an SU-internal break within an incomplete utterance automatically implies the use of "/–". When the SU-internal break occurs in a very short introductory phrase such as *"víte /, že (you know /, that)"*, it is possible to annotate it as a DelReg (if other necessary conditions are met).

5.3.3.13 Turns with Missing Onsets

Turns whose onsets are missing in the verbatim transcripts, or whose onsets are transcribed within the immediately preceding overlapping speech section, are annotated in the same way as if their onsets were present. Note that the overlapping speech regions in the verbatim transcripts were not used for MDE annotation, and thus they do not contain any MDE symbols. An example of such a turn follows. The first line in the example corresponds to an overlapping speech region (both A and B speak), in the second line, the speaker B continues the utterance that was started in the overlapping region.

```
A: Překvapil vás.  B: Na druhou stanu překvapil ∼ (OVERLAP)
B: ∼ i mě /.  Já jsem k tomu už názor vyjádřil /.

A: He surprised you.  B: On the other hand he surprised ∼ (OVERLAP)
B: ∼ me as well /.  I have already stated my opinion on this /.
```

[11] In the original MDE standard, incomplete SUs are only recognized if "a speaker is interrupted or when the speaker trails off, failing to complete the utterance within a turn". Thus, incomplete SUs can only occur at the end of a speaker's turn.

5.3.3.14 Direct Responses Expressing Agreement or Disagreement

Direct responses to questions expressing speaker's agreement or disagreement such as *ano (yes), ne (no), jo (yeah), m-hm (uh-huh)* typically form a complete SU.

```
A: Bude to hotové do příštího týdne /?
B: Ano /.  Pevně v to doufám /.
```

```
A: Will it be finished by next week /?
B: Yes /.  I strongly hope so /.
```

If a subordinate clause having an explanatory function is attached to words expressing agreement or disagreement, a clausal break is used.

```
A: Zkusíte to /?
B: Ne /, protože už je příliš pozdě /.
```

```
A: Will you try that /?
B: No /, because it's too late now /.
```

One must also be aware of the fact that words that often express agreement or disagreement may also function as discourse markers. The discriminative rule for these ambiguities says that both agreement and disagreement words must *always* be preceded by a question. Otherwise, a DM is recognized. Although this simplification is not absolutely accurate in terms of discourse analysis, it was introduced in this simplified form in order to support IAA. The use of this rule is illustrated in the following example presenting a part of a fictitious dialog.

A: Myslím /, že se to stane /.
B: Ano /?
A: Ano /. Ten příkaz už je podepsaný /.
B: *Ano tak* to je problém /.
A: *Ano* ⊘ je to opravdu nepříjemné /.
B: *Takže* oni přijdou /? a *jo* ⊘ odnesou všechno /?
A: Ano /. Je mi to líto /.

A: I guess /, that this will happen /.
B: Yes /?
A: Yes /. The order has already been signed /.
B: *Yes so* it's a problem /.
A: *Yes* ⊘ it is really bothersome /.
B: *So* will they come /? and *yeah* ⊘ take everything /?
A: Yes /. I am sorry /.

Note that nonverbal sounds such as *uh-huh* may also function as direct responses to yes/no questions. In that case, they also form a complete SU.

```
A: Je to v pořádku /?
B: HM /.  Pojďme dál /.
```

```
A: Is it ok /?
B: Uh-huh /.  Let's move on /.
```

5.3.3.15 Subordinate Clauses within Complex Sentences

Dealing with complex and compound sentences represents one of the most important parts of the MDE annotation. This section describes annotation of complex sentences that contain some kind of subordination. Subordinate clauses cannot themselves constitute a complete SU because they depend on the rest of the sentence and thus may not stand on their own; they are semantically linked to their main clauses. Subordinate clauses are separated by clausal breaks within MDE.

```
Já ti tu adresu dám /, když mi zavoláš /.
I will give you the address /, if you call me /.
```

If there is not just a single subordinate clause but two subordinate clauses dependent on the same independent clause and joined by coordination, a coordination break is used to separate these two dependent clauses. An SU-external break cannot be applied since neither of these subordinate clause can stand on its own without changing meaning of the whole statement.

```
Já ti tu adresu dám /, když mi zavoláš /& nebo pošleš email /.
I will give you the address /, if you call me /& or send me an email /.
```

Unlike English, relative clauses are separated by clausal breaks in the Czech MDE. This adjustment reflects Czech syntax which requires to separate relative clauses by commas, regardless whether they are restrictive or not. If we did not use clausal breaks for relative clauses, the MDE transcripts would be less transparent for the annotators.

```
Daniel /, který se narodil v Praze /, miluje Karlův most /.
Daniel /, who was born in Prague /, loves the Charles bridge /.
```

5.3.3.16 Compound Sentences

Compound sentences consist of two ore more main (independent) clauses joined by coordination. As described above, the goal is to divide the compound sentences within a spoken discourse into “minimal meaningful units” functioning to express a complete idea. It means that we split independent clauses into two complete SUs every time they can stand alone (i.e. they do not depend on each other for completion of an idea). The potential break point is the interword boundary right before the coordinating conjunction as shown in the following example.

```
Adam hraje tenis /. a Robert cvičí jógu /.
Adam plays tennis /. and Robert practices yoga /.
```

However, not all cases are that clear as the one in the example above. In some cases, coordinated main clauses cannot be split into two independent SUs. In such cases, a coordination break is used instead of an SU-external symbol. In English MDE, this situation most frequently arise when the second coordinate clause has a dropped subject. In English, subject dropping is only allowed in the second clause of a compound sentence when both clauses share the same subject. It implies that such compound sentences cannot be divided into two SUs because coordinated main clauses with dropped subjects do not form syntactically encapsulated units, and thus cannot stand alone. See the difference in the following illustrative example.

```
I love volleyball /. but I hate playing with beginners /.
vs.
I love volleyball /& but hate playing with beginners /.
```

However, this rule cannot be applied to Czech. In contrast with English, Czech subjects (pronouns) can be dropped every time they are "understood" from context and/or from the form of a conjugated verb (predicate). Thus, since the conjugation of the verb includes both person and number of the subject, it is possible to say for instance just "*Běžím /.*" which means "*(I am) running /.*" This phenomenon of subject dropping is typical for highly inflective languages.

For the above stated reason, subject dropping in the coordinated clause does not imply the use of the coordinating break alone, as is the case for English. Instead, we separate the coordinated clauses with an SU-external break, even if the subject is present in the first clause and dropped in the second clause:

```
Robert do práce šel pěšky /. ale domů jel vlakem /.

Robert walked to work /. but (he) took the train home /.
```

In the Czech MDE, a coordination break is used for separation of coordinated main clauses in the following cases:

1. The compound sentence is structured with a non-continuous expression such as *sice – ale (though – but), buď – nebo (either – or)*, or *nejen – ale i (not only – but also).*

   ```
   Ona je nejenom krásná /& ale také je laskavá /.
   Not only she is beautiful /& but also she is kind /.
   ```

2. The second coordinate clause is elliptical and cannot stand alone.

   ```
   Katka miluje kosatce /& ale Eva tulipány /.
   Katka loves irises /& but Eva tulips /.
   ```

3. There exists a subordinate clause that is dependent on both main clauses.

   ```
   Když byl hotov /, zavřel okno /& a sedl si na postel /.
   When he was finished /, he closed the window /& and sat on the bed /.
   ```

4. Main clauses are joined by the syntactically primarily coordinating yet semantically often rather subordinating conjunction *neboť (for).*

   ```
   Šli jsme se koupat /& neboť bylo krásné počasí /.
   We went swimming /& for the weather was great /.
   ```

The rule No. 1 was adopted from the English MDE. The rules No. 2 and 3 are not explicitly mentioned in the English guidelines, however, the correct annotation of these phenomena should be the same as for Czech. The last rule in the list is specific for Czech.

5.3.3.17 Coordinate Questions

When two questions are coordinated within a compound sentence, both of them receive the question label.

```
Půjde Robert do divadla /?  a Adam zůstane doma /?
Will Robert go to the theater /?  and will Adam stay at home /?
```

However, it is very common to drop an auxiliary verb in the second interrogative clause, which, in contrast, induces the use of a coordination break.

```
Bude Robert v divadle /& a Adam doma /?
Will Robert be in the theatre /& and Adam at home /?
```

5.3.3.18 Compound Predicates

Another important fact influencing the Czech MDE is that Czech syntax discriminates between compound sentences sharing a single common subject and simple sentences with compound predicates (i.e. compound predication in a simple sentence). The compound predicate is defined as a "tight unit" of two or more predicate verbs predicating on the same subject. On the other hand, if the predicate verbs do not form such a "tight unit", a compound sentence is recognized. Unfortunately, there is not absolute agreement in the literature on the exact borderline between compound predicates and compound sentences. For the MDE purposes, I only recognize those compound predicates that can be identified based on unambiguous features. Within Czech MDE, the compound predicate is recognized if:

1. The predicate verbs share a common constituent (e.g., object).

   ```
   Nacpal /& a zapálil si dýmku /.
   He filled /& and lit up his pipe /.
   ```

2. The predicate verbs joined by a copulative conjunction have the same or very similar meaning.

   ```
   Naši hosté často slaví /& a radují se /.
   Our guests often rejoice /& and celebrate /.
   ```

While compound predicates did not motivate any SU breaks according to the initial version of the annotation guidelines, the current version instructs annotators to separate parts of compound predicates by a coordination SU break. A preliminary analysis showed that the redefined annotation rule supported IAA.

5.3.4 Technical Aspects of MDE Annotation

The two Czech MDE corpora were annotated just by two annotators. Since Czech syntax is quite complex, naive annotators could not be employed; at least some linguistic education is necessary and such annotators are quite difficult to find. The small number of labelers slowed down the annotation process, but, on the other hand, it supported annotation consistency. Moreover, the submitted annotations were checked by the author of this thesis. The MDE annotations of the easier BN corpus were checked on the basis of a random sample, while for the more difficult RF data, all submitted annotations were carefully revised.

To ease the annotation process, a new annotation software has been developed. It was designed to reflect the particulars of the Czech annotation task. As with LDC's MDE Annotation Toolkit [114], the Czech tool allows annotators to highlight relevant spans of text, play corresponding audio segments, and then record annotation decisions. The screen of this tool is shown in Appendix A. The screenshot also shows a longer stretch of the MDE annotated Czech data.

5.4 Analysis of Czech Structural Metadata Corpora

This section presents and discusses some interesting statistics about structural metadata in the two MDE annotated Czech corpora. It is organized as follows. The first subsection is devoted to IAA on the Czech MDE annotation task. The following three subsections report particular corpus statistics relating to fillers, disfluencies, and SUs, respectively. The numbers are compared with available corresponding numbers relating to English MDE corpora (mostly taken from [115]). However, note that while these comparisons are interesting, they do not allow to draw fundamental conclusions about structural differences between spoken Czech and English since there is not a perfect match in genre and speaking style. The Czech BN corpus is compared with the English BN MDE corpus, and the RF corpus with the English CTS corpus. Quite a good match is expected for BN data, but when comparing Czech broadcast conversations with English telephone speech, one must also take into account significant differences in the speaking styles. Broadcast conversations are more formal and less interactive.

5.4.1 Inter-Annotator Agreement on Czech MDE Annotation

This section focuses on testing annotation reliability in terms of IAA. If we employ good annotators and the annotation task is well-defined in the guidelines, different annotators should consistently generate similar annotations. To estimate annotation consistency, we typically use a random corpus sample that is independently labeled by two or more human annotators and measure the degree of IAA on this sample [116].

In general, it is not convenient to directly measure the percentage of judgments on which the annotators agree when coding the same data independently. The absolute agreement numbers do not yield values that give a good notion about the quality of annotation since some agreement is always due to chance. Therefore, we should measure agreement above chance in order to receive meaningful IAA numbers. To this end, we can employ the K (kappa) statistic [117] which is considered to be a standard measure of agreement in many annotation tasks related to language processing. This agreement measure is defined as

$$K = \frac{A_o - A_e}{1 - A_e} \tag{5.1}$$

where A_o denotes the observed agreement and A_e the expected (chance) agreement. The interpretation of K is not completely straightforward. We must be aware of the fact that measuring IAA is not equivalent to hypothesis testing, so that there is no theoretic cut-off value as well as no clear probabilistic interpretation. K evaluates the magnitude of agreement rather than compares two hypotheses. In the original paper presenting this method [117], Carletta claims that for tasks like content analysis, $K > 0.8$ is considered to be good reliability, and $0.67 < K < 0.8$ allows to draw tentative conclusions.

For Czech MDE, we tested IAA on three recordings from the more difficult RF corpus. These recordings were dually-annotated by two experienced annotators. The total duration of these recordings was 86 minutes, the total number of tokens was 13k. For SU breaks, we

Table 5.2: Frequencies of filled pauses

	BN	**RF**
% of words followed by FPs	0.5%	3.8%
Proportion of *EEs*	N/A	93.1%
Proportion of *MMs*	N/A	6.9%

got $K = 0.88$ taking into account all SU types (both internal and external). For the key task annotation – "SU-boundary" vs. "no SU-boundary" – we got $K = 0.92$. For filler and disfluency labels, we got overall $K = 0.85$. Given the complexity of this annotation task, these numbers seem to be very well acceptable. Unfortunately, we could not compare our IAA with IAA for the English MDE since the numbers for English are not publicly available.

We also measured IAA for the additional Czech MDE annotation subtask – labeling of prosodic strength of SU boundaries. Measuring IAA on "/. vs. //." on the same dually-annotated recordings, the following consistency values were achieved. If the two annotators agreed on using a statement break after a particular word, they also used the identical symbol (/. or //.) in 86.9 % of the cases. In terms of the kappa statistics, we got $K = 0.69$ when only taking into account the words followed by a statement break (i.e. ignoring all words with no statement label). For comparison, for "/. vs. //. vs. other", we obtained $K = 0.85$.

5.4.2 Statistics about Fillers

Table 5.2 reports numbers relating to occurrences of FPs. As expected, FPs are significantly more frequent in conversational than in broadcast news data. Another important observation is that *EE* FPs are much more frequent than *MMs* – they represent more than 93 % of FPs. Note that this comparison is only available for the RF corpus since the Czech BN corpus uses different rules for transcription of FPs (cf. Section 5.3.1.1). For comparison, 2.2 % of words is followed by an FP in the English CTS corpus, and 1.4 % in the English BN corpus. A relatively smaller number of FPs in English CTS data might be explained by three different factors. First, transcribers of the English database could have missed a number of FPs since some of them are less audible and telephone data are more noisy. Second, Czech syntax is more complex than English, thus speaking in Czech represents a more complex mental process which may cause a higher number of hesitations. Third, talkers may hesitate by voice more when speaking in public. In private conversations, people often do not care about being grammatically correct, which makes speech planning easier. On the other hand, the larger percentage of FPs in English BN data is caused by the fact that these data contain a larger proportion of speech having a relatively higher level of spontaneity. Commercial TVs and radios are in minority in the Czech BN corpus, and broadcast news on Czech public radio and TV channels has significantly less interactive style than typical American broadcast news.

Table 5.3 shows numbers of words labeled as DMs or DRs. As with FPs, this kind of fillers is more common in conversational speech – just 0.1 % of words is labeled as a DM or a DR in the BN data. An interesting statistic to observe is the proportion of DMs and DRs. In the RF corpus, "non-DR" DMs prevail, whereas the DR subtype covers over 53 % of all DMs in the BN corpus. The explanation for this observation may be the following. DMs are not frequently used by anchors in the studio, they are more typical for local reporters referring on actual events directly from their venues. These reporters typically react on questions coming from the studio and their interactive replies contain a number of DRs. English MDE data contain a higher number of DMs than Czech data – 4.4 % in CTS and 0.5 % in BN speech.

Table 5.3: Proportions of discourse markers and discourse responses

	BN	RF
% of words in DMs and DRs	0.1%	1.6%
Proportion of DMs	46.7%	73.3%
Proportion of DRs	53.3%	26.7%

Table 5.4: Most frequent discourse markers (overall, in the RF corpus, and in the BN corpus)

DM	Overall	BN	RF	DM	Overall	BN	RF
tak	17.0%	32.8%	15.8%	*tedy*	5.7%	4.1%	5.8%
no	13.2%	25.4%	12.2%	*jako*	5.0%	7.8%	4.8%
prostě	12.9%	4.5%	13.6%	*čili*	4.8%	0.8%	4.8%
vlastně	7.1%	3.7%	7.4%	*ano*	3.8%	2.9%	3.9%
jaksi	7.0%	2.5%	7.4%	*teda*	2.9%	2.0%	2.9%

Table 5.4 shows ten most frequent DMs in the Czech data. The most frequent DM was *tak (so)* followed by *no (well)* and *prostě (simply)*. The DM *tak* was first in both corpora (but more dominant in the BN corpus), while *no* came second in the spontaneous corpus, and *prostě* came second in the news corpus. Note that all frequent DMs consist of just one word. The table indicates that most significant difference are in DMs as *prostě*, *vlastně (actually)*, and *jaksi (somehow)*, which are more frequent in the RF corpus. The reason is that DMs of the DR subtype prevail in the BN corpus, while the three mentioned DMs typically occur as turn-internal. Another interesting observation is that DMs containing a verb are much less frequent in Czech than in English. Although there exist some Czech equivalents of the popular English DM *you know* (such as *víte* or *víte co*), only a minority of speakers really use them.

Table 5.5 displays frequencies and average lengths of A/Ps and EETs. A/Ps are relatively frequent in the RF corpus, containing 1.5 % of all uttered words. On the other hand, A/P structures are quite rare in planned speech; they cover just 0.2 % of words. For comparison, English CTS data contain only 0.3% of A/Ps, which supports the hypothesis that A/Ps are more frequent in conversational Czech than in conversational English. The statistics also indicate that A/Ps in conversational speech are on average approximately one word longer.

As expected, EETs were really rare. In total, they include just 0.08 % of words in the RF data and 0.01 % in the BN data. These numbers are pursuant to English MDE corpora where EETs represent 0.05 % and 0.02 % of words, respectively. Since average lenghts of EETs are 1.2 and 1.1 words, respectively, it is possible to ratiocinate that one word EETs are strongly dominant. Table 5.6 shows the most frequent Czech EETs. For both corpora, we can observe that by far the most frequent EET is *nebo (or)*; it represents more than two thirds of all EET instances. The second most frequent EET is *respektive (let's say, respectively)*.

The total proportions of filler words (i.e. sum of all FPs, DMs, A/Ps, and EETs) significantly differ between the two Czech corpora. While they represent just 0.79 % of all words in the BN corpus, they cover 6.97 % of words in the spontaneous RF corpus. For English MDE corpora, this total filler percentage is 9.2 % for CTS, and 2.1 % for BN data.

5.4.3 Statistics about Edit Disfluenscies

Statistics relating to edit disfluencies are presented in Table 5.7. As with fillers, disfluencies were much more frequent in the spontaneous corpus, where 2.8 % of words were labeled as

Table 5.5: Statistics about A/Ps and EETs

	BN	RF
% of words in A/Ps	0.2%	1.6%
Average length of an A/P (in words)	5.4	6.2
% of words in EETs	0.01%	0.08%
Average length of an EET (in words)	1.1	1.2

Table 5.6: Most frequent explicit editing terms (overall, in the RF corpus, and in the BN corpus)

EET	Overall	BN	RF	EET	Overall	BN	RF
nebo	69.6%	66.7%	69.9%	*vlastně*	1.9%	0.0%	2.1%
respektive	3.7%	6.7%	3.4%	*prostě*	1.9%	0.0%	2.1%
tedy	3.1%	0.0%	2.7%	*pardon*	1.9%	6.7%	1.4%
teda	2.5%	0.0%	2.7%	*ne*	1.9%	0.0%	2.1%

within a DelReg. The corresponding percentage in Czech BN speech was just 0.2 %. Likewise, edit IPs were ten times more frequent in the RF corpus than in the BN corpus. In the compareable English corpora, edit disfluencies were more frequent. DelRegs covered 5.4 % of words in CTS, and 1.5 % in the English BN data. Again, the explanation for this is in different speaking styles.

Another interesting numbers refer to occurrences of corrections and EETs within edit disfluencies. The proportion of DelRegs having a correction is dependent on the frequency of restart disfluencies since this disfluency type does not contain a correction. Because restarts are typical for spontaneous speech, the relative number of DelRegs having corrections is accordingly smaller in the RF corpus. As expected, EETs were very rare in both corpora. Only approximately 4 % of disfluencies contain an EET. The next statistics in Table 5.7 express average lengths of DelRegs and their corrections. It is possible to see that short disfluencies predominate; their average length is around 1.5 word in both corpora. Another interesting observation that should be pointed out is that corrections have almost the same average length as DelRegs.

Furthermore, notable statistics are those referring to the total portion of data marked for the potential automatic cleanup. This number is obviously correlated with the complexity of the MDE task for a particular corpus. Hence, I summed words in DelRegs and fillers and compared the two Czech MDE corpora. As expected, the results were largely unequal – 9.8 % for the RF corpus while just 1.1 % for the BN corpus. For comparison, it was 17.7 % for the English CTS and 3.8 % for the English BN corpus.

Unlike previous work, I also analyzed DelRegs and corrections in terms of which parts of speech they typically contain. To my best knowledge, this is the first analysis studying the POS content of speech disfluencies in any language. Both Czech corpora were tagged using a state-of-the-art automatic morphological tagger based on the averaged perceptron method [118]. Of the positional Czech morphological tags[12], I only used the first positions that correspond to the POS information. For either Czech corpus, I computed three POS distributions corresponding to the whole corpus, DelRegs, and corrections, respectively.

The relative frequencies of particular POSs are shown in Figure 5.1. The top chart represents the BN corpus, the bottom chart the RF corpus. POS on the x-axis are sorted according

[12] The positional Czech tagset is described in Section 10.4.1.3 on page 121.

Table 5.7: Statistics about edit disfluencies

	BN	**RF**
% of words followed by Edit IPs	0.2%	2.0%
% of words within DelRegs	0.3%	2.8%
% of DelRegs having Correction	94.6%	83.8%
% of DelRegs having EET	3.5%	4.0%
Average length of DelRegs (in words)	1.4	1.6
Average length of Corrections (in words)	1.5	1.6

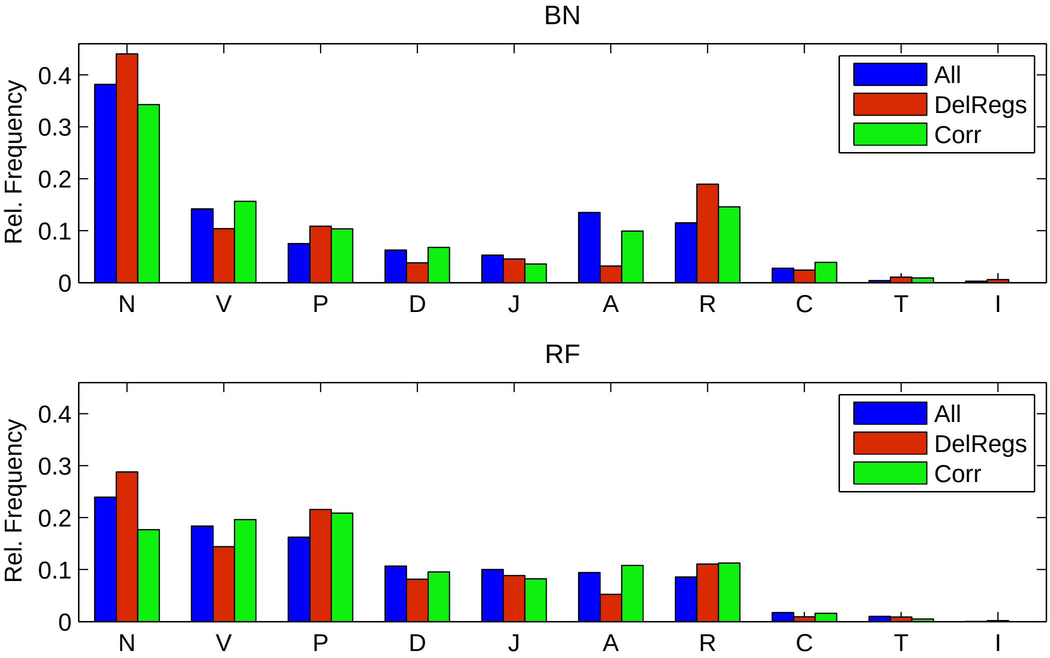


Figure 5.1: Relative frequencies of Czech POS in all data, DelRegs, and corrections in the BN (top) and the RF corpus (bottom). (POS legend: N – Nouns, V – Verbs, P – Pronouns, D – Adverbs, J – Conjunctions, A – Adjectives, R – Prepositions, C – Numerals, T – Particles, I – Interjections)

to their frequencies in the RF corpus. The blue bars show that the corpora differ in overall POS distributions. The BN corpus contains significantly more nouns and adjectives, while the conversational corpus shows distinctively higher relative numbers of pronouns, adverbs, and conjunctions, and a slightly higher proportion of verbs. These observations may be clarified by differences in speaking styles. Broadcast news data consist of sentences that were prepared to be as informative as possible, and thus contain a lot of nouns and adjectives. On the other hand, conversational language is characterized by a more complex way of locution. A higher number of complex and compound sentences logically implies a higher number of conjunctions, while numerous discourse markers having the form of adverbs cause the higher proportion of that POS type.

Despite the differences in overall POS distribution, both corpora show similar changes in the distribution when only the words in DelRegs are taken into account. The proportion of nouns, pronouns, and prepositions is increased, while verbs, adverbs, and adjectives are less frequent in comparison with the general distribution. The increased number of nouns in DelRegs is not surprising. Speech disfluencies more frequently occur in more informative

Table 5.8: Relative frequencies of SU symbols (both SU-internal and SU-external)

SU break	BN	RF	SU break	BN	RF
/.	6.7%	15.1%	/-	0.0%	0.4%
//.	**60.2%**	28.9%	/∼	0.6%	2.7%
/?	0.3%	0.7%	/&	2.9%	6.7%
//?	1.3%	3.4%	/,	28.1%	**42.2%**

regions of utterances and nouns usually carry information more densely than, for instance, verbs. A higher frequency of prepositions is consequent because prepositions are dependent on nouns.

Interesting variances may also be observed between DelRegs and their corrections. The most prominent difference is in the higher rate of adjectives within corrections. This fact indicates that speakers often put in the adjectives omitted in DelRegs. We can also observe slightly higher frequencies of verbs and adverbs. On the contrary, nouns, conjunctions, and prepositions are less common than in DelRegs. The difference in the proportion of prepositions is only visible in the BN corpus.

5.4.4 Statistics about SUs

Relative frequencies of all SU symbols (both SU-external and SU internal) are displayed in Table 5.8. This table shows that both corpora significantly differ in SU distributions. The symbol "/," is most frequent in the RF corpus, whereas "//." is strongly dominant in the BN corpus. This indicates that complex and compound sentences are more common in spontaneous conversations, while prearranged broadcast news typically consist of statements with simpler clause syntax.

Another substantial findings relate to differences between relative frequencies of single- and double-slash SU-symbols. Overall, the double-slash symbols are more frequent. The contrast is more distinctive in BN data, where "//." is almost ten times more frequent than "/.". This fact is not surprising since sentence boundaries are usually attentively prosodically marked by professional newscasters. In the RF corpus, "//." is approximately twice as frequent as "/.".

The statistics about incomplete SUs indicate that incompletes are much more common in conversational speech. Furthermore, incomplete SUs interrupted by another speaker (/∼) are more frequent than arbitrarily abandoned statements (/–). The latter type of incompletes almost never appears in broadcast news.

Table 5.9 shows average lengths of all, complete, and incomplete SUs in both Czech corpora. The numbers indicate that SUs in the conversational corpus are slightly longer. For comparison, the English CTS corpus has mean SU length 7.0 words and the English BN corpus 12.5 words. The average SU length for broadcast news corpora is similar for both languages, whereas the average SU lengths in the the compared spontaneous corpora strongly differ. The average segment length in conversational data is largely affected by backchannels. There is a large number of short backchannel SUs in the English telephone corpus, while backchannel SUs are not taken into account for the single-channel Czech RF corpus.

Table 5.10 reports average lengths of particular SU subtypes. These numbers indicate several interesting findings. First, statement SUs are longer than interrogative SUs. Furthermore, double slash SUs (i.e., SUs with a strong prosodic marking at their boundaries) are significantly longer than corresponding one slash SUs. Note that this difference in length is more prominent for the BN data.

Table 5.9: Average length (in number of words) of complete and incomplete SUs

	BN	RF
All SUs	13.0	14.5
Complete SUs	13.0	14.7
Incomplete SUs	10.2	11.9

Table 5.10: Average length (in number of words) of individual SU subtypes

SU type	BN	RF	SU type	BN	RF
/.	7.8	12.6	//?	9.9	12.6
//.	13.7	16.1	/–	11.3	15.2
/?	8.7	11.5	/∼	10.2	11.4

In addition, I also analyzed what happens with the average SU length when single slash symbols are not considered to be SU-external but rather SU-internal boundaries. This hypothetical modification of the SU boundary definition lead to increase of average "//." SU length from 16.1 to 21.9 words for the BN data, and from 13.7 to 14.6 words for the RF corpus. The mean length of double slash interrogative SUs rose from 12.6 to 15.7 words, and from 11.3 to 13.2 words, respectively. Note that these increases in length are more prominent for spontaneous speech data where double slash boundaries are less frequent.

5.5 Chapter Summary and Conclusions

In this chapter, I have described the creation of Czech speech corpora annotated with structural metadata. Two corpora from two different domains were created – one in the domain of broadcast news (mostly read-aloud speech) and the other in the domain of broadcast conversations (mostly spontaneous speech). The first corpus was created by enriching an existing corpus with the structural metadata annotation, while the second was built from scratch – it had to be recorded and manually transcribed first.

The structural metadata annotation was based on the LDC's "Simple Metadata Annotation Specification", originally defined for English. The original guidelines were adjusted to accommodate specific phenomena of Czech syntax. Moreover, I proposed and used a novel approach to transcribing and annotating filled pauses in Czech, distinguishing vowel-like (*EE*) and consonant-like (*MM*) sounds. In addition to the necessary language-dependent modifications, I applied some language-independent modifications refining the original annotation scheme. The refinements included limited prosodic labeling at sentence unit boundaries and distinction of two types of incomplete units.

Furthermore, I have presented a comparison of Czech broadcast news and broadcast conversations in terms of MDE statistics relating to fillers, edit disfluencies, and SUs. The comparison is useful for evaluation of complexity of the Czech MDE task in the two genres. Moreover, it provides interesting data about domain-specific speaking styles. It also shows some cross-linguistic differences because I not only compared the two Czech corpora with each other, but also with the available numbers relating to existing English MDE corpora.

Among others, the comparison shows that the total proportion of filler words (i.e., the sum of all FPs, DMs, A/Ps, and EETs) is significantly higher in the RF corpus (6.97 % of words) than in the BN corpus (0.79 %). Likewise, edit disfluencies are much more frequent in the RF corpus (2.8 % of words within DelRegs in the RF and 0.2 % in the BN). I also found that DelRegs and their corrections show differences in POS distributions in comparison with the general POS distribution. Regarding SU symbols, I observed that clausal breaks are more frequent in the RF corpus, which indicates that complex sentences are more common in talk shows than broadcast news. Furthermore, I found that SUs in the conversational broadcast data are on average longer by 1.5 words than SUs in broadcast news.

In order to have more data for training and testing automatic systems for spontaneous Czech, the RF corpus is currently being extended by 20 additional talk show recordings. After this extension, the RF corpus should yield over 33 hours of transcribed and MDE annotated broadcast conversation speech. I should also mention that both Czech MDE corpora are planned to be made publicly available – their publication at the LDC is currently being prepared.

Speech corpus development is a very time-consuming and labor-intensive process which cuts down the time available for other research tasks. However, I hope that all the effort put into it paid off and the corpora created as part of this work will be of benefit for the spoken language processing community. Besides their importance to automatic structural metadata extraction research, the two new MDE corpora should also be useful for training ASR systems as well as for linguistic analysis of read-aloud and spontaneous Czech. Finally, note that the corpora described in this chapter were used for automatic SU segmentation experiments that are described in Chapter 10.

Chapter 6

Prosodic Features for Classification

The power of speech is not in words, but in the oration.
Srijit Prabhakaran

The previous chapter was devoted to the creation of data resources. This chapter begins the thesis part concerning the automatic sentence segmentation system development. First of all, I take a closer look at the use of prosodic features. Some important qualitative aspects of prosody were described in Section 2.2. This chapter focuses on methods for extraction of prosodic features for automatic classification. Thus, prosody is viewed in terms of a quantitative description herein. In this chapter, rather than giving an exhaustive enumeration of all employed prosodic features, I describe motivations and general techniques used for their design and extraction. The overall number of designed features is quite high – a complete list of all implemented features, along with their brief descriptions, is given in Appendix B.

The prosodic features for sentence segmentation of speech reflect breaks in temporal, intonational, and loudness contours, and are inspired by linguistic knowledge. To use prosody in automatic systems effectively, all of the prosody processing must be performed automatically. In the employed approach, prosodic information is obtained from a combination of the speech signal and speech recognition output, which is used to provide word and phone alignments. All employed features are designed to be automatically extractable, without need for any human labeling. First, a huge set of potentially useful features was extracted, which, after initial investigations, was pared down to a smaller set by eliminating redundant features.

Most of the implemented prosodic features were inspired by [49, 119, 120]. The prosodic features can be grouped into broad feature classes based on the prosodic quality they are designed to capture – *pause*, *pitch*, *duration*, and *energy*. In addition to the pure prosodic features, the automatic classifiers also have access to a limited number of "other" features, capturing important phenomena such as speaker change or overlap. Since these features are usually incorporated into the prosody model for machine learning purposes, their description is also included in this chapter.

The remainder of this chapter is structured as follows. Section 6.1 depicts feature extraction regions. Section 6.2 describes pause features. Section 6.3 is devoted to pitch preprocessing and pitch features. Section 6.4 presents duration features, Section 6.5 describes energy features, and Section 6.6 overviews "other" features. Section 6.7 presents the method used for feature space reduction. Finally, Section 6.8 summarizes the chapter.

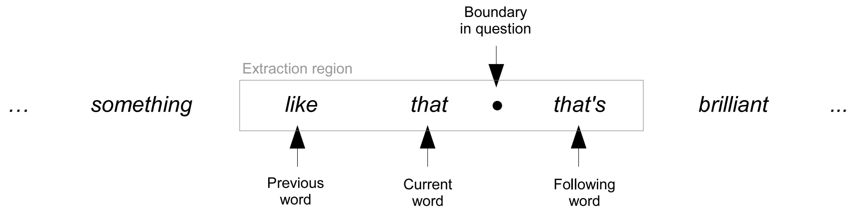


Figure 6.1: Prosodic feature extraction region spanning the previous, the current, and the following word

6.1 Feature Extraction Regions

As in previous work on automatic sentence segmentation, I have only used local prosodic features. The reasons for this decision are quite straightforward. As mentioned in Section 3.1, a number of linguistic studies based on listening tests showed that most of important prosodic information is coded locally. Moreover, the local features are much easier to extract. Although in principle one may consider longer regions, it is not obvious how such long range prosodic features should be designed. This problem exceeds the scope of this thesis, however, it is definitely an interesting direction for possible future research.

All prosodic features employed in this work were extracted on the word level, i.e. the features are associated with particular interword boundaries. However, the classifiers do not only have access to features relating to the boundary after the current word, but to capture local prosodic dynamics, the classifiers also use features associated with boundaries after the previous and the following word. The feature extraction region is depicted in Fig. 6.1. Typically, one feature describes prosodic information contained in the word before the referential boundary, or captures prosodic differences between the two words across the boundary. Word boundary timestamps are obtained using an automatic speech recognizer. Throughout this thesis, I use the following notation. Names of the features relating to *previous* words are prefixed with "*p.*", names of the features relating to *following* words with "*f.*", and the features relating to *current* words have no prefix.

6.2 Pause Features

Pauses signal breaks in prosodic continuity, so they are very important indicators of sentence unit boundaries. Intuitively, if the pause between the current word and the following word is long, it is more likely that the current word is followed by a sentence boundary. Pause features are quite robust in the face of speech recognition errors since, in principal, their value is only dependent on correct speech/non-speech segmentation. Pause durations can easily be extracted from automatic word alignments. If there is no pause at the boundary, which is the most frequent case in continuous speech, the pause duration feature is output as zero. I do not only use the pause duration at the actual boundary; the pauses at the preceding and the following interword boundary may be important as well. Their duration may indicate short backchannels or reflect whether speech right before the boundary was just starting up or continuous in a longer region. The question is whether the pause features should be normalized per speaker or not. My preliminary experiments showed that raw durations perform better. This finding is in agreement with [49].

6.3 Pitch Features

As mentioned in Section 2.2, perceived pitch is a psychoacoustic quantity that cannot be directly measured from the speech signal. Instead, we measure the fundamental frequency of the glottal tone (F_0), which is its correlate. It is the frequency at which vocal chords oscillate when producing voiced sounds. On the other hand, unvoiced regions of speech do not contain this periodic component.

6.3.1 Pitch Tracking

Several methods for measuring F_0 from the speech signal have been proposed. One of the popular methods is the RAPT algorithm (Robust Algorithm for Pitch Tracking) [121]. This method employs Normalized Cross Correlation Function (NCCF) to generate pitch period candidates. This function is more robust to quick F_0 changes than the formerly frequently used autocorrelation function. The NCCF is defined as

$$\phi_{i,k} = \frac{\sum_{j=m}^{m+n-1} s_j s_{j+k}}{\sqrt{e_m e_{m+k}}}, \qquad k = 0, K-1;\ m = iz;\ i = 0, M-1 \tag{6.1}$$

where m is the sample number, i is the index of the current frame, k is the index of the delay, z the frame shift, M the total number of frames, and n is the size of the analyzed window. The normalization factor e_j is defined as

$$e_j = \sum_{l=j}^{j+n-1} s_l^2 \tag{6.2}$$

$\phi_{i,k}$ ranges in $\langle -1, 1 \rangle$, values close to 1 signal that the delay kT is an integer multiple of the period $\frac{1}{F_0}$.

The idea of the RAPT method is the following. First, the input signal is downsampled and the peaks (areas of interest) are found in this coarse version of the signal. The values of the NCCF are then computed from the original signal only in these areas of interest. All peaks found in this stage are then viewed as F_0 candidates for the current frame. Finally, by using a dynamic programming method, we either select the most probable candidate value or mark the frame as unvoiced. The selection of the most probable candidate also takes into account typical properties of F_0 curves.

6.3.2 Octave Error Correction

Despite the fact that the RAPT algorithm is quite robust, it is, as well as other pitch tracking methods, prone to *octave errors*. For instance, noise and/or a creaky or gaspy voice may cause tracking errors. In such cases, the algorithm often outputs halved or doubled F_0 values. Reasons for the halving (subharmonic) errors are the following. If the signal is T-periodic, it is also $2T$, $3T$, etc. -periodic. The NCCF values for the multiples of T (subharmonic components) may be high and confuse the tracking algorithm. These errors are frequently caused by a creaky voice quality.

The doubling (harmonic) errors have the following reasons. If the component M dominates in the energy of the speech signal, the correlation score for the period T/M becomes high. For example, if this harmonic component is equal to the first formant frequency, it causes a resonance effect and this component is amplified. The doubling errors have $M = 2$.

Most of these errors are successfully detected during the RAPT's dynamic programming stage, especially if the speech signal is not noisy. However, the pitch tracking algorithm output

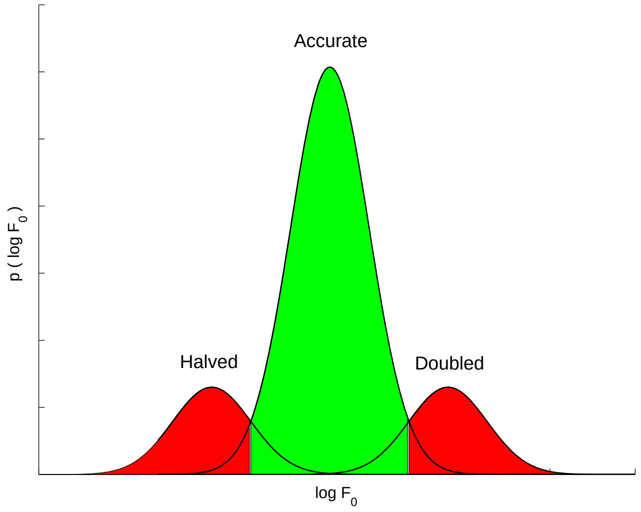


Figure 6.2: LTM model for $\hat{F_0}$

usually still contains a number of halved/doubled values so that postprocessing techniques should additionally be applied. For example, a postprocessing algorithm called *de-step* filter was introduced by Bagshaw [122]. This simple method is based on the assumption that F_0 values of two contiguous frames may differ at most by 75%, while the first value of a voiced region may change arbitrarily with respect to the last value of the previous voiced region. Since I have not employed this technique, I do not describe it here in detail, but interested readers may consult the above cited publication.

For this work, I have adopted a pitch postprocessing technique that is based on using a speaker's longterm F_0 distribution. Sönmez et al. showed that correct F_0 values approximately have a lognormal distribution [123]. The measured pitch containing octave errors can be modeled using the Lognormal Tied Mixture (LTM) model with three Gaussian components. Individual components correspond to halved, accurate, and doubled F_0 values, respectively. It is assumed that all three components have the same variance and their mean values are shifted by $\log 2$ in the log domain. The model is illustrated in Fig. 6.2. In a mathematical notation, the LTM model can be expressed as

$$\begin{aligned} \log(\hat{F_0}) \sim LTM(\mu,\sigma,\lambda_1,\lambda_2,\lambda_3) = \quad & \lambda_1 \cdot \mathcal{N}(\mu - \log 2, \sigma^2) + \lambda_2 \cdot \mathcal{N}(\mu, \sigma^2) + \\ & + \lambda_3 \cdot \mathcal{N}(\mu + \log 2, \sigma^2) \end{aligned} \tag{6.3}$$

In addition, the model must satisfy the constraints $\lambda_i \geq 0,\ i = 1,2,3$ and $\sum_{i=1}^{3} \lambda_i = 1$. The parameters of the model are estimated using the Expectation-Maximization (EM) algorithm [13]. The F_0 values marked as halved or doubled may be replaced by interpolated values, or, alternatively, multiplied by 2 or $\frac{1}{2}$, respectively. The borderline between the halved and the accurate component corresponds to speaker's baseline F_0 [49]. This value is important for normalization of F_0 values, as will be described below.

6.3.3 Automatic Pitch Contour Stylization

F_0 contours consist of two components – *macroprosodic* and *microprosodic*. The macroprosodic component represents F_0 changes intented by the speaker. On the contrary, the microprosodic component is not purposely controlled by the speaker, but is influenced by the local segmental (phonetic) content. Since we are only interested in the intentional prosody, it is advisable to remove microprosody before analyzing the pitch contour. First, the measured pitch is usually filtered by a median filter. The median filtering is similar to low pass filtering, but unlike ordinary low pass filters, the median filter preserves the integrity of sudden transitions to a different level in the signal. A typical size of the median window for pitch filtering is 5 or 7.

The median filtered contours are typically further processed – *stylized*. The pitch stylization methods are inspired by the 't Hart's idea of the "close copy". The basic idea is to replace the measured pitch contour by a contour consisting of line fits perceptually undistinguishable from the original contour. For TTS systems, D'Alessandro and Mertens introduced a perceptually-based stylization method [124]. Their method was based on so-called *glissandos*, which are defined as audible pitch differences. In another approach, Hirst and Espesser interpolated pitch contours by quadratic splines [125]. Later work of the same authors extended the approach and introduced a more complex algorithm called MOMEL [126].

The stylization by quadratic splines may be convenient for some TTS-related tasks. However, for tasks related to sentence segmentation, it is more convenient to smooth the contour by lines since slopes of the smoothing lines may themselves be important features for classification. Within the VERBMOBIL project, Batliner and his colleagues interpolated all measured F_0 values within a voiced region by a single line using linear regression in the log domain [11]. In this thesis, I have adopted a more sophisticated method in which the pitch contour is stylized by a piece-wise linear (PWL) function [127].

In order to get the stylized contour, it is necessary to determine coordinates of the nodes connecting the line fits in each voiced region. The number of nodes may either be determined from the duration of the voiced region, or chosen to warrant local smoothness of the stylized contour. In this work, I used the latter option, with a fixed minimum length of a region. For a voiced region to be modeled by K line segments, the free parameters are $(x_k, y_k)_{k=0}^{K}$, excluding the x-coordinates x_0 and x_k which are given by the edges of the voiced region. The stylization function is then given by

$$g(x) = \sum_{k=1}^{K} (a_k x + b_k) \boldsymbol{I}_{[x_{k-1} < x \leq x_k]} \tag{6.4}$$

where a_k is the slope b_k the intercept of the line defined by the node (x_k, y_k). The coordinates of the nodes are determined by minimizing the Mean Square Error between the stylized contour and the measured F_0 values

$$MSE\left((x_k, y_k)_{k=0}^{K}\right) = \frac{1}{T} \sum_{t=1}^{T} (F_0(t) - g(t))^2 \tag{6.5}$$

The minimization can be performed numerically, e.g. using the Nelder-Mead simplex method [128]. An example of the stylized pitch contour is displayed in Fig. 6.3. A diagram of the whole pitch preprocessing procedure is shown in Fig. 6.4.

6.3.4 Pitch Normalization

Raw, unnormalized F_0 values do not convey information about their relative positions within the pitch register of the current speaker. For instance, the same F_0 value may represent a relatively high pitch value for one speaker (a male speaker with a deep voice), but a relatively low pitch value for another (a female speaker with a high voice). Likewise, absolute F_0 differences do not convey accurate information about linguistically meaningful pitch rises and falls since the magnitudes are dependent on the pitch range of the actual speaker. Thus, some form of pitch normalization is necessary.

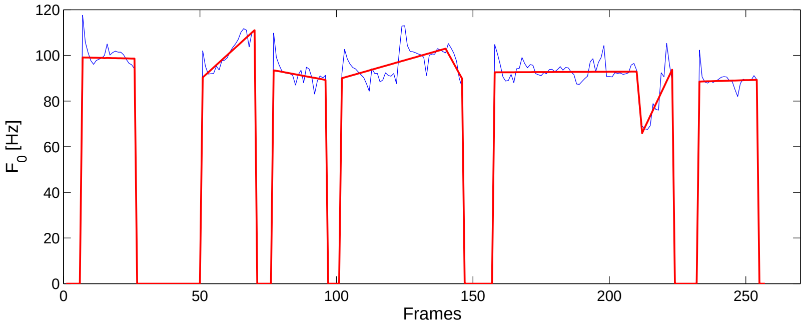


Figure 6.3: Raw and stylized F_0 contours

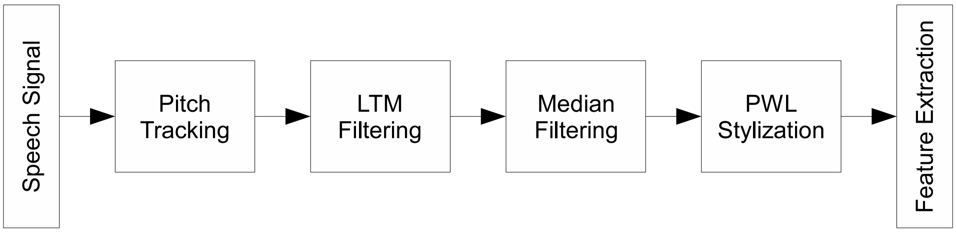


Figure 6.4: Scheme of pitch preprocessing for feature extraction

A convenient reference value for the pitch normalization is the F_0 baseline, mentioned at the end of Section 6.3.2. The baseline for a particular speaker corresponds to the lowest pitch value in the non-halved mode. It is the point at which the probability of an accurate pitch is equal to the probability of halving within the LTM model. Using the notation from Section 6.3.2, the baseline (BL) value can be computed as

$$BL = \exp\left(\mu - \frac{1}{2}\log 2 + \frac{(\log\lambda_1 - \log\lambda_2)\,\sigma^2}{\log 2}\right) \quad [Hz] \tag{6.6}$$

Past work also investigated using F_0 toplines and means for normalization, but these values proved to be by far less convenient [49]. The use of baselines for normalization is also in agreement with the findings of Rietveld and Vermillion [129], who showed that listeners best estimate the speaker's pitch register from low tones. The low tones show more stable frequencies than high tones.

F_0 values can be normalized with respect to the baseline value using various approaches. I have implemented the following normalizations – linear difference, linear ratio, log difference, and log ratio. The use of logarithms, in addition to linear differences and ratios, is motivated by the fact that humans perceive pitch changes logarithmically. All four types of normalized pitch features are available to the machine learning algorithms, the most useful normalization form is determined in the feature selection phase (described below in Section 6.7).

Note that this form of pitch normalization assumes that speakers are recurring and the identity of the actual speaker is known at the time of feature computation. For some applications, this condition does not represent an additional problem (e.g., meetings with participants recorded by head-worn microphones), while for others, employment of an automatic speaker identification and tracking system is required. Since the state-of-the-art speaker diarization systems perform with diarization error rates (defined as percentage of missed speech + percentage of false alarm speech + percentage of speech by mislabeled speakers) in the range of 10–20% (depending on the target domain), a slight degradation of the normalized pitch features may be expected.

6.3.5 Pitch Feature Types

The implemented pitch features may be grouped into three broad classes according to prosodic phenomena they aim to capture – *pitch range*, *pitch reset*, and *pitch slope*. The group of range features reflects the pitch range of a single with respect to the speaker's global F_0 statistics. These features are based on computation of minimum, maximum, mean, first, and last F_0 values within a single word. While some raw pitch features were also implemented, to achieve good results with a speaker-independent system, the above mentioned F_0 normalization is necessary.

The second set of features was designed to capture the pitch reset phenomenon. Typically, we compare F_0 values in the last voiced region before the interword boundary with the corresponding values in the first voiced region after the boundary. The computed features comparing the current and the following word included the ratio and the difference between the two values, both in the linear and the log domain.

The final group of pitch features looks at the slopes of the stylized F_0 segments. It either captures melodic trends in the word before the interword boundary, or it checks trend continuity across the boundary. For the latter subgroup, it is expected that discontinuous ("broken") trajectories would tend to indicate boundaries, regardless of the absolute difference in F_0 values across adjacent words. Implemented slope features include the first and the last slope in the word, and the difference between the last slope before the boundary and the first slope in the word following the boundary.

6.4 Duration Features

Another important prosodic cues to sentence boundaries in speech are changes in the speaking rate. Duration features primarily aim to capture the phenomenon of preboundary lengthening, which was mentioned in Section 3.1. A variety of duration features can be computed using hypothesized time marks from the speech recognizer. The implemented duration features relate to three different basic units – phonemes, rhymes, and words.

6.4.1 Duration Normalization

Similarly to pitch features, feature normalization techniques have to be applied. It is obvious that raw phoneme durations are inapplicable since every phoneme has a different standard duration. For example, Czech consonant *r* is on average almost twice shorter than another Czech consonant *c*. Likewise, Czech "short"[1] vowel *a* is on average 1.8 times shorter than its "long" counterpart *á*.

Two basic normalization techniques were applied. The first was using the z-score, which transforms a random variable into a random variable having normalized Gaussian distribution with zero mean and unit variance. Thus, z_{dur} for an unit u is defined as

$$z_{dur(u)} = \frac{dur(u) - \mu_{dur(u)}}{\sigma_{dur(u)}} \tag{6.7}$$

[1] Czech phonology discriminates between short and long vowels, which form minimal pairs. The length is an important distinctive feature since it differentiates various word meanings. The vowel length is independent of the stress.

where $\mu_{dur(u)}$ is the mean and $\sigma_{dur(u)}$ the standard deviation of the duration of the phoneme u.

The alternative method only uses the mean durations for normalization, avoiding potential errors caused by noisy estimates of standard deviations. The normalized duration is then expressed as

$$norm_{dur(u)} = \frac{dur(u)}{\mu_{dur(u)}} \tag{6.8}$$

If the unit u represents a single phoneme, it is possible to apply the above stated formulas directly. On the other hand, if we are interested in the duration normalization of longer units, such as rhymes or words, it is not possible since there is not enough data to reliably estimate statistics μ_{dur} and σ_{dur} for the larger units. Thus, it is necessary to capture the normalized duration using duration statistics of particular phonemes contained in the larger unit. Disregarding the influence of coarticulation, it is possible to approximate $\mu_{dur(w)}$ of a larger unit $w = (p_1, p_2, \ldots, p_N)$ consisting of phonemes p_i as

$$\mu_{dur(w)} = \sum_{i=1}^{N} \mu_{dur(p_i)} \tag{6.9}$$

Another issue is the computation of larger unit z-scores. One possibility how to perform it is to make a simplifying assumption that phoneme durations are independent random variables and approximate the standard deviations using the well-known formula

$$\sigma^2_{dur(p_1+p_2)} = \sigma^2_{dur(p_1)} + \sigma^2_{dur(p_2)} \tag{6.10}$$

I have chosen an alternative approach which does not estimate $\sigma_{dur(w)}$, but computes z-score of a larger unit as an average of z-scores of individual phones. This approach seems to be more robust. Then, the formula for the computation of the z-score becomes

$$z_{dur(w)} = \frac{1}{N} \sum_{i=1}^{N} \frac{dur(i) - \mu_{dur(i)}}{\sigma_{dur(i)}} \tag{6.11}$$

For all normalized duration features, both speaker-independent and speaker-specific normalized versions were extracted. The former versions use data from the whole speech corpus, whereas the latter versions only use data of the particular speaker for normalization.

6.4.2 Duration Feature Types

On the phone level, only vowel duration features were implemented. Vowels contribute much more to the overall speaking rate than consonants. In addition, they are also more robust against ASR errors. A number of various vowel features was computed. Intuitively, the most important feature is the normalized duration of the last vowel before the boundary. However, other vowel features were also extracted. For instance, the longest normalized vowel in the words reflects lengthening of prefinal syllables in multisyllabic words. In addition, vowel duration maxima, minima, and averages were computed, as well as a number of other duration statistics for individual vowels in the word.

Rhyme duration features are motivated by the fact that preboundary lengthening particularly affects the nucleus and coda of syllables. Moreover, unlike syllable-based features, which represent a possible alternative, the rhyme-based features do not require an automatic syllabifier. Similarly to phoneme-based features, both speaker-independent and speaker-dependent normalizations were performed. The normalizations were carried out using the method for normalization of larger unit durations described in the previous section.

Word-level features reflect durations of whole words. Again, the same normalization techniques were applied. An interesting issue is the use of raw, unnormalized word durations. We should be very careful when employing them. Although they usually aid performance of the prosodic classification, they correlate with lexical features that should be modeled in a language model. For example, certain frequent short sentences (especially backchannels) have small set of words, so raw durations may capture those words rather than prosody.

6.5 Energy Features

Energy features aim to capture loudness patterns. It is expected that talkers tend to begin their utterances aloud and gradually taper off. Since loudness is a psychoacoustic quantity, which cannot be directly measured from the speech signal, we use the short term RMS (Root Mean Square) energy instead. The RMS is defined as

$$RMS = \sqrt{\frac{1}{K}\sum_{k=1}^{K} s_k^2} \tag{6.12}$$

where s_k denotes the signal samples and $k = 1, \ldots, K$ indices of the samples in the window.

The problem with the energy-based features is that they are less reliable because of the channel variability. Moreover, there is also some redundancy because of correlation with pitch features – they are generated using the same physiological mechanisms during speech production. Thus, it is often very difficult to get some gain from using them.

The used energy features represent mean, minimum, and maximum RMS values in a single word. These values were extracted using two approaches. The first approach only uses values from the voiced frames, while the second uses all RMS values. The normalized variants of these features were computed by dividing the raw values by mean RMS values for the current turn.

6.6 Other Features

Some additional automatically extracted features are also included. These features describe phenomena such as turn-taking or speaker overlaps. Although they are not inherently prosodic, they are put into the group of prosodic features for modeling purposes. Since they may influence prosodic quantities, it can be advantageous for prosodic classifiers to have access to them in order to model possible interactions.

For example, turn-related features include the flags indicating whether the current word is the first or the last word in the current turn. This information is extremely important for the prosodic classifiers since pause durations as well as other important features are not defined at turn boundaries. It is obvious that speaker changes are strong indicators of sentence boundaries. Another turn feature records the time elapsed from the start of the turn.

When dealing with multi-channel data, it is also possible to add speaker overlap features. Overlaps are frequent in natural conversations. In a study of overlap in two-party and multiparty conversations, it was shown that 30% to 50% of all speech spurts (regions of speech in which a particular speaker does not pause for more than half a second) include at least one frame of concurrent speech by another speaker [130]. Since talkers not holding the floor predict the end of the current speaker's turn and often start speaking before the current speaker finishes, overlap information might be helpful in automatic sentence segmentation.

The implemented overlap features include the number of speakers on other channels overlapping with the current word as well as the number of spurt-initial, spurt-internal, and spurt-final overlaps, where the spurt position is meant with respect to speakers on other than the current channel. The spurt-position-based features aim to capture information that says in which phase of their spurt the overlapping speakers are at the moment. All overlap features are implemented as integer-valued (number of overlapping features) but may easily be converted into binary features indicating whether an overlap occurs or not.

6.7 Feature Selection

As mentioned above, the overall feature set was designed to be as exhaustive as possible. A number of features are highly correlated, differing only in the normalization approach. Because of the greedy nature of some statistical classifiers, using large feature sets may yield suboptimal results. Furthermore, redundant features negatively affect computational efficiency and result interpretability. Thus, a feature space reduction is desirable. The feature selection method used in this work is motivated by prosodic knowledge. To reduce the feature space, I first combined similar features into groups, and then selected the features from each group that were most frequently used in a first set of decision trees.

Individual features were grouped according to their prosodic category (pause, pitch, duration, energy, other) and the reference word (previous, current, following). Moreover, the feature capturing pause duration after the current word was included in each group. Past experience has indicated that this feature is essential and always selected for the final feature set. It also functions as a catalyst during the selection process because without it, duration, pitch, and energy features are much less discriminative. In addition, all F_0 subsets were analyzed in two versions. The first one contained just normalized features, while the second contained both normalized and raw values. This bifurcation was proposed in order to avoid masking effects of some good normalized features by the greedy raw features. Thus, for example, one feature subset included pause after the current word plus all normalized pitch features relating to the word following the boundary of interest. Another example of a feature group could be the set comprised of pause plus all duration features referring to the current word.

Subsequently, a set of decision tree classifiers was trained from each subset using an ensemble sampling approach (described below in Section 7.1.4.3) and average relative feature usage across the trees was inspected. All features that showed greater usage than an (empirically estimated) threshold were passed on to form a new feature set containing prosodic features of all categories. This subset, already much smaller than the huge original set, was further pared down by eliminating redundant features using the leaving-one-out approach.

6.8 Chapter Summary

This chapter has described prosodic features employed in this thesis. I use local features extracted from the window spanning the previous, the current, and the following word. The prosodic features can be divided into groups based on what quantities they capture – pause, pitch, duration, energy, and "other". Pause features are essential for automatic sentence segmentation since pauses are the strongest indicators of sentence boundaries. They are also very robust since their extraction only relies on speech/non-speech segmentation.

On the other hand, pitch-related features are largely dependent on accurate F_0 contour preprocessing. The preprocessing steps involve a removal of halved and doubled values based on an LTM model, median filtering, and pitch contour stylization by a piece-wise linear function.

The implemented F_0 features capture pitch ranges, pitch resets, and slopes of the stylized contour.

Duration features primarily aim to capture the phenomenon of preboundary lengthening. The implemented duration features relate to three referential units – phonemes, rhymes, and words. The features are normalized with respect to phone duration means and standard deviations. Energy features aim to capture loudness patterns – talkers often tend to gradually taper off toward utterance unit boundaries. The features are computed based on the short term RMS energy and normalized to the mean RMS value of the actual turn. The group of "other" features involves features capturing phenomena such as turn-taking or speaker overlaps. Although these features are not prosodic, they are put into the group of prosodic features for modeling purposes.

Because the overall set of implemented features is really huge, a feature space reduction is needed before using the prosodic feature set in statistical classifiers. The feature selection method used in this work is based on searching for the best features in smaller groups of features capturing similar prosodic phenomena. The best features from each group are passed on to form a new feature set containing prosodic features of all categories. This subset is further pared down by eliminating redundant features on the basis of the leaving-one-out approach.

Chapter 7

Statistical Models for Sentence Segmentation of Speech

As we must account for every idle word,
so must we account for every idle silence.
Benjamin Franklin

In this thesis, I examine three statistical approaches to sentence segmentation of speech – a hidden Markov model (HMM), a maximum entropy (MaxEnt) model, and a boosting-based model called BoosTexter. All three approaches rely on both textual and prosodic information. The approaches are interesting to compare. They not only employ different machine learning methods but also combine the two basic knowledge sources in different ways.

The HMM-based approach uses two independent models that are combined on the score level during testing. The BoosTexter approach builds one integral model that combines the two information sources on the feature level already during training. The MaxEnt-based approach, in the variant used here, lies somewhere in between the previous two. During training, the machine learning algorithm combines textual features with thresholded prosodic posteriors obtained from an independent prosodic classifier. The different views on the knowledge source combination are not only interesting to compare but also may be of benefit for a subsequent model output combination. The different models are likely to be at least partly complementary, and thus their combination may yield superior performance.

This chapter provides a detailed description of the three above mentioned modeling approaches. Its remainder is organized as follows. Section 7.1 presents the HMM approach, Section 7.2 overviews the MaxEnt approach, and Section 7.3 describes the boosting-based approach. Section 7.4 gives a brief summary of the whole chapter.

7.1 HMM-Based Approach

This approach to sentence segmentation, which was introduced by Shriberg and Stolcke, combines lexical and prosodic model within the hidden Markov model (HMM) framework. Lexical information is modeled by an N-gram language model, prosodic information is represented by posteriors output by an independent prosodic classifier. During testing, both knowledge sources are combined within an HMM.

This section is structured as follows. Subsection 7.1.1 summarizes N-gram modeling techniques, Subsection 7.1.2 overviews fundamentals of hidden Markov models, Subsection 7.1.3

describes the HMM-based hidden event language model, Subsection 7.1.4 presents the prosodic model based on decision trees, and Subsection 7.1.5 explains the combination of prosodic and language models.

7.1.1 N-gram Language Models

In many language processing problems, we need to estimate the probability $P(W)$ of a word string W consisting of words $w_1, w_2, \dots, w_T$. These probabilities are typically estimated using a statistical language model (LM). In general, the LM distribution $P(W)$ should depend on syntactic, semantic, and pragmatic properties of the target language, however, it is usually approximated by so-called *N-gram models* in real-world applications.

The N-gram models can be formalized as follows. Using the Bayes' chain rule, the probability $P(W)$ may be computed as

$$P(W) = P(w_1, w_2, \dots, w_T) = P(w_1) \prod_{i=2}^{T} P(w_i|w_1^{i-1}) = P(w_1) \prod_{i=2}^{T} P(w_i|w_1, \dots, w_{i-1}) \qquad (7.1)$$

However, such a model would have too many parameters and it would be basically impossible to estimate them robustly. That is why we use simplified models that assume that the occurrence of the word at position i only depends on the $N-1$ previous words. This simplification may be formalized as follows

$$P(w_i|w_1, w_2, \dots, w_{i-1}) \approx P(w_i|w_{i-N+1}, w_{i-N+2}, \dots, w_{i-1}) \qquad (7.2)$$

For example, if we assume that the current word depends on two preceding words, we use so-called *trigram* probabilities $P(w_i|w_{i-2}, w_{i-1})$. By analogy, we can define *bigram* $P(w_i|w_{i-1})$ and *unigram* $P(w_i)$ probabilities.

N-gram probabilities of orders greater than one can be estimated using the maximum likelihood approach as

$$\hat{P}(w_i|w_{i-1}^{i-N+1}) = \frac{C(w_{i-N+1}^{i})}{C(w_{i-N+1}^{i-1})} \qquad (7.3)$$

where the function $C(\cdot)$ returns the number of occurrences of its argument in training data.

However, this approach fails when $C(w_{i-N+1}^{i}) = 0$ or even $C(w_{i-N+1}^{i-1}) = 0$. In the former case, the probability[1] $P(w_i|w_{i-1}^{i-N+1})$ would be zero, in the latter case, it would be undefined. To overcome this problem, we use so-called *smoothing* techniques. The term "smoothing" refers to the fact that after its application we get a flatter distribution in which zero probabilities are replaced by non-zeros and non-zero probabilities are lowered. The smoothing methods not only prevent zero probabilities but also increase overall model robustness. There are two basic smoothing strategies: *interpolation* and *back-off.*

The interpolation smoothing methods mix the highest order N-gram distribution with lower order N-gram distributions that suffer less from data sparseness. A weighted sum of the N-gram distributions is used. In the mixture, the lowest order distribution, *zerogram*, has a specific uniform distribution in which all words from the vocabulary have a probability equal to the reciprocal of the vocabulary size $\frac{1}{|V|}$. For example, for a trigram model, we get

$$P_{IP}(w_i|w_{i-2}, w_{i-1}) = \lambda_0 \frac{1}{|V|} + \lambda_1 P(w_i) + \lambda_2 P(w_i|w_{i-1}) + \lambda_3 P(w_i|w_{i-2}, w_{i-1}) \qquad (7.4)$$

[1] To simplify the notation I will use $P(\cdot|\cdot)$ instead of $\hat{P}(\cdot|\cdot)$ in the remainder of this thesis.

The interpolation weights $\lambda_i > 0, \ i = 0, 1, 2, 3$ satisfying the condition $\sum_{i=0}^{3} \lambda_i = 1$ are estimated using so-called *held-out* data, i.e. data that we hold out from the training set from which we extract the N-gram counts. The λs are usually computed via the Expectation-Maximization (EM) algorithm or Powell search [128]. Note that the interpolation do not necessarily have to be fixed numbers; some interpolation-based smoothing methods define the weights as functions of the word history.

While the interpolation approach always uses probabilities from lower order N-gram models, the back-off approach exploits information from the lower order N-gram estimator only if it does not have non-zero count for the higher order N-gram. For instance, if the model has a corresponding (smoothed) trigram probability, it solely relies on it. By contrast, if this trigram probability is zero, the model backs-off to bigrams. The central idea of back-off smoothing is to distribute the overall probability mass between seen and unseen events. Most back-off methods use for this purpose the *Good-Turing* estimate.

According to the Good-Turing estimate, it is necessary to distribute among unseen N-grams the same portion of the probability mass as corresponds to N-grams that have only been seen once (so-called *singletons*). Hence,

$$P_0 = \frac{n_1}{N} \tag{7.5}$$

where n_1 is the number of singletons and N the overall number N-grams. Furthermore, for N-grams occurring exactly r-times, we pretend they occurred r^*-times

$$r^* = (r+1)\frac{n_{r+1}}{n_r} \tag{7.6}$$

Using the probability notation, we get

$$P_{GT}(x) = \frac{r^*}{N} \tag{7.7}$$

where x is an N-gram occurring r-times. The basic back-off smoothing (so-called *Katz* back-off) may be for trigrams written as follows

$$P_{Katz}(w_i|w_{i-2}, w_{i-1}) = \begin{cases} \frac{C^*(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})} & \text{for } C(w_{i-2}, w_{i-1}, w_i) > 0 \\ \alpha(w_{i-2}, w_{i-1}) P_{Katz}(w_i|w_{i-1}) & \text{for } C(w_{i-2}, w_{i-1}, w_i) = 0 \end{cases} \tag{7.8}$$

where $\alpha(w_{i-2}, w_{i-1})$ is a normalization factor satisfying the constraint that all probabilities in the distribution sum up to 1.

Nowadays, the most popular smoothing technique is the *Kneser-Ney* method [131]. It is based on a simple technique called absolute discounting. In absolute discounting, the higher order distribution is created by subtracting a fixed discount $D \leq 1$ from each non-zero count

$$P_{absolute}(w_i|w_{i-2}, w_{i-1}) = \begin{cases} \frac{C(w_{i-2}, w_{i-1}, w_i) - D}{C(w_{i-2}, w_{i-1})} & \text{if } C(w_{i-2}, w_{i-1}, w_i) > 0 \\ \alpha(w_{i-2}, w_{i-1}) P_{absolute}(w_i|w_{i-1}) & \text{if } C(w_{i-2}, w_{i-1}, w_i) = 0 \end{cases} \tag{7.9}$$

The Kneser-Ney method extends the absolute discounting model by a more sophisticated way of handling the back-off distribution. The intuition behind it is that words appearing in a higher number of different contexts are also more probable to occur in an previously unseen context. Nowadays, the Kneser-Ney method is usually not used based on the formula introduced in the original paper. Chen and Goodman [132] showed that modified, interpolated version of the smoothing algorithm, which uses three different discount parameters D_1, D_2, D_{3+} rather than a single discount D for all non-zero counts, performs better. The modified smoothing formula may be expressed as follows:

$$P_{KN}(w_i|w_{i-N+1}^{i-1}) = \frac{C(w_{i-N+1}^{i}) - D(C(w_{i-N+1}^{i}))}{\sum_{w_i} C(w_{i-N+1}^{i})} + \lambda(w_{i-N+1}^{i-1})P_{KN}(w_i|w_{i-N+2}^{i-1}) \tag{7.10}$$

where

$$D(c) = \begin{cases} 0 & \text{if } c = 0 \\ D_1 & \text{if } c = 1 \\ D_2 & \text{if } c = 2 \\ D_{3+} & \text{if } c \geq 0 \end{cases} \tag{7.11}$$

To make the sum of the distribution equal to 1, we use

$$\lambda(w_{i-N+1}^{i-1}) = \frac{D_1 N_1(w_{i-N+1}^{i-1}, \cdot) + D_2 N_2(w_{i-N+1}^{i-1}, \cdot) + D_{3+} N_{3+}(w_{i-N+1}^{i-1}, \cdot)}{\sum_{w_i} c(w_{i-N+1}^{i})} \tag{7.12}$$

where

$$N_1(w_{i-N+1}^{i-1}, \cdot) = \left|\left\{w_i : C(w_{i-N+1}^{i-1}, w_i) = 0)\right\}\right| \tag{7.13}$$

$N_2(w_{i-N+1}^{i-1}, \cdot)$ and $N_{3+}(w_{i-N+1}^{i-1}, \cdot)$ are defined by analogy.

Another useful technique is referred to as *Witten-Bell smoothing* [133]. While empirical results by Chen and Goodman [132] suggest that this method does not perform as well as other common smoothing techniques (especially for small training sets), in some specific cases it represents a good option. For instance, if no singletons appear in training data, we cannot use the methods based on the Good-Turing discounting. Such a situation often comes up when we use class-based language models. The Witten-Bell method also proved to be robust to various irregularities in training data since it is more conservative in subtracting the probability mass. The idea of the Witten-Bell discounting is to calculate the probability of seeing a new word based on the number of different words that follow a certain word history. The smoothed model is defined recursively as

$$P_{WB}(w_i|w_{i-N+1}^{i-1}) = \lambda(w_{i-N+1}^{i-1})P(w_i|w_{i-N+1}^{i-1}) + (1 - \lambda(w_{i-N+1}^{i-1}))P_{WB}(w_i|w_{i-N+2}^{i-1}) \tag{7.14}$$

To calculate the parameters $\lambda(w_{i-N+1}^{i-1})$, we use the number of unique words that occur following the history w_{i-N+1}^{i-1}. This count, which is denoted as $N_{1+}(w_{i-N+1}^{i-1}, \cdot)$, is formally defined as

$$N_{1+}(w_{i-N+1}^{i-1}, \cdot) = \left|\left\{w_i : C(w_{i-N+1}^{i-1}, w_i) > 0)\right\}\right| \tag{7.15}$$

Using this notation, the probability mass corresponding to unseen N-grams is given by

$$1 - \lambda(w_{i-N+1}^{i-1}) = \frac{N_{1+}(w_{i-N+1}^{i-1}, \cdot)}{N_{1+}(w_{i-N+1}^{i-1}, \cdot) + \sum_{w_i} C(w_{i-N+1}^{i})} \tag{7.16}$$

A more detailed survey of LM smoothing techniques may be found in [132, 134, 135, 13, 136], among others.

7.1.2 Hidden Markov Models

Hidden Markov Model (HMM) is a sequence classifier representing a probabilistic function of a Markov process. HMMs are very popular classifiers in many different tasks ranging from part-of-speech tagging and speech recognition to bioinformatics and musical score following. One of the reasons for the HMM popularity is the fact that there exist very efficient methods of HMM training and testing.

The general task of sequence classification is to assign a label to every element of the observed sequence. The HMM, given the sequence of elements, computes a probability distribution over all possible labels, and consequently finds the most probable label sequence. The model has the following components:

- $O = o_1, o_2, \ldots, o_N$ — an observation sequence (discrete or continuous valued);
- $S = s_1, s_2, \ldots, s_N$ — an underlying sequence of states;
- $S = s_0, s_{N+1}$ — special start and end states not associated with observations;
- $A = a_{01}, a_{02}, \ldots, a_{nn}$ — a transition probability matrix where a_{ij} represents a probability of moving from state i to state j, satisfying the condition $\sum_{j=1}^{N} = 1$;
- $B = b_i(o_t)$ — a set of observation likelihoods expressing the probability that an observation o_t is generated by state i.

The word "hidden" in the name of the model refers to the fact that we do not directly observe the state sequence that the model passes when generating the observation sequence, but only its probabilistic function. A first-order HMM makes two strong simplifying assumptions. First, the probability of the following state is only dependent on the directly preceding state

$$P(s_i|s_1, s_2, \ldots, s_{i-1}) \approx P(s_i|s_{i-1}) \tag{7.17}$$

Second, the probability of the output observation o_i depends only on the particular state s_i that generated the observation.

$$P(o_i|s_1, s_2, \ldots, s_i, \ldots, s_n, o_1, o_2, \ldots, o_i, \ldots, o_n) \approx P(o_i|s_i) \tag{7.18}$$

There are three fundamental methods for HMMs. First, we need a method to compute the likelihood of an observed sequence O given an HMM and its parameteres (A, B). This likelihood can be efficiently computed using the *Forward algorithm* (or alternatively, using its reversed version, the *Backward algorithm*) which is based on dynamic programming. A combination of the two algorithms, the *Forward-Backward algorithm*, can be used to estimate probabilities of the hidden state values given the output sequence.

Second, we need a method to find the best hidden state sequence S given an observation sequence O, and an HMM and its parameters (A, B). The best state sequence is usually decoded using the *Viterbi algorithm.* Finally, we need a method to estimate HMM parameters (A, B) given an observation sequence O (i.e., training data). For this purpose, we can use the *Baum-Welch algorithm*, which is a special case of the *Expectation-Maximization* (EM) algorithm and enables training of HMM parameters in an unsupervised approach. More details on these HMM algorithms are given in [134, 136], among others.

7.1.3 Hidden Event Language Model

For modeling of textual information within the HMM-based segmentation approach, the Hidden Event Language Model (HELM) [137] has been proposed. In most tasks, the role of the language model is to predict the next word given the word history. In contrast, the goal of language modeling in the sentence segmentation task is to estimate the probability that a sentence boundary occurs in an observed word context. Because the sentence boundaries are not explicitly present in the speech signal, they are called *"hidden events"*. Let W denote the given word sequence $w_1, w_2, ..., w_i, ..., w_n$ and E the sequence of interword events $e_1, e_2, ..., e_i, ..., e_n$. Then, the HELM describes the joint probability of words and hidden events $P(W, E)$ in an HMM. In this case, the HMM hidden variable is the type of the event (including "no-event"). The states of the model correspond to word/event pairs, the observations are the words (with the possibility of adding another observations such as prosodic feature vectors as discussed below). Note that, in contrast to HMM taggers, words appear in both the states and the observations.

Because we typically have training data annotated with event labels, the HMM model can be trained using a supervised approach; using Baum-Welch algorithm is not necessary here. For training, words and event labels are merged into a single data stream. Then, standard N-gram techniques are used to estimate HMM transition probabilities. In other words, event labels are treated in the same way as words during training. The absence of any hidden event in a interword boundary may be marked either explicitly (by a special label) or implicitly (by the absence of any event label). The latter approach seems to be more convenient since explicit "no-event" labels shorten the considered word context. As opposed to the typical way of N-gram model estimation, we do not split the training data into sentences before using them in training. Such splitting would evidently hurt the sentence boundary prediction ability since the model would not be aware of typical word sequences occurring across the sentence boundaries.

The most probable event sequence is identified with respect to individual word boundary classifications, rather than by finding the highest probability sequence of events [88]. Thus, we use

$$\hat{e_i} = \underset{e_i}{\operatorname{argmax}}\, P(e_i|W) \tag{7.19}$$

To obtain hidden event posteriors, standard HMM algorithms may be employed. The forward-backward algorithm is typically used for decoding, the alternative is to use the Viterbi algorithm. Note that the forward-backward algorithm is more accurate here since we are looking for the most likely event at each interword location, rather than finding the most probable event sequence. An implementation of the HELM is available as part of the SRILM toolkit [138].

7.1.4 Prosodic Model Based on Decision Trees with Ensemble Bagging

The aim of the prosodic model in the sentence unit segmentation task is to use prosodic features to provide sentence boundary probabilities for each interword boundary. The prosodic posteriors $P(E|X)$ may either immediately be used to make final decisions about the sentence boundaries, or, more typically, may later be combined with posteriors from a language model. To obtain the prosodic posteriors, CART-style decision trees [139] are employed.

In general, we can employ any statistical classifier, such as neural network or Gaussian mixture model, but decision trees offer several advantages. Not only that they yield good results, but also can handle both continuous and categorical features, as well as features with undefined values. The undefined feature values frequently occur at the edges of acoustic segments since many prosodic features refer to preceding or following words. An additional advantage of decision trees is that the trained models are easy to interpret by humans. This facility helps us understand how individual prosodic features are used to make a particular decision. Another advantage is that decision trees do not require data normalization or scaling since they do not make any assumptions about the data distribution.

7.1.4.1 CART-style Decision Trees

The decision tree is a classifier in the form of a tree structure in which each node is either a leaf node indicating the value of the target class and its probability, or a decision node specifying a test to be carried out on a single attribute-value, with one branch and sub-tree for each possible outcome of the test. The tree model is "grown" by splitting the source data set into subsets by asking one question at a time of the available features. The feature queried in each question, as well as the threshold value in the question (e.g., "*Is pause at the boundary in question longer than 250 ms?*"), is that which best discriminates the target classes at that node in the tree. The splitting process is repeated on each derived subset in a recursive manner until a stopping condition (the maximum depth of the tree or the minimum number of samples in the leaf node) is met.

In the CART algorithm, the predictive values of single features at a particular node are measured using the Gini index of diversity which is based on squared probabilities of membership for each target category in the node. It is defined as

$$I_G(i) = 1 - \sum_{j=1}^{m} f(i,j)^2 = \sum_{j \neq k} f(i,j) f(i,k) \tag{7.20}$$

where m is the number of target classes, and $f(i,j)$ denotes the probability of getting the value j in the node i. That is, $f(i,j)$ is the proportion of data samples assigned to the node i for which the correct class is j. The criterion is equal to zero when all samples in the node fall into a single target category. To avoid overfitting, various pruning and smoothing techniques, such as cost-complexity pruning, may be applied. In the testing phase, for each sample X, the decision tree estimates the posterior probability of each of the events E, yielding $P(E|X)$. An example of a decision tree is shown in Fig. 7.1.

7.1.4.2 Bagging

A drawback of decision trees is their instability [140]. It means that small changes in input training data may cause large changes in output classification rules. This problem may be mitigated by using aggregating methods. The aggregating methods generate a number of versions of classifiers, which are then combined to make final predictions. A popular aggregating method, which is frequently used with decision trees, is *bagging*[2] [141]. It is a powerful machine learning technique that takes advantage of the instable behavior of classifiers such as decision trees. The method is based on averaging predictions of multiple classifiers that are trained on a number of datasets obtained by sampling with replacement from the original training set (so-called bootstrap sampling).

Using a formal notation, this aggregating scheme may be described as follows. Let $Y = \{y_1, \ldots, y_M\}$ be set of target classes and T a training set consisting of data $T = \{(\boldsymbol{x_n}, y_n), n = 1, \ldots, N\}$ where $y_n \in Y$ corresponds to a class label. First, the bagging procedure creates a sequence of K training datasets (bags) $\{T_k\}, k = 1, .., K$ that are obtained by random sampling with replacement from T, and contain the same number of samples as T.

[2] The word *bagging* is an acronym for *B*ootstrap *Agg*regat*ing*.

```
pause.after < 4.5:
|   word.dur < 0.265:
|   |   p.pause.after < 124:
|   |   |   word.dur < 0.215:  0.909 0.09104 N
|   |   |   word.dur >= 0.215:  0.7797 0.2203 N (non_leaf)
|   |   p.pause.after >= 124:
|   |   |   vowel.75_dur < 0.055:  0.7442 0.2558 N (non_leaf)
|   |   |   vowel.75_dur >= 0.055:  0.4762 0.5238 S (non_leaf)
|   word.dur >= 0.265:
|   |   f.word.dur.norm < 0.625:
|   |   |   f.f0.slope.first < -0.035:  0.5686 0.4314 N (non_leaf)
|   |   |   f.f0.slope.first >= -0.035:  0.4028 0.5972 S (non_leaf)
|   |   f.word.dur.norm >= 0.625:
|   |   |   f0.ratio.last_min__baseline < 1.275:  0.5514 0.4486 N (non_leaf)
|   |   |   f0.ratio.last_min__baseline >= 1.275:  0.692 0.308 N (non_leaf)
pause.after >= 4.5:
|   pause.after < 70.5:
|   |   pause.after < 18.5:
|   |   |   word.dur < 0.255:  0.5749 0.4251 N (non_leaf)
|   |   |   word.dur >= 0.255:  0.369 0.631 S (non_leaf)
|   |   pause.after >= 18.5:
|   |   |   pause.after < 36.5:  0.3235 0.6765 S (non_leaf)
|   |   |   pause.after >= 36.5:  0.1928 0.8072 S
|   pause.after >= 70.5:  0.03217 0.9678 S
```

Figure 7.1: Example of a CART-style decision tree for sentence segmentation (only top 4 levels listed)

Then, let $\varphi(\boldsymbol{x}, T)$ be a procedure that builds a decision tree from data T and outputs posterior probabilities of target classes. Then, the final "bagged" posteriors are computed as an average of $\varphi(\boldsymbol{x}, T_k)$ over k

$$\varphi_A(\boldsymbol{x}, T) = \frac{1}{K} \sum_{k=1}^{K} \varphi(\boldsymbol{x}, T_k) \tag{7.21}$$

For a large N, the bags T_k are expected to have 63.2% of the examples of T, while the remaining samples are expexted to be duplicates.

Besides addressing the instability problem, bagging also decreases classifier variance and makes the resulting predictor more robust to noise in training data. In addition, since training of a individual decision tree is independent of each other tree training, it can be efficiently implemented using parallel machines to decrease the overall training time.

7.1.4.3 Ensemble Bagging

When training a sentence boundary classifier, we have to deal with the problem of imbalanced data. Sentence boundaries occur much less frequently than "non-boundaries"; their proportion is ranging approximately from 7 to 20 %, depending on the particular domain and language. The skewed distribution of training data may cause decision trees to miss out on inherently valuable features that are dwarfed by data priors. One solution to this problem is to train classifiers on data randomly downsampled to equal class priors [49]. During testing on (the imbalanced) test data, the resulting posteriors are adjusted to take into account the original class priors. Note that this adjustment is only necessary when using the prosody model alone because the "downsampled" posteriors can be used directly when combined with the HELM, as described below in Section 7.1.5.

The problem of the downsampling approach is that it does not utilize all available data from the majority class. However, it is possible to perform the downsampling in a smarter way – in combination with bagging. To take advantage of all available data, we can apply ensemble sampling instead of simple downsampling. Ensemble sampling is performed by randomly splitting the majority class into int(R) non-overlapping subsets, where R is the ratio between the number of samples in the majority and minority classes. Each subset is joined with all minority class samples to form int(R) balanced sets. Then, we apply bagging on each of these newly formed balanced sets. This combination of bagging and ensemble sampling makes up a method called *ensemble bagging*.

It was shown that this technique is very powerful in addressing the imbalanced dataset problem in the sentence segmentation task [142]. Hence, ensemble bagging was employed in all our experiments with decision trees. I used 25 bags per ensemble which was found to be a reasonable trade-off between the training time and performance. The number of ensembles varies according to ratios between the event priors in individual tasks. In the datasets used in this work, this number ranges from 5 to 13. For my experiments with CARTs, I used the tree growing algorithm as implemented in the IND package [143] along with a set of my own wrapper scripts performing higher-level machine learning algorithms.

7.1.5 Combination of Language and Prosodic Model

Shriberg and Stolcke proposed to combine prosodic and lexical features in the HMM framework under assumption of conditional independence of word identities and prosodic features [49]. The integrated HMM then models the joint probability distribution $P(W, X, E)$, where W denotes observed words, X observed prosodic features, and E hidden events. During classification, we try to find the event sequence $\hat{E}$ with the maximal posterior probability given W and X

$$\hat{E} = \operatorname*{argmax}_{E} P(E|W, X) \tag{7.22}$$

For $P(E|W, X)$, it holds

$$P(E|W, X) = \frac{P(X|W, E)P(E|W)}{P(X|W)} \tag{7.23}$$

If we make a simplifying assumption that prosodic features only depend on events E, and not on word identities W, we may substitute $P(X|W, E)$ by $P(X|E)$. It is necessary to note that this assumption is not always fully true because of several reasons. First, the prosodic features are partly influenced by particular phonetic content of individual words. This problem can be mitigated by proper feature normalization, however, can hardly be completely eliminated. Second, there also exists some indirect, and difficult to capture, relation between prosody and meaning [68]. Moreover, strictly speaking, even after applying the assumption of independence on word identities, real prosodic features are still dependent on word alignment since some prosodic features depend on phone or word boundary information for extraction or normalization. However, despite these facts, the simplifying assumption of independence is considered to be acceptable. Equation (7.23) may thus be rewritten as

$$\begin{aligned} P(E|W, X) &\approx \frac{P(X|E)P(E|W)}{P(X|W)} \\ &= \frac{P(E|X)P(E|W)}{P(E)} \cdot \frac{P(X)}{P(X|W)} \end{aligned} \tag{7.24}$$

Because the fraction $\frac{P(X)}{P(X|W)}$ does not depend on E, we may search for $\hat{E}$ using the following proportion

$$P(E|W, X) \propto \frac{P(E|X)P(E|W)}{P(E)} \tag{7.25}$$

Furthermore, if the prosodic classifier was trained on data downsampled to equal priors $P(E)$ (cf. Section 7.1.4), we can use

$$P(E|W,X) \propto P(E|X)P(E|W) \tag{7.26}$$

Assuming that prosodic observations are conditionally independent of each other given the event type, $P(E|X)$ can be computed as

$$P(E|X) = \prod_i P(e_i|x_i) \tag{7.27}$$

As a result, $\hat{E}$ can be found by treating prosodic features as state emissions with the probability $P(x_i|e_i)$ incorporated into the HELM. Thus, we obtain the following formula for $\hat{E}$

$$\hat{E} = \operatorname*{argmax}_E P(E|W,X) = \operatorname*{argmax}_E \left(P(W,E) \left(\prod_i \frac{P(e_i|x_i)}{P(e_i)} \right)^{\lambda} \right) \tag{7.28}$$

Note that for searching for the maximum, $P(W,E)$ and $P(E|W)$ are equivalent since $P(W)$ is constant. Moreover, $P(e_i)$ in the denominator may be omitted if the prosodic model was trained on data with equal priors (as proposed in Section 7.1.4.3). λ is an exponential scaling factor estimated using held-out data, which allows us to weight relative contributions from the two models. The use of the scaling factor is advisable since we combine scores from two probabilistic models of a different type.

7.2 Maximum Entropy Approach

As described above, the HMM-based model is generative. Its supervised training method maximizes the joint word/event pair sequence likelihood $P(W,E)$ on the training text; prosodic likelihoods are obtained from an independent classifier and integrated into the model during testing. Thus, the HMM-based model training algorithm does not guarantee that the correct event posteriors needed for classification are maximized. There is a mismatch between the training criterion and the use of the model for testing. On the contrary, Maximum Entropy (MaxEnt) is a discriminative model, which is trained to directly maximize the posterior boundary label probabilities. On the other hand, a drawback of MaxEnt is that it only makes local decisions, whereas HMM classifies the entire input sequence.

MaxEnt belongs to the exponential (or log-linear) family of classifiers, i.e. the features extracted from the input are combined linearly and then used as an exponent. The MaxEnt framework enables a natural combination of features relating to different data streams within a single model. MaxEnt models are very popular in many NLP tasks, such as machine translation, POS tagging, chunking, or word sense disambiguation.

7.2.1 General Method

Generally speaking, the classification problem is defined as a prediction of class $y \in \mathcal{Y}$ in a context $x \in \mathcal{X}$. The classifier $k : \mathcal{X} \rightarrow \mathcal{Y}$ is implemented using a conditional probability distribution $P(y|x)$. Furthermore, we define a context predicate[3] as a logic function

$$cp : \mathcal{X} \rightarrow \{0{,}1\} \tag{7.29}$$

[3] Context predicates are often called "features" within the MaxEnt framework.

returning 1 if "useful information" occurs in the context $x \in \mathcal{X}$. The set of context predicates $cp_1, \ldots, cp_m$ has to be defined by the system designer in advance. Furthermore, a set of binary indicator functions is defined

$$f : \mathcal{X} \times \mathcal{Y} \rightarrow \{0{,}1\} \tag{7.30}$$

for which it holds

$$f_{cp,x'}(x, y) = \begin{cases} 1 & \text{if } x = x' \wedge cp(y) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{7.31}$$

In the context of the sentence segmentation task, one such feature function might be

$$f_{cp,x'}(x, y) = \begin{cases} 1 & \text{if } \text{ word}_i = \text{“}\textit{today}\text{”} \quad \wedge \quad \text{word}_{i+1} = \text{“}\textit{I}\,\text{”} \quad \wedge \quad y_i = SU \\ 0 & \text{otherwise} \end{cases} \tag{7.32}$$

where SU denotes a sentence unit boundary.

The idea of MaxEnt is that the model should follow empirical constraints we impose on it, but beyond these constraints, it should make as few assumptions as possible. The empirical constraints are given by the training data, i.e.

$$E(f_i(x, y)) = E'(f_i(x, y)) \tag{7.33}$$

where $E(\cdot)$ denotes expectation and $E'(\cdot)$ its empirical estimate. The model finds a probability distribution that satisfies these constraints and has the maximum conditional entropy

$$H(P) = -\sum_x P(x)P(y|x) \log P(y|x) \tag{7.34}$$

Berger et al. [144] showed that the solution to this constrained optimization has an exponential form corresponding to a multinomial logistic regression model

$$P(y|x) = \frac{1}{Z(x)} \exp\left(\sum_i \alpha_i f_i(x, y)\right) \tag{7.35}$$

where

$$Z(x) = \sum_y \exp\left(\sum_i \alpha_i f_i(x, y)\right) \tag{7.36}$$

is the normalization factor ensuring that $\sum_y p(y|x) = 1$.

The MaxEnt model is trained by finding parameters α_i^* that maximize the likelihood product over the training data

$$\alpha_i^* = \underset{\alpha_i}{\operatorname{argmax}} \prod_{j=1}^{N} P(y_j|x_j) \tag{7.37}$$

where N denotes the number of samples in the labeled training set. A number of various numerical optimization algorithms has already been employed to solve this weight estimation problem. In all my experiments with the MaxEnt models, I employ the Conjugate Gradient Ascent method as implemented in the `MegaM` model optimization toolkit written by Hal Daumé III [145]. This method is very efficient for binary classification tasks, as is automatic sentence segmentation. Note that this optimization technique cannot be used for multiclass problems. The reason is that for multiclass classification, explicit construction and inversion of the Hessian matrix, which is an inherent part of the optimization algorithm, becomes impossible. Thus, for multiclass tasks such as automatic punctuation, one has to use an alternative method. For instance, L-BFGS (Limited Memory Broyden-Fletcher-Goldfarb-Shanno) method [146] uses an iteratively built approximation to the true Hessian.

An important feature of MaxEnt models is that they are prone to overfitting. To overcome this drawback, we typically use smoothing with Gaussian priors that penalize large weights. This technique aims to force weights to have Gaussian distribution with the mean $\mu = 0$ and the variance σ^2. The value of σ^2 is typically empirically optimized on development data. The application of the smoothing method changes the optimized likelihood function from (7.37) to

$$\alpha_i^* = \operatorname*{argmax}_{\alpha_i} \prod_{j=1}^{N} P(y_j|x_j) \prod_{\alpha_i} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{\alpha_i^2}{2\sigma_i^2}\right) \tag{7.38}$$

In log space, where the optimization is usually performed, the objective function becomes

$$\alpha_i^* = \operatorname*{argmax}_{\alpha_i} \sum_{j=1}^{N} P(y_j|x_j) - \sum_{\alpha_i} \frac{\alpha_i^2}{2\sigma_i^2} \tag{7.39}$$

More detail on general MaxEnt models may be found in [147, 148, 136], among others.

7.2.2 Textual Features for MaxEnt

To encode textual features for use in the MaxEnt model, it is necessary define some feature templates. The textual features I used correspond to N-grams available to the HELM plus an additional binary feature indicating whether the word before the boundary of interest is identical with the following word. This additional feature aims to capture word repetitions.

Given the word sequence $w_{i-3} \dots w_i \dots w_{i+3}$ where w_i refers to the word before the boundary in question, the N-gram templates were the following:

- Unigrams – U0: w_i, U1: w_{i+1}
- Bigrams – B0: $w_{i-1}w_i$, B1: w_iw_{i+1}
- Trigrams – T0: $w_{i-2}w_{i-1}w_i$, T1: $w_{i-1}w_iw_{i+1}$, T2: $w_iw_{i+1}w_{i+2}$
- Fourgrams – F0: $w_{i-3}w_{i-2}w_{i-1}w_i$, F1: $w_{i-2}w_{i-1}w_iw_{i+1}$, F2: $w_{i-1}w_iw_{i+1}w_{i+2}$, F3: $w_iw_{i+1}w_{i+2}w_{i+3}$

In other words, all N-grams containing the word right before the current boundary (w_i) are included. The only exception is the unigram capturing the word right after the boundary (w_{i+1}). If some words from the analyzed window were undefined because of coinciding with text boundaries, they were replaced by a special symbol for the feature extraction purposes. Note that although I also list fourgram features here, these have not been found to be useful in any of the tasks described in this thesis. Given the amount of data available for training in the individual tasks, they have never yielded improvement over the trigram model.

As already mentioned above, unlike the HELM, the MaxEnt model has the ability to handle mutually dependent or overlapping textual features. A single MaxEnt model does not have to only rely on words, but we can also add parallel features relating to other textual knowledge sources. For instance, we can run the training text through a part-of-speech tagger, generate parallel features based on those tags, and pass them to the learning algorithm together with the word-based features. Note that particular textual features used in individual classification tasks are described in special sections of the respective chapters, for example in Section 8.5.1 in the following chapter.

7.2.3 Prosodic Features for MaxEnt

A problem associated with using MaxEnt models for sentence segmentation is that it is not straightforward how to efficiently use continuous (prosodic) features. Although the MaxEnt features f_i may generally be real-valued, continuous are typically not used directly since combination of real- and binary-valued features in a single MaxEnt model may cause many problems. To this end, continuous features are usually converted into binary features via thresholding.

In the approach used here, prosodic features are not embedded into the MaxEnt model directly. The MaxEnt model is powerful in combining different sources of textual knowledge, but not expected to achieve superior performance when dealing with many (originally) real-valued features coming from a single knowledge source. This hypothesis was supported by experiments performed in [149]. Thus, it looks more natural to estimate prosodic posteriors using an independent classifier (in our case the bagged decision trees) and then to encode the posteriors via thresholding, as proposed in [88].

Since the presence of each feature in a MaxEnt model raises or lowers the final probability by a constant factor, it is reasonable to encode the prosodic posteriors in a cumulative way. This approach is more robust than using interval-based bins since small changes in prosodic scores may influence at most one feature. I have experimented with various gaps between adjacent thresholds and found that 0.1 is a convenient value. Thus, I got the following sequence of binary features – $p > 0.1$, $p > 0.2$, $p > 0.3$, $\ldots$, $p > 0.9$.

The prosodic posteriors for training samples were estimated using a cross-validation method. My preliminary experiments showed that using decision trees trained on the same data for which we generate the posteriors led to biased estimates and consequently hurt MaxEnt performance. Hence, the training set was divided into 5 non-overlapping subsets and CARTs for each of the subsets were trained only using other 4 subsets. In testing, models trained on all training data are used to generate the posteriors.

Note that some binary non-prosodic features that because of the HMM architecture had to be grouped with prosodic features in the HMM approach may naturally be handled separately within the MaxEnt framework. The speaker change feature is a good example of such features.

For illustration, the following example shows three data samples corresponding to three consecutive words from the training data. The MaxEnt features shown in the example capture word N-grams and thresholded prosodic posteriors. The first value in each sample corresponds to a class label (S – sentence boundary, N – no boundary). The following space-separated character strings correspond to binary features written in the so-called Bernoulli format, i.e. all present features are assumed to have value one and any non-present feature is assumed to have value zero.

```
N U0_anything U1_like B0_or_anything B1_anything_like T0_syllables_or_anything
T1_or_anything_like T2_anything_like_that

N U0_like U1_that B0_anything_like B1_like_that T0_or_anything_like
T1_anything_like_that T2_like_that_but pros_gt_0_1

S U0_that U1_but B0_like_that B1_that_but T0_anything_like_that
T1_like_that_but T2_that_but_but pros_gt_0_1 pros_gt_0_2 pros_gt_0_3
pros_gt_0_4 pros_gt_0_5 pros_gt_0_6 pros_gt_0_7 pros_gt_0_8
```

7.3 Boosting-Based Approach (BoosTexter)

Besides their positive properties, the two above presented approaches also have some disadvantages. In the HMM approach, the combination of prosodic and lexical models makes strong independence assumptions, which are not fully met in actual language data. Moreover, the HMM training method maximizes the joint probability of data and hidden events, but a criterion more closely related to classification error would be the posterior probability of the correct hidden variable assignment given the observations. The drawbacks of HMM may partly be eliminated by using the MaxEnt model, however, the MaxEnt model itself also shows some setbacks. As described in the previous section, lexical and prosodic features are not combined directly but via thresholding of prosodic posteriors generated by an independent prosodic classifier. Thus, I have explored yet another approach in which prosodic and textual features are integrated into a single model based on boosting.

7.3.1 General Method

Similarly to bagging, the principle of boosting is to combine many weak learning algorithms to produce an accurate classifier. However, whereas in bagging, weak classifiers are trained independently of each other, in boosting, each weak classifier is built based on the outputs of previous classifiers, focusing on the samples that were formerly classified incorrectly. The algorithm generates weak classification rules by calling the weak learners repeatedly in series of rounds. The method maintains a set of importance weights over training data samples and labels. The weights are used by the weak learning algorithm to find a weak hypothesis with "moderately low" error by forcing the weak learner to focus on "more difficult" samples from the training data. The general boosting method can basically be combined with any classifier.

In the sentence segmentation approach described here, an algorithm called BoosTexter [150] was employed.[4] This machine learning technique was initially designed for the task of text categorization. It combines weak classifiers having a basic form of one-level decision trees (stumps) using confidence-rated predictions. The test at the root of each tree can check for the presence or absence of an N-gram, or for a value of a continuous feature. Hence, the approach allows a straightforward combination of lexical and prosodic features in a single statistical model. Moreover, another positive property of BoosTexter is that it is very robust to overtraining. In parallel with [92], the BoosTexter method was newly applied to the sentence segmentation task as part of this work (as published in [151]).

The particular boosting method implemented in BoosTexter is based on a boosting algorithm called *AdaBoost.MH*. Using the same notation as in Section 7.1.4.2, its basic principle may be described as follows [150, 152]. The learning algorithm assigns M weights (i.e. one weight for one of the target classes, in the sentence segmentation task $M = 2$) to each instance of training data $(\boldsymbol{x_n}, y_n)$. Let this distribution over training samples and labels be called D_i and the set of labels $L_n = \{l_{n1}, \ldots, l_{nM}\}$. In the first round, the distribution is uniform. On each iterative step i, the weak learner uses D_i to generate a hypothesis $h_{i,n}(\boldsymbol{x_n}, l_n)$. The sign of $h_{i,n}(\boldsymbol{x_n}, l_n)$ indicates whether the label l_n is assigned to $\boldsymbol{x_n}$ or not, and the magnitude $|h_{i,n}(\boldsymbol{x_n}, l_n)|$ is the confidence of the prediction. The distribution D_i is updated to increase the weight of misclassified sample-label pairs

$$D_{i+1}(n, l) = \frac{D_i(n, l) \exp\left(-\alpha_i C_n(l) h_{i,n}(\boldsymbol{x_n}, l)\right)}{Z_i} \tag{7.40}$$

[4] An open source reimplementation of the original BoosTexter algorithm called `icsiboost` is available from `http://code.google.com/p/icsiboost/`.

where

$$Z_i = \sum_{n=1}^{N} \sum_{l \in Y} D_i(n,l) \exp(-\alpha_i C_n(l) h_{i,n}(\boldsymbol{x_n}, l)) \tag{7.41}$$

is the normalization factor, $C_n(l)$ is a function

$$C_n(l) = \begin{cases} 1 & \text{if } l = y_n \\ -1 & \text{if } l \neq y_n \end{cases} \tag{7.42}$$

indicating whether l is a correct label for $\boldsymbol{x_n}$, and α_i is the weight of a weak classifier. In general, $\alpha_i \in \mathcal{R}$. For binary problems, the original AdaBoost implementation use

$$\alpha_i = \frac{1}{2} \log \left(\frac{1 - e_i}{e_i} \right) \tag{7.43}$$

where e_i is the weighted error of classifier h_i. However, the AdaBoost.MH with real predictions employed in this thesis uses

$$\alpha_i = 1 \tag{7.44}$$

Finally, the resulting hypothesis is output as

$$f(\boldsymbol{x_n}, l_n) = \sum_{i=1}^{I} \alpha_i h_{i,n}(\boldsymbol{x_n}, l_n) \tag{7.45}$$

where I denotes the total number of training iterations. The resulting BoosTexter scores may be converted into posterior probabilities using a probability calibration method. I employ Friedman's logistic correction [153]

$$p(y_n|\boldsymbol{x_n}) = \frac{1}{1 + \exp\left(-2 \cdot f(\boldsymbol{x_n}, l_n)\right)} \tag{7.46}$$

which showed good results when used with boosted stumps. In the particular BoosTexter implementation, where the output scores are divided by sum $\sum_{i=1}^{I} \alpha_i$, the equation (7.46) becomes

$$p(y_n|\boldsymbol{x_n}) = \frac{1}{1 + \exp\left(-2I \cdot f_{BT}(\boldsymbol{x_n}, l_n)\right)} \tag{7.47}$$

7.3.2 Prosodic and Textual Features for BoosTexter

For training the classifier, I use exactly the same set of prosodic features as for prosodic classification by decision trees. Also the form of all features was the same, i.e. no feature scaling or any other preprocessing operation was necessary. Similarly to CARTs, BoosTexter is also able to handle features that may take undefined values.

On the other hand, textual features for BoosTexter were extracted in the same way as for the MaxEnt approach (cf. Section 7.2.2). Thus, all N-grams containing the word before the boundary of interest, plus the unigram right after the boundary, and a binary flag indicating whether the two words across the boundary are identical or not, were extracted. Similarly to MaxEnt, BoosTexter may handle dependent and overlapping features so that textual features used in a single learning procedure do not have to rely only on a single data stream but parallel knowledge sources may naturally be combined.

7.4 Chapter Summary

In this chapter, I have described three modeling approaches to sentence segmentation of speech – HMM, MaxEnt, and BoosTexter. The models differ in the ways they combine textual and prosodic cues. The HMM-based approach uses two independent models that are combined on the score level during testing. The MaxEnt-based approach combines binary textual features with thresholded prosodic posteriors obtained from an independent prosodic classifier. The BoosTexter approach builds one integral model that combines the two information sources on the feature level during training. In addition to the used machine learning techniques, I have also described how individual approaches encode lexical and prosodic features. The three modeling approaches presented herein are evaluated in experiments reported in the following chapters.

Chapter 8

Dialog Act Segmentation of Multiparty Meetings in English

Never interrupt me when I'm trying to interrupt you.
Winston Churchill

Well-timed silence hath more eloquence than speech.
Martin Fraquhar Tupper

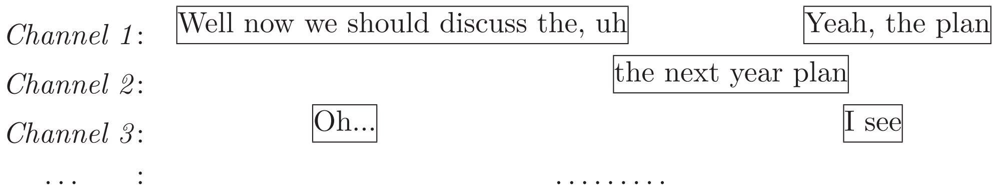
An area of growing interest in the spoken language technology community is the automatic processing of multiparty meetings. Important tasks in this domain include automatic meeting browsing, summarization, information extraction and retrieval, and machine translation [154, 155]. As in other domains, the downstream applications require input segmented into meaningful units. The goal of the work described in this chapter is to develop a segmentation system for the meeting domain, investigate usefulness of various textual and prosodic features, and compare performance of the three modeling approaches described in the preceding chapter. Unlike previous work, which has generally examined the use of prosody for DA segmentation of meetings using only pause information, I also explore the use of prosodic features beyond pauses, including duration, pitch, and energy features. The ICSI meeting corpus is used for all experiments herein.

This chapter is organized as follows. Section 8.1 describes the used corpus, Section 8.2 surveys related work in the meeting domain, Section 8.3 defines the particular task, and Section 8.4 presents the experimental setup. Sections 8.5, 8.6, and 8.7 report results of the experiments based on using only textual information, only prosodic information, and a combination of both information sources, respectively. Section 8.8 presents a system combining all three modeling approaches, and Section 8.9 summarizes all experiments and draws conclusions.

8.1 Speech Data

8.1.1 Particularities of Meeting Speech

Natural meetings represent very difficult data for automatic processing. They typically contain regions of high speaker overlaps, emotional speech, abandoned or interrupted utterances, and complicated speaker interactions. These frequent phenomena pose a number of challenges for researchers in the speech processing community. The following example illustrates overlapping speech as occurs in meetings.



8.1.2 ICSI Meeting Corpus

In this work, I use the ICSI Meeting Corpus [156] which is publicly available from LDC. It is a collection of data from 75 natural meetings that occurred at the International Computer Science Institute (ICSI) in Berkeley, CA, USA. In total, it yields approximately 72 hours of multichannel conversational speech data sampled at 16 kHz with 16-bit resolution. All meetings were recorded using both head-worn wireless microphones and several desktop microphones. In all experiments described in this chapter, I only used data from the close-talking microphones.

Most of the recorded face-to-face meetings were regularly scheduled weekly group appointments. Meetings of the following types were recorded:

- *Even Deeper Understanding* (Bed) – discussing NLP and neural theories of language (15 meetings)
- *Meeting Recorder* (Bmr) – on the ICSI Meeting Corpus (29)
- *Robustness* (Bro) – discussing robust ASR methods (23)
- *Network Services & Applications* (Bns) – internet architectures and standards (3)
- *One time only meetings* (varies) – miscellaneous other meetings (5)

In the whole corpus, there appear 53 unique speakers – 13 females and 40 males, 28 native speakers of English and 25 nonnative speakers. Note that the number of nonnative speakers of English is quite high, however, many of them are fluent when speaking in English. Average number of participants per meeting was about 6. A unique five-character tag was assigned to each speaker in the corpus. First letter stands for sex (*f* for female or *m* for male), the second letter is either *e* for native speakers or *n* for nonnative speakers. The first two characters are followed by a unique three digit number. For instance, one of the native male speakers is *me013*.

Each of the near-field channels was manually transcribed on the word level. In addition to the full words, transcripts also contain other information such as filled pauses, backchannels, contextual comments (e.g. while whispering) and non-lexical events such as laughter, breath, lip smack etc. Overall, the corpus transcripts yield 773k words.

8.1.3 Dialog Act Markup

In addition to the word transcripts, the ICSI corpus was hand-annotated for Dialog Acts (DAs) [157, 158]. This companion set of DA-level annotations is called the Meeting Recorder Dialog Act (MRDA) corpus. The labeling included marking of DA segment boundaries, marking of DA types, and marking of correspondence between DAs. The MRDA annotation scheme defines five main dialog act types as well as a number of their subtypes. The main DA classes are the following:

- *statement* (59.0% of DAs in the corpus) – such as "I agree.".
- *question* (6.4%)– such as "Do you agree?".
- *backchannel* (13.3%) – is a short positive comment, such as "uh-huh" or "yeah", to the other speaker to encourage further talk or to confirm that one is listening.
- *floor-grabber/holder* (7.2%) – indicates that the talker wants to start ("okay", "yes, but") or keep talking ("so", "uh").
- *disruption* (14.1%) – is an incomplete statement or question such as "but there's a –".

DA segmentation rules within the MRDA guidelines were designed to separate speech regions having different discourse functions. The annotators should not only pay attention to textual features but also to pauses and intonational grouping. However, note that DA units as defined in MRDA differ from the SUs used in the MDE corpora described in Chapter 5. The MRDA segmentation rules instruct annotators to split utterances on a clausal level to maximize the amount of information derived from DAs. This means that sometimes even grammatically subordinate clauses may form a complete DA, which is in contrast with the SU definition. Typical examples of such grammatically subordinate DAs are clauses beginning with conjunctions as *because* or *although.* An illustrative example of a DA-segmented utterance follows (DA boundaries are marked by "/." tags).

```
well so I wouldn't be too concerned about it with respect to that /.
although we should clear with American down course /.
but these results are based on data which haven't had be uh haven't had the
chance to be reviewed by the subject /.
so I don't know how that stands /.
```

Besides standard DA boundaries, a pipe bar (|) was used to distinguish utterances that are prosodically one unit but contain multiple DAs. An example of such case may be the utterance "yeah | that's right" when pronounced as one prosodic unit. This setting enables researchers to decide whether to split or not split at the pipes in dependence on the task of their research.

8.2 Related Work on Segmentation of Meetings

The first sentence segmentation study in the meeting domain was conducted by Baron et al. [159]. The authors focused on disfluency detection and automatic punctuation in meetings using the standard HMM approach. Since they used an older version of the ICSI meeting corpus and a different definition of sentence-like unit boundaries, the results cannot be directly compared with more recent work on the same data. For a three-way classification task (sentence boundary/IP/other), their experimental results showed that in the meeting domain, similarly to other spontaneous speech domains, a combination of lexical and prosodic models outperforms a lexical model alone.

The study by Ang et al. [160] mainly aimed to provide baseline performance rates for DA segmentation and classification in the meeting domain. For DA segmentation, the authors combined lexical and pause features in the HMM framework. As expected, it was shown that prosodic information (they only used pause duration) is less degraded by ASR errors. For DA classification, a MaxEnt approach was employed.

Zimmermann et al. [161] got inspired by [80] and used an A*-based algorithm to perform DA segmentation and classification simultaneously. Their method was based solely on lexical features. The heuristic search used a probabilistic framework based on DA-specific N-gram models. During A* search, an optimal path through the input word sequence was found. The nodes of the search graph corresponded to inter-word boundaries, whereas the edges carried labels indicating the DA type and spanned one or more consecutive words. Even though the system did not use pause information, it outperformed [160] in classification of backchannels and questions.

In a more recent paper [162], the same authors employed a combination of a word-based hidden event HMM and a MaxEnt model jointly modeling words and pause duration. In contrast to previous work [49, 160] which modeled pause duration independently from surrounding words, they modeled word boundary types based on both the pause duration and surrounding words. The novel approach resulted in a modest error rate reduction.

Cuendet et al. [163] focused on adaptation of sentence segmentation models trained on conversational telephone speech to meeting style conversations. Their models relied on lexical and pause features. They used the ICSI meeting corpus, but only focused on meetings of one type (“Bed”). Several different adaptation approaches including data concatenation, logistic interpolation, boosting adaptation, and out-of-domain confidences as an extra feature, were tested. The authors also analyzed adaptation performance in dependence on the amount of used in-domain data. The experimental results showed that boosting adaptation and logistic interpolation were the best performing approaches when only small adaptation data were used. For larger adaptation data sets, the logistic interpolation approach showed the best results.

8.3 Segmentation Task for Meetings

This section defines the sentence-like units into which we segment the meeting data. Although the original manual transcripts of the ICSI corpus do contain punctuation, and thus sentence boundaries, the punctuation is highly inconsistent. Transcribers were instructed to focus on transcribing words as quickly as possible; there was not a focus on consistency or conventions for marking punctuation. As a result, different transcribers used different approaches to punctuation annotation.

Hence, rather than using the inconsistent first-pass punctuation, we decided to employ special DA segmentation marks from the MRDA annotation project. In this annotation pass, labelers carefully annotated both dialog acts and their boundaries, using using a set of segmentation conventions for the latter (as described in Section 8.1.3). Thus, we define the target units as *dialog acts*. Consequently, the task is *automatic dialog act segmentation of multiparty meetings*.

Since the class of DA boundaries involves boundaries of all five MRDA types, it includes boundaries of both complete and incomplete DAs. Also note that, in line with [160], we decided to split at the pipe bar boundaries, creating a slightly larger number of total DA segments. Note that splitting here makes it slightly more difficult to see significant gain from using prosodic cues since such splits occur within prosodic units.

Formally speaking, DA segmentation can be viewed as a two-way classification problem with “DA-boundary” and “Non-DA-boundary” as the target classes. For a given word sequence $w_1 w_2 ... w_i ... w_n$, the task of DA segmentation is to determine which interword boundaries correspond to a DA boundary.

8.4 Experimental Setup

For the DA segmentation experiments herein, I used 73 out of the total 75 available meetings. The remaining two meetings were excluded because of their very different character from the rest of the data. The 73 meetings were split into a training set (51 meetings, 539k words), a development set (11 meetings, 110k words), and a test set (11 meetings, 102k words). In the experiments to follow, classification models are trained on the training set, tuned on the development set, and evaluated on the test set. The test set contains unseen speakers, as well as speakers appearing in the training data as it is typical for the real world applications.

For all experiments, two different test conditions were considered: human-generated reference transcripts (REF) and speech recognition output (ASR). Recognition results were obtained using the state-of-the-art SRI CTS system [164], which was trained using no acoustic data or transcripts from the analyzed meeting corpus. To represent a fully automatic system, I also used automatic speech/non-speech segmentation, although manual segmentation was available as well. Word error rates for this difficult data are still quite high; the employed speech-to-text system performed at $WER = 38.2\%$ (on the whole corpus).

To allow performance evaluation on the ASR hypotheses, it is necessary to have some "reference" DA boundaries for the ASR words. Their generation is not straightforward since true words and ASR hypotheses often differ in the number of words they contain. To this end, the reference setup was aligned to the ASR hypotheses based on the minimum edit distance. DA boundaries for the ASR words were then taken from the corresponding aligned reference words with the constraint that two aligned words could not occur further apart than a fixed time threshold. Since the ASR hypotheses tend to miss short backchannels that are usually followed by a DA boundary, the boundaries are less frequent in the ASR hypotheses (13.9% of words) than in the reference transcripts (15.9%). The DA boundary alignment for an erroneous ASR hypothesis is illustrated in the following example.

REF:

```
you could do it about you /. right /. sure /. I mean different among classes /.
because it’s it has a   high rate energy /. somewhere around sixty must be /.
```

ASR:

```
    could what about you /.           sure /. or mean different him  pluses /.
because it’s it doesn’t high rate angie /. there     are  sixteen  mostly /.
```

I use BER (defined in Eq. (3.1) on page 18) as the main performance measure for evaluation of my experiments. In addition, I always show chance error rate which corresponds to BER achieved when all interword boundaries in test data are classified as within-DA boundaries. To ease a performance comparison across test conditions, I also report $NIST$ error rate (3.2) and F-measure (3.5) for the best model in each section.

Differences between individual models are tested for statistical significance using the Sign test [50]. For all statistical tests in this thesis, I take $p < 0.05$ as a standard level of significance. However, I do not only report whether a tested difference is significant or not, but also present p-values of all significant improvements. In addition, I also explicitly present p-values falling between 0.05 and 0.10 since differences with p-values close to 0.05 may be interpreted as "marginally significant". All differences with $p > 0.10$ are referred to as statistically insignificant.

8.5 DA Segmentation Based on Textual Information

In this section, I focus on an effective utilization of information contained in the recognized words. Well-tuned language models (LMs) are not only important for applications where they are combined with a prosody model, but also for the applications in which we do not have access to, or cannot exploit, prosodic information. The LM evaluation is twofold; I search both for a convenient representation of textual knowledge (i.e., convenient textual features) and a suitable modeling approach.

I do not only explore simple word-based models, but also utilize textual information beyond word identities, as captured by word classes and part-of-speech (POS) tags. In general, it is also possible to use chunking (or even full-parsing) features. Chunking features may slightly increase performance for well-structured English speech such as broadcast news [149], but my preliminary investigations showed that, because of poor chunking performance on meeting data, these features rather hurt DA segmentation accuracy on meeting speech. Hence, I did not use them in this work.

8.5.1 Textual Features

The following sections describe individual groups of employed features. Since I do not only use LMs based on a single knowledge source but also LMs based on their combinations, it should be noted how the parallel features are combined. While MaxEnt and BoosTexter allow us to combine overlapping features from parallel textual knowledge sources directly, the current implementation of HELM is not able to build the HMM trellis based on two parallel token sequences, just as are words and part-of-speech tags. To overcome this problem, I have used each LM separately and then combined the outputs via posterior probability interpolation. The interpolation weights were tuned using the development set.

8.5.1.1 Words

Word features simply capture word identities around possible sentence boundaries. They represent a baseline feature set for the experiments herein.

8.5.1.2 Automatically Induced Classes

In language modeling, we always have to deal with data sparseness. In some tasks, it is possible to mitigate this problem by grouping words with similar properties into word classes. The grouping reduces the number of model parameters to be estimated during training. We assume a mapping function

$$r:\ V \rightarrow C \tag{8.1}$$

assigning a class $c_i \in C$ to every word w_i from the vocabulary V, i.e., $c_i = r(w_i)$. In general, one word could belong to more than one class, however, such mapping would cause problems for decoding. Thus, I assume that every word w_i only belongs to a single class c_i.

Automatically induced classes (AIC) are derived in a data-driven way. Data-driven methods typically perform a greedy search procedure to find the best fitting class for each word given an objective function. The clustering algorithm I used [165] minimizes perplexity of the induced class-based N-gram with respect to provided word bigram counts. The DA boundary token is excluded from class merging, however, it affects the clustering procedure since it occurs in bigram histories.

Initially, each word is placed into its own class. Then, the classes are iteratively merged until the desired number of clusters is reached. The resulting classes are mutually exclusive, i.e., each word is only mapped into a single class. In every step of the algorithm, the overall perplexity is minimized by joining the pair of classes maximizing the mean mutual information of adjacent classes

$$I(c_1, c_2) = \sum_{c_1, c_2 \in C} P(c_1, c_2) \log \frac{P(c_1, c_2)}{P(c_1)P(c_2)} \tag{8.2}$$

The crucial parameter of the word clustering algorithm is the number of target classes. The optimal number was empirically estimated on the development data by evaluating performance of models with a different granularity. I started with a 500-class model and then was gradually decreasing the number of classes by 25 in each iteration. The optimal number of classes was estimated as 100.

I have also tested a model that mixes AICs and frequent words. In this approach, the frequent words are excluded from class merging; they form their own one-word classes. This approach can be viewed as a form of back off; we back off from words to classes for rare words but keep word identities for frequent words. However, I have tested various numbers of the left out words in combination with individual class granularities, but have never achieved better results than for the 100 classes with no excluded words.

8.5.1.3 Parts of Speech

The AICs reflect word usage in our datasets, but do not form clusters with a clearly interpretable linguistic meaning. In contrast, part-of-speech (POS) tags describe grammatical features of words. POS tagging is one of the most explored tasks in NLP. In many cases, several different POS categories may correspond to a particular word form. For instance, English word *monitor* may either be a noun (as in "*The monitor is battery operated.*"), or a verb ("*Brokers monitor emerging markets everyday.*"). The goal of POS tagging is to automatically disambiguate such ambiguities. The task can formally be defined as follows:

$$\phi : W \to T,\ \phi(w_i) = t_i;\ t_i \in \tau,\ i = 1, \ldots, N_w \tag{8.3}$$

where W denotes input text consisting of N_w words w_i, T is a sequence of tags t_i, and τ a set of all possible tags.

For the tests herein, the POS tags were obtained by using the TnT tagger [166] with a POS model tailored for conversational English. The tagger was trained using hand-labeled data from the Switchboard Treebank corpus. To achieve a better match with speech recognition output used in testing, punctuation and capitalization information was removed before using the data for tagger training [149].

Besides pure POS-based models, I also tested mixed POS models (POSmix). In contrast with AICs, mixing of frequent words and POS of infrequent words yielded an improvement. The reason is that while the automatic clustering algorithm takes into account bigram counts containing the sentence boundary token and thus is aware of strong sentence boundary indicators, POS classes are purely grammatical. By keeping the frequent words we also keep some strong sentence indicators. Optimizing the model on our development data, I ended up with 500 most frequent words being kept and not replaced by their POS tags. In terms of a relative frequency, the cutoff value for not replacing the word by its POS was $1.87 \cdot 10^{-4}$ (i.e., 101/539k).

8.5.2 Experimental Results

Table 8.1 presents experimental results for all three models (HMM, Maxent, and BoosTexter), all feature sets (words, AIC, POS, and POSmix, and training and test conditions. The models

Table 8.1: DA segmentation error rates for LMs with various textual features [BER %]. REF=Reference human transcripts, ASR=Automatic transcripts, AIC=Automatically Induced Classes, POS=pure POS-based model, POSmix=infrequent words replaced by POS tags and frequent words kept. The best results for each model are displayed in boldface.

Model	**Used Features**	**Train/Test Setup**		
		REF/REF	**REF/ASR**	**ASR/ASR**
Chance	—	15.92%	13.85%	13.85%
HMM	Words	7.45%	9.41%	9.50%
	AIC	7.58%	9.70%	9.78%
	POS	10.62%	12.06%	11.85%
	POSmix	7.65%	9.57%	9.59%
	Words+AIC	7.11%	9.25%	9.18%
	Words+POSmix	7.23%	9.25%	9.31%
	Words+AIC+POSmix	**7.02**%	**9.12**%	9.12%
MaxEnt	Words	7.50%	9.38%	9.38%
	AIC	7.42%	9.44%	9.37%
	POS	10.52%	11.79%	11.80%
	POSmix	7.26%	9.23%	9.25%
	Words+AIC	7.19%	9.25%	9.21%
	Words+POSmix	7.27%	9.27%	9.25%
	Words+AIC+POSmix	**7.15**%	9.24%	**9.16**%
BoosTexter	Words	7.70%	9.52%	9.49%
	AIC	7.61%	9.50%	9.53%
	POS	10.87%	12.03%	11.13%
	POSmix	7.68%	9.45%	9.46%
	Words+AIC	7.50%	9.42%	9.40%
	Words+POSmix	7.66%	9.44%	9.45%
	Words+AIC+POSmix	**7.46**%	**9.40**%	9.40%

for segmentation of human transcripts were trained on reference words. For testing on ASR data, I tried to use both true and recognized words for training, and compared performance of the respective models.

8.5.2.1 Results in Reference Conditions

In reference conditions, the best models based on a single textual feature group were MaxEnt for mixed POS and AICs, and HMM for words. On the other hand, the models only using POS information performed poorly. A comparison of POS and POSmix shows that grammatical properties of words are not sufficient indicators of sentence boundaries and information provided by some frequent cue words is necessary to achieve satisfactory performance. In terms of a modeling approach comparison, it is interesting to observe that the generative HMM model is better in dealing with word information, while the discriminative MaxEnt model better captures class information (both AIC and POSmix). The BoosTexter model always performed worse than the other two models.

The results also indicate that an improvement is achieved when word information is combined with class information. For all models, the best results are obtained when all three information sources are combined. Using the Sign test, the improvements over baseline word-based models are statistically significant at the levels of $p < 10^{-23}$, $p < 10^{-18}$, and $p < 10^{-5}$ for HMM, MaxEnt, and BoosTexter, respectively. Of the three statistical models, HMM was the best performing, achieving $BER = 7.02\%$. Using other evaluation metrics, this result corresponds to $NIST = 44.09\%$ and $F = 76.88\%$. The difference between HMM and Boostexter is significant at $p < 10^{-13}$, and the difference between HMM and MaxEnt at $p < 0.02$. Of the other two models, MaxEnt outperformed BoosTexter (BER : 7.15% vs. 7.46%). This difference is significant at $p < 10^{-9}$.

The interpolation weights of the combined HMM model were estimated as 0.36, 0.39, and 0.25 corresponding to words, AIC, and POSmix, respectively. Thus, the largest weight is given to the AIC-based model. However, when interpreting the weight distribution, we must take into account that some word-based features are also contained in the POSmix model. Hence, the word-based feature weights are distributed between the two submodels. We can also observe that when only two types of features are employed, words plus AIC perform better than words plus POSmix. Once again, these results were expected since word and POSmix features are partly overlapping and thus the combination of word and AIC features is more complementary.

8.5.2.2 Results in ASR Conditions

As in reference conditions, MaxEnt for mixed POS was the best single model in the ASR-based tests. On the other hand, unlike reference conditions, MaxEnt was also the best model for capturing word information. The combination of all three knowledge sources was helpful once again, the best performing combined model was HMM ($BER = 9.12\%$, $NIST = 65.81\%$, and $F = 62.78\%$), while BoosTexter was the worst. Both HMM and MaxEnt show a significant outperformance of the BoosTexter model ($p < 10^{-5}$ and $p < 10^{-4}$). In contrast, the difference between HMM and MaxEnt is not significant. Improvements of individual models containing all available features over the baseline word-based models are significant at $p < 10^{-10}$, $p < 10^{-17}$, and $p < 0.02$ for HMM, MaxEnt, and BoosTexter, respectively.

A comparison of models trained on clean and erroneous data shows the following. While for HMM and BoosTexter, the error rates were almost the same, for the MaxEnt model, I got better results when training on automatic transcripts. However, even for the MaxEnt model, the difference in BER (9.16% vs. 9.24%) is only significant at $p < 0.08$.

The interpolation weights of the combined HMM model in ASR conditions were estimated as 0.33, 0.35, and 0.32. As in REF conditions, AICs had the largest weight. In comparison with REF conditions, the POSmix submodel got a slightly larger weight, while the other two submodels got accordingly lower weights.

An interesting finding was that the MaxEnt approach was slightly better than HMM for all individual feature types, but slightly worse when all feature types were combined. Hence, I decided to test whether it could be caused by the model combination approach. In HMM, we interpolate three independent models, while for MaxEnt, all feature types are accessible during a single model learning procedure. Thus, I trained three independent MaxEnts for words, AICs, and POSmix and interpolated them. However, I did not get better results than for the original MaxEnt model.

8.6 DA Segmentation Based on Prosodic Information

The preceding section described experiments with language models, whereas in this section, I focus on the prosodic model. I explore the use of prosodic features including pause, duration, pitch, and energy features when lexical information is not accessible. The original database of prosodic features (described in Chapter 6 and listed in Appendix B) contained over 300 different features. Using the approach described in Section 6.7, the huge database was reduced to a smaller set of 40 features. Note that the target number of features was not determined beforehand but resulted from the used feature selection method.

Besides reporting overall accuracy of prosodic models with the "rich" prosodic feature set, I also investigate whether there is gain from using the richer set of prosodic features in comparison with using pause information alone. The alternative pause-only feature set contains just three features capturing pause duration after the previous, the current, and the following word. I experiment with this feature set since there was a valid suspicion that prosodic features beyond pause duration might not be helpful; prosodic marking of boundaries in meetings seems to be largely irregular. Since extraction of prosodic features requires a lot of effort, it is purposeful to test whether the additional features yield a significant improvement in meeting applications.

Only two modeling approaches are evaluated in this section. Although I investigate three modeling approaches (HMM, MaxEnt, and BoosTexter) in this work, two of them (HMM and MaxEnt) use prosodic scores obtained from the same prosodic classifier (CART-style decision tree with ensemble bagging). Hence, I only compare CART and BoosTexter in this section.

8.6.1 Experimental Results for Prosodic Models

Table 8.2 shows BERs achieved by particular prosody models in both reference and ASR test conditions.

8.6.1.1 Results in Reference Conditions

In reference conditions, the best result ($BER = 8.06\%$) was achieved by the CART model using the rich prosodic feature set. This result corresponds to $NIST = 50.60\%$ and $F = 70.60\%$. I also should point out that in reference conditions, the best prosodic model on its own performs significantly worse that the best language model on its own ($BER = 8.06\%$ vs. $BER = 7.02\%$).

We can see that prosody beyond pause is helpful to a great extent. The p-values of the Sign test for the comparison of rich prosody vs. pause-only are 10^{-99} and 10^{-24} for CART and BoosTexter, respectively. A comparison of CART and BoosTexter indicates that the boosting-based model is better when only three pause duration features are employed. On the other hand, bagged decision trees are better in handling the larger prosodic feature set. This superiority is significant at $p < 0.001$.

8.6.1.2 Results in ASR Conditions

As in reference conditions, the best result in ASR conditions ($BER = 8.30\%$) was achieved by the CART model with the rich feature set. Using the other two metrics, this best result corresponds to $NIST = 59.89\%$ and $F = 63.58\%$. In contrast with the results on clean reference data, the best prosodic model significantly outperforms the best language model ($BER = 8.30\%$ vs. $BER = 9.12\%$) in ASR conditions. This comparison supports the hypothesis that prosodic features are less degraded by word errors than lexical features.

Table 8.2: DA segmentation error rates for prosodic models [BER %]

Model	**Used Features**	**Train/Test Setup**		
		REF/REF	**REF/ASR**	**ASR/ASR**
Chance	—	15.92%	13.85%	13.85%
CART w. Ens. Bag.	Pause	8.96%	8.97%	9.03%
	All Prosody	**8.06%**	**8.30%**	8.32%
BoosTexter	Pause	8.81%	8.85%	8.82%
	All Prosody	**8.22%**	**8.31%**	8.35%

The gap between CART and BoosTexter was really tiny and statistically insignificant in ASR conditions. I infer that the reason for this is that the simpler BoosTexter model (which only uses stumps instead of fully grown trees as weak learners) is more robust to errors in test data and this robustness may balance more accurate modeling shown by the CART model. Note that although prosodic features are more robust than lexical features, it does not mean that they are not affected by word errors at all since a number of them depend on phone or word boundary information for extraction or normalization.

As in reference conditions, the models benefit from using the rich prosody feature set. The gain over the pause feature set is a bit smaller than for reference data but still clearly significant – at $p < 10^{-41}$ for CART and $p < 10^{-18}$ for BoosTexter. The results also indicate that the models trained on clean data slightly outperform their counterparts trained on ASR data. However, note that the differences between them are not statistically significant.

8.6.2 Prosodic Feature Usage

To explore which prosodic features were useful in this task, I analyzed feature usage in prosodic decision trees; the decision trees are much easier to interpret than the BoosTexter model. The metric called *feature usage* [49] reflects the number of times a feature is queried in a tree, weighted by the number of samples it affects at each node. Thus, features that are used higher up in the decision tree have higher usage values. Total feature usage within a tree sums to 1. The results reported here are based on averaging results over multiple trees as generated during ensemble bagging.

For illustration purposes, prosodic features were grouped into five non-overlapping groups: pause at the boundary in question, duration, pitch, energy, and “near pause” (describing pauses associated with the previous and the following word boundary positions). The statistics for each group as well as the most used features from each group are listed in Table 8.3. The numbers show that the most frequently queried feature group was “duration” followed by “pitch”. However, if we sum “pause” and “near pause”, we can conclude that pause-related features are overall more important than F_0-related features. As expected, the most used individual feature is duration of the pause after the current word. The next most queried features were duration of the current word, pause after the preceding word, and normalized duration of the last rhyme in the current word. On the other hand, I observed that overlap features were not helpful at all, which was a bit surprising. It suggests that turn-taking in this multichannel data is well captured by pause-related features.

I also looked at the feature usage for the pause-only feature set. In this feature set, pause after the current word has usage 39.2%, whereas pauses after the previous and the following word have 29.9% and 30.9%, respectively. Thus, as expected, the pause at the boundary in question was the most frequently queried feature in this feature set.

Table 8.3: Feature usage for individual groups of prosodic features

Group	Tot. usage	Two most used features from the group
Pause	16.1%	pause.after(16.1%), —
Duration	48.9%	word.dur(9.3%), word.dur.last_rhyme.snorm(5.5%)
Pitch	21.4%	f.f0.slope.first(3.6%), f0.ratio.last_min___baseline(3.2%)
Energy	4.9%	f.RMS.mean(2.3%), RMS.min.voiced.norm (1.5%)
NearPause	8.7%	p.pause.after(5.7%), f.pause.after(3.0%)

Table 8.4: DA segmentation error rates for models combining textual and prosodic features [BER%]

Model	Used Features	Train/Test Setup		
		REF/REF	REF/ASR	ASR/ASR
Chance	—	15.92%	13.85%	13.85%
HMM	LM+Pause	5.60%	6.86%	6.93%
	LM+Prosody	**5.42%**	6.85%	**6.83%**
MaxEnt	LM+Pause	5.59%	6.82%	6.64%
	LM+Prosody	**5.42%**	6.64%	**6.55%**
BoosTexter	LM+Pause	5.84%	6.85%	6.67%
	LM+Prosody	**5.72%**	6.73%	**6.60%**

8.7 DA Segmentation Using Both Textual and Prosodic Information

In this section, I report results of models that rely both on prosodic and textual information. The approaches to knowledge source combination differ among individual models; particular combination methods were already described in Section 7. I compare two sets of models; the first combines the best LMs (using word, AIC, and POSmix features) with the pause-only models, whereas the second combines the same LMs with the prosody models based on the richer feature set. The results for both test conditions are presented in Table 8.4.

Note that in contrast with the experiments reported in the previous section, prosodic trees trained on REF data were used to generate prosodic posteriors also for both ASR-based setups. Thus, in the ASR/ASR experiments, HMM and MaxEnt combine textual features extracted from ASR data with prosodic posteriors obtained from trees trained on REF data. This decision was motivated by the results reported in the previous section; they showed that CARTs trained on clean data outperformed their ASR-trained counterparts.

8.7.1 Results in Reference Conditions

The best result in REF conditions was achieved by the HMM model combining the LM with the rich prosodic model ($BER = 5.42\%$, $NIST = 34.03\%$, $F = 83.17\%$). However, the MaxEnt model performed nearly as well, and, in total, made only two more errors in 102k test samples. Thus, the difference between these two models is evidently not statistically significant. On the other hand, the differences between HMM and BoosTexter, and MaxEnt and BoosTexter are significant at $p < 10^{-5}$ using the Sign test.

The models with the richer prosodic features sets outperform models with pause information only in all three approaches. The improvements are significant at $p < 10^{-4}$ for HMM and

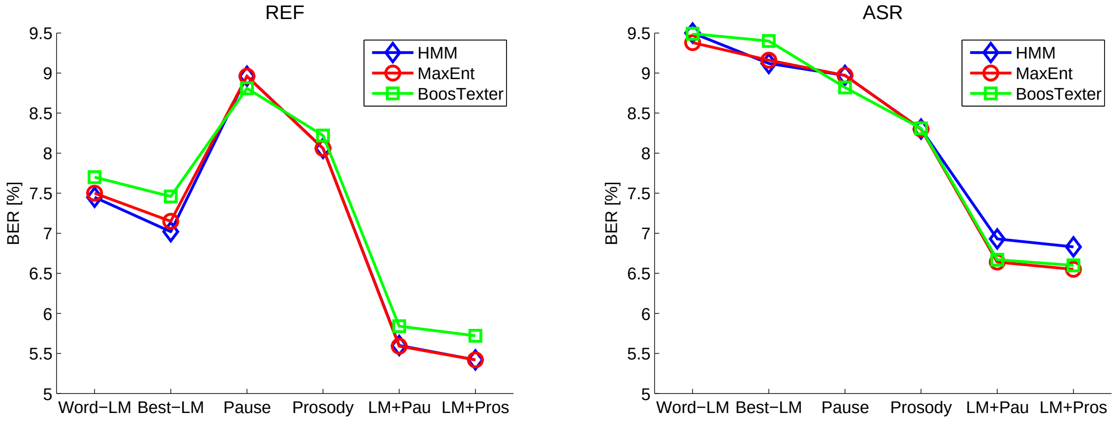


Figure 8.1: DA segmentation error rates for individual models in dependence on used knowledge sources [BER %]

MaxEnt, and at $p < 0.005$ for BoosTexter. Despite the fact that models using rich prosody were always better than models with pauses, the results indicate that the gain from better prosodic modeling is diminished when combined with an LM. For example, the relative error rate reduction for prosody only conditions is over 10% in the HMM approach, whilst it is just 3.2% for the combined prosodic-textual model in the same approach.

Furthermore, we can observe that combined models notably outperform both language and prosody models alone. For comparison, the chance error rate is 15.92%, the best LM alone performed at $BER = 7.02\%$, the best prosody model at $BER = 8.06\%$, and the best combined model at $BER = 5.42\%$. The left hand graph in Fig. 8.1 confronts BERs according to available information sources for all three modeling approaches. In the figure, lines connect points corresponding to the same model for the sake of readability.

8.7.2 Results in ASR Conditions

The best performing model in ASR conditions was the MaxEnt model that was using rich prosody and was trained on ASR data ($BER = 6.55\%$, $NIST = 47.29\%$, $F = 73.90\%$). Of the three modeling approaches, BoosTexter came second and HMM (which was the best in REF conditions) was the worst. The Sign test showed that MaxEnt is significantly better than HMM ($p < 10^{-4}$), but the difference between MaxEnt and BoosTexter ($BER = 6.55\%$ vs. $BER = 6.60\%$) is not significant. The gap between BoosTexter and HMM is significant at $p < 0.001\%$.

For all three approaches, models trained on ASR data outperform models trained on REF data. While the difference is not significant for HMM, it is significant at $p < 0.05$ and $p < 0.01$ for MaxEnt and BoosTexter, respectively. These results indicate that training on ASR data is more beneficial for the discriminative models.

As in the tests on reference data, the models with the richer prosodic feature sets outperform models only using pause information. However, the gains are relatively smaller in ASR conditions. The improvement is clearly insignificant for the HMM model, the p-values for MaxEnt and BoosTexter are 0.02 and 0.07, respectively. Again, we can observe that the gain from rich prosody is much smaller when combined with LMs. Whereas the relative error rate reduction for prosody only conditions is 7.5% using decision trees, it is just slighltly over 1% in the best performing MaxEnt approach (which also uses decision trees for prosody modeling).

The right-hand graph in Fig. 8.1 shows BERs according to available information sources

Table 8.5: DA segmentation error rates for a combination of HMM, MaxEnt, and BoosTexter [BER%]

Combination Approach	**Test Data**	
	REF	**ASR**
Chance	15.92%	13.85%
Best Single Approach	5.42%	6.55%
Majority Voting	5.20%	6.29%
Linear Interpolation	**5.18%**	**6.19%**

for all three modeling approaches in ASR conditions. The chance error rate is 13.85%, the best LM performed on its own at $BER = 9.12\%$, the best prosody model at $BER = 8.30\%$, and the best combined model at $BER = 6.55\%$. Although the shapes of the performance curves are different from those for the REF-based tests displayed in the left-hand graph (prosody models outperform LMs for ASR), the combined models are by far the best models again.

8.8 Combination of All Three Modeling Approaches

In this section, I evaluate models that not only combine knowledge sources but also statistical modeling techniques. As was already mentioned in Chapter 7, the three machine learning approaches used in this work train segmentation models in different ways. In machine learning tasks, it is often the case that superior results are achieved by a combination of several different submodels. Models of a different nature usually make different errors and thus may be at least partly complementary. Hence, I also tried to combine the three modeling approaches to see whether the combination could provide improvement over the best single-approach model. I examined two combination methods – simple majority voting and linear interpolation of posterior probabilities.

Table 8.5 presents the overall error rates for both test conditions. The two combination approaches are compared with the best individual models – HMM for REF and MaxEnt for ASR conditions. The results indicate that the model combination boosts DA segmentation performance. Overall, BER was reduced from 5.42% to 5.18% for REF, and from 6.55% to 6.19% for ASR. These error reductions are significant at $p < 10^{-17}$ and $p < 10^{-7}$, respectively. In the other two metrics, the best results are $NIST = 32.49\%$ and $F = 83.18\%$ for REF, and $NIST = 44.70\%$ and $F = 75.85\%$ for ASR. Of the two combination approaches, linear interpolation worked better, however, the gain over the majority voting approach was only significant for the ASR-based tests ($p < 0.005$). The interpolation weights for HMM, MaxEnt, and BoosTexter were estimated on the development data as 0.36, 0.33 and 0.31 for REF, and 0.32, 0.32 and 0.36 for ASR. None of the models has a dominant position and the weight distributions are not far from uniform. Examples of meeting transcripts automatically segmented by the best performing system are shown in Appendix C (page 153).

I also compared overall classification accuracy across the test conditions. Note that this comparison cannot be carried out in terms of absolute BER because this metric does not take into account event priors. Since the DA boundary priors differ between the two compared datasets, a comparison in terms of absolute BERs would be skewed. However, we can compare model performance across corpora in terms of F-measure, NIST error rate, or relative reductions of BER over the chance error. As expected, the comparisons on the meeting data clearly show that the final segmentation models are relatively more successful when dealing with human-generated test data. For instance, the relative BER reduction over chance is 67.5% for REF and 55.3% for ASR.

8.9 Chapter Summary and Conclusions

In this chapter, I explored automatic DA segmentation of the ICSI meeting corpus. Three different modeling approaches were examined – HMM, MaxEnt, and BoosTexter. For language modeling, I did not only use simple word-based models, but also utilized textual information beyond word identities, as captured by automatically induced classes and POS tags. In prosody modeling, I experimented with two distinct feature sets – the first contained just pause-based features, while the other was a richer set also containing features relating to duration, pitch, and energy. All experiments were evaluated both on manual and automatic transcripts.

First, I explored DA segmentation based on textual information. The results of language models only using a single textual knowledge source indicate that both word-based and AIC-based features show good results, while POS information is helpful only when the most frequent words are kept and not replaced by POS classes. For both test conditions, the best results were achieved when all textual information sources were combined. The most successful language modeling approach was HMM, while BoosTexter was the worst.

The next set of experiments focused on performance of prosodic models. These experiments compared decision tree and BoosTexter classifiers since both HMM and MaxEnt use prosodic posteriors obtained from the decision trees. The results show that decision trees outperform BoosTexter. The difference between the two models was statistically significant in reference conditions but insignificant in ASR conditions. The results also indicate that prosodic features are less degraded by speech recognition errors than lexical features.

Prosody beyond pause was helpful to a great extent. The rich feature set significantly outperformed pause in both test conditions. The analysis of feature usage in decision trees revealed that the most frequently queried group of features was the pause group followed by the duration group and the pitch group. Thus, information from pauses is extremely important for successful DA segmentation of meetings, but adding duration, pitch, and energy features yields a further significant improvement.

Models combining prosodic and lexical information clearly outperform language and prosody models alone. Similarly to prosody-only models, the models combining textual information with rich prosody feature set were better than the models combining it only with pause-based features. However, the relative gain from additional prosodic features was smaller than for the models not employing textual features.

The comparison of individual modeling approaches shows that HMM and MaxEnt prevails in reference conditions, and MaxEnt is the best in ASR conditions. The results of all models suggest that a tighter integration of prosodic and lexical knowledge during training is helpful for DA segmentation in ASR conditions. The best overall results for both test conditions were achieved by a combination of HMM, MaxEnt, and BoosTexter based on linear interpolation of posterior probabilities. This result indicates that the three modeling approaches are at least partly complementary.

Another comparison of automatic segmentation accuracy shows that the best DA segmentation system is relatively more successful in reference than in ASR conditions. The relative BER reduction over the chance error is 67.5% for reference and 55.3% for ASR conditions. The overall results in ASR conditions also indicate that models trained on ASR data slightly outperform their counterparts trained on manual transcripts. For the generative HMM approach, the difference is not significant, whereas the discriminative models (MaxEnt and BoosTexter) get a modest yet significant gain from training on ASR data. Thus, we can conclude that for corpora in which ASR performance is still rather poor, it may be useful to train models on recognized rather than reference words. However, we must take into account that automatic recognition of large amounts of training data may be computationally expensive.

Chapter 9

Speaker-Dependent Modeling for Dialog Act Segmentation of Meetings

Society exists only as a mental concept;
in the real world there are only individuals.
Oscar Wilde

The previous chapter focused on automatic DA segmentation of meeting speech in a speaker-independent fashion. In this chapter, by contrast, I take a closer look at *speaker-dependent* prosodic and language modeling for DA segmentation of meetings. Speaker adaptation methods are widely used for speech recognition, but much less is known about speaker-specific variation in prosodic and lexical patterns that are important for detection of linguistic boundaries in speech.

The meeting domain is a good test bed for speaker-dependent modeling. In many real-life meeting applications, the speakers are often known beforehand and recorded on a separate channel. Moreover, many meetings have recurring participants, presenting the opportunity for adapting models to the individual talkers. In the remainder of this chapter, I explore speaker-specific issues for both prosody and language modeling. Section 9.1 discusses speaker-dependent prosodic modeling, Section 9.2 is devoted to speaker-dependent language modeling, and Section 9.3 provides an overall summary and conclusions.

9.1 Speaker-Dependent Prosodic Modeling

9.1.1 Motivation and Goals

The general idea of speaker adaptation is well known. Speaker adaptation methods were first successfully used in the cepstral domain for speech recognition [167, 168]. On the other hand, only little is known about speaker-specific variation in prosodic patterns, beyond basic F_0 normalization. Studies in speech synthesis and automatic speaker recognition have used prosodic variation successfully, but to my best knowledge, modeling stylistic prosodic variability for sentence boundary recognition has to date been mentioned only anecdotally in the literature [49, 169, 170].

The experiments reported in this section try to find answers to two questions about speaker variation in prosodic marking of DA boundaries, using 20 frequent speakers from the ICSI meeting corpus. The first question asks whether individual speakers benefit from modeling rich prosodic features other than simple pause information. The second question inquires whether speakers differ enough from an overall (speaker-independent) model of prosody to benefit from a model trained (or adapted) on only their speech. The rest of this section is organized as follows. Subsection 9.1.2 describes used prosodic features, classifiers, modeling approaches, and the experimental setup. Subsection 9.1.3 provides results and discussion showing answers to the two questions. The results are presented separately for individual speakers as well as summed over all speakers to enable an overall model comparison.

9.1.2 Method

9.1.2.1 Prosodic Features

For the speaker-dependent experiments, I used a feature set comprising 32 prosodic features. This set was derived from the 40 feature set used for the speaker-independent experiments by removing all raw duration features. Although unnormalized duration features aid performance of prosody-based segmentation, they could correlate with lexical features that should be modeled in a language model. Certain frequent DAs (esp. backchannels) have small set of words, so raw durations may capture those words rather than prosody. Hence, in order to focus on the aspect of the speaker-dependent prosodic characteristics, these raw features were taken out to avoid interference with language modeling. In addition to the rich prosodic feature set, the comparison in Section 9.1.3.1 also evaluates models only using pause features.

9.1.2.2 Classifiers

As in past work on DA segmentation, CART-style decision trees with ensemble bagging were employed as prosodic classifiers. Since the trees were trained on bagged ensembles downsampled to equal class priors, when applying the classifiers on (the imbalanced) test data, the output posteriors were adjusted to take into account the original class priors. The advantage of decision trees is not only in that they yield good results, but they are also easy to interpret in terms of feature usage. These classifiers have been described in detail in Section 7.1.4.

9.1.2.3 Evaluated Prosodic Models

This study compares three types of prosodic models differing in speaker-dependency. Speaker-independent (SI) models are trained using all available data, whereas speaker-dependent (SD) models are only trained on speech of the target talker. Models of the last type, speaker-adapted (SI+SD), are created via posterior probability interpolation of the two previous models

$$P_{SI+SD}(X;\lambda) = \lambda P_{SI}(X) + (1-\lambda)P_{SD}(X) \tag{9.1}$$

where X denotes the observed prosodic feature vector, $P_{SI}(X)$ the speaker-independent and $P_{SD}(X)$ speaker-dependent posterior, and λ is a weighting factor estimated using the jackknife approach, as described below in Section 9.1.2.4.

9.1.2.4 Data and Experimental Setup

As in the previous chapter, the DA segmentation experiments reported here were performed using the ICSI meeting corpus. However, the experimental setup was different because the speaker-dependent experiments required a special data split. The models were evaluated on

Table 9.1: Data set sizes for individual speakers. ID=Speaker ID, Train=Trainig set size, Test-R=Test size for REF, Test-A=Test size for ASR. All data sizes are presented as numbers of words.

No.	ID	Train	Test-R	Test-A	No.	ID	Train	Test-R	Test-A
1.	me013	115.2k	51.2k	43.4k	11.	mn052	10.7k	3.8k	3.5k
2.	me011	50.6k	24.8k	22.9k	12.	mn021	9.6k	4.1k	4.1k
3.	fe008	50.6k	22.6k	19.5k	13.	me003	9.3k	3.6k	3.2k
4.	fe016	32.0k	15.4k	13.9k	14.	mn005	7.7k	3.1k	3.0k
5.	mn015	31.9k	14.7k	13.7k	15.	me045	8.1k	2.4k	2.1k
6.	me018	31.8k	14.7k	13.3k	16.	me025	7.7k	2.4k	1.6k
7.	me010	26.1k	12.6k	11.3k	17.	me006	6.9k	1.5k	1.3k
8.	mn007	27.2k	10.1k	8.4k	18.	me026	5.2k	2.5k	2.3k
9.	mn017	21.0k	7.1k	6.0k	19.	me012	5.3k	2.1k	1.9k
10.	mn082	13.3k	4.2k	3.7k	20.	fn002	5.9k	1.5k	1.4k

the top 20 speakers in terms of total words. This speaker set contains 17 males and 3 females; 12 speakers are native English speakers and 8 are nonnative speakers.

Each speaker's data was split into a training set (∼70% of data) and a test set (∼30%), with the caveat that a speaker's recording in any particular meeting appeared in only one of the sets. Because of data sparsity, especially for the less frequent speakers, I did not use a separate development set, but rather jackknifed the test set for my experiments. In this approach, one half of speaker's test data is used to tune weights for the other half, and vice versa. As in the previous experiments, two different test conditions are used for evaluation: reference transcripts (REF) and automatic speech recognition output (ASR). BER (defined in Eq. (3.1) on page 18) is used for performance evaluation as in the preceding segmentation experiments.

The total training set for speaker-independent models (comprising the training portions of the 20 analyzed speakers, as well as all data from 32 other less-frequent speakers) contained 567k words. The total test set contained 204k words for reference test conditions and 180k for ASR test conditions. Data set sizes for individual speakers are shown in Table 9.1; size of training sets available for training of the speaker-dependent models ranges from 5.2k to 115.2k words. Note that the test set sizes for REF and ASR conditions differ since the number of words in ASR outputs is usually smaller than that in the corresponding reference. Speaker identity is described using the official corpus speaker IDs. The first letter of the ID denotes the sex of the speaker ("f" or "m"); the second letter indicates whether the speaker is a native ("e") or nonnative ("n") speaker of English.

9.1.3 Results and Discussion

9.1.3.1 Pause-only vs. Rich Prosody Features for Individual Speakers

Before analyzing the SD prosodic models, I took a look on prosodic differences among individual speakers from a different perspective. The experiments in Chapter 8 showed that features capturing prosodic information beyond pause help prosody-based sentence segmentation of meetings. However, here our goal is to explore whether there is a gain from using rich prosody for all speakers, or only for some. Hence, I used an SI model based on the rich feature set and an SI model based only on pause information and compared their results for individual speakers.

Table 9.2: DA segmentation error rates by pause-only (Pause) and rich prosody models (RichPr) models for individual speakers in reference conditions [BER %]. RER denotes relative BER reduction by RichPr with respect to Pause. The best result for each speaker is shown in boldface.

ID	Pause	RichPr	RER	ID	Pause	RichPr	RER
me013	8.93%	**8.36%**	6.29%	mn052	8.93%	**8.29%**	7.17%
me011	7.47%	**6.61%**	11.50%	mn021	8.23%	**8.01%**	2.64%
fe008	8.92%	**8.53%**	4.37%	me003	6.18%	**5.83%**	5.79%
fe016	10.15%	**9.62%**	5.18%	mn005	8.74%	**7.73%**	11.57%
mn015	8.69%	**7.99%**	8.06%	me045	7.95%	**7.20%**	9.37%
me018	8.30%	**7.74%**	6.72%	me025	8.74%	**8.32%**	4.85%
me010	9.25%	**8.30%**	10.21%	me006	10.72%	**9.86%**	7.98%
mn007	11.53%	**10.71%**	7.10%	me026	**7.94%**	**7.94%**	0.00%
mn017	8.67%	**8.03%**	7.35%	me012	8.85%	**8.66%**	2.11%
mn082	9.76%	**9.00%**	7.82%	fn002	11.26%	**9.79%**	13.09%

Table 9.2 shows the results using the two feature sets for each speaker. The speakers displayed in the table are sorted according to the total number of words they have in the corpus. As shown, the richer prosodic feature set (RichPr) yields a better performance than the pause-only model (Pause), for 19 of the 20 speakers, and the other speaker (me026) showed the same BER using both models.

The relative error rate reduction by RichPr with respect to Pause is also provided. It shows that differences across speakers on this measure, interestingly, do not appear to be correlated with the amount of training data. They may thus reflect differences in speaking styles, although other factors such as robustness of feature extraction or production of different rates of DA types, may also play a role.

9.1.3.2 Speaker-Independent vs. Speaker-Dependent Models for Individual Speakers

The main problem I investigate is whether some speakers may benefit from speaker-dependent training, despite significantly less data for SD than for SI models. All the experiments reported in this section were performed with the rich prosodic feature set. I will present results for both reference transcripts and ASR conditions, but I will mainly analyze results using the reference transcripts, because I was more interested in looking at the effects of speaker differences than effects of recognition errors.

Table 9.3 compares performance of SI, SD, and SI+SD models in REF conditions. The results indicate that the SD model is better than the SI model for 4 of the 10 most frequent speakers, and for 5 of the 20 speakers. The SD or SI+SD model is better than the SI model for 5 of the top 10 speakers, and for 7 of all 20 speakers. Note that it is possible for the SI+SD to perform worse than the SI model, because weights are estimated on fairly small amounts of data that are separate from the data on which the model is tested. I used the Sign test for statistical significance measurement. Four speakers (me011, mn007, fe016, mn005) showed improvements significant at $p < 0.05$ or better; one speaker (fn002) was marginally significant at $p < 0.10$. It is also interesting that the error reduction with respect to SI varies across speakers, but in a manner uncorrelated with training set size.

Although only some speakers show these improvements (while some others show rather poor results from SD modeling), the finding is important. If a speaker shows significantly

Table 9.3: DA segmentation performance comparison of SI, SD, and SI+SD prosodic models in REF conditions [BER %]. The best result for each speaker is shown in boldface, * indicates that the improvement of SI+SD over SI is significant at $p < 0.05$ using the Sign test.

ID	SI	SD	SI+SD	ID	SI	SD	SI+SD
me013	**8.36%**	8.47%	8.39%	mn052	**8.29%**	8.64%	8.32%
me011*	6.61%	6.60%	**6.41%**	mn021	**8.01%**	9.27%	8.08%
fe008	**8.53%**	8.55%	8.55%	me003	**5.83%**	6.57%	5.83%
fe016*	9.62%	9.55%	**9.52%**	mn005*	7.73%	**7.15%**	7.18%
mn015	7.99%	8.47%	**7.96%**	me045	**7.20%**	7.62%	7.29%
me018	**7.74%**	8.09%	7.74%	me025	**8.32%**	8.95%	8.32%
me010	8.30%	**8.20%**	8.30%	me006	**9.86%**	10.65%	9.99%
mn007*	10.71%	10.47%	**10.19%**	me026	**7.94%**	8.99%	7.94%
mn017	**8.03%**	8.06%	8.03%	me012	**8.66%**	8.76%	8.66%
mn082	**9.00%**	9.62%	9.02%	fn002	9.79%	10.52%	**9.32%**

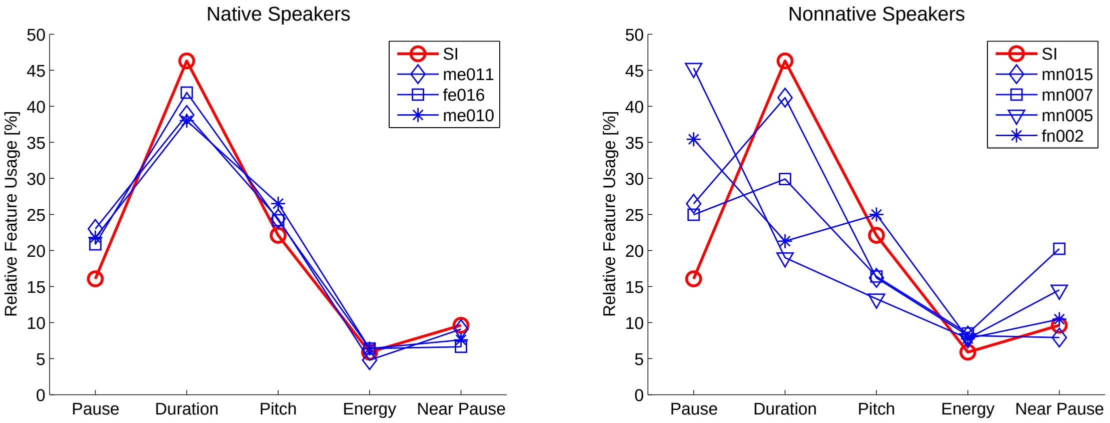


Figure 9.1: Relative usage of prosodic feature groups for native (left) and nonnative (right) speakers who improved using SD information

improved results using a model trained on far less data than the SI model, this suggests that the speaker's prosodic marking of DA boundaries differs from that of the SI model. That a number of speakers do not benefit from SD modeling is consistent with their being well described by the SI model. That is, there are most likely some consistent ways that people behave prosodically, but for some speakers who deviate from these norms, speaker-dependent modeling can be of value.

Figure 9.1 displays relative *feature usage* statistics for those speakers for whom there is an improvement from using SD information. As in the speaker-independent feature usage analysis presented in Section 8.6.2, the prosodic features were grouped into five nonoverlapping groups: pause at the boundary in question, duration, pitch, energy, and "near pause". The figure compares the SD feature usage distribution with the SI distribution, for native speakers (the left-hand graph) and nonnative speakers (the right-hand graph).

The three natives show very similar usage to each other and to the SI model. However, as we saw earlier, SD models improve their results. This suggests that even when general feature usage patterns for a talker are similar to those of the SI model, specific features and/or feature thresholds may still be better modeled by training on the specific speaker. Given only three

Table 9.4: DA segmentation performance comparison of SI, SD, and SI+SD prosodic models in ASR conditions [BER %]. The best result for each speaker is shown in boldface, * indicates that the improvement of SI+SD over SI is significant at $p < 0.05$ using the Sign test.

ID	SI	SD	SI+SD	ID	SI	SD	SI+SD
me013	8.41%	8.47%	**8.37%**	mn052	8.96%	9.30%	**8.70%**
me011	7.01%	6.80%	**6.77%**	mn021	**7.56%**	8.54%	7.56%
fe008	9.04%	8.98%	**8.90%**	me003	6.40%	6.49%	**6.27%**
fe016	9.40%	**9.39%**	9.41%	mn005	6.78%	**5.87%**	5.90%
mn015	8.09%	8.47%	**7.96%**	me045	7.93%	8.13%	**7.89%**
me018	**8.16%**	8.30%	8.18%	me025	11.14%	**10.77%**	10.83%
me010	8.62%	8.34%	**8.30%**	me006	10.94%	10.78%	**10.63%**
mn007*	10.74%	10.06%	**10.01%**	me026	7.91%	8.51%	**7.61%**
mn017	**8.04%**	8.16%	8.04%	me012	**8.99%**	9.31%	8.99%
mn082	8.03%	8.38%	**8.00%**	fn002	9.74%	**9.24%**	9.46%

native speakers showing improvements here; it is possible that not all native speakers show the same pattern, but this is a question for further research on a larger data set.

Feature usage for nonnative speakers, on the other hand, looks quite different. Speakers differ from each other, as well as from the SI pattern. Although more research is needed before drawing conclusions, this finding is nevertheless consistent with stylistic differences between nonnative speakers and an overall SI model, in prosodic marking of DA boundaries. Obvious next question would be whether improvement depends on native language, proficiency in English, or degree of perceived accent. The sample of nonnative speakers is too small to examine these questions, however, I do note that of three native German speakers, all highly proficient in English, one speaker improved from individual modeling while two others did not. Of three Spanish speakers, all moderately proficient, two improved and one did not.

Table 9.4 presents results for ASR conditions. We can see that the distribution of the best performing models differ from that for REF. In contrast to REF, most of speakers improved by using SD information. While only 7 of 20 speakers improved in REF, 16 speakers improved in the ASR-based tests. Also note that all 7 speakers who improved in REF also improved in ASR conditions. On the other hand, the magnitudes of improvement were relatively smaller. Using the Sign test, only one speaker (mn007) showed a gain significant at $p < 0.05$. One other speaker (mn005) showed a marginally significant gain ($p < 0.09$).

I suppose that some of the differences between REF and ASR are related to backchannels, often missed by the ASR system. The backchannels are rather different from other speech material, and thus their smaller number may considerably influence overall results. This supposition suggests to test prosodic adaptation on other data also having recurring speakers, yet a smaller number of backchannels and more regular prosody.

9.1.3.3 Overall Results of Speaker-Independent and Speaker-Dependent Models

After presenting results for individual speakers, I also present overall results summed over all tested speakers. These are shown for both test conditions in Table 9.5, along with chance error rate. In REF tests, the overall best performance showed the speaker-adapted SI+SD model followed by the SI model while the SD model was the worst. The improvement of SI+SD over SI in reference conditions is statistically significant at $p < 0.001$ using the Sign test. Given the absolute difference in BER, the relatively high level of significance may seem

Table 9.5: Overall DA segmentation error rates for SI, SD, and SI+SD prosodic models [BER%]

Method	**REF**	**ASR**
Chance	15.02%	13.19%
SI (Baseline)	8.25%	8.41%
SD	8.40%	8.41%
SI+SD	**8.19%**	**8.26%**

a little surprising, but there exist two explanatory reasons. First, the overall test set is really huge (204k words). Second, if the SD information does not benefit a particular speaker, it often get zero weight in the interpolation model. Thus, both SI+SD and SI show identical output for many unimproved speakers which decreases variance between overall output and consequently supports statistical significance in the Sign test. The difference between SI+SD and SD is significant at $p < 0.01$ and between SI and SD at $p < 0.05$.

In ASR conditions, the SI+SD model was also the best performing one. The superiority of SI+SD over SI and SD used alone was significant at $p < 0.05$. The relative gain by using SI+SD instead of SI was higher than in reference conditions. Interestingly, the SD model performed almost as well as the SI model. In total, SD made only 4 more errors than SI in the 180k test samples.

9.2 Speaker-Dependent Language Modeling

While the previous section dealt with speaker-dependent prosodic modeling, this section focuses on speaker-dependent language modeling for DA segmentation.

9.2.1 Motivation and Goals

Section 9.1 indicated that the prosody-based automatic DA segmentation system may for some speakers show good results with prosodic models trained only on a small amount of speech from the same speaker. However, similar results cannot be expected for LMs since LMs trained only on such small amounts of data as we have available for individual talkers would apparently suffer from large out-of-vocabulary rates and yield inferior performance. For speaker-dependent language modeling, nevertheless, we can get inspired by the good results achieved by the SI+SD prosodic model that combines speaker-independent and speaker-dependent information. Thus, speaker-dependent modeling is viewed in terms of speaker LM adaptation in this study. The goal is to try to adapt the LM to capture speakers' idiosyncratic lexical patterns associated with DA boundaries.

Speaker adaptation of LMs is not meaningful in all domains. First, the target domain must show spontaneous speech. Evidently, it would not be reasonable to try speaker LM adaption on, for example, broadcast news where anchors just read prompted text prepared by another person. In addition, the possible target domain should also have recurring speakers since it seems to be very difficult to perform effective speaker LM adaptation to previously unseen speakers. Both conditions are met for multiparty meeting data, which presents the opportunity for investigating speaker LM adaptation.

General LM adaptation has been studied rather extensively in speech recognition and other language processing tasks, using both supervised [171, 172] and unsupervised [173, 174, 175] approaches. A useful survey of LM adaptation techniques is given in [176]. The typical approach is to use the test words to determine the topic for the test document, and then use the topic-specific LM or its combination with the generic LM.

Topic- and domain-based LM adaptation approaches have received significant attention in the literature, but much less is known about LM adaptation for individual talkers. First of all, Besling and Meier [177] improved an automatic speech dictation system by speaker LM adaptation based on the LM fill-up method. Akita and Kawahara [178] showed improved recognition performance using LM speaker adaptation by scaling the N-gram probabilities with the unigram probabilities estimated via probabilistic latent semantic analysis. Tur and Stolcke [179] demonstrated that unsupervised within-speaker LM adaptation significantly reduced word error rate in meeting speech recognition.

Unlike previous work in LM adaptation (mostly topic-based adaptation in the task of speech recognition), the goal of the study presented in this section is to investigate whether speaker adaptation of LMs may help in automatic DA segmentation of meetings. The remainder of this section is structured as follows. Subsection 9.2.2 describes used LMs, the speaker adaptation approach, and the experimental setup. Subsection 9.2.3 presents results and discussion; the results are presented separately for individual speakers as well as summed over all speakers to enable an overall model comparison.

9.2.2 Method

9.2.2.1 Language Models

HELMs (described in detail in Section 7.1.3) that showed good results in previous work are used to automatically detect DA boundaries in the unstructured word sequence. The model is trained by explicitly including the DA boundary token in the vocabulary in word-based N-gram LM. I use trigram LMs with modified Kneser-Ney smoothing. During testing, the HELM performs forward-backward decoding in the HMM framework.

9.2.2.2 Speaker Adaptation Approach

To adapt the generic speaker-independent LM to a particular speaker, I used the simple yet powerful interpolation approach again. The speaker-adapted model is obtained from a linear combination of the speaker-independent model SI and a speaker-dependent model SD as follows:

$$P_{SI+SD}(t_i|h_i;\lambda) = \lambda P_{SI}(t_i|h_i) + (1-\lambda)P_{SD}(t_i|h_i) \tag{9.2}$$

where t_i denotes a token (word or DA boundary) and h_i its history of $n-1$ tokens within an N-gram LM. λ is a weighting factor that is empirically optimized on held-out data. Note that the SD data is already contained in the SI data for LM training; therefore, this interpolation does not help reduce out-of-vocabulary rate, it rather gives a larger weight to N-grams observed in the data corresponding to a particular speaker and is expected to be better suitable to this speaker.

9.2.2.3 Data and Experimental Setup

I use the same data split as for the prosodic adaptation experiments, as described in Section 9.1.2.4. Thus, all data set sizes presented in Table 9.1 are also valid for the language-modeling experiments. As in previous work, I test the method using both manual and automatic transcripts; DA segmentation performance is measured using BER. Likewise, I also employ the jackknife approach, instead of using a separate development test. The only notable

Table 9.6: DA segmentation error rates for speaker-independent (SI) and speaker-adapted LMs for individual speakers in REF conditions [BER %]. AdInW=Adaptation with individual weights, AdGlW=Adaptation with global weights. The best result for each speaker is shown in boldface, * and ** indicate that the improvement over SI is significant by the Sign test at $p < 0.05$ for one or both methods, respectively.

ID	SI	AdInW	AdGlW	ID	SI	AdInW	AdGlW
me013**	6.75%	6.55%	**6.52%**	mn052	7.33%	**7.28%**	7.28%
me011*	7.40%	7.38%	**7.25%**	mn021**	6.68%	**5.41%**	5.65%
fe008**	7.51%	**7.12%**	7.16%	me003	8.78%	**8.45%**	8.56%
fe016*	7.35%	7.22%	**7.18%**	mn005**	7.83%	7.01%	**6.92%**
mn015**	8.05%	**7.75%**	7.80%	me045	**8.90%**	8.94%	8.90%
me018*	6.64%	**6.43%**	6.45%	me025	8.06%	8.02%	**7.85%**
me010**	7.24%	6.96%	**6.84%**	me006	9.53%	10.32%	**9.47%**
mn007*	7.59%	7.36%	**7.31%**	me026	5.80%	**5.76%**	5.80%
mn017**	7.02%	**6.44%**	6.44%	me012**	6.85%	**6.29%**	6.29%
mn082	6.33%	6.28%	**6.21%**	fn002	10.92%	**10.79%**	11.33%

difference compared to the prosodic adaptation setup is that I test two approaches to estimate the interpolation weights for LM adaptation.

A robust estimation of the interpolation weight λ in (9.2) may be a problem because of data sparsity. In the jackknife approach, one half of speaker's test data is used to estimate λ for the other half, and vice versa. I test two methods for estimating λs. First, λs are estimated individually for each speaker. In the second method, I set λ as the average value of the interpolation weights across all the speakers. Note that in the latter approach, however, I still use only data from the first halves of individual test sets to estimate λ for the second halves, and vice versa. This approach eliminates having two significantly different values of λ for a single speaker, which did occur for some of the 20 analyzed speakers. It indicated that for those speakers, there is a mismatch in the two halves of the test data used in jackknife, and thus the weights were not optimized properly for the test set.

9.2.3 Results and Discussion

9.2.3.1 Results for Individual Speakers

Table 9.6 shows a comparison of DA segmentation performance for the baseline speaker-independent LM and speaker-adapted LMs for individual speakers, using reference transcripts. The speakers displayed in the table are sorted according to the total numbers of words they have in the corpus. The results indicate that for 17 of 20 speakers, performance improved using both individual and global weights, and two other speakers improved only for one of the two interpolation methods. However, the degree of the improvement varies across particular speakers. For 8 talkers, the improvement was statistically significant at $p < 0.05$ using the Sign test for both methods. For 4 others it was significant for only one of the methods.

Table 9.7 reports the corresponding results for the ASR conditions. The results show that 15 speakers improved using both interpolation methods, while 4 other speakers improved just for one of the methods. Again, for 8 talkers, the improvement was significant at $p < 0.05$ for both methods, and for 4 others the improvement was significant for one method. An interesting observation is that for both testing conditions, the relative error reduction achieved by speaker

Table 9.7: DA segmentation error rates of speaker-independent and speaker-adapted LMs for individual speakers in ASR conditions [BER %]. AdInW=Adaptation with individual weights, and AdGlW=Adaptation with global weights. The best result for each speaker is shown in boldface, * and ** indicate that the improvement over SI is significant by the Sign test at $p < 0.05$ for one or both methods, respectively.

ID	SI	AdInW	AdGlW	ID	SI	AdInW	AdGlW
me013**	8.29%	**8.16%**	8.18%	mn052*	10.87%	10.49%	**10.17%**
me011**	8.81%	8.59%	**8.51%**	mn021*	8.20%	7.83%	**7.73%**
fe008*	9.19%	9.05%	**8.89%**	me003	9.36%	9.36%	**9.33%**
fe016	8.42%	8.40%	**8.31%**	mn005**	11.47%	**8.94%**	10.42%
mn015**	10.16%	9.90%	**9.84%**	me045	**11.08%**	11.42%	11.27%
me018**	8.12%	**7.83%**	7.90%	me025	14.36%	**14.23%**	14.36%
me010**	8.39%	7.96%	**7.91%**	me006*	10.94%	**10.01%**	10.40%
mn007**	11.28%	**10.73%**	10.78%	me026	7.35%	7.01%	**6.88%**
mn017**	8.92%	8.01%	**7.84%**	me012	8.68%	**8.15%**	8.31%
mn082	10.37%	10.45%	**10.10%**	fn002	13.40%	13.40%	**12.89%**

Table 9.8: Overall DA segmentation error rates of speaker-independent and speaker-adapted LMs in reference and ASR conditions [BER %]

Method	REF	ASR
Chance	15.02%	13.19%
SI (Baseline)	7.30%	9.06%
SI+SD: Individual weights	7.02%	8.79%
SI+SD: Global weights	**6.99%**	**8.76%**
SI+SD: Adapt. with ASR data	N/A	8.97%

adaptation is not correlated with the amount of adaptation data. This finding suggests that speakers differ inherently in how similar they are to the generic speaker-independent LM. Some talkers probably differ more and thus show more gain, even with less data.

9.2.3.2 Overall Results of Speaker-Independent and Speaker-Dependent Models

An overall comparison of performance of baseline speaker-independent and speaker-adapted LMs is presented in Table 9.8. The test set contains 204k words for REF and 180k words for ASR conditions. These results show that for both conditions, speaker-adapted LMs – with either global interpolation weights or individual weights – outperform the baseline. The overall improvements by LM speaker adaptation for both conditions are statistically significant at $p < 10^{-15}$, using the Sign test. Of the two weight options, global interpolation results in better performance; however, the difference between the two approaches is only marginally significant at $p < 0.1$.

In ASR conditions, I also tried interpolating the speaker-independent model trained on reference transcriptions with a speaker-dependent model trained on the recognizer output. The idea was to allow the model also to adapt for error patterns typical for an individual talker. However, this adaptation performed less well than using reference transcriptions as the training data, which indicates that, at least with the amount of data available for our experiments, it is preferable to adapt LMs using clean data. In consequence it also suggests that prospective unsupervised approaches to LM speaker adaptation will perform less well than the supervised approach.

9.3 Chapter Summary and Conclusions

In this chapter, I have investigated speaker-specific prosodic and language modeling for DA segmentation in meetings. The method was evaluated on 20 frequent speakers with a wide range of total words available for speaker-dependent modeling. First, it was found that overall, prosodic features beyond pause provide benefit over the pause-only features for 19 of the 20 speakers studied. Further, it was found that interpolating the large, speaker-independent prosodic model with a much smaller prosodic model trained only on that talker's speech yielded improvements for 6 of the 20 speakers in reference conditions, and for 16 of 20 the speakers in ASR conditions. The ASR conditions showed a higher number of improved speakers, but the improvements were relatively smaller than those in the reference conditions. Overall results, summed over all 20 speakers, indicate modest yet significant improvement with respect to the SI-model for both test conditions.

Feature analysis, while preliminary given the number of speakers, suggests that nonnative speakers may differ from native speakers in overall feature usage patterns associated with DA boundaries. An important question for future work is to explore what factors predict whether speaker-dependent modeling will benefit a particular speaker since it did not benefit all speakers. The absolute amount of data did not appear to be a predictor in our experiments, although data is certainly necessary for robustness.

In the second set of experiments, I explored speaker adaptation of hidden event language models for the same task. Analogous to prosodic model adaptation, the speaker LM adaptation was based on a linear combination of the generic speaker-independent and speaker-dependent LMs. Improvements were found for 17 of the 20 speakers using reference transcripts, and for 15 of the 20 speakers using automatic transcripts. Overall, I achieved a statistically significant improvement over the baseline LM for both test conditions. It can be concluded that speaker adaptation of LMs aids DA segmentation, and that future work should investigate the potential of speaker-specific modeling for other spoken language understanding tasks.

For both types of speaker adaptation, improvements were achieved even for some talkers who had only a relatively small amount of data available for adaptation. In addition, the relative error reduction achieved by speaker adaptation was not correlated with the amount of adaptation data. This finding suggests that speakers differ inherently in how similar they are to the generic models. Some talkers probably differ more and thus show more gain, even with less data.

An obvious question is whether the speakers who benefit from prosodic model adaptation also benefit from language model adaptation. Of the 6 speakers who improved by prosodic adaptation in reference conditions, 5 also improved by language model adaptation. The number of improved speakers in common may seem to be high, but it was not higher than the chance agreement based on the counts of improved speaker in both sets (5.1 speakers). Similarly, of the 16 speakers who improved by prosodic adaptation in ASR conditions, only 10 improved also by language model adaptation, while the chance agreement was 12 speakers. These numbers indicate that that there was no apparent correlation between speakers' idiosyncrasy in prosodic and lexical patterns associated with DA boundaries.

Chapter 10

Sentence Unit Segmentation of Czech Corpora

Change your language and you change your thoughts.
KARL ALBRECHT

While the experiments presented in the two preceding chapters focused on sentence segmentation of conversational English, this chapter focuses on sentence segmentation of spoken Czech using the two MDE corpora described in Chapter 5. The same three modeling approaches as in the experiments with the meeting data (HMM, MaxEnt, and BoosTexter) are compared herein. In addition, I analyze gains from using various textual and prosodic features.

The remainder of this chapter is organized as follows. Section 10.1 summarizes most prominent differences between Czech and English, Section 10.2 surveys related work on sentence segmentation of spoken Czech, and Section 10.3 defines the task and experimental setup. Sections 10.4, 10.5, and 10.6 report results of the experiments based on using only textual information, only prosodic information, and a combination of both information sources, respectively. Section 10.7 presents a system combining all three modeling approaches, and Section 10.8 summarizes all experiments and draws conclusions.

10.1 Differences between Czech and English

When dealing with Czech data, we must cope with some specific issues that are absent in English. Czech belongs to the family of Slavic languages, which are highly inflectional and derivational. The languages with a highly inflectional morphology typically use an extremely large number of distinct word forms. For example, Czech nouns have seven cases in the singular number and another seven cases in the plural number. Even though some of the inflected forms are identical, the number of distinct word forms relating to lemmas of inflected parts of speech is high. Hence, the ubiquitous problem of data sparseness is even more painful for Czech.

Additional problems arise when we must deal with colloquial Czech [180]. Colloquial Czech deviates from standard Czech as defined by orthographic, morphological, lexical and syntactic rules by the Czech normative bodies. With respect to pronunciation variation, Czech is different from English and many other languages in that spelling rules for Czech are phonetically based. Therefore, colloquial Czech words have well-defined but different spellings than their standard variants. In other words, colloquial Czech has an orthographic written form. The difference between colloquial and standard Czech is most prominently displayed in the morphology – prefixes and endings are often changed in the former.

Another problem is a relatively free word order in Czech. Word order flexibility is generally correlated with rich inflection in a language. In highly inflective languages, syntactic roles of sentence members are often disambiguated based on morphological information, rather than a position in a sentence. Although the word order flexibility does not imply that the word (or phrase) order may be chosen absolutely arbitrarily, it evidently affects the predictive power of statistical models based on N-gram contexts.

On the prosody side, the largest difference between Czech and American English is probably in typical sentence melody. While sentence-final pitch falls/rises are present in both languages, intrasentential pitch movements (e.g., at prosodic phrase boundaries) are typically less steep in Czech than in English. Another distinction is caused by a different function of stress in the two languages. Stress has a lexical function in English, i.e. it may distinguish word identities. Stressed syllables in English are typically louder, as well as being longer and having a higher pitch than non-stressed syllables. On the other hand, stress in Czech has only a delimitative function. Stress is fixed to the first syllable of a foot and its only function is to acoustically delimit foot boundaries. This delimitative stress is usually less strongly marked than the lexical stress in English. Furthermore, preboundary lengthening in Czech is also less emphatic than in English. The reason for this is that the length of vowels also serves a lexical function in Czech, which offers less opportunity for prosodically motivated lengthening. All these differences make Czech prosody sound “flatter” than English; foreigners often find spoken Czech rather monotonous.

10.2 Related Work on Czech

At the time I started to work on my thesis, there was no published research relating to sentence segmentation or automatic punctuation of spoken Czech. Thus, our (Kolář, Švec, and Psutka) paper from 2004 [181] was the first published work focusing on this language. Since the early experiments from this study are not included in this thesis, I briefly summarize their results here. The pilot paper focused on automatic punctuation (insertion of commas and periods) of Czech broadcast news data. Although the same Czech BN corpus was used, we did not have the MDE annotation available at that time. Thus, instead of the MDE symbols, we had to use the slightly inconsistent punctuation from the first transcription pass.

We used the HMM approach and tested two prosodic classifiers – CART and MLP. Of the two classifiers, CART worked slightly better. On the language modeling side, a baseline word-based trigram was outperformed by a POSmix model in which infrequent word forms were replaced by their morphological tags. The methods were only tested on automatically aligned reference transcripts and the best performing combined system (HELM+CART) achieved $BER = 4.8\%$ and $F = 78.2\%$ for all punctuation marks. For sentence boundary detection, the results were only reported in F-measure ($F = 88.1\%$).

One year later, Kolorenč presented an automatic system for punctuation of automatically recognized Czech broadcast news [182]. He used a simple approach utilizing automatically learned rules, separate for periods and commas. The period insertion rules were based on replacing automatically recognized “noises”, such as long pauses or audible breaths, by periods. For induction of the replacement rules, a genetic algorithm based on a grammatical evolution approach was employed. By contrast, the rules for commas were induced from a newspaper corpus. These rules were only relying on two words following the interword boundary of interest and looked for positions of conjunctions, pronouns, adverbs, and prepositions (i.e., potential clause-joining words).

After applying the learned rules to speech recognizer output, a postprocessing step based on a simple morphological analysis was performed. In this step, all punctuation marks were removed from all segments not containing nouns, numbers or pronouns in the first case (possible subjects), or verbs in a finite form (possible predicates). The author reports the performance rates as $F = 81.9\%$ for periods and $F = 83.2\%$ for commas. Overall, my view on this paper is that while the presented approach is interesting, it seem to be suboptimal. For period insertion, it only looks for noise flags instead of using prosodic and lexical features. Furthermore, the comma insertion rules do not take into account the words occuring before the classified boundary. Moreover, the rules learned in that way do not seem to be robust against word recognition errors.

10.3 Data and Experimental Setup

10.3.1 Segmentation Task for Czech Data

As mentioned above, this chapter uses the two MDE-annotated Czech corpora described in Chapter 5. The diversity of MDE symbols allows us to define the sentence-unit segmentation task for these corpora in a variety of ways. For example, we can only recognize sentence-like units delimited by the double slash symbols or, on the other hand, we can segment speech into smaller syntactic units with boundaries marked by both sentence-external and sentence-internal SU breaks. However, since this part of my work aimed to perform an initial set of experiments with SU segmentation of spoken Czech, I decided to follow the original SU definition as used in the EARS project. Thus, I defined a two-way classification task in which a sentence-like unit boundary was recognized when the current word was followed by any SU-external symbol (both complete and incomplete – "/.", "//.", "/?", "//?", "/-", "/∼"). On the other hand, words followed by SU-internal symbols ("/&", "/,") or no symbol were treated as "non-boundary".

10.3.2 Experimental Setup

The two Czech corpora were evaluated separately. As usual, each corpus' data were split into a training set, a development set, and a test set. The splits were performed based on the dates of broadcast in order to ensure that the training data only contain older recordings than those on which the models are tested. Particular data set sizes in terms of a total number of words are shown in Table 10.1. Note that, in line with the setup of the previous English experiments, pseudo-words, such as filled pauses or "uh-huhs", are treated as word tokens during testing and evaluation. On the other hand, background noises or mouth noises, such as loud breaths or coughs, are not taken into account during evaluation and therefore not counted in the table.

The numbers in Table 10.1 show that the automatically generated transcripts contain slightly more words than the manual transcripts, indicating that the employed ASR system (described in the following section) was making more insertion errors than deletion errors. Also note that both Czech corpora are significantly smaller than the ICSI meeting corpus used in the two previous chapters. For example, both Czech training sets are more than three times smaller than the training set of the meeting corpus.

In terms of the number of broadcast programs, the BN corpus setup used 244 programs for training, 39 for development, and 59 for testing. The seemingly unbalanced distribution of the broadcast programs among the sets was caused by largely varying lengths of the recordings in the corpus. The development and the test set were designed to contain an approximately equal number of words, regardless of the number of shows. On the other hand, the RF corpus comprises talk shows of a similar length. Hence, the RF split – 40 shows for training, 6 shows for development, and 6 shows for testing – was balanced in this respect. The evaluation sets

Table 10.1: Data set sizes for Czech corpora (numbers of words)

Data Set	**BN**		**RF**	
	REF	**ASR**	**REF**	**ASR**
Training	174.8k	N/A	159.1k	N/A
Development	28.2k	28.6k	24.1k	24.2k
Test	31.2k	31.6k	24.6k	24.7k

contain unseen speakers as well as speakers appearing in the training data, as it is typical for the real world applications. For example, the talk shows usually have recurring interviewers, whereas the invited guests typically occur only in a single show.

To generate "reference" SU boundaries for ASR transcripts, I used the same approach as described in Section 8.4. In contrast with previous experiments on the meeting data, I did not examine models trained on ASR data. This decision was made because the Czech ASR system was trained on the same data as I use for training of my segmentation models. Consequently, it was expected that the training portion of the data would be recognized with a higher accuracy, causing an apparent mismatch between training and test data.

As in all preceding experiments, BER is used as the main performance measure. In addition, chance error rate is always reported. To ease a performance comparison across the experiments, I also present the $NIST$ error rate and F-measure for best results in individual sections. All experiments are evaluated using both human-generated and automatic speech transcripts.

10.3.3 Speech Recognition System for Czech

Although building ASR models for Czech data was not one of the thesis goals, I briefly mention the speech recognition system here since I had to build my own ASR models as a necessary preliminary step for my work on sentence segmentation. For automatic recognition of the Czech speech data, I used the LVCSR system developed at UWB [183]. The system was designed for real-time recognition of highly inflected languages.

In the UWB recognizer, each individual basic speech unit (triphone) is represented by a three-state HMM with a continuous output probability density function assigned to each state. Feature vectors are computed at the rate of 100 frames per second using a parametrization with 12 PLP cepstral coefficients plus corresponding delta and delta-delta subfeatures. For decoding, the system employs a time synchronous Viterbi search with token passing within a lexical-tree recognition network. In the first pass, the recognizer uses a bigram language model to generate word lattices, which are rescored in the second pass using a higher order language model.

The acoustic models for either corpus were trained in HTK only using data from the MDE training sets of the same corpora. For training of language models, I have used a database provided by a Czech media-monitoring company. This database contained manual transcripts of radio and TV broadcasts spanning the years from 1996 to 2006. First, the pieces of text had to be preprocessed to normalize non-standard tokens, such as numerals, dates, or abbreviations. For this purpose, a text normalization system tailored for languages with rich inflection was employed [184]. Then, the transcripts were manually divided into five groups according to the program type: news, discussions, economics, investigative journalism, and sports. Since the target domain does not include sports, this subpart of the training corpus was not used, nor all the data corresponding to the dates on which our development and test sets were recorded.

The remaining four subcorpora yielded 107M words in total.

For each of the subcorpora, I trained a separate trigram language model with modified Kneser-Ney smoothing in the SRILM toolkit [138]. Subsequently, the four LMs plus a model trained only on the MDE training set transcripts were interpolated to form a mixture LM with five components. The interpolation weights for the two target corpora were estimated on corresponding development sets using the EM algorithm. Thus, the LM for recognition of broadcast news was different from the LM for broadcast conversation recognition. The recognizer's vocabulary contained 200k words. The words for the ASR vocabulary were selected using the vocabulary optimization method described in [185]. A relative weight of the LM for a combination with the acoustic model and the word insertion penalty value were tuned using the development data. The overall word error rates were 12.4% for the BN and 29.3% for the RF corpus. The same automatic system was also used to generate forced alignments of manual transcripts.

10.4 SU Segmentation Based on Textual Information

10.4.1 Textual Features

10.4.1.1 Words

In this set of experiments, the term "word features" is used to refer to word-based features only extracted from the training portions of the two Czech MDE corpora. Before I started to evaluate the word-based segmentation experiments, I had focused on the issue whether the word-based LMs for either corpus may benefit from also using training data from the other corpus. The answer to this question was not straightforward. While both corpora have been MDE annotated in the same way, they largely differ in speaking styles. My initial experiments showed that using data from the other corpus was not helpful for either corpus. Besides employing the data from the Czech MDE corpora, I also investigated using data from other sources. The exploitation of auxiliary textual resources is discussed below in Subsection 10.4.1.4.

10.4.1.2 AICs

Automatic word classes for the Czech data were induced in the same way as for the meeting data (as described in Section 8.5.1.2). The optimization of the number of target classes was performed for each corpus separately. It was based on evaluation of models with various class granularities on corresponding development data. The optimal numbers of classes were estimated as 300 for the BN corpus and 275 for the RF corpus. These numbers are higher than the number for the ICSI meeting corpus, which only used 100 classes. The differences are not very surprising since we must take into account that, although the Czech corpora are smaller, their vocabularies contain a higher number of word forms. While the training portion of the ICSI corpus contains 11,034 distinct word forms, the training set of the BN corpus contains 26,805 word forms and the training set of the RF corpus 20,919 word forms. Thus, the corpora with larger vocabularies also show a higher number of resulting word clusters.

10.4.1.3 POS for Czech

Since the use of POS-based features helped in a variety of Czech speech processing tasks, such as speech recognition [186, 187] or semantic parsing [188], employment of POS information for sentence segmentation of spoken Czech is explored here as well. In contrast to languages with poor inflection (such as English), highly inflected languages (such as Czech) often use

Table 10.2: Description of individual tag positions in PDT tagset

No.	**Category**	**No.**	**Category**
1	POS	9	Tense
2	Detailed POS	10	Degree of comparison
3	Gender	11	Negation
4	Number	12	Voice
5	Case	13	Reserve 1
6	Possessor's Gender	14	Reserve 2
7	Possessor's Number	15	Special Usage
8	Person		

structured morphological tagsets. In addition to labeling words with a POS category, these structured tagsets use tags comprising of "subtags" providing information about morphological categories.

For Czech, the most popular tagset is the positional tag system from the Prague Dependency Treebank (PDT) [189].[1] In this tagset, every tag is represented as a string of 15 columns (positions). 13 of the 15 positions correspond to individual morphological categories, which approximately fit the formal Czech morphology. The two remaining positions are currently unused and kept as reserves for a possible future use. The description of individual positions is presented in Table 10.2. Values in each position are represented by a single character, mostly an uppercase letter. The values that are not applicable for a particular word (e.g., Gender for prepositions) are denoted by a single hyphen (-). For example, the word form "*rezignoval*", lit. "*(he) resigned*", is tagged as VpYS---XR-AA---.

It is evident that the structured tagsets for morphologically rich languages are much larger than the compact tagsets for languages with a poor morphology. While the *Penn Treebank Tagset* [191] used for our English data contains just 36 POS tags plus 12 tags for punctuation, the rich tagsets usually contain more than 1,000 distinct tags. For example, there is about 1,500 different tags in the PDT corpus. A theoretical number of possible tags for Czech is even higher.

Several automatic taggers have been developed for Czech. In this work, I used automatic tags obtained from the Morče tagger[2] [118], which is based on the averaged perceptron method. Because the Czech tagset is so rich, it was worth exploring whether the sentence segmentation systems could benefit from using features based on some reduced tags. I explored various reduced tags including: (1) tags containing just the second position - Detailed POS; (2) tags containing just first five positions; (3) the pair comprising of POS plus Case (or Detailed POS instead of Case for POSs for which Case is not defined) often used for parsing of Czech; (4) the reduced tag found optimal in [181] - Detailed POS, Case (reduced to nominative, genitive, accusative, and "other"), person, tense, and grade. Of the four tested reductions, only option (4) achieved performance close to that using the full tags. However, since none of the reduced tagsets yielded improved results on the development data, I chose to use the full tags in all the following POS-based experiments.

Besides the pure POS-based models, I also tested the models combining tags with frequent

[1] Besides the PDT tagset, there also exists an alternative tag system called *Ajka* [190]. This tagset is not described herein since I do not use it in this work.

[2] The tagger is available from `http://ufal.mff.cuni.cz/morce/`.

words (POSmix). Optimizing the model on my development data, I ended up with 1600 most frequent word forms being kept for the BN corpus, and 2000 word forms being kept for the RF corpus. In terms of relative frequency, the cutoff values for not replacing words by their POS was $8.57 \cdot 10^{-5}$ (i.e. 15/175k) for the BN data and $5.66 \cdot 10^{-5}$ (i.e. 9/159k) for the RF corpus. These boundary values are lower than the cutoff value for the English meeting corpus, which was $1.87 \cdot 10^{-4}$.

10.4.1.4 Auxiliary Words

In my experiments with Czech data, I did not only employ textual data from the MDE corpora, but also investigated benefits from using an additional text corpus. In comparison with English, the need for additional text data is larger for Czech, since the rich morphology sharpens the data sparseness problem. Some additional textual data are often available, but we must take into account two important facts. First, such data are typically not annotated for SU boundaries in terms of the MDE guidelines but only contain standard punctuation. This may cause a mismatch between training and test data, and it is not clear beforehand whether the additional data would improve or rather hurt segmentation performance.

The second problem is that for the auxiliary textual data, we do not have available any prosodic features associated with the words. In the HMM approach, the auxiliary LM can easily be incorporated by interpolation with the baseline LM, but both MaxEnt and Boostexter, which do not have a separate LM, assume that all features are available during training. In addition, these two methods are not very suitable for large N-gram-based training data. Therefore, I used a trick similar to the trick that was used for incorporating prosodic features into the MaxEnt model (cf. Section 7.2.3). The HELM model is used to estimate posterior event probabilities based on the auxiliary LM, and these posteriors are subsequently used as an extra feature during training of the models from the primary data. In the BoosTexter model, the auxiliary posteriors are employed directly, while for the MaxEnt-based model, the posteriors are thresholded to yield binary features.

The auxiliary LM was trained on the same data as I used to build my LMs for the Czech ASR system (cf. Section 10.3.3). To generate reference SU boundaries for this data, the 107M word corpus of broadcast transcripts was automatically split into sentences using a set of heuristic rules based on punctuation and capitalization information. Then, a HELM model could be trained in a standard way.

10.4.2 Experimental Results

Table 10.3 displays experimental results for all three models (HMM, Maxent, and BoosTexter) and all textual feature sets (words, AIC, POS, POSmix, and Auxiliary words). For either of the two Czech corpora, each model is evaluated using both manual and automatically generated transcripts.

10.4.2.1 Results for BN Corpus

In both REF and ASR-based tests on the BN data, the best result by a model only relying on a single knowledge source was achieved by the auxiliary word (AuxWord) HMM. These results indicate that there is a good match between the textual database of broadcast transcripts and the BN corpus. The second most successful single-source feature set was POSmix. Similar to the meeting experiments, POSmix outperforms the pure POS model, however, the gap is much smaller on the Czech data. The different results between the two languages can be explained by a different number of tags in the two tagsets. The Czech tagset is much larger so that it

Table 10.3: SU segmentation error rates for LMs with various textual features [BER %]. AIC=Automatically Induced Classes, POS=pure POS-based model, POSmix=infrequent words replaced by POS tags and frequent words kept, AuxWords=Auxiliary words, REF=Reference conditions, ASR=ASR conditions, BN=Broadcast News corpus, RF=Radioforum corpus (Broadcast conversations). The best results for each model are displayed in boldface.

Model	**Used Features**	**BN**		**RF**	
		REF	**ASR**	**REF**	**ASR**
Chance	—	8.11%	8.01%	6.81%	6.89%
HMM	Words	6.58%	6.72%	5.38%	6.18%
	AIC	7.18%	7.31%	6.12%	6.55%
	POS	6.74%	7.25%	6.17%	6.98%
	POSmix	6.13%	6.39%	5.27%	5.94%
	AuxWords	5.23%	5.47%	6.25%	6.28%
	Words+AIC	6.50%	6.63%	5.43%	6.11%
	Words+POSmix	5.24%	6.23%	5.24%	5.91%
	Words+AuxWords	4.68%	5.05%	5.07%	**5.24%**
	Words+AIC+POSmix	5.98%	6.19%	5.16%	5.86%
	Words+POSmix+AuxWords	**4.53%**	**4.96%**	**4.88%**	5.27%
	Words+AIC+POSmix+AuxWords	4.56%	4.99%	4.89%	5.34%
MaxEnt	Words	6.65%	6.78%	5.76%	6.30%
	AIC	7.11%	7.23%	6.21%	6.76%
	POS	6.70%	7.00%	6.21%	6.65%
	POSmix	6.37%	6.64%	5.77%	6.21%
	Words+AIC	6.82%	6.93%	5.92%	6.56%
	Words+POSmix	6.21%	6.50%	5.64%	6.14%
	Words+AuxWords	4.71%	4.99%	5.16%	5.55%
	Words+AIC+POSmix	6.42%	6.65%	5.99%	6.48%
	Words+POSmix+AuxWords	4.65%	4.93%	**5.16%**	5.72%
	Words+AIC+POSmix+AuxWords	**4.51%**	**4.85%**	5.31%	**5.60%**
BoosTexter	Words	6.77%	7.07%	5.31%	6.10%
	AIC	7.09%	7.29%	6.11%	6.73%
	POS	7.01%	7.31%	6.08%	6.65%
	POSmix	6.72%	6.91%	5.42%	6.22%
	Words+AIC	6.95%	7.13%	5.96%	6.47%
	Words+POSmix	6.67%	6.87%	5.43%	6.22%
	Words+AuxWords	**4.70%**	**4.93%**	**5.12%**	**5.78%**
	Words+AIC+POSmix	6.85%	7.10%	5.88%	6.53%
	Words+POSmix+AuxWords	4.88%	5.15%	5.28%	5.82%
	Words+AIC+POSmix+AuxWords	4.91%	5.00%	5.47%	6.02%

forms a class-based model with a much finer granularity. As a result, it suffers less from the absence of the important cue words. Finally, we can see that the least effective textual feature set was AIC.

Note that since MaxEnt and BoosTexter did not use a separate AuxWord model but AuxWords posteriors from the HMM model, it is not possible to make a comparison across the three modeling approaches for the AuxWord feature set. However, we can compare the results of the modeling approaches with other single-source feature sets. The numbers show that the HMM approach was superior also for the second best feature set (POSmix), as well as for the third best feature set (Words).

In general, lower error rates were achieved by feature sets relying on more than one information source. The best performing feature sets differ across the modeling approaches. The globally best result for both test conditions was achieved by the MaxEnt model combining all four knowledge sources. This best result was $BER = 4.51\%$ ($NIST = 55.64\%$, $F = 70.59\%$) for BN REF and $BER = 4.85\%$ ($NIST = 60.52\%$, $F = 67.27\%$) for BN ASR. On the other hand, HMM worked best with Words+POSmix+AuxW and BoosTexter run best when only Words and AuxWords were used.

Overall, it seems to be very difficult to get some gain from the AIC information. Only the MaxEnt model improved by adding AICs to the overall feature set. For BN REF, the gain was significant at $p < 0.03$ using the Sign test. For BN ASR, the p-value of the same test was 0.11.

For all models, the differences between the results achieved by the baseline word-based feature set and the best performing combined feature sets were large and statistically significant at p-values close to zero. Using the Sign test, I have also evaluated the gaps between the results of the three modeling approaches. In reference conditions, the difference between MaxEnt and HMM is not significant, whereas the differences MaxEnt–BoosTexter and HMM–BoosTexter are significant at $p < 0.02$. In ASR conditions, the gaps are much smaller; only the supremacy of MaxEnt over HMM has a p-value smaller than 0.1. But even for that case, the significance level is just $p < 0.08$.

10.4.2.2 Results for RF Corpus

Results for the RF corpus are also shown in Table 10.3. In both REF and ASR-based tests on the RF data, the best result by a model only relying on a single knowledge source was achieved by the HMM using POSmix features. By contrast, the best single-source BoosTexter was using the Word features for both conditions, and the MaxEnt model worked best with Words in REF and with POSmix in ASR conditions.

Unlike BN, the HMM with AuxWords was not performing very well on the RF data when used alone. It indicates that the differences between the training broadcast text database and spontaneous conversations are rather big. The differences are not only in the colloquiality of the used language but also in sentence definitions. SUs in the RF corpus differ from written sentences to a larger extent than SUs from the BN corpus. In addition, AuxWords suffer from the absence of pseudo-words such as filled pauses or “uh-huhs”. These pseudo-words are frequent in the corpus transcripts but absent in the auxiliary training text. However, as shown by the results of combined feature sets discussed, the AuxWord information help improve the overall performance when AuxWords are not used alone but in a combination with other feature sets.

The models using feature sets relying on more than one information source displayed generally better results than the models with single-source features. The best performing feature sets differ across modeling approaches and test conditions. The globally best performing approach was HMM. For reference conditions, the best result was achieved with Words+POSmix+AuxWords – $BER = 4.88\%$, $NIST = 70.19\%$, $F = 55.14\%$. For ASR conditions, the best result was achieved with Words+AuxWords – $BER = 5.24\%$, $NIST = 75.95\%$, $F = 49.04\%$. BoosTexter performed at its best in both conditions when combining Words and AuxWords. On the other hand, MaxEnt worked best with Words+POSmix+AuxW on manual transcripts and with Words+AIC+POSmix+AuxW on automatic transcripts. This was the only top feature set that included the AIC information. However, even here the improvement by adding AICs was only marginally significant ($p < 0.06$), as shown by the Sign test.

I also checked statistical significance of gains from adding the AuxWord information to the overall feature sets. The comparison of the best results achieved with the Auxword information (HMM with Words+POSmix+AuxWords for RF REF and with Words+AuxWords for RF ASR) and without the AuxWord information (HMM Words+AIC+POSmix for both conditions) showed that these improvements were significant at $p < 0.001$ for REF and at $p < 10^{-6}$ for ASR conditions.

The gaps between the baseline models only employing Words and the models with best performing feature sets were smaller for the RF corpus than for the BN corpus. For HMM, the statistical significance levels of the improvements were $p < 10^{-5}$ for REF and $p < 10^{-14}$ for ASR. Likewise for Maxent, the significance levels were $p < 10^{-5}$ for REF and $p < 10^{-8}$ for ASR. On the other hand, the improvement of BoosTexter showed in reference conditions a p-value of just 0.09. In ASR conditions, the improvement of BoosTexter was significant at $p < 0.01$. Thus, it can be concluded that both HMM and MaxEnt display greatly significant improvements by adding the textual information beyond the RF training set words, while the BoosTexter model was not that largely successful in getting gain from the additional textual features.

Finally, I measured statistical significance of the differences between the results of the three modeling approaches. For reference conditions, the supremacy of HMM was significant – over MaxEnt at $p < 0.005$ and over Boostexter at $p < 0.03$. The gap between BoosTexter and MaxEnt was not significant. For ASR conditions, the dominance of HMM over the other models was significant at $p < 10^{-5}$. The prevalence of MaxEnt over BoosTexter was significant at $p < 0.05$.

10.4.2.3 Result Comparison across Corpora

The most distinctive difference between the results for BN and RF is in the importance of AuxWords. The auxiliary training text resembles more the BN corpus than the RF corpus. Accordingly, it is more beneficial for sentence segmentation of the former corpus. When AuxWords are used alone, they reduce the chance BER for BN REF and BN ASR by 35.5% and 37.1% relative, whereas for RF REF and RF ASR, they reduce it just by 8.2% and 8.5% relative, respectively. In addition, a comparison of the best results achieved with AuxWord features shows difference in improvements. The addition of AuxWords reduce BER by 13.9% relative for BN REF, by 22.2% for BN ASR, by 5.4% for RF REF, and by 10.6% for RF ASR.

Performance of individual modeling approaches also differ between the two corpora. MaxEnt was the most successful for the BN corpus, while HMM was the best performing method for the RF corpus. The superiority of HMM for RF was greater than the superiority of MaxEnt for BN. On the other end, BoosTexter was the worst for three of the four test conditions. The only exception was RF REF, where it slightly outperformed MaxEnt. The dominance of HMM for the RF data indicates that the smoothing method used in the HELM is more robust to lexical irregularities frequent in colloquial Czech data than those used in other two approaches.

We can also compare relative error reductions with respect to chance achieved by the best models in individual test sets. The relative reductions are 44.4% for BN REF, 39.5% for BN ASR, 28.3% for RF REF, and 23.9% for RF ASR. It is possible to see that language modeling was more succesful in the BN corpus. The differences between the corpora are also in the performance degradation caused by ASR errors. SU segmentation performance in ASR tests is relatively more degraded in the RF corpus, which was recognized with a much higher WER.

Interesting insights are also offered by the comparison with the LM results on English meeting data. For that corpus, the relative error reductions with respect to chance were 55.9% a 49.3% for REF and ASR, respectively. Despite not using any additional textual data, language modeling for the English meeting corpus was more successful than for the Czech corpora. Although such a comparison may be a bit imprecise beacuse of easier-to-detect backchannel DAs frequent in the English meeting corpus, it yet illustrates that language modeling is more difficult for Czech.

10.5 SU Segmentation Based on Prosodic Information

In this section, I evaluate performance of prosodic models. The prosodic feature selection for Czech corpora was performed in the same way as for the English meeting data (cf. Section 8.6.2). In contrast with meetings, the number of selected features was lower. The feature reduction algorithm ended up with 11 features for the BN corpus and 17 features for the RF corpus. The varying number of features may not only be explained by different characteristics of individual corpora (number of channels, level of spontaneity, rate of backchannels), but also by different amount of training data. The meeting corpus is much larger than the Czech corpora, and thus it could allow a model with a higher number of features to be robustly estimated.

As in the prosody-based experiments on the ICSI meeting corpus, two modeling approaches are examined here (CART and BoosTexter) since the overall MaxEnt approach only bins prosodic scores obtained from CARTs. Besides reporting the overall accuracy of individual prosodic models, I also investigate whether there is some gain from using a richer set of prosodic features in comparison with using pause information only. In contrast with meetings, the alternative pause feature set not only contains the three features capturing pause duration after the previous, the current, and the following word, but also one “other” feature indicating speaker change. As described below, this non-prosodic feature significantly improves classification accuracy for the broadcast data. Therefore, its absence in one of the two evaluated feature sets would skew the comparison.

10.5.1 Experimental Results for Prosodic Models

Table 10.4 shows BERs achieved by particular prosody models in both reference and ASR test conditions. The two following subsections report results for individual corpora.

10.5.1.1 Results for BN Corpus

In the BN corpus, the best results were achieved by bagged CARTs with the rich prosodic feature set. In reference test conditions, this model achieved $BER = 1.81\%$ (corresponding to $NIST = 22.48\%$ and $F = 88.52\%$). In ASR conditions, it achieved $BER = 2.11\%$ ($NIST = 26.41\%$, $F = 86.37\%$). Using the Sign test, the differences between CART with rich prosody and BoosTexter with rich prosody were significant at $p < 10^{-11}$ and $p < 10^{-10}$ for the two

Table 10.4: SU segmentation error rates for prosodic models [BER %]

Model	Used Features	BN		RF	
		REF	**ASR**	**REF**	**ASR**
Chance	—	8.11%	8.01%	6.81%	6.89%
CART with Ens. Bag.	Pause	2.96%	3.75%	4.88%	5.29%
	All Prosody	**1.81%**	**2.11%**	**4.48%**	**4.82%**
BoosTexter	Pause	2.76%	3.40%	4.89%	5.30%
	All Prosody	2.28%	2.55%	4.60%	5.13%

test conditions. Moreover, the gains over pause-only models were significant at $p < 10^{-45}$ and $p < 10^{-99}$, respectively.

On the other hand, the boosting-based model was better than CART in modeling pause information alone. These results were conformable with the results for the English meeting corpus, where BoosTexter also handled the pause feature set better. The differences were significant at $p < 0.001$ and $p < 10^{-7}$ for reference and ASR-based tests, respectively.

10.5.1.2 Results for RF Corpus

For the RF corpus, the best results were also achieved by the CART model with the rich prosodic feature set. In reference test conditions, the CART model achieved $BER = 4.48\%$ ($NIST = 66.27\%$ and $F = 56.81\%$). In ASR conditions, this model achieved $BER = 4.82\%$ ($NIST = 70.21\%$, $F = 53.77\%$). Using the Sign test, the difference between rich-prosody CART and BoosTexter was significant at $p < 0.001$ in ASR conditions. On the other hand, the p-value of the same test in reference conditions was just 0.06. The two modeling approaches showed similar results when only using the pause feature set.

The gaps between rich prosody and pause only CART models showed p-values 10^{-6} and 10^{-8} for reference and ASR conditions, respectively. For BoosTexter, these p-values were lower: 0.001 and 0.05.

10.5.1.3 Result Comparison across Corpora

A comparison of relative error reductions with respect to the chance error rate indicates that prosody-only models perform much better in the BN corpus than in the RF corpus. Expressed in percentages, these error reductions are 77.7% for BN REF, 73.7% for BN ASR, 34.2% for RF REF and 30.0% for RF ASR. Since broadcast news speech has more regular prosody, more accurate prosody-based predictions were expected for that data. In addition, I compared these error reductions with those obtained on the English meeting corpus. There, the chance error was reduced by 49.4% for REF and 40.1% for ASR. Thus, the prosody model was most successful for the Czech BN corpus, second most successful for the ICSI meetings, and least successful for the Czech RF corpus.

I also analyzed the improvements achieved by the rich prosody feature set in comparison with the pause-only feature set. Gain by adding prosody beyond pause is much more substantial for the BN corpus. The relative error reductions over using pauses alone are 38.8% and 43.7% for human- and ASR-generated BN transcripts, while the relative reductions are just 8.2% and 8.9% for the RF transcripts. These numbers also indicate that using rich prosody is more helpful in ASR conditions.

We can also compare the results achieved by prosodic models with the earlier presented

Table 10.5: Prosodic feature usage for the BN corpus ("—" indicates that less than two features from the group appear in the overall feature set.)

Group	Tot. Usage	Two most used features in the group
Pause	26.9%	pause.after (26.9%), —
Duration	43.9%	vowel.dur.last_1st.snorm (9.4%), word.dur (8.4%)
Pitch	12.0%	f0.logratio.pwl_last_baseline (12.0%), —
Energy	6.3%	f.RMS.max.norm (6.3%), —
Near Pause	0.0%	—
Other	10.8%	turn.is_end (10.8%), —

results of LMs (Table 10.3). Unlike the ICSI meetings, for which the prosody models outperformed the language models only in ASR conditions, prosody models outperformed language models in all four Czech test sets. However, the magnitudes of the differences varied across the test sets. In comparison with the BERs by the LM, the BER by the prosody model was lower by 59.8% relative for BN REF, 56.5% for BN ASR, 8.2% for RF REF, and 8.0% for RF ASR.

10.5.2 Prosodic Feature Usage

10.5.2.1 Prosodic Feature Usage for Czech Corpora

Similarly to Section 8.6.2, I also explored prosodic feature usage in decision trees. The usage rates for each group as well as the most used individual features are presented in Table 10.5 (BN) and Table 10.6 (RF). In addition to the five feature groups from Chapter 8 (pause, duration, pitch, energy, near pause), the group "other" is also represented here. While no feature from this group was selected for multi-channel meeting data, the feature called *turn.is_end* indicating speaker change was found to be very important for the single-channel broadcast data.

The numbers for the BN corpus show that the most frequently queried feature group was "duration" followed by "pause". The difference between these two groups is quite high – 17% absolute. The third most queried group was "pitch", the fourth was "other", and the fifth was the group of energy features. As contrated to meetings, near pause features were not selected at all. The most used individual feature was pause after the current word. Other heavily queried features were the log ratio of the last stylized F_0 value and the speaker's F_0 baseline, the speaker change flag, and the normalized duration of the last vowel in the current word.

For the RF corpus, the most frequently used feature groups were also "duration" and "pause". The gap between the two groups was more prominent in the RF data – it was over 38% absolute. The third most queried group was "energy", the fourth was "other", and the fifth was the group of pitch-related features. Although the near pause group had nonzero usage in this corpus, it was the least represented group again, taking up just 2.2%. The most used individual feature was duration of the current word, surpassing the pause after the current word. Other heavily queried features were the speaker change flag and the feature capturing speaker-normalized lengthening of the longest vowel in the current word.

A comparison of feature group usage between the two corpora is visualized in Fig. 10.1. The group of duration features prevails in both corpora, however, the two distributions vary. The difference is most prominently displayed in pausing features. The feature capturing pause duration at the boundary of interest was more frequently queried in the BN corpus. Pausing in news speech is much more regular than in conversational speech; therefore, pause features are

Table 10.6: Prosodic feature usage for the RF corpus ("—" indicates that less than two features from the group appear in the overall feature set.)

Group	Tot. Usage	Two most used features in the group
Pause	14.7%	pause.after (14.7%), —
Duration	53.0%	word.dur (16.5%), vowel.max_dur.snorm (6.0%)
Pitch	9.1%	f0.last.min (5.0%), f0.logratio.pwl_last_baseline (4.1%)
Energy	11.5%	f.RMS.max.norm (5.8%), f.RMS.min.norm (5.8%)
Near Pause	2.2%	p.pause.after (2.2%), —
Other	9.4%	turn.is_end (9.4%), —

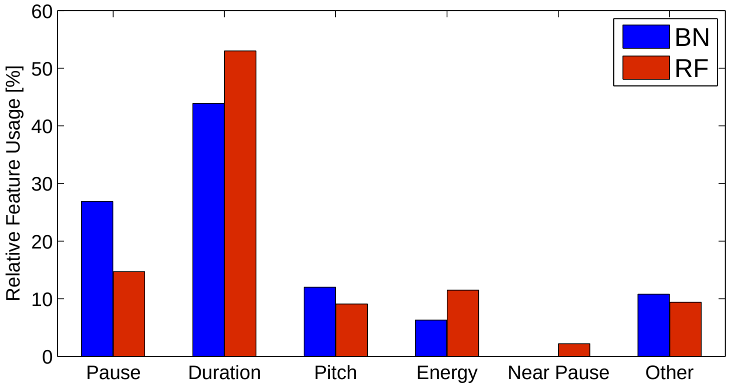


Figure 10.1: Prosodic feature group usage for Czech corpora (BN and RF)

more reliable for that corpus. Another difference between the two distributions is the usage proportion of pitch and energy features. The BN corpus prefers pitch features, while the RF corpus prefers energy features.

Variances in usage of individual features were the following. From the duration group, a feature that aims to capture segmental lengthening is the most important for the BN corpus, while the raw word duration feature is dominant for the RF corpus. The pitch group of the BN corpus heavily use a feature reflecting the ratio between the last F_0 value and the speaker's F_0 baseline. It suggests that radio anchors regularly mark statement boundaries with expressive pitch falls. This pitch feature was also important for the RF corpus, but an unnormalized feature capturing minimal F_0 value in the last voiced region was used more. In both corpora, energy features were represented by normalized RMS values computed from the word following the boundary in question. These features reflect the fact that speakers typically start sentences loudlier than they finish them.

I have also explored feature usage in the "pause-only" feature set. For the BN corpus, the pause after the current word had usage of 67.5%, while the pauses after the previous and the following word had 1.9% and 1.0%, respectively. The remainder (29.6%) corresponded to the speaker change feature. For the RF corpus, the pause after the current word takes in 48.4%. In contrast with BN, the pause after the preceding word was also largely used – 21.6% of the total usage. The pause after the following word showed usage 6.0% and the speaker change flag 24.0%. A quick comparison of the two corpora indicates that in the BN corpus, we can do well with just using pause at the current boundary, while in the RF corpus, the other two pause features are also important when other prosodic features are not available.

10.5.2.2 Prosodic Feature Usage Comparison across Languages

It is not easy to make a direct comparison of prosodic feature usage between our Czech corpora and some similar English corpora because authors of parallel sentence segmentation studies usually do not present relative usage of their prosodic features. An exception is [149], where Liu shows the best performing features for HMM-based sentence segmentation of English broadcast news and telephone conversations. For broadcast news, the most important feature was the pause duration at the word boundary in question. As expected, the most used feature is identical with the most used feature for Czech BN, however, their relative usage percentages differ. Liu reports the pause usage as 44.3%, while my Czech experiments showed the pause usage only as 26.9%.

The next best prosodic features[3] for English BN were reported as the normalized duration of the last rhyme in the word (17.7%), and the difference between last PWL F_0 value and the speaker's baseline (4.5%). These two features capture the same phenomena as my most used duration and F_0 features, but their order is reversed in the feature ranking and their relative percentages largely vary. The comparison of particular numbers indicate that features capturing final lengthening are more important for English, while features capturing final pitch fall are more important for Czech. This finding is in agreement with the fact that Czech offers less opportunity for final lengthening because length also serves a lexical function in Czech.

In the comparison of prosodic feature usage between BN and CTS data, Liu concludes that pitch plays a more important role for BN than for CTS, whereas phone and word duration is more important for CTS. This observation is in line with my findings regarding differences between read-aloud and conversational Czech data.

Shriberg et al. reported a comparison of feature usage between English broadcast news and Switchboard conversations only based on feature groups [49]. The differences between read-aloud and conversational speech in terms of feature group usage observed in that study were once again similar to those I report here for Czech data – pitch and pause features[4] were more used in broadcast news, whereas duration features dominated in Switchboard.

While broadcast news speech have been studied for prosody-based sentence segmentation extensively, much less is in this respect known about broadcast conversations since this genre has drawn attention of the speech technology community rather recently. To my best knowledge, the only published study analyzing prosodic features for sentence segmentation of broadcast discussions is [192] by Cuendet et al. However, the way in which this study evaluates prosodic features does not allow a direct comparison with my feature usage statistics.

The authors do not present feature usage of individual features in the overall set but compare segmentation accuracy of systems using individual prosodic feature groups – duration, energy, pitch and pause. For energy and pitch features, the authors also looked at feature subgroups – range, reset and slope. The error rates of individual feature subsets were compared across three genres, namely natural meetings, broadcast news, and broadcast conversations. The results revealed that in terms of sentence boundary cues, broadcast conversations stand closer to broadcast news than meetings. This finding cannot be confirmed nor refuted here because the variance between meetings and broadcast conversations observed in my thesis may not only be caused by genre but also by language differences.

[3] In this cross-lingual comparison, I only take into account "pure" prosodic features, leaving out "other" features in the top feature rankings. Since the non-prosodic features cover different proportions in the compared feature sets, and these proportions are usually not reported, precise conclusions about "pure" prosodic features cannot be drawn based on comparisons of the absolute usage numbers.

[4] Energy-based features were not used in this early study by Shriberg et al.

Table 10.7: SU segmentation error rates for models combining textual and prosodic features [BER %]

Model	Used Features	BN		RF	
		REF	**ASR**	**REF**	**ASR**
Chance	—	8.11%	8.01%	6.81%	6.89%
HMM	LM+Pause	2.18%	2.59%	3.61%	4.25%
	LM+Rich Prosody	1.44%	1.73%	**3.44%**	**4.14%**
MaxEnt	LM+Pause	2.40%	2.78%	3.57%	4.27%
	LM+Rich Prosody	1.76%	2.07%	3.45%	4.16%
BoosTexter	LM+Pause	1.85%	2.28%	3.58%	4.42%
	LM+Rich Prosody	**1.42%**	**1.72%**	3.56%	4.21%

10.6 SU Segmentation Using Both Textual and Prosodic Information

In this section, I present results of models that rely both on prosodic and textual information. For each modeling approach, I compare two types of models. The first combines the best LMs (different across the three approaches) with the pause-only models, whereas the second type of models combines the same LMs with prosody models based on the richer feature set. The error rates for both Czech corpora are displayed in Table 10.7.

10.6.1 Results for BN Corpus

For the BN corpus, the best results were achieved by the BoosTexter model combining textual information with the rich prosody feature set. These results were $BER = 1.42\%$, $NIST = 17.49\%$, $F = 91.27\%$ for REF test conditions and $BER = 1.72\%$, $NIST = 21.44\%$, $F = 89.19\%$ for ASR conditions. However, the gap between BoosTexter and HMM using the same feature sets was really tiny and statistically insignificant for both test conditions. On the other hand, MaxEnt performed significantly worse than the other two approaches. Using the Sign test, the difference between MaxEnt and BoosTexter was significant at $p < 10^{-6}$ for both test conditions. The difference between MaxEnt and HMM was significant at $p < 10^{-7}$ for REF and at $p < 10^{-8}$ for ASR.

The models with richer prosodic feature sets consistently outperform the models with pause information only. The improvements by adding rich prosody information are statistically significant for all three employed modeling approaches and both test conditions. The significance levels for REF conditions are $p < 10^{-28}$, $p < 10^{-17}$ and $p < 10^{-13}$ for HMM, MaxEnt, and Boostexter, respectively. The corresponding significance levels for ASR conditions are $p < 10^{-35}$, $p < 10^{-20}$ and $p < 10^{-19}$.

Although the combined models with rich prosody outperform the combined models with pause-only prosody, the relative gains from using rich prosody are diminished in comparison with the gains for the prosody-only classification presented in Table 10.4. The decrease of improvement is smaller for Czech BN than for the English meetings, but it is still visible. Whereas the addition of prosodic features beyond pause reduces BER of the prosody-only classification by 34.4% relative for BN REF and by 37.9% relative for BN ASR, the classification by combined models is only improved by 23.2% and 32.6%, respectively.

10.6.2 Results for RF Corpus

In the RF corpus, the best results were achieved by the combined HMM model employing the rich prosodic feature set. The best model performed at $BER = 3.44\%$, $NIST = 49.50\%$, $F = 74.48\%$ for REF and $BER = 4.14\%$, $NIST = 60.06\%$, $F = 68.59\%$ for ASR test conditions. The second best results were obtained from the MaxEnt model. The MaxEnt results were basically identical with those of HMM. Even the BoosTexter model, which was the worst of the three modeling approaches, showed just insignificantly ($p > 0.1$) higher error rates than the best HMM.

Same as in the BN corpus, the models with the rich prosodic features sets outperformed the models only using pause information for prosody modeling. However, the improvements by adding rich prosody information were much smaller for the RF corpus. For REF conditions, only HMM showed a clearly significant improvement ($p < 0.01$). The gain for MaxEnt was only "marginally" significant at $p < 0.08$, while the gain for BoosTexter was even clearly insignificant. On the other hand, BoosTexter was the only model displaying a clearly significant improvement ($p < 0.005$) in ASR conditions. The improvements by the other two approaches only showed p-values between 0.08 and 0.10.

Again, I observed that relative gains from using rich prosody for models relying both on textual and prosodic cues are smaller than the gains for a prosody-only classification. This relative decrease of improvement for the RF corpus is much larger than for the BN corpus. While the addition of prosodic features beyond pause reduced BER of the prosody-only classification by 8.2% relative for RF REF and by 8.9% for RF ASR, the classification by the models combining textual and prosodic features was only improved by 4.9% and 2.7% relative, respectively.

10.6.3 Result Comparison across Corpora

The presented numbers relating to models combining textual and prosodic information show some differences between BN and RF. BoosTexter was the best combined model for BN, while HMM was the best for RF. However, the gaps among the three explored approaches were rather small for both corpora. The only exception was MaxEnt for BN, which showed significantly worse results than the other approaches. These results of individual modeling approaches suggest that tight integration of prosodic and textual features is important for BN, while more efficient language modeling as shown by the HMM approach is important for RF.

The results also indicate differences in the importance of using prosodic features beyond pause in the combined models. The additional prosodic features decreased BER by 23.2% relative for BN REF and by 32.6% relative for BN ASR, but they only reduced the error by 4.9% relative for RF REF and by 2.7% relative for RF ASR.

Fig. 10.2 visualizes BERs in dependence on used knowledge sources for all three modeling approaches. Individual graphs in the figure correspond to four Czech test sets on which the models were evaluated (BN REF, BN ASR, RF REF, and RF ASR). In the graphs, lines connect points corresponding to the same model in order to increase readability. The graphs show some interesting facts. For instance, two of the three models for the BN corpus show a better performance when only using rich prosody than when using an LM plus pause information. This finding is quite surprising. It shows how important the prosodic features beyond pause are for sentence segmentation of broadcast news data. Unlike BN, the LM plus pause models consistently outperform the rich prosody models used on their own in the RF corpus. In general, lexical cues are more important for spontaneous speech corpora, which display less regular prosodic marking of boundaries than planned speech corpora do.

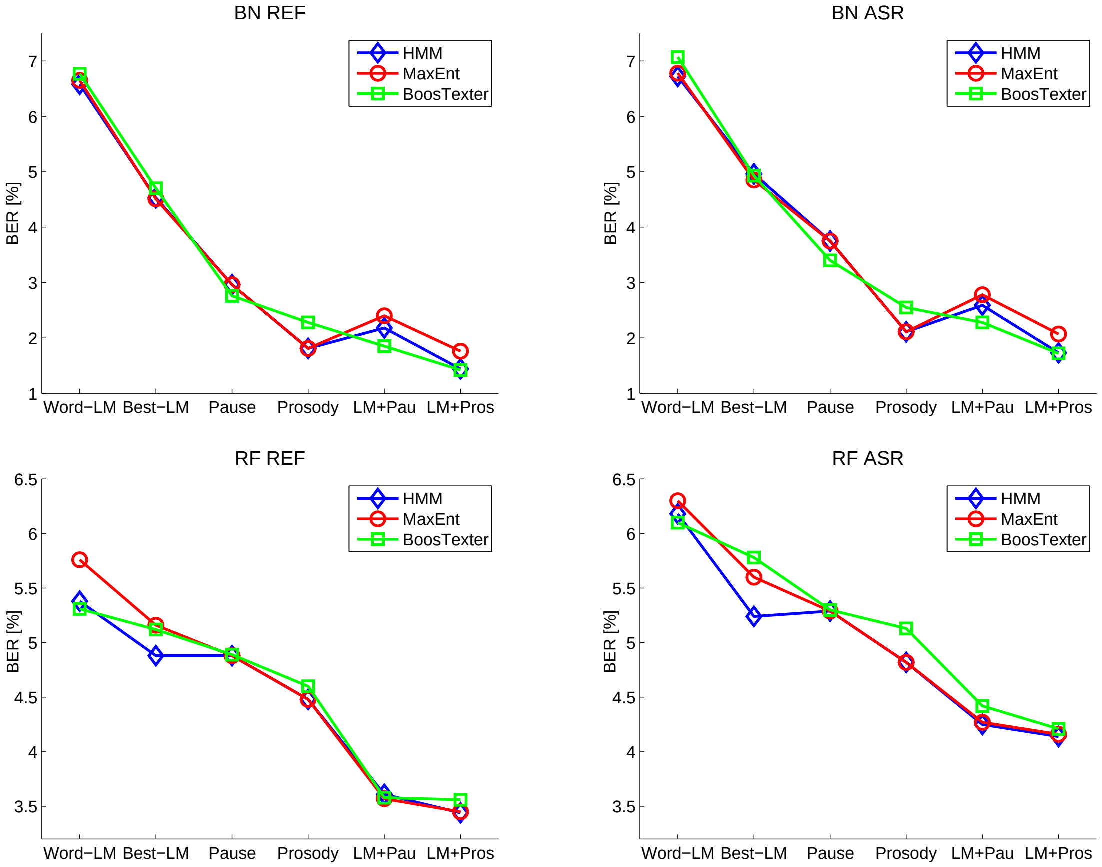


Figure 10.2: SU segmentation error rates for individual models in dependence on used knowledge sources [BER %]

The differences among the results of the three modeling approaches are more visible in the RF graphs. The hugest deviation is displayed in the LM-only tests in which the HMM approach clearly outperforms the two others. Another notable deviation is in the results of prosody-only tests in which we can see that the BoosTexter model is not good in handling the rich prosody feature set on its own. This inferiority is more visible in ASR conditions.

10.7 Combination of All Three Modeling Approaches

Just as for the meeting data (cf. Section 8.8), I tried to combine HMM, MaxEnt, and BoosTexter models into a single SU segmentation system. Thus, the resulting system combines not only various knowledge sources but also various machine learning techniques. I only combine the best performing models, i.e. those using both lexical and prosodic features. Again, two model combination methods were examined – simple majority voting and linear interpolation. The results for both Czech corpora are presented in Table 10.8. For a better illustration, the table compares the two combination approaches with the best results achieved by single-approach models – BoosTexter for BN REF and BN ASR, and HMM for RF REF and RF ASR.

Table 10.8: SU segmentation error rates for a combination of HMM, MaxEnt, and BoosTexter [BER %]

Combination Approach	**BN**		**RF**	
	REF	**ASR**	**REF**	**ASR**
Chance	8.11%	8.01%	6.81%	6.89%
Best Single Approach	1.42%	1.72%	3.44%	4.14%
Majority Voting	1.35%	1.65%	3.23%	4.00%
Linear Interpolation	**1.27%**	**1.61%**	**3.22%**	**3.98%**

10.7.1 Results for BN Corpus

The results on the BN corpus indicate that the model combination improves SU segmentation accuracy. Of the two combination options, the linear interpolation worked better, however, the gain over the voting approach was only significant for tests on manual transcripts ($p < 0.01$). Overall, BER was decreased from 1.42% to 1.27% for BN REF, and from 1.72% to 1.61% for BN ASR. The Sign test showed that these improvements over the best single-approach model are significant at $p < 0.001$ and at $p < 0.02$, respectively. In our other metrics, the best results of the combined models were evaluated as $NIST = 15.63\%$ and $F = 92.30\%$ for REF, and $NIST = 20.09\%$ and $F = 90.02\%$ for ASR conditions.

The interpolation weights for HMM, MaxEnt, and BoosTexter were estimated from development data using the EM algorithm. I got 0.33, 0.38, and 0.29 for BN REF, and 0.36, 0.36, and 0.28 for BN ASR. The estimated weights do not largely vary across the test conditions and both weight distributions are not far from uniform. For illustration of the automatic sentence segmentation accuracy achieved by the best performing system, an example of an automatically segmented BN transcript is shown in Appendix C.2.

10.7.2 Results for RF Corpus

The results for the RF corpus also indicate an improvement over the best single-approach model. Of the two combination options, the linear interpolation approach showed better results. However, the gain over the simple majority voting approach was statistically insignificant for both test conditions. The interpolation decreased BER from 3.44% to 3.22% for RF REF, and from 4.14% to 3.98% for RF ASR. The Sign test indicated that the gaps between the best combined system and the best single-approach model were significant at $p < 0.005$ and at $p < 0.02$ for the two test conditions, respectively. The best results correspond to $NIST = 46.35\%$ and $F = 75.65\%$ for RF REF, and $NIST = 57.65\%$ and $F = 69.12\%$ for RF ASR.

The optimal interpolation weights for HMM, MaxEnt, and BoosTexter were determined by the EM algorithm as 0.30, 0.39, and 0.31 for RF REF, and as 0.34, 0.32, and 0.34 for RF ASR. Again, the estimated weight distributions did not diverge far from the uniform distributions. An example of an automatically segmented RF transcript is presented in Appendix C.3.

10.7.3 Result Comparison across Corpora

The model combination approach helped relatively more for the BN corpus than for the RF corpus. The relative error reductions by model interpolation were 10.6% for BN REF, 6.4% for BN ASR, 6.4% for RF REF, and 3.9% for RF ASR. Since the results by the interpolated models are the best for all Czech test sets, I also used them to compare overall classification accuracy across the corpora. In terms of a relative BER reduction with respect to chance error rate, my segmentation system was relatively most successful on BN REF (BER reduction by 84.3%) followed by BN ASR (79.9%), RF REF (52.7%), and RF ASR (42.2%). Analogous comparisons based on NIST error rates and F-measures show the results in the same order.

Furthermore, we can compare these relative BER reductions with those achieved for the ICSI meeting corpus – 67.5% for ICSI REF and 55.3% for ICSI ASR. Thus, it is possible to conclude that of the three corpora explored in this thesis, the best results were achieved for the Czech broadcast news, second best for the English multiparty meetings, and the relatively worst results were obtained for the Czech broadcast conversations. However, this order does not necessarily imply that meetings represent an easier domain for automatic sentence segmentation of speech than broadcast conversations. There are two important facts that should be taken into account when interpreting the above presented numbers. First, the ICSI meeting data contain a large number of backchannel DAs whose boundaries are relatively easier to detect. Second, the meetings are in English, which is an easier language for modeling than Czech.

Finally, my results for BN can be confronted with those presented in [115] for SU boundary detection in English BN data. Therein, the results are reported using the NIST error rate, which allows to make a performance comparison across tasks and corpora having different event priors. The lowest NIST error rates for English BN achieved by a combination of HMM, MaxEnt, and CRF were reported as 47.44% for REF and 57.23% for ASR ($WER = 11.7\%$) conditions. These NIST error rates are much higher than those achieved for the Czech BN herein – $NIST = 15.63\%$ and $NIST = 20.09\%$, respectively. Nevertheless, this comparison does not allow to claim that my Czech sentence segmentation system is better. The comparison of the Czech BN MDE corpus and the English MDE corpus presented in Chapter 5 revealed that the English data are more difficult since they contain a higher number of fillers and edit disfluencies.

10.8 Chapter Summary and Conclusions

In this chapter, I have explored automatic sentence unit segmentation of spoken Czech from two different domains – broadcast news and broadcast conversations. As for the experiments with English meeting data, I have examined three different modeling approaches: HMM, MaxEnt, and BoosTexter, and evaluated them on manual and automatic transcripts of the two corpora.

In language modeling, I not only employ simple word-based models, but also textual information beyond word identities, as captured by automatically induced word classes and part-of-speech tags. In addition, I investigated the possibility of using additional text resources that were not annotated for SUs but only contained standard punctuation. Prosody models were evaluated with two distinct feature sets – the first contained just pause-based features, while the other was a richer set also comprising features relating to duration, pitch, and energy. Features for the richer feature set were selected for each corpus separately.

The experiments with language models showed that the HMM model trained on the auxiliary text corpus was by far the best single-source model for the BN corpus. On the other hand, this feature set did not perform that well for the RF corpus, for which the best feature set was POSmix. However, when AuxWords were combined with other textual information sources, they significantly improved segmentation performance also for the RF corpus.

In general, language models combining several textual knowledge sources worked better than models using just a single information source. For BN REF and BN ASR, the best language model was MaxEnt with Words+AIC+POSmix+AuxWords. For RF REF, the best was HMM with Words+POSmix+AuxWords. The HMM model was also the best for RF ASR, but it performed slightly better when only using Words+AuxWords.

The next set of experiments focused on performance of prosodic models. For all four test sets, the best results for prosody-only classification were achieved by the CART-based model using the rich prosodic feature set. The prosodic features beyond pause were more helpful for the BN corpus which has a more regular prosodic marking of sentence boundaries. I also observed that relative gain from the additional prosodic features was slightly larger in ASR conditions.

Prosodic feature usage analysis revealed that duration and pause feature groups were most important for both corpora. However, the feature group usage distributions differed between the two corpora. The difference was most prominently displayed in pausing features, which were more frequently queried in the BN corpus. Another difference between the two distributions was in the proportion of pitch and energy features. The BN corpus preferred pitch features, while the RF corpus preferred energy features. For the BN data, I also compared my feature usage with those reported for English BN. As expected, the most used feature (pause duration at the boundary) was the same for both languages, however, its relative usage differed. It was higher for English. A comparison of other most used features demonstrated that features capturing final lengthening were more important for English, while features capturing final pitch fall were more important for Czech. This finding is in agreement with the fact that Czech offers less opportunity for final lengthening because length also serves a lexical function in Czech.

Further, models relying on both lexical and prosodic cues were examined. BoosTexter was the best modeling approach for BN, while HMM was the best for RF. The results of individual modeling approaches suggest that tight integration of prosodic and textual features is important for broadcast news data, while more efficient language modeling as shown by the HMM approach is important for broadcast conversation data. However, the gaps among the three explored approaches were not very large. The results indicated differences in the importance of using prosodic features beyond pause in the combined textual-prosodic models. Similarly to sentence segmentation only based on prosodic models, the additional prosodic features were much more helpful for the BN corpus.

Overall, the best results for all test sets were achieved by a model that combines HMM, MaxEnt, and BoosTexter models via posterior probability interpolation. This result is in line with my results for English multiparty meetings, for which the interpolation model also showed superior performance. Finally, I compared relative BER reductions with respect to chance across all three corpora used in this thesis. This comparison indicates that the highest chance error reduction was achieved for the Czech broadcast news corpus, the second highest for the ICSI meeting corpus, and the relatively worst results were obtained for the Czech broadcast conversation corpus.

Chapter 11

Conclusion

Success is not final, failure is not fatal:
it is the courage to continue that counts.
Winston Churchill

This final chapter summarizes contributions of this thesis (Section 11.1), presents a thesis summary and main conclusions (Section 11.2), and proposes possible extensions and research directions for future work (Section 11.3).

11.1 Contributions

This section explicitly lists contributions of this thesis. They are categorized according to the list of objectives presented in Chapter 4.

1. Two Czech speech corpora with structural annotation have been created – one in the domain of broadcast news and the other in the domain of broadcast conversations. The employed annotation scheme creates boundaries between natural breakpoints in the flow of speech, flags non-content words for optional removal, and identifies sections of disfluent speech. The original structural metadata annotation guidelines for English have been adjusted to accommodate specific phenomena of Czech syntax. In addition to the necessary language-dependent modifications, I have proposed some language-independent modifications refining the original annotation scheme. Finally, I have performed a detailed analysis of the metadata annotated corpora.

2. Automatic dialog act segmentation of English multiparty meetings has been investigated. Various textual and prosodic features have been explored for usefulness, and three modeling approaches (HMM, Maximum Entropy, and BoosTexter) have been compared. In addition to experiments in a speaker-independent fashion, I have also explored speaker adaptation of both prosodic and language models in this domain. Speaker adaptation for sentence (or dialog act) segmentation of speech is a novel idea proposed in this thesis.

3. The first sentence segmentation system for spoken Czech has been developed. The system has been evaluated on the two Czech corpora created in Objective 1. Again, various prosodic and textual features have been examined, and the three modeling approaches have been compared.

11.2 Summary and Main Conclusions

The work presented in the thesis can be divided into two major parts based on the type of work – *Creation and analysis of data resources* (Objective 1) and *Development and evaluation of automatic systems for segmentation of speech into sentence-like units* (Objective 2 and 3). The work on these two major parts is summarized in the two following subsections.

11.2.1 Creation and Analysis of Czech Data Resources

11.2.1.1 Corpus Design and Creation

Objective 1 is about creation of Czech speech corpora with appropriate annotation of sentence-like unit boundaries. The employed annotation scheme was based on the LDC's "Simple Metadata Annotation Specification", which was originally defined for English. In order to make this standard applicable to Czech, the original annotation guidelines have been adjusted to accommodate specific phenomena of Czech syntax. I have also proposed a novel approach to transcribing and annotating filled pauses in Czech, distinguishing vowel-like (*EE*) and consonant-like (*MM*) sounds. In addition to the necessary language-dependent modifications, I have proposed and applied some language-independent modifications refining the original annotation scheme. The refinements included limited prosodic labeling at sentence unit boundaries and distinction of two types of incomplete units.

Two Czech corpora with structural metadata annotation were created – one in the domain of broadcast news (BN) and the other in the domain of broadcast conversations (RF). The first corpus has been created by enriching an existing corpus with the structural metadata annotation, while the second has been built from scratch – it had to be recorded and manually transcribed first. Besides their importance to automatic structural metadata extraction research, these corpora are also useful for training ASR systems as well as for linguistic analysis of read-aloud and spontaneous Czech.

11.2.1.2 Corpus Analysis

I have conducted a detailed comparison of the two Czech corpora in terms of structural metadata statistics. The comparison revealed that, as expected, Czech broadcast conversations represent a more difficult data for automatic metadata extraction than Czech broadcast news since they contain significantly more disfluencies and fillers. I also found that reparanda and their corrections show differences in POS distributions in comparison with the general POS distribution. For example, the relative proportion of nouns, pronouns, and prepositions is higher in reparanda than in all data. On the other hand, verbs, adverbs, and adjectives are relatively less frequent in reparanda.

Regarding SU symbols, I observed that clausal breaks were more frequent in the broadcast conversation corpus, which indicates that complex sentences are more common in talk shows than broadcast news. Furthermore, the comparison showed that SUs in conversational data were on average longer by 1.5 words.

Moreover, I also reported most frequent filler words in the two corpora. Statistics of filled pauses showed that *EEs* were much more frequent than *MMs*; more than 90% of all filled pauses were *EEs*. Another interesting observation was that discourse markers containing a verb (such as English *you know*) are much less frequent in Czech than in English. The most frequent Czech discourse markers were *tak* (lit. *so*), *no* (*well*), and *prostě* (*simply*).

11.2.2 Development and Evaluation of Automatic Sentence Segmentation Systems

The other two objectives referred to the development of automatic systems for segmentation of speech into sentence-like units. The objective No. 2 was focused on DA segmentation of multiparty meetings from the ICSI meeting corpus and the objective No. 3 on sentence segmentation of the two Czech MDE corpora. Although the systems for Czech and English were trained on different data, they share a common modeling groundwork. All sentence segmentation experiments were conducted on two types of speech transcripts – manual transcripts (reference conditions) and automatically-generated transcripts (ASR conditions). Furthermore, I have evaluated sentence segmentation accuracy achieved by models only relying on textual information, models only relying on prosodic information, and combined models relying on both sources of information.

Since the experimental part of the thesis was manifold, its summary is divided into several sections. The first section provides a brief description of the general segmentation system. The following three sections summarize results of experiments – speaker-independent experiments on English meetings, speaker-dependent experiments on the same corpus, and experiments on Czech corpora. The final section summarizes general findings learned from the described experiments.

11.2.2.1 General System Description

For either language, three modeling approaches have been examined – HMM, MaxEnt, and BoosTexter. All these approaches rely on both textual and prosodic features. The textual features describe lexical patterns associated with sentence-external and sentence-internal interword boundaries. I use features capturing word identities, parts of speech, and automatically induced word classes. The prosodic features for sentence segmentation of speech reflect breaks in temporal, intonational, and loudness contours in an utterance. In the approach I employ in this thesis, prosodic features for automatic classification are extracted directly from the speech signal based on time alignments from automatic speech recognition, without any need for hand-labeling of prosodic events. A toolkit for the direct extraction of prosodic features has been implemented as part of this work.

11.2.2.2 Speaker-Independent Experiments on Multiparty Meetings in English

The work conducted on DA segmentation of meetings from the ICSI corpus can be divided into two parts – speaker-dependent and speaker-independent. The speaker-independent experiments had the following results. The best performing language modeling approach for text-based DA segmentation was HMM combining words, POS, and AICs. The best performing model for DA segmentation only based on prosodic information was bagged CART using a rich prosodic feature set. An analysis of feature usage in decision trees revealed that the most frequently queried group of features was pause, followed by duration-related features and pitch-related features.

Models combining prosodic and lexical information clearly outperformed language and prosody models used on their own. A comparison of individual modeling approaches showed that HMM and MaxEnt were superior in reference conditions, and MaxEnt was the best in ASR conditions. The best overall results for both test conditions were achieved by a combination of HMM, MaxEnt, and BoosTexter based on linear interpolation of posterior probabilities. The overall results in ASR conditions also indicate that models trained on ASR data slightly outperform their counterparts trained on human-generated transcripts. The best system achieved $BER = 5.18\%$, $NIST = 32.49\%$, $F = 83.18\%$ in reference conditions, and $BER = 6.19\%$, $NIST = 44.70\%$, $F = 75.85\%$ in ASR conditions.

11.2.2.3 Speaker-Dependent Experiments on Multiparty Meetings in English

These experiments examined speaker-specific prosodic and language modeling for DA segmentation of meetings. The method was evaluated on 20 frequent speakers with a wide range of total words available for speaker-dependent modeling. For prosodic models, it was found that interpolating the large, speaker-independent prosodic model with a much smaller prosodic model trained only on that talker's speech yielded improvements for 6 of the 20 speakers in reference conditions, and for 16 of 20 the speakers in ASR conditions. The ASR conditions showed a higher number of improved speakers, but the improvements were relatively smaller. Overall results, summed over all 20 speakers, indicate a modest, yet statistically significant improvement with respect to the speaker-independent model for both test conditions. Feature analysis, while preliminary given the number of speakers, suggests that nonnative speakers may differ from native speakers in overall feature usage patterns associated with DA boundaries.

For speaker-adapted language models, improvements were found for 17 of the 20 speakers using reference transcripts, and for 15 of the 20 speakers using automatic transcripts. Overall, a statistically significant improvement over the baseline LM was achieved for both test conditions. For both types of speaker adaptation, improvements were achieved even for some talkers who had only a relatively small amount of data available for adaptation. In addition, the relative error reduction achieved by speaker adaptation was not correlated with the amount of adaptation data.

11.2.2.4 Experiments on Czech Corpora

The last group of experiments focused on sentence segmentation of the two Czech corpora. The experimental setup was the same as for the speaker-independent experiments on the meeting data. In contrast to the meeting experiments, I have also examined the possibility of using auxiliary text resources not annotated for sentence-like unit boundaries but only containing standard punctuation. The experiments with language models showed that the HMM model trained on the auxiliary text corpus was by far the best single-source model for the BN corpus. On the other hand, this feature set did not perform that well for the RF corpus. However, when auxiliary word features were combined with other textual information sources, they significantly improved segmentation performance also for the RF corpus.

For both Czech corpora, same as for the meeting corpus, the best results for prosody-only segmentation were achieved by the CART-based model with a rich prosodic feature set. A prosodic feature usage analysis showed that the feature group usage distributions differed between the two Czech corpora. The difference was most prominently displayed in pausing features, which were more frequently queried in the BN corpus.

Of the models relying on both lexical and prosodic cues, BoosTexter was the best modeling approach for BN, while HMM was the best for RF. Overall, the best results for all test sets were achieved by a model that combined HMM, MaxEnt, and BoosTexter models via posterior probability interpolation. For the BN corpus, the best performing system achieved $BER = 1.27\%$, $NIST = 15.63\%$, $F = 92.30\%$ in reference conditions, and $BER = 1.61\%$, $NIST = 20.09\%$, $F = 90.02\%$ in ASR conditions. For the RF corpus, the best system achieved $BER = 3.22\%$, $NIST = 46.35\%$, $F = 75.65\%$ in reference conditions, and $BER = 3.98\%$, $NIST = 57.65\%$, $F = 69.12\%$ in ASR conditions. The results indicate that, as expected, broadcast conversations are more difficult for automatic sentence segmentation than broadcast news.

11.2.2.5 General Findings

The performed experiments indicate several important findings. First, prosody represents a valuable knowledge source for automatic sentence segmentation of speech, both in English and Czech. Nevertheless, the importance of prosodic features varies across corpora. They were relatively most beneficial for the Czech BN corpus, where professional newscasters usually pay attention to proper prosodic marking of sentence boundaries. However, prosodic information largely reduced error also for other two corpora. Similarly, the gains from using rich prosodic feature sets instead of only pause features differed corpus to corpus; they were the largest for the BN corpus. Furthermore, feature analysis reveals that English and Czech slightly differ in overall prosodic feature usage patterns. The experimental results also show that prosodic features are less degraded by word recognition errors than textual features.

Textual features based on N-gram contexts are also important for sentence segmentation of speech. An increase in performance was achieved when word-based N-grams were combined with N-grams based on automatically induced classes and POS tags. Automatic class information was more helpful for English, while POS information was more helpful for Czech. As expected, language models were relatively more successful on English data since Czech represents a more complex language for language modeling. In addition, it was shown that textual features are complementary to prosodic features. The models relying on both prosodic and textual features outperformed prosody-only and language-only models in all test sets.

An important part of this work was a comparison of three statistical modeling approaches to sentence segmentation of speech – HMM, MaxEnt, and BoosTexter. The experimental results indicate that there is no clear overall winner among the approaches since each of them was superior in some of the tests. On the other hand, a clear conclusion is that the best performance is achieved when all these approaches are combined via posterior probability interpolation. This model was superior in all tests in both languages, and all the improvements over the best single approach model were statistically significant.

Of the individual modeling approaches, HMM showed most consistently good results. In the vast majority of tests, it produced best or close to best results. The only test in which this method did not perform very well was DA segmentation of meetings in ASR conditions. In this test with the highest WER, the best model was MaxEnt. A tighter integration of prosodic and lexical knowledge was helpful for this particular data since BoosTexter also performed well. On the other hand, the MaxEnt model had problems with the RF corpus because its language model did not work well on this data. Apparently, its smoothing method was not able to efficiently deal with a higher out-of-vocabulary rate and lexical irregularities frequent in the RF corpus. The last approach, BoosTexter, was relatively most successful on Czech BN data where it was the best of the three approaches for both types of transcripts. On the contrary, it worked rather poor on manual transcripts of the meeting corpus. As indicated by the results of individual language models, this shortcoming was probably caused by the fact that the simpler BoosTexter's language model was less effective on errorless manual English transcripts. By contrast, the same model performed reasonably well on ASR hypotheses of the same transcripts, showing that a simpler lexical model may also be robust in the face of word recognition errors.

The results of speaker-dependent experiments indicate that prosodic features beyond pause information are helpful for a vast majority of speakers. In my tests, they provided benefit for 19 of the 20 speakers studied. On the other hand, not all speakers benefited from speaker adaptation of prosodic and language models. Interestingly, improvements were achieved even for some talkers with relatively little data available for adaptation. The relative error reduction achieved by speaker adaptation was not correlated with the amount of adaptation data. Hence, I infer that speakers differ inherently in how similar they are to the generic speaker-independent model. Some talkers differ more and thus show more gain, even with less data.

11.3 Future Work

The work presented in this thesis suggests a number of extensions and future research directions. One obvious direction is to examine possible gain from using other machine learning techniques. Among others, support vector machines and conditional random fields have shown good success in many similar applications. In addition, the statistical models might be combined with some rules (either hand-crafted or induced from data) imposing constraints on the form of the extracted sentence-like segments. However, there arises a question how robust such rules would be in the face of ASR errors.

Another possible direction is to take into account ASR word confidences. Their inclusion would allow to dynamically adjust relative weights of language and prosodic models. It might be helpful since it was shown that prosodic models are less degraded by ASR errors. For example, the language model weight could automatically be decreased for speech regions recognized with low confidence scores. Such a setup could especially be useful in domains for which ASR accuracy is still rather poor.

Furthermore, the importance of additional classification features ought to be investigated. In prosody modeling, an interesting direction would be to explore longer-range features since only local features have been employed so far. On the other hand, language models might benefit from using parsing features. These features should be generally helpful, but the problem is that the quality of speech parsing is largely affected by ASR errors, missing punctuation, and disfluencies. Moreover, there is a kind of a circular problem: parsing features are important for sentence segmentation and sentence segmentation is important for parsing. Roark et al. [72] have solved this problem by using a hypothesis reranking approach, however, their approach also has some limitations, such as impracticability in real-time applications. If reliable, parsing features would probably benefit sentence segmentation for Czech more than English since the standard N-gram models are less powerful for Czech because of the more flexible word order.

Moreover, employment of information sources beyond prosody and recognized words should be examined. Given the recent progress in audio-visual speech processing, using visual cues to sentence segmentation represents a promising research direction. For example, TV broadcast data offer a good opportunity for testing multimodal segmentation models.

The results of the speaker-dependent modeling experiments also motivate further research. An important unanswered question about prosodic adaptation is what factors predict whether speaker-dependent prosodic modeling will benefit a particular speaker. Prosodic adaptation did not benefit all speakers and the absolute amount of data did not appear to be a good predictor. Additional areas for further research include development of other adaptation methods, exploration of unsupervised adaptation approaches, and exploration of clustering of speakers similar in behavior, for greater model robustness. Future work should also investigate the potential of speaker-specific modeling for other tasks in spoken language understanding.

For Czech, this thesis has only focused on automatic sentence boundary detection. However, the rich MDE markup available for the two Czech corpora also enables research of automatic detection of other structural events. A variety of structural event detection tasks beyond SU segmentation can be defined based on the MDE annotation scheme. A meaningful extension of the segmentation task would be to not only look for SU boundaries, but also to automatically label their types (statement/question/incomplete). Likewise, we can also try different definitions of the detected boundary. For instance, we might only recognize sentence-like units delimited by the double slash SU symbols, or, on the contrary, segment speech into smaller units with boundaries delimited by both sentence-external and sentence-internal SU breaks.

SU boundary detection is one of the four subtask defined in the EARS MDE task. The three other subtasks defined by MDE are *Filler word detection* (identification of words used as FPs, DMs, or EETs), *Edit word detection* (identification of all the words within DelRegs), and *Interruption point detection* (identification of the interword locations at which fluent speech becomes disfluent). All these subtasks can be evaluated using the two newly created Czech corpora. Future research on Czech metadata extraction may also include exploration of automatic detection of events that are not among the MDE subtasks, such as Asides/Parentheticals.

The automatic detection of disfluencies and filler words mentioned in the previous paragraph is important for spontaneous speech where these events are frequent. On the other hand, the task for the more regularly structured BN data may be extended in another way. In addition to automatic sentence segmentation, we could focus on automatic punctuation from speech, especially on automatic insertion of commas. Commas not only enhance human readability, but also help downstream automatic processes, as was shown by some recent work [193, 194]. I have already experimented with automatic punctuation for Czech broadcast news [181], but this task should be explored more thoroughly for Czech. Note that Czech punctuation is slightly different from that of English since Czech grammar has different (and more strict) rules for using commas.

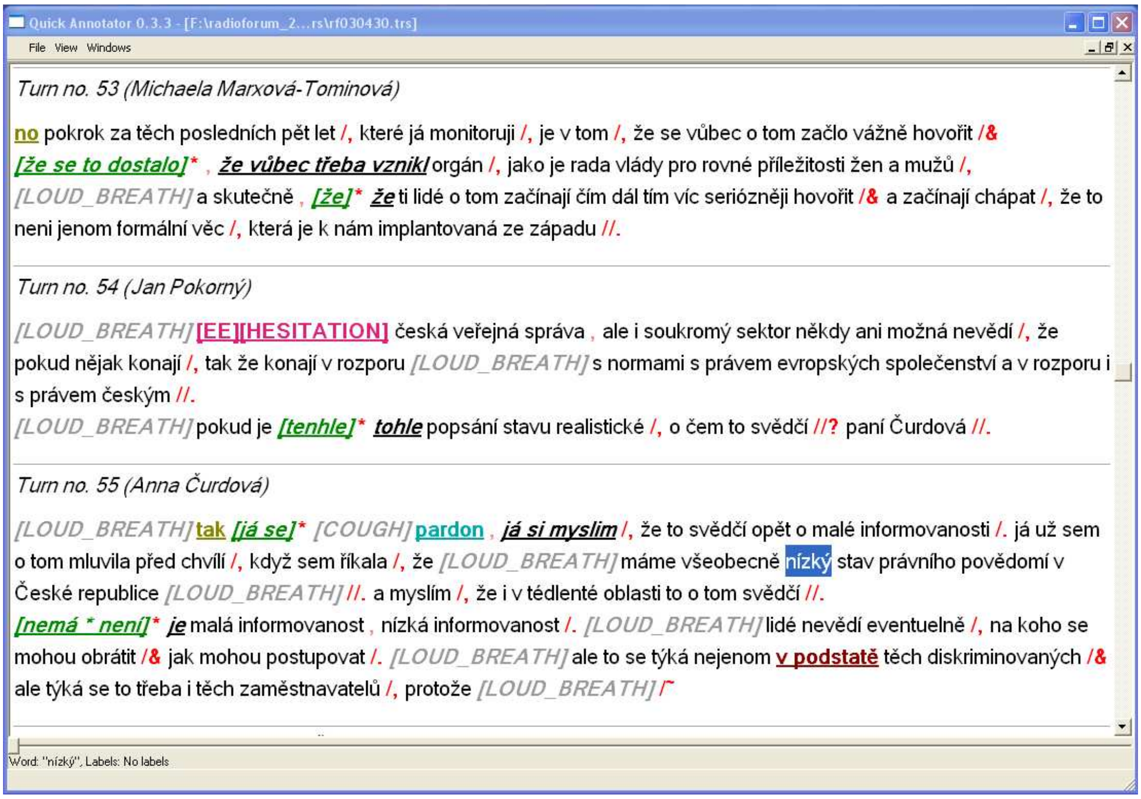
Finally, the impact of sentence segmentation on downstream automatic processes, such as POS tagging, named entity tagging, speech summarization, and machine translation, should be evaluated. This is very important since different applications may require input segmented in a different way. The very recent work of Hillard [113] argues that when providing automatic sentence segmentation to downstream applications, performance can be improved when the automatic sentence segmentation is optimized for downstream process performance rather than for sentence segmentation performance itself.

Appendix A

Annotation Tool for Czech MDE

In order to ease the MDE annotation process, a software tool – QAn (Quick Annotator) – was developed for the Czech MDE annotation task. Analogous to the English MDE tool [114], it enables to highlight relevant spans of text, play corresponding audio segment, and then record annotation decisions with few mouse clicks or keystrokes. Moreover, it respects particularities of the Czech MDE task (data format, new MDE symbols, etc.). The tool was implemented by Jan Švec.

The following screenshot not only displays the tool itself, but also serves as an example of MDE annotated Czech data. In the screenshot, mouth noises are displayed in gray, DelRegs in green, corrections in black, EETs in blue green, DMs in dark brown, DRs in light brown, FPs in magenta, and SUs in red. Asterisks denote IPs.



Turn no. 53 (Michaela Marxová-Tominová)

no pokrok za těch posledních pět let /, které já monitoruji /, je v tom /, že se vůbec o tom začlo vážně hovořit /&
[že se to dostalo]* , že vůbec třeba vznikl orgán /, jako je rada vlády pro rovné příležitosti žen a mužů /,
[LOUD_BREATH] a skutečně , [že]* že ti lidé o tom začínají čím dál tím víc seriózněji hovořit /& a začínají chápat /, že to
neni jenom formální věc /, která je k nám implantovaná ze západu //.

Turn no. 54 (Jan Pokorný)

[LOUD_BREATH] [EE][HESITATION] česká veřejná správa , ale i soukromý sektor někdy ani možná nevědí /, že
pokud nějak konají /, tak že konají v rozporu [LOUD_BREATH] s normami s právem evropských společenství a v rozporu i
s právem českým //.
[LOUD_BREATH] pokud je [tenhle]* tohle popsání stavu realistické /, o čem to svědčí //? paní Čurdová //.

Turn no. 55 (Anna Čurdová)

[LOUD_BREATH] tak [já se]* [COUGH] pardon , já si myslim /, že to svědčí opět o malé informovanosti /. já už sem
o tom mluvila před chvílí /, když sem říkala /, že [LOUD_BREATH] máme všeobecně nízký stav právního povědomí v
České republice [LOUD_BREATH] //. a myslím /, že i v tédlenté oblasti to o tom svědčí //.
[nemá * není]* je malá informovanost , nízká informovanost /. [LOUD_BREATH] lidé nevědí eventuelně /, na koho se
mohou obrátit /& jak mohou postupovat /. [LOUD_BREATH] ale to se týká nejenom v podstatě těch diskriminovaných /&
ale týká se to třeba i těch zaměstnavatelů /, protože [LOUD_BREATH] /~

Word: "nízký", Labels: No labels

Appendix B

List of Prosodic Features

This appendix presents a complete list of implemented prosodic features. Note that this is a list of all available features, not a list of features that were found to be useful for the sentence segmentation task. Only the features referring to the current word are listed in the following table for the sake of brevity. In addition, all feature names may be prefixed by “p.” or ”c.” to indicate that they refer to the previous or the following word, rather than to the current word. Thus, thrice as many features may be generated in total. The horizontal lines in the table delimit individual feature groups (pause, duration, pitch, energy, and “other”).

Feature Name	**Description**
pause.after	Pause duration after current word
vowel.avg_dur	Average duration of vowels in current word
vowel.avg_dur.norm	Normalized average duration of vowels in current word
vowel.avg_dur.snorm	Speaker-normalized average duration of vowels in current word
vowel.avg_dur.var	Mean variance of vowel duration
vowel.dur.first_1st	Duration of first vowel in current word
vowel.dur.first_1st.norm	Normalized duration of first vowel in current word
vowel.dur.first_1st.snorm	Speaker-normalized duration of first vowel in current word
vowel.dur.first_1st.word_norm	Duration of first vowel in current word normalized by mean duration of all vowels in current word
vowel.dur.first_2nd	Duration of second vowel in current word
vowel.dur.first_2nd.norm	Normalized duration of second vowel in current word
vowel.dur.first_2nd.snorm	Speaker-normalized duration of second vowel in current word
vowel.dur.first_2nd.word_norm	Duration of first vowel in current word normalized by mean duration of all vowels in current word

Feature Name	**Description**
vowel.dur.last_1st	Duration of last vowel in current word
vowel.dur.last_1st.norm	Normalized duration of last vowel in current word
vowel.dur.last_1st.snorm	Speaker-normalized duration of last vowel in current word
vowel.dur.last_1st.word_norm	Duration of last vowel in current word normalized by mean duration of all vowels in current word
vowel.dur.last_2nd	Duration of penultimate vowel in current word
vowel.dur.last_2nd.norm	Normalized duration of penultimate vowel in current word
vowel.dur.last_2nd.snorm	Speaker-normalized duration of penultimate vowel in current word
vowel.dur.last_2nd.word_norm	Duration of penultimate vowel in current word normalized by mean duration of all vowels in current word
vowel.max_dur	Duration of the longest vowel in current word
vowel.max_dur.norm	Normalized duration of the longest vowel
vowel.max_dur.snorm	Speaker-normalized duration of the longest vowel
vowel.max_dur.z	Z-score of vowel.max_dur
vowel.med_dur	Median duration of vowels in current word
vowel.med_dur.norm	Normalized median duration of vowels
vowel.med_dur.snorm	Speaker-normalized median duration of vowels
vowel.min_dur	Duration of the shortest vowel in current word
vowel.min_dur.norm	Normalized duration of the shortest vowel
vowel.min_dur.snorm	Speaker-normalized duration of the shortest vowel in current word
vowel.min_dur.z	Z-score of vowel.min_dur
vowel.75_dur	75% fractile of duration of vowels in current word
vowel.75_dur.norm	75% fractile of normalized duration of vowels
vowel.75_dur.snorm	75% fractile of speaker-normalized duration of vowels
word.dur	Raw word duration
word.dur.norm	Normalized word duration
word.dur.snorm	Speaker-normalized word duration
word.dur.last_rhyme	Duration of last rhyme
word.dur.last_rhyme.norm	Normalized duration of last rhyme
word.dur.last_rhyme.snorm	Speaker-normalized duration of last rhyme
f0.baseline	Speaker's F_0 baseline
f0.max	Max value of F_0 in word
f0.mean	Mean value of F_0 in word
f0.min	Min value of F_0 in word
f0.first	First value of raw F_0 in word
f0.last	Last value of raw F_0 in word

Feature Name	**Description**
f0.last.max	F_0 maximum in last voiced region
f0.last.min	F_0 minimum in last voiced region
f0.pwl_first	First value of PWL F_0 in word
f0.pwl_last	Last value of PWL F_0 in word
f0.diff.first___baseline	Difference between f0.first and speaker's baseline
f0.diff.last___baseline	Difference between f0.last and speaker's baseline
f0.diff.last___first	Difference between f0.last of current and f0.first of following word
f0.diff.pwl_first___baseline	Difference between f0.pwl_first and speaker's baseline
f0.diff.pwl_last___baseline	Difference between f0.pwl_last and speaker's baseline
f0.diff.pwl_last___pwl_first	Difference between f0.pwl_last of current and f0.pwl_first of following word
f0.logdiff.first___baseline	Log of f0.diff.first___baseline
f0.logdiff.last___baseline	Log of f0.diff.last___baseline
f0.logdiff.last___first	Log of f0.diff.last___first
f0.logdiff.pwl_first___baseline	Log of f0.diff.pwl_first___baseline
f0.logdiff.pwl_last___baseline	Log of f0.diff.pwl_last___baseline
f0.logdiff.pwl_last___pwl_first	Log of f0.diff.pwl_last___pwl_first
f0.ratio.first___baseline	Ratio of f0.first and speaker's baseline
f0.ratio.last___baseline	Ratio of f0.last and speaker's baseline
f0.ratio.last___first	Ratio of f0.last of current and f0.first of following word
f0.ratio.pwl_first___baseline	Ratio of f0.pwl_first and speaker's baseline
f0.ratio.pwl_last___baseline	Ratio of f0.pwl_last and speaker's baseline
f0.ratio.pwl_last___pwl_first	Ratio of f0.pwl_last of current and f0.pwl_first of following word
f0.ratio.first_avg___baseline	Ratio of average F_0 in first voiced region and baseline
f0.ratio.first_beg___baseline	Ratio of F_0 onset in first voiced region and baseline
f0.ratio.first_end___baseline	Ratio of F_0 offset in first voiced region and baseline
f0.ratio.first_max___baseline	Ratio of F_0 maximum in first voiced region and baseline
f0.ratio.first_min___baseline	Ratio of F_0 minimum in first voiced region and baseline
f0.ratio.last_avg___baseline	Ratio of average F_0 in last voiced region and baseline
f0.ratio.last_beg___baseline	Ratio of F_0 onset in last voiced region and baseline
f0.ratio.last_end___baseline	Ratio of F_0 offset in last voiced region and baseline
f0.ratio.last_max___baseline	Ratio of F_0 maximum in last voiced region and baseline
f0.ratio.last_min___baseline	Ratio of F_0 minimum in last voiced region and baseline
f0.logratio.first___baseline	Log of f0.ratio.first___baseline
f0.logratio.last___baseline	Log of f0.ratio.last___baseline
f0.logratio.last___first	Log of f0.ratio.last___first
f0.logratio.pwl_first___baseline	Log of f0.ratio.pwl_first___baseline
f0.logratio.pwl_last___baseline	Log of f0.ratio.pwl_last___baseline

Feature Name	**Description**
f0.logratio.pwl_last___pwl_first	Log of f0.ratio.pwl_last___pwl_first
f0.slope.first	First slope PWL F_0 in word
f0.slope.last	Last PWL F_0 in word
f0.slope.diff.last___first	Difference between f0.slope.last of current and f0.slope.first of following word
RMS.max	Max RMS in current word
RMS.max.norm	Max RMS divided by mean RMS in current turn
RMS.mean	Mean RMS in current word
RMS.mean.norm	Mean RMS divided by mean RMS in current turn
RMS.min	Min RMS in current word
RMS.min.norm	Min RMS divided by mean RMS in current turn
RMS.voiced.max	Max voiced RMS in current word
RMS.voiced.max.norm	Max voiced RMS divided by mean voiced RMS in current turn
RMS.voiced.mean	Mean voiced RMS in current word
RMS.voiced.mean.norm	Mean voiced RMS divided by mean voiced RMS in current turn
RMS.voiced.min	Min voiced RMS in current word
RMS.voiced.min.norm	Min voiced RMS divided by mean voiced RMS in current turn
turn.is_begin	Is word turn-initial? (speaker change before the word)
turn.is_end	Is word turn-final? (speaker change after the word)
turn.speaker	Speaker ID
overlap.n_overlap	Number of words overlapping with current word
overlap.spurt_end25	Number of "spurt ends" overlapping with current word - boundary 25ms
overlap.spurt_end50	Number of "spurt ends" overlapping with current word - boundary 50ms
overlap.spurt_intern25	Number of "spurt-internal" words overlapping with current word - boundary 25ms
overlap.spurt_intern50	Number of "spurt-internal" words overlapping with current word - boundary 50ms
overlap.spurt_start25	Number of "spurt starts" overlapping with current word - boundary 25ms
overlap.spurt_start50	Number of "spurt starts" overlapping with current word - boundary 50ms
word.start	Elapsed time from the beginning of current turn

Appendix C

Examples of Automatically Segmented Speech Transcripts

In this appendix, I present illustrative examples of speech transcripts automatically segmented into sentence-like units. Even though there is no audio available, I believe that readers may get a good notion how successful the automatic methods are. To present objective illustrations, I tried to find such regions of the test data for which the automatic system performs at similar error rates as on the whole test set.

This appendix is organized as follows. There are three one-page sections with examples. Each of the sections represents one of the three corpora used in this thesis (English ICSI meetings, Czech BN, and Czech RF) and shows a pair of examples. In each pair, the first example illustrates an automatically segmented manual transcript (REF), and the second example illustrates an automatically segmented automatic transcript (ASR) that corresponds to the same region of speech as the manual transcript. For each corpus, I also recapitulate overall performance rates using all three metrics (BER, NIST, and F). In the examples, automatic decisions are displayed as follows:

- Correctly placed boundaries are displayed in green: /.
- Insertion errors (false alarms) are displayed in red: /.
- Deletion errors (misses) are displayed in blue: /.

For the sake of readability, the first example starts at the beginning of the next page.

C.1 ICSI Meeting Data

REF:
(Overall performance rates – $BER = 5.18\%$, $NIST = 32.49\%$, $F = 83.18\%$)

the the worst system still reduced the error rate by thirty three percent or something in development set /. so /. so you know sort of everybody is doing things between /. well roughly a third of the errors and half the errors being eliminated /. uh and varying on different test sets /. and so forth /. so i think /. um /. it's probably a good time to look at what's really going on /. and seeing if there's a there's a way to combine the best ideas /. while at the same time not blowing up the amount of uh resources used /. because that's that's critical for this this test /. um uh /. the uh the- there were two systems that were put forth by a combination of of uh french telecom and alcatel /. and um /. they they differed in some respects /. but they e- them- one was called the french telecom alcatel system the other was called the alcatel french telecom system /. uh which is the biggest difference /. i think /. but but there're there're there're some other differences too /. uh and and uh they both did very well /. you know /. so um my impression is they also did very well on on the the uh evaluation set /. but um /. i i- we haven't seen /. you've- you haven't seen any final results for that /. yeah /. there is a couple pieces to it /. there's a spectral subtraction style piece it was basically you know wiener filtering /. and then then there was some p- some modification of the cepstral parameters where they /. yeah /. but some people have done exactly that sort of thing /. of of and the i mean it's not to to look in speech only to try to m- to measure these things during speech /. that's p- that's not that uncommon /. but i- it- it /. so it looks like they did some some uh reasonable things /.

ASR:
(Overall performance rates – $BER = 6.19\%$, $NIST = 44.70\%$, $F = 75.85\%$)

the the worst system still reduce their race by thirty three percent or something /. so it's it's /. so /. so you know so everybody is doing things between uh roughly thirty years /. and the errors being eliminated uh and bearing on different s. /. and so forth /. so i think /. um /. it's probably a good time to look at what's really going on /. and see if there's a there's a way to combine the best ideas /. swell same time not blowing up the amount of uh resources used /. because that's that's critical for this this test /. um uh of /. the uh they they were two systems that were put forth by combination of of uh france telecom and now could tell /. and um /. they they differed in some respects /. but they in one was called french telecom elk tell system it was called up and tell francetelecom still kind of system uh what's the biggest difference i think /. but but there there there are some other differences too /. uh /. and and uh they both did very well /. you know /. so um /. my impression is they also did very well on on the the uh evaluation set /. but um /. i i we haven't seen /. if you haven't seen any find ourselves /. there's a couple pieces to it /. to suspect this attraction southeast is basically you know wiener filter /. and then then there was some some modification of the capsule parameters /. but some people have done exactly that sort of thing /. of i mean it's not to to look in speech only try to measure these things during speeches /. but that's not that uncommon /. but it it still looks like he did some some uh reasonable things /.

C.2 Czech BN Data

REF:
(Overall performance rates – $BER = 1.27\%$, $NIST = 15.63\%$, $F = 92.30\%$)

zprávy /. jednání mezi mosteckou uhelnou a společností shd peel /. o osudu dolu kohinoor v mariánských radčicích trvalo dlouho do noci /. konkrétní informace o možnosti prodeje dolu ale budou zveřejněny až dopoledne /. uzavření kupní smlouvy je přitom hlavním požadavkem sedmačtyřiceti horníků kteří zůstávají už čtrnáctý den pod zemí z obavy před uzavřením dolu a ztrátou zaměstnání /. poslanci by mohli už odpoledne zahájit volbu členů nové rady české televize /. z předběžné dohody mezi poslaneckými kluby vyplývá že kandidáti čssd a ods zřejmě obsadí v tomto mediálním orgánu sedm míst z devíti přičemž sociálním demokratům připadnou čtyři místa a ods tři /. lidovci a unionisté by pak podle této dohody mohli do rady české televize dosadit po jednom kandidátovi /. zatímco komunisté jsou z této hry pravděpodobně vyšachováni /. podle předsedy klubu ksčm vojtěcha filipa se ale nevzdávají /. a své dva kandidáty i přesto navrhnou /. neveřejným zasedáním budou ve své schůzi pokračovat senátoři /. projednají vládní materiál který vyhodnocuje činnost armádních jednotek v zahraničí /. poté se horní parlamentní komora bude zabývat třemi zákony které souvisejí s reformou veřejné správy /. v praze budou demonstrovat lidé kterým dluží peníze zkrachovalé kampeličky /. protest na hradčanském náměstí pořádá celostátní koordinační centrum postižených střadatelů družstevních záložen /. v praze také proběhne demonstrace za dodržování lidských práv v čečensku /. vystoupí na ní zástupci nevládních organizací politici i čečenští uprchlíci /. večer se pak uskuteční ekumenická bohoslužba za ukončení násilí na severním kavkaze /.

ASR:
(Overall performance rates – $BER = 1.61\%$, $NIST = 20.09\%$, $F = 90.02\%$)

z trávy ER /. jednání mezi mosteckou uhelnou a s společností s hady peel o osudu dolu kohinoor v mariánských radčicích trvalo dlouho do noci /. konkrétní informace o možnosti prodeje dolu ale budou zveřejněny až dopoledne /. uzavření kupní smlouvy je přitom hlavním požadavkem sedmačtyřiceti horníků kteří zůstávají už čtrnáctý den podzimní z obavy před uzavřením dolu a ztrátou zaměstnání /. a poslanci by mohli už odpoledne zahájit volbu členů nové rady české televize /. předběžné dohody mezi poslaneckými kluby vyplývá že kandidáti čssd a ods zřejmě obsadí v tomto mediálním orgánu se do míst z devíti přičemž sociálním demokratům připadnout čtyři místa a ods tři /. lidovci a unionisté by pak podle této dohody mohli do rady české televize dosadit po jednom kandidátovi /. zatímco komunisté jsou z této hry pravděpodobně vyšachování /. podle předsedy klubu ksčm vojtěcha filipa se ale nevzdávají /. a své dva kandidáty přesto navrhnou /. neveřejným zasedáním budou ve své schůzi pokračovat senátoři /. projednají vládní materiál který vyhodnocuje činnost armádních jednotek v zahraničí /. poté se horní parlamentní komora bude zabývat třemi zákony které souvisejí s reformou veřejné správy /. vzp /. v praze budou demonstrovat lidé kterým dluží peníze zkrachovalé kampeličky /. protest na hradčanském náměstí pořádá celostátní koordinační centrum postižených střadatelů družstevních záložen /. v praze také proběhne demonstrace za dodržování lidských práv v čečensku /. vystoupí na ní zástupci nevládních organizací politici i čečenští uprchlíci /. večer se pak uskuteční ekumenická bohoslužba za ukončení násilí na severním kavkaze /.

C.3 Czech RF Data

REF:
(Overall performance rates – $BER = 3.22\%$, $NIST = 46.35\%$, $F = 75.65\%$)

A: pojďme se zastavit u jiné věci /. ministr práce a sociálních věcí místopředseda čssd zdeněk škromach v sobotním právu vyjádřil názor že kdyby bylo po jeho bohatí by měli platit do systému víc /. a i těch třicet sedum procent v pátém daňovém pásmu které se nepodařilo prosadit by pro pana ministra škromacha byla jen symbolika protože by tak státní rozpočet získal jen necelou miliardu /. co říkáte takovému názoru jako předseda rozpočtového výboru a taky jako křesťanský demokrat jemuž EE so- solidarita sou není cizí /.

B: ale oni přece ti bohatí platí víc /. třicet dva procent z velkého příjmu je samozřejmě mnohem víc peněz než patnáct procent z malého příjmu /. a také daň z příjmu fyzických osob je jediná daň u nás která je progresivní /. která skutečně stoupá podle toho čím víc máte tím více platíte podle toho v jakém ste daňovém pásmu /. ale ta progresivita musí mít svojí míru /. kdybych MM extrapoloval vyjádření pana ministra škromacha tak bych samozřejmě ře- tak by se dalo říct že kdo bude že když se bude platit sto procent v nejvyšším daňovém pásmu tak se vybere nejvíc daní /. a to není pravda /. pak vám samozřejmě ti nejschopnější kteří jsou schopni EE vydělávat nejvíc a tím pádem také nejvíc přispívat /. prostě z takového systému utečou /. EE my sme přesvědčeni že současná míra progrese která je mezi patnácti až třiceti dvěma procenty je míra progrese která je optimální /. a nechceme tu progresi zvyšovat právě proto aby ti schopní neutíkali a nebyli do- demotivováni k tomu aby vytvářeli hodnoty /.

ASR:
(Overall performance rates – $BER = 3.98\%$, $NIST = 57.65\%$, $F = 69.12\%$)

A: pojďme zase jiné věci /. ministra práce a sociálních věcí místopředseda čssd zdeněk škromach v sobotním právu vyjádřil názor že kdyby bylo po jeho bohatí by měli platit do systému HM s /. i těch třicet sedm procent v pátém daňovém pásmu které se nepodařilo prosadit by pro pana ministra škromacha dva jen symbolika protože by tak státní rozpočet získal jen necelou miliardu /. co říkáte takovému názoru jako předseda rozpočtového výboru a tak jako křesťanský demokrat jemuž jsou solidarita soud není cizí /.

B: ale MM oni přece ti bohatí platí víc /. třicet dva procent z velkého příjmu je samozřejmě mnohem víc peněz než patnáct procent z malého příjmu /. a také daň z příjmu fyzických osob je jediná liberální u nás která je progresivní které skutečně stoupá podle toho čím víc o té tím více platíte podle toho v jakém jste daňovém pásmu /. ale ta progresivita musí mít svoji míru /. kdybych šel extrapolovat vyjádření pana ministra škromacha tak bych samozřejmě že teď by se iluzí že výrobu nepodílí se bude platit sto procent nejvyšším daňovém pásmu kde se vybere víc víc /. denního /. to není pravda /. pak nám samozřejmě ti nejschopnější kteří jsou schopni /. a EE nevyděláváte nejvíc a tím pádem taky nejvíc přispívat prostě z takového systému utečou /. my EE jsme přesvědčeni že současná míra progrese která je mezi patnácti až třiceti dvěma procenty je EE míra progrese která je optimální /. a nechceme tu progresi zvyšovat právě proto aby ti schopní neutíkali nebyly demotivováni který by tvořili hodnoty /.

Bibliography

[1] D. Jones, F. Wolf, E. Gibson, E. Williams, E. Fedorenko, D. Reynolds, and M. Zissman, "Measuring the readability of automatic speech-to-text transcripts," in *Proc. Eurospeech*, Geneva, Switzerland, 2003.

[2] J. G. Kahn, M. Ostendorf, and C. Chelba, "Parsing conversational speech using enhanced segmentation," in *Proc. HLT-NAACL'04*, Boston, MA, USA, 2004.

[3] S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori, "Speech-to-text and speech-to-speech summarization of spontaneous speech," *IEEE Transaction on Speech and Audio Processing*, vol. 12, no. 4, pp. 401–408, 2004.

[4] E. Matusov, D. Hillard, M. Magimai-Doss, D. Hakkani-Tür, M. Ostendorf, and H. Ney, "Improving speech translation with automatic boundary prediction," in *Proc. INTERSPEECH 2007*, Antwerp, Belgium, 2007.

[5] Z. Palková, *Fonetika a fonologie češtiny (Phonetics and Phonology of Czech).* Karolinum, 1997.

[6] I. Lehiste, "Perception of sentence and paragraph boundaries," in *Frontiers of Speech Communication Research*, 1979, pp. 191–201.

[7] M. Swerts, A. Wichmann, and R.-J. Beun, "Filled pauses as markers of discourse structure," in *Proceedings of ICSLP 96*, vol. 2, Philadelphia, USA, 1996.

[8] M. Swerts, "Prosodic features at discourse boundaries of different strength," *Journal of Acoustic Society of America*, vol. 101, no. 1, pp. 514–521, 1997.

[9] P. Hansson, "Prosodic phrasing in spontaneous Swedish," *Travaux de l'institut de linguistique de Lund*, vol. 43, 2003.

[10] A. Fox, *Prosodic Features and Prosodic Structure.* Oxford University Press, 2000.

[11] A. Batliner, A. Buckow, H. Niemann, E. Nöth, and V. Warnke, "The prosody module," in *VERBMOBIL: Foundations of Speech-to-speech Translations*, W. Wahlster, Ed. New York, Berlin: Springer, 2000, pp. 106–121.

[12] J. Psutka, L. Müller, J. Matoušek, and V. Radová, *Mluvíme s počítačem česky (Speaking with Computer in Czech).* Prague: Academia, 2006.

[13] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing.* Prentice Hall PTR, 2001.

[14] I. Shafran, M. Ostendorf, and R. Wright, "Prosody and phonetic variability: Lessons learned from acoustic model clustering," in *Proceedings of Workshop on Prosody in ASR.* ISCA, 2001.

[15] M. Swerts, E. Strangert, and M. Heldner, "F_0 declination in read-aloud and spontaneous speech," in *Proceedings of ICSLP 96*, Philadelphia, USA, 1996.

[16] J. Vaissiere, "Language-independent prosodic features," in *Prosody: Models and Measurments.* Springer, 1983, pp. 53–66.

[17] J. Hirschberg and C. Nakatani, "A prosodic analysis of discourse segments in direction-giving monologues," in *Proceedings ACL*, Santa Cruz, 1996, pp. 286–293.

[18] M. Oliveira, "Pitch reset as a cue for narrative segmentation," in *Proceedings of Prosodic Interfaces 2003*, 2003.

[19] J. 't Hart, R. Collier, and A. Cohen, *A perceptual study of intonation.* Cambridge University Press, 1990.

[20] J. Ohala, "Prosody and phonology," in *Proceedings of International conference Speech Prosody*, Nara, Japan, 2004.

[21] F. Daneš, *Intonace a věta ve spisovné češtině (Intonation and sentence in standard Czech).* Nakladatelství ČSAV, 1957.

[22] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard scheme for labeling prosody," in *Proceedings ICSLP'92*, Banff, Canada, 1992.

[23] D. Oppermann and S. Berger, "What makes speech data spontaneous?" in *Proceedings of ICPhS*, San Francisco, 1999.

[24] E. Shriberg, "Preliminaries to a theory of speech disfluencies," Ph.D. dissertation, University of California, Berkeley, 1994.

[25] M. Honal and T. Schulz, "Correction of disfluencies in spontaneous speech using a noisy-channel approach," in *Proceedings of EUROSPEECH*, Geneva, Switzerland, 2003, pp. 2781–2784.

[26] E. Shriberg, "Disfluencies in Switchboard," in *Proceedings of ICSLP96*, vol. addendum, Philadelphia, USA, 1996, p. 11.

[27] M. Swerts, "Filled pauses as markers of discourse structure," *Journal of Pragmatics*, vol. 30, pp. 485–496, 1998.

[28] K. Bailey and F. Ferreira, "Do non-word disfluencies affect syntactic parsing?" in *Proceedings of ITRW on Disfluency in Spontaneous Speech (DiSS'01)*, Edinburgh, Scotland, UK, 2001.

[29] H. H. Clark and J. F. Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, pp. 73–111, 2002.

[30] J. Fox Tree, "The effect of false starts and repetitions on the processing of subsequent words in spontaneous speech," *Journal of Memory and Language*, vol. 34, pp. 709–738, 1995.

[31] H. H. Clark and T. Wasow, "Repeating words in spontaneous speech," *Cognitive Psychology*, vol. 37, pp. 201–242, 1998.

[32] M. Swerts and R. Geluykens, "Prosody as a marker of information flow in spoken discourse," *Language and speech*, vol. 37, no. 1, pp. 21–43, 1994.

[33] A. Kiessling, R. Kompe, A. Batliner, H. Niemann, and E. Nöth, "Classification of boundaries and accents in spontaneous speech," in *Proc. of the 3rd CRIM / FORWISS Workshop*, Montreal, Canada, 1996.

[34] E. Shriberg, "Phonetic consequences of speech disfluency," in *Proceedings of ICPhS '99*, San Francisco, CA, USA, 1999.

[35] B. Megyesi and S. Gustafson-Čapková, "Production and perception of pauses and their linguistic context in read and spontaneous speech in Swedish," in *Proceedings of ICSLP*, Denver, USA, 2002.

[36] J. Vaissiere, "Phonetic explanations for cross–linguistic prosodic similarities," *Phonetica*, vol. 52, pp. 123–130, 1995.

[37] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1707–1717, 1992.

[38] M. Horne, E. Strangert, and M. Heldner, "Prosodic boundary strength in Swedish: Final lengthening and silent interval duration," in *Proceedings of the XIIIth ICPhS*, Stockholm, Sweden, 1995.

[39] M. Heldner and B. Megyesi, "Exploring the prosody-syntax interface in conversations," in *Proceedings of 15th ICPhS*, Barcelona, Spain, 2003.

[40] E. Strangert, “Speech chunks in conversation: Syntactic and prosodic aspects,” in *Proceedings of Speech Prosody 2004*, Nara, Japan, 2004.

[41] L. Yang, “Duration and pauses as phrasal and boundary marking indicators in speech,” in *Proceedings of 15th ICPhS*, Barcelona, Spain, 2003, pp. 1791–1794.

[42] M. van Donzel and F. Koopmans, “Pausing strategies in discourse in Dutch,” in *Proceedings of ICSLP96*, Philadelphia, USA, 1996.

[43] M. van Donzel and F. J. Koopmans, “Perception of discourse boundaries and prominence in spontaneous Dutch speech,” Working papers 46, Lund University, 1997.

[44] F. Grosjean, “How long is the sentence? Prediction and prosody in the on-line processing of language,” *Linguistics*, vol. 21, pp. 501–529, 1983.

[45] F. Grosjean and C. Hirt, “Using prosody to predict the end of sentences in English and French: Normal and brain-damaged subject,” *Language and cognitive processes*, vol. 11, no. 1-2, pp. 107–134, 1996.

[46] R. Carlson and M. Swerts, “Perceptually based prediction of upcoming prosodic breaks in spontaneous Swedish speech materials,” in *Proceedings of 15th ICPhS*, Barcelona, Spain, 2003.

[47] R. Carlson, J. Hirschberg, and M. Swerts, “Prediction of upcoming Swedish prosodic boundaries by Swedish and American listeners,” in *Proc. od Speech Prosody 2004*, Nara, Japan, 2004.

[48] M. Fach, “A comparison between syntactic and prosodic phrases,” in *Proceedings of EUROSPEECH*, Budapest, Hungary, 1999.

[49] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, “Prosody-based automatic segmentation of speech into sentences and topics,” *Speech Communication*, vol. 32, no. 1-2, pp. 127–154, 2000.

[50] G. K. Kanji, *100 Statistical Tests*, 3rd ed. Sage Publications Ltd, 1999.

[51] M. Ostendorf and D. Hillard, “Scoring structural MDE: Towards more meaningful error rates,” in *Proc. of EARS Rich Transcription Workshop*, 2004.

[52] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, “Performance measures for information extraction,” in *Proc. DARPA Broadcast News Workshop*, Herndon, 1999.

[53] M. Haase, W. Kriechbaum, G. Möhler, and G. Stenzel, “Deriving document structure from prosodic cues,” in *Proc. of EUROSPEECH 2001*, Aalborg, Denmark, 2001.

[54] D. Wang, L. Lu, and H.-J. Zhang, “Speech segmentation without speech recognition,” in *ICASSP 2003*, Hong Kong, 2003.

[55] D. Wang and S. Narayanan, “A multi–pass linear fold algorithm for sentence boundary detection using prosodic cues,” in *Proceedings of ICASSP 2004*, Montreal, Canada, 2004.

[56] M. Gavalda, “High performance segmentation of spontaneous speech using part of speech and trigger word information,” in *Proc. of 5th Conference on Applied Natural Language Processing*, Washington D.C., USA, 1997.

[57] D. D. Palmer and M. A. Hearst, “Adaptive multilingual sentence boundary disambiguation,” *Computational Linguistics*, vol. 23, no. 2, pp. 241–267, 1997.

[58] D. Beeferman, A. Berger, and J. Lafferty, “Cyberpunc: A lightweight punctuation annotation system for speech,” in *Proceedings of ICASSP*, 1998, pp. 689–692.

[59] M. Stevenson and R.Gaizauskas, “Experiments on sentence boundary detection,” in *Proc. NAACL*, Seattle, USA, 2000.

[60] N. Gupta, S. Bangalore, and M. Rahim, “Extracting clauses for spoken language understanding in conversational systems,” in *Proceedings of ICSLP 2002*, 2002, pp. 361–364.

[61] S. Shieber and X. Tao, “Comma restoration using constituency information,” in *Proceedings of HLT-NAACL 2003*, vol. Main papers, 2003, pp. 142–148.

[62] J. Mrozinski, E. Whittaker, P. Chatain, and S. Furui, "Automatic sentence segmentation of speech for automatic summarization," in *Proc. IEEE ICASSP*, Toulouse, France, 2006, pp. 981–984.

[63] Y. Lee, Y. Al-Onaizan, K. Papineni, and S. Roukos, "IBM spoken language translation system," in *Proc. TC-Star Workshop on Speech-to-Speech Translation*, Barcelona, Spain, 2006.

[64] K. Takanashi, T. Maruyama, K. Uchimoto, and H. Isahara, "Identification of "sentences" in spontaneous Japanese," in *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, Japan, 2003.

[65] C. Chelba and F. Jelinek, "Structured language modeling," *Computer Speech and Language*, vol. 14, pp. 283–332, 2000.

[66] S. Mori, N. Itoh, and M. Nishimura, "An automatic sentence boundary detector based on a structured language model," in *Proceedings of ICSLP*, Denver, Colorado, 2002, pp. 921–924.

[67] T. Oba, T. Hori, and A. Nakamura, "Sentence boundary detection using sequential dependency analysis combined with CRF-based chunking," in *Proc. INTERSPEECH 2006 - ICSLP*, Pittsburgh, PA, USA, 2006.

[68] E. Shriberg and A. Stolcke, "Direct modeling of prosody: An overview of applications in automatic speech processing," in *Proceedings of International Conference Speech Prosody 2004*, Nara, Japan, 2004.

[69] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 469–481, 1994.

[70] S. Ananthakrishnan and S. Narayanan, "Combining acoustic, lexical, and syntactic evidence for automatic unsupervised prosody labeling," in *Proc. INTERSPEECH 2006 - ICSLP*, Pittsburgh, PA, USA, 2006.

[71] N. Veilleux, M. Ostendorf, and C. Wightman, "Parse scoring with prosodic information," in *Proceedings of ICSLP*, Banff, Canada, 1992.

[72] B. Roark, Y. Liu, M. Harper, R. Stewart, M. Lease, M. Snover, I. Shafran, B. Dorr, J. Hale, A. Krasnyanskaya, and L. Yung, "Reranking for sentence boundary detection in conversational speech," in *Proc. IEEE ICASSP*, Toulouse, France, 2006.

[73] M. Dreyer and I. Shafran, "Exploiting prosody for PCFGs with latent annotations," in *Proc. Interspeech'07*, Antwerp, Belgium, 2007.

[74] A. Stolcke, E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tür, and Y. Lu, "Automatic detection of sentence boundaries and disfluencies based on recognized words," in *Proceedings of ICSLP98*, Sydney, Australia, 1998.

[75] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, B. Peskin, and M. Harper, "The ICSI-SRI-UW metadata extraction system," in *Proc. of ICSLP*, Jeju, Korea, 2004.

[76] D. Hillard, M. Ostendorf, A. Stolcke, Y. Liu, and E. Shriberg, "Improving automatic sentence boundary detection with confusion networks," in *HLT/NAACL 04*, Boston, USA, 2004.

[77] J. H. Kim and P. Woodland, "A combined punctuation generation and speech recognition system and its performance enhancement using prosody," *Speech Communication*, vol. 41, no. 4, pp. 563–577, 2003.

[78] A. Batliner, R. Kompe, A. Kiessling, M. Mast, H. Niemann, and E. Nöth, "M = Syntax + Prosody: A syntactic–prosodic labelling scheme for large spontaneous speech databases," *Speech Communication*, vol. 25, pp. 193–222, 1998.

[79] E. Schukat-Talamazzini, F. Gallwitz, S. Harbeck, and V. Warnke, "Rational interpolation of maximum likelihood predictors in stochastic language modeling," in *Proc. Eurospeech '97*, Rhodes, Greece, 1997, pp. 2731–2734.

[80] V. Warnke, R. Kompe, H. Niemann, and E. Nöth, "Integrated dialog act segmentation and classification using prosodic features and language models," in *Proc. Europeech 97*, Rhodes, Greece, 1997.

[81] A. Batliner, E. Nöth, J. Buckow, R. Huber, V. Warnke, and H. Niemann, “Duration features in prosodic classification: Why normalization comes second, and what they really encode,” in *ITRW on Speech recognition and understanding*, Red Bank, NJ, USA, 2001, pp. 23–28.

[82] F. Gallwitz, *Integrated Stochastic Models for Spontaneous Speech Recognition*, ser. Studien zur Mustererkennung. Berlin, Germany: Logos Verlag, 2002.

[83] F. Gallwitz, H. Niemann, E. Nöth, and V. Warnke, “Integrated recognition of words and prosodic phrase boundaries,” *Speech Communication*, vol. 36, no. 1–2, 2002.

[84] C. Chen, “Speech recognition with automatic punctuation,” in *Proc. of EUROSPEECH*, Budapest, Hungary, 1999, pp. 447–450.

[85] Y. Gotoh and S. Renals, “Sentence boundary detection in broadcast speech transcripts,” in *Proceedings of ISCA workshop Challenges for the New Millenium (ASR-200)*, Paris, France, 2000.

[86] H. Christensen, Y. Gotoh, and S. Renals, “Punctuation annotation using statistical prosody models,” in *Proc. of ISCA Workshop on Prosody in Speech Recognition and Understanding*, Red Bank, NJ, USA, 2001.

[87] A. Srivastava and F. Kubala, “Sentence boundary detection in Arabic speech,” in *Proceedings of EUROSPEECH 2003*, Geneva, Switzerland, 2003.

[88] Y. Liu, A. Stolcke, E. Shriberg, and M. Harper, “Comparing and combining generative and posterior probability models: Some advances in sentence boundary detection in speech,” in *Proc. EMNLP*, Barcelona, Spain, 2004.

[89] ——, “Using conditional random fields for sentence boundary detection in speech,” in *Proc. ACL*, Ann Arbor, MI, USA, 2005.

[90] M. Tomalin and P. C. Woodland, “Discriminatevely trained Gaussian mixture models for sentence boundary detection,” in *Proc. IEEE ICASSP*, Toulouse, France, 2006, pp. 549–552.

[91] S. Young, G. Evermann, M. J. F. Gales, T. Hain, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, “The HTK book (for HTK version 3.4),” http://htk.eng.cam.ac.uk, 2006.

[92] M. Zimmermann, D. Hakkani-Tur, J. Fung, N. Mirghafori, L. Gottlieb, E. Shriberg, and Y. Liu, “The ICSI+ multilingual sentence segmentation system,” in *Proc. INTERSPEECH 2006 - ICSLP*, 2006, pp. 117–120.

[93] R. Schapire, “The boosting approach to machine learning: An overview,” in *Proc. of MSRI Workshop on Nonlinear Estimation and Classification*, 2002.

[94] E. Matusov, A. Mauser, and H. Ney, “Automatic sentence segmentation and punctuation prediction for spoken language translation,” in *Proc. International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006.

[95] F. Batista, D. Caseiro, N. Mamede, and I. Trancoso, “Recovering punctuation marks for automatic speech recognition,” in *Proc. INTERSPEECH*, Antwerp, Belgium, 2007.

[96] K. Shitaoka, K. Uchimoto, T. Kawahara, and H. Isahara, “Dependency structure analysis and sentence boundary detection in spontaneous Japanese,” in *Proceedings of the 20th international conference on Computational Linguistics*, Geneva, Switzerland, 2004.

[97] Y. Akita, M. Saikou, H. Nanjo, and T. Kawahara, “Sentence boundary detection of spontaneous Japanese using statistical language model and support vector machines,” in *Proc. INTERSPEECH 2006 - ICSLP*, Pittsburgh, PA, USA, 2006.

[98] V. Radová, J. Psutka, L. Müller, W. Byrne, J. V. Psutka, P. Ircing, and J. Matoušek, “Czech Broadcast News Speech and Transcripts,” Linguistic Data Consortium, CD-ROM LDC2004S01 and LDC2004T01, Philadelphia, PA, USA, 2004.

[99] J. Psutka, V. Radová, L. Müller, J. Matoušek, P. Ircing, and D. Graff, "Large broadcast news and read speech corpora of spoken Czech," in *Proceedings of EUROSPEECH.* Aalborg, Denmark: ISCA, 2001, pp. 2067–2070.

[100] J. Psutka, J. Hajič, and W. Byrne, "The development of ASR for Slavic languages in the MALACH project," in *IEEE ICASSP 2004*, Montreal, Canada, 2004.

[101] Linguistic Data Consortium, "Guidelines for RT-04 transcription (version 3.1)," http://projects.ldc.upenn.edu/Transcription/rt-04/RT-04-guidelines-V3.1.pdf.

[102] M. Meeter, "Dysfluency annotation stylebook for the Switchboard corpus," ftp://ftp.cis.upenn.edu/pub/treebank-/swbd/doc/DFL-book.ps, 1995.

[103] P. Heeman, "Speech repairs, intonational boundaries and discourse markers: Modeling speakers' utterances in spoken dialogs," Ph.D. dissertation, University of Rochester, New York, 1997.

[104] O. Müllerová, *Mluvený text a jeho syntaktická výstavba (The Syntax of Spoken Text).* Praha: Academia, 1994.

[105] S. Strassel, "Simple metadata annotation specification V6.2," http://www.ldc.upenn.edu/Projects/MDE/Guidelines/SimpleMDE_V6.2.pdf, 2004.

[106] S. Strassel, D. Miller, K. Walker, and C. Cieri, "Shared resources for robust speech-to-text technology," in *Proc. EUROSPEECH 2003*, Geneva, Switzerland, 2003.

[107] Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, J. Ang, D. Hillard, M. Ostendorf, M. Tomalin, P. Woodland, and M. Harper, "Structural metadata research in the EARS program," in *Proc. IEEE ICASSP*, Philadelphia, USA, 2005.

[108] K. Svoboda, *Souvětí spisovné češtiny (Compound Sentences in Standard Czech).* SPN Praha, 1970.

[109] M. Grepl and P. Karlík, *Skladba češtiny (Syntax of Czech).* Votobia, 1998.

[110] M. Grepl, Z. Hladká, M. Jelínek, P. Karlík, M. Krčmová, M. Nekula, Z. Rusínová, and D. Šlosar, *Příruční mluvnice češtiny (Handbook of Czech grammar)*, P. Karlík, M. Nekula, and Z. Rusínová, Eds. Lidové noviny, 2003.

[111] M. Erard, "Just like, er, words, not, um, throwaways," *New York Times*, vol. (Jan 3, 2004), 2004, available from http://www.speech.sri.com/press/nyt-jan03-2004.html.

[112] P. Kaderka and Z. Svobodová, "Manuál pro přepisovatele televizních diskusních pořadu (Guidelines for annotators of broadcast discussions)," *Jazykovědné aktuality*, no. 3-4, pp. 18–51, 2006.

[113] D. Hillard, "Automatic sentence structure annotation for spoken language processing," Ph.D. dissertation, University of Washington, Seattle, WA, USA, 2008.

[114] K. Maeda and S. Strassel, "Annotation tools for large-scale corpus development: Using AGTK at the Linguistic Data Consortium," in *Proc. LREC 2004*, Lisbon, Portugal, 2004.

[115] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[116] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," *Computational Linguistics*, To Appear.

[117] J. Carletta, "Assessing agreement on annotation tasks: The kappa statistic," *Computational Linguistics*, vol. 22, no. 2, pp. 249–254, 1996.

[118] D. Spoustová, J. Hajič, J. Votrubec, P. Krbec, and P. Květoň, "The best of two worlds: Cooperation of statistical and rule-based taggers for Czech," in *Proc. of the ACL Workshop on Balto-Slavonic Natural Language Processing*, Prague, Czech Republic, 2007.

[119] D. Baron, "Prosody-based automatic detection of punctuation and interreption events in the ICSI Meeting Recorder corpus," Research project, University of California, Berkeley, USA, 2002.

[120] J. Buckow, V. Warnke, R. Huber, A. Batliner, E. Nöth, and H. Niemann, “Fast and robust features for prosodic classification,” in *Proceedings of TSD’99 Mariánské Lázně*. Berlin: Springer, 1999, pp. 193–198.

[121] D. Talkin, “A robust algorithm for pitch tracking (RAPT),” in *Speech Coding and Synthesis*. Amsterdam, Netherlands: Elsevier Science, 1995, pp. 495–518.

[122] P. C. Bagshaw, “Automatic prosodic analysis for computer aided pronunciation teaching,” Ph.D. dissertation, The University of Edinburgh, Scotland, UK, 1994.

[123] K. Sönmez, L. Heck, M. Weintraub, and E. Shriberg, “A lognormal tied mixture model of pitch for prosody-based speaker recognition,” in *Proceedings of EUROSPEECH’97*, Rhodes, Greece, 1997, pp. 1391–1394.

[124] C. d’Alessandro and P. Mertens, “Automatic pitch contour stylization using a model of tonal perception,” *Computer Speech and Language*, vol. 9, pp. 257–288, 1995.

[125] D. Hirst and R. Espesser, “Automatic modelling of fundamental frequency using a quadratic spline function,” *Travaux de l’Institut de Phonétique d’Aix*, vol. 15, pp. 75–85, 1993.

[126] D. Hirst, A. di Cristo, and R. Espesser, “Levels of representation and levels of analysis for the description of intonation systems,” in *Prosody: Theory and Experiment*, M. Horne, Ed. Kluwer Academic Press, 2000.

[127] K. Sönmez, E. Shriberg, L. Heck, and M. Weintraub, “Modeling dynamic prosodic variation for speaker verification,” in *Proceedings of ICSLP*, Sydney, Australia, 1998, pp. 3189–3192.

[128] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling, *Numerical Recipes in C*. Cambridge University Press, 1992.

[129] T. Rietveld and P. Vermillion, “Cues for perceived pitch register,” *Phonetica*, vol. 60, pp. 261–272, 2003.

[130] E. Shriberg, A. Stolcke, and D. Baron, “Observations on overlap: Findings and implications for automatic processing of multi-party conversation,” in *Proc. EUROSPEECH’01*, Aalborg, Denmark, 2001.

[131] R. Kneser and H. Ney, “Improved backing-off for M-gram language modeling,” in *Proc. ICASSP*, Detroit, MI, 1995.

[132] S. Chen and J. Goodman, “An empirical study of smoothing techniques for language modeling,” Harvard University, USA, Tech. Rep., 1998.

[133] I. Witten and T. Bell, “The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression,” *IEEE Transactions on Information Theory*, vol. 37, no. 4, pp. 1085–1094, 1991.

[134] C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, 2nd ed. MIT Press, 2000.

[135] F. Jelinek, *Statistical Methods for Speech Recognition*. The MIT Press, 1998.

[136] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 2nd ed. Prentice Hall, 2008.

[137] A. Stolcke and E. Shriberg, “Automatic linguistic segmentation of conversational speech,” in *Proc. ICSLP*, Philadelphia, PA, USA, 1996.

[138] A. Stolcke, “SRILM – An extensible language modeling toolkit,” in *Proc. ICSLP’02*, Denver, CO, USA, 2002.

[139] L. Breiman, J. Friedman, R. Ohlsen, and C. Stone, *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth and Brooks Inc., 1984.

[140] R.-H. Li and G. Belford, “Instability of decision tree classification algorithms,” in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Canada, 2002.

[141] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[142] Y. Liu, N. Chawla, M. Harper, E. Shriberg, and A. Stolcke, "A study in machine learning from imbalanced data for sentence boundary detection in speech," *Computer Speech and Language*, vol. 20, pp. 468–494, 2006.

[143] W. Buntime and R. Caruana, "Introduction to IND version 2.1 and recursive partitioning," Moffet Fields, CA, USA, 1992.

[144] A. Berger, S. A. Della Pietra, and V. J. Della Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[145] H. Daumé III, "Notes on CG and LM-BFGS optimization of logistic regression," 2004, paper available at http://pub.hal3.name\#daume04cg-bfgs, implementation available at http://hal3.name/megam/.

[146] D. C. Liu and J. Nocedal, "On the limited memory method for large scale optimization," *Mathematical Programming B*, vol. 45, no. 3, pp. 503–528, 1989.

[147] A. Ratnaparkhi, "Maximum entropy models for natural language ambiguity resolution," Ph.D. dissertation, University of Pennsylvania, Philadelphia, PA, USA, 1998.

[148] A. Berger and R. Miller, "Just-in-time language modelling," in *Proc. ICASSP98*, Seattle, WA, USA, 1998.

[149] Y. Liu, "Structural event detection for rich transcription of speech," Ph.D. dissertation, Purdue University, W. Lafayette, IN, USA, 2004.

[150] R. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2–3, pp. 135–168, 2000.

[151] J. Kolář, E. Shriberg, and Y. Liu, "Using prosody for automatic sentence segmentation of multi-party meetings," *Lecture Notes in Computer Science (Proc. TSD'06)*, vol. 4188, pp. 629–636, 2006.

[152] S. Cuendet, "Model adaptation for sentence unit segmentation from speech," IDIAP, Martigny, Switzerland, Tech. Rep. IDIAP-RR 06-64, 2006.

[153] A. Niculescu-Mizil and R. Caruana, "Obtaining calibrated probabilities from boosting," in *Proc. 21st Conference on Uncertainty in Artificial Intelligence (UAI '05)*, Edinburgh, Scotland, UK, 2005.

[154] S. Armstrong, A. Clark, G. Coray, M. Georgescul, V. Pallotta, A. Popescu-Behs, D. Portabella, M. Rajman, and M. Starlander, "Natural language queries on natural language data: A database of meeting dialogues," in *Proc. NLDB*, Burg/Cottbus, Germany, 2003.

[155] A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner, "Advances in automatic meeting record creation and access," in *Proc. IEEE ICASSP*, Salt Lake City, UT, USA, 2001.

[156] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *IEEE ICASSP 2003*, Hong Kong, 2003.

[157] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proc. 5th SIGdial Workshop on Discourse and Dialogue*, Boston, MA, USA, 2004, pp. 97–100.

[158] R. Dhillon, S. Bhagat, H. Carvey, and E. Shriberg, "Meeting recorder project: Dialog act labeling guide," ICSI, Berkeley, CA, USA, Tech. Rep. TR-04-002, 2004.

[159] D. Baron, E. Shriberg, and A. Stolcke, "Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues," in *Proceedings of ICSLP*, Denver, USA, 2002, pp. 949–952.

[160] J. Ang, Y. Liu, and E. Shriberg, “Automatic dialog act segmentation and classification in multiparty meetings,” in *Proc. IEEE ICASSP 2005*, Philadelphia, USA, 2005.

[161] M. Zimmermann, Y. Liu, E. Shriberg, and A. Stolcke, “A* based joint segmentation and classification of dialog acts in multi-party meetings,” in *Proceedings of ASRU’05*, San Juan, Puerto Rico, 2005.

[162] M. Zimmermann, A. Stolcke, and E. Shriberg, “Joint segmentation and classification of dialog acts in multiparty meetings,” in *Proc. IEEE ICASSP 2006*, Toulouse, France, 2006.

[163] S. Cuendet, D. Hakkani-Tür, and G. Tur, “Model adaptation for sentence segmentation from speech,” in *Proc. IEEE SLT 2006*, Aruba, 2006.

[164] A. Stolcke, B. Chen, H. Franco, V. R. R. Gadde, M. Graciearena, M.-Y. Hwang, K. Kirchhoff, A. Mandal, N. Morgan, X. Lei, T. Ng, M. Ostendorf, K. Sönmez, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng, and Q. Zhu, “Recent innovations in speech-to-text transcription at SRI-ICSI-UW,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, 2006.

[165] P. Brown, V. D. Pietra, P. de Souza, J. Lai, and R. Mercer, “Class-based n-gram models of natural language,” *Computational Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[166] T. Brants, “TnT – A statistical part-of-speech tagger,” in *Proc. ANLP2000*, Seattle, WA, USA, 2000.

[167] J.-L. Gauvain and C.-H. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture obsevation of Markov chains,” *IEEE Transaction on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[168] M. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[169] M. Ostendorf and N. Veilleux, “A hierarchical stochastic model for automatic prediction of prosodic boundary locations,” *Computational Linguistics*, vol. 20, no. 1, pp. 27–54, 1994.

[170] A. Hirst and A. D. Cristo, Eds., *Intonation Systems.* Cambridge University Press, 1998.

[171] R. Kneser, J. Peters, and D. Klakow, “Language model adaptation using dynamic marginals,” in *Eurospeech 97*, Rhodes, Greece, 1997.

[172] K. Seymore and R. Rosenfeld, “Using story topics for language model adaptation,” in *Proc. Eurospeech 97*, Rhodes, Greece, 1997.

[173] R. Gretter and G. Riccardi, “On-line learning of language models with word error probability distributions,” in *Proc. ICASSP’01*, Salt Lake City, UT, USA, 2001.

[174] T. Niesler and D. Willett, “Unsupervised language model adaptation for lecture speech transcription,” in *ICSLP*, Denver, CO, USA, 2002.

[175] H. Nanjo and T. Kawahara, “Unsupervised language model adaptation for lecture speech recognition,” in *Proc. ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, Japan, 2003.

[176] J. R. Bellegarda, “Statistical language model adaptation: Review and perspectives,” *Speech Communication*, vol. 42, pp. 93–108, 2004.

[177] S. Besling and H.-G. Meier, “Language model speaker adaptation,” in *Proc. EUROSPEECH ’95*, Madrid, Spain, 1995.

[178] Y. Akita and T. Kawahara, “Language model adaptation based on PLSA of topics and speakers,” in *Proc. INTERSPEECH 04-ICSLP*, Jeju, Korea, 2004.

[179] G. Tur and A. Stolcke, “Unsupervised language model adaptation for meeting recognition,” in *Proc. IEEE ICASSP*, Honolulu, HI, 2007.

[180] J. Psutka, P. Ircing, J. Hajič, V. Radová, J. V. Psutka, W. J. Byrne, and S. Gustman, “Issues in annotation of the Czech spontaneous corpus in the MALACH project,” in *Proceedings of LREC 2004*, Lisbon, Portugal, 2004.

[181] J. Kolář, J. Švec, and J. Psutka, "Automatic punctuation annotation in Czech broadcast news speech," in *Proc. Intl. Conf. Speech and Computer - SPECOM´2004*, St. Petersburg, Russia, 2004.

[182] J. Kolorenč, "Automatic punctuation of automatically recognized speech," in *Proc. ESSV 2005*, ser. Studientexte zur Sprachkommunikation, vol. 36, Prague, Czech Republic, 2005, pp. 291–297.

[183] A. Pražák, L. Müller, J. V. Psutka, and J. Psutka, "Live TV subtitling: Fast 2-pass LVCSR system for online subtitling," in *Proc. 2nd International Conference on Signal Processing and Multimedia Applications, SIGMAP 2007*, Barcelona, Spain, 2007.

[184] J. Zelinka, J. Kanis, and L. Müller, "Automatic transcription of numerals in inflectional languages," *Lecture Notes in Computer Science*, vol. 3658, pp. 326–333, 2005.

[185] A. Venkataraman and W. Wang, "Techniques for effective vocabulary selection," in *Proc. Eurospeech'03*, Geneva, Switzerland, 2003.

[186] P. Krbec, P. Podveský, and J. Hajič, "Combination of a hidden tag model and a traditional n-gram model: A case study in Czech speech recognition," in *Proc. Eurospeech'03*, Geneva, Switzerland, 2003.

[187] I. Shafran and K. Hall, "Corrective models for speech recognition of inflected languages," in *Proc. EMNLP*, Sydney, Australia, 2006.

[188] J. Švec, F. Jurčíček, and L. Müller, "Input parameterization of the HVS semantic parser," *Lecture Notes in Computer Science*, vol. 4629, pp. 415–422, 2007.

[189] J. Hajič, J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, and M. Mikulová, "Prague Dependency Treebank 2.0," Linguistic Data Consortium, CD-ROM LDC2006T01, Philadelphia, PA, USA, 2006.

[190] R. Sedláček, "Morphemic analyser for Czech," Ph.D. dissertation, Masaryk University, Brno, Czech Republic, 2004.

[191] A. Taylor, M. Marcus, and B. Santorini, *The Penn Treebank: An Overview.* Kluwer, 2003, ch. 1, pp. 5–22.

[192] S. Cuendet, E. Shriberg, B. Favre, J. Fung, and D. Hakkani-Tür, "An analysis of sentence segmentation features for broadcast news, broadcast conversations, and meetings," in *Proc. SIGIR 2007 SSCS workshop*, Amsterdam, The Netherlands, 2007.

[193] D. Hillard, Z. Huang, R. Grishman, D. Hakkani-Tür, M. Harper, M. Ostendorf, and W. Wang, "Impact of automatic comma prediction on POS/name tagging of speech," in *Proc. of IEEE/ACL Workshop on Spoken Language Technology (SLT)*, Aruba, 2006.

[194] B. Favre, R. Grishman, D. Hillard, H. Ji, D. Hakkani-Tür, and M. Ostendorf, "Punctuating speech for information extraction," in *Proc. IEEE ICASSP'08*, Las Vegas, NV, USA, 2008.

List of Publications

Publications in English:

1. KOLÁŘ, J.: “A comparison of language models for dialog act segmentation of meeting transcripts,” In *Lecture Notes in Computer Science (TSD 2008)*, 2008, Vol. 5246, pp. 117–124, ISSN 0302-9743.
2. KOLÁŘ, J.; ŠVEC, J.: “Structural metadata annotation of speech corpora: Comparing broadcast news and broadcast conversations,” In *Proc. of the 6th International Conference on Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco, 2008, ISBN 2-9517408-4-0.
3. KOLÁŘ, J.; LIU, Y.; SHRIBERG, E.: “Speaker adaptation of language models for automatic dialog act segmentation of meetings,” In *Proc. INTERSPEECH 2007*, Antwerp, Belgium, 2007, pp. 1621–1624, ISSN 1990-9772.
4. KOLÁŘ, J.; SHRIBERG, E.; LIU, Y.: “On speaker-specific prosodic models for automatic dialog act segmentation of multi-party meetings,” In *Proc. INTERSPEECH 2006*, Pittsburgh, PA, USA, 2006, pp. 2014–2017, ISSN 1990-9772.
5. KOLÁŘ, J.; SHRIBERG, E.; LIU, Y.: “Using prosody for automatic sentence segmentation of multi-party meetings,” In *Lecture Notes in Computer Science (TSD 2006)*, 2006, Vol. 4188, pp. 629–636, ISSN 0302-9743.
6. KOLÁŘ, J.; ŠVEC, J.; STRASSEL, S.; WALKER, CH.; KOZLÍKOVÁ, D.; PSUTKA, J.: “Czech spontaneous speech corpus with structural metadata,” In *Proc. INTERSPEECH 2005*, Lisbon, Portugal, 2005, pp. 1165–1168, ISSN 1018-4074.
7. STRASSEL, S.; KOLÁŘ, J.; SONG, Z.; BARCLAY, L.; GLENN, M.: “Structural metadata annotation: Moving beyond English,” In *Proc. INTERSPEECH 2005*, Lisbon, Portugal, 2005, pp. 1545–1548, ISSN 1018-4074.
8. KOLÁŘ, J.; ŠVEC, J.; PSUTKA, J.: “Automatic punctuation annotation in Czech broadcast news speech,” In *Proc. Speech and Computer - SPECOM´2004*, Saint-Petersburg, Russia, 2004, pp. 319–325, ISBN 5-7452-0110-X.
9. KOLÁŘ, J.; ROMPORTL, J.; PSUTKA, J.: “The Czech speech and prosody database both for ASR and TTS purposes,” In *Proc. EUROSPEECH 2003*, Geneva, Switzerland, 2003, pp. 1577–1580, ISSN 1018-4074.
10. KOLÁŘ, J.; MÜLLER, L.: “The application of Bayesian Information Criterion in acoustic model refinement,” In *Proc. ECMS 2003*, Liberec, Czech Republic, 2003, pp. 44–48, ISBN 807083708X.

Publications in Czech:

11. KOLÁŘ, J.: “Anotace strukturálních metadat ve spontánní mluvené češtině (Annotation of structural metadata in spontaneous spoken Czech),” In *Proc. Čeština v mluveném korpusu 2007*, Praha, Czech Republic, 2008, ISBN 978-80-7106-982-9 [In Press].

Citations:[1]

The publication No. 3 was cited in:

- H. op den Akker, C. Schulz, "Exploring Features and Classifiers for Dialogue Act Segmentation," in *Lecture Notes in Computer Science (MLMI'08)*, vol. 5237, 2008, pp. 196–207, ISSN 0302-9743.

The publication No. 5 was cited in:

- L. Xie, "Discovering salient prosodic cues and their interactions for automatic story segmentation in Mandarin broadcast news," in *Multimedia Systems*, vol. 14, no. 4, 2008, pp. 237–253, ISSN 0942-4962.

The publication No. 8 was cited in:

- Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, 2006, pp. 1526–1540, ISSN 1558-7916.
- M. Zimmermann, D. Hakkani-Tür, J. Fung, N. Mirghafori, L. Gottlieb, E. Shriberg, and Y. Liu, "The ICSI+ multilingual sentence segmentation system," in *Proc. INTERSPEECH 2006 - ICSLP*, Pittsburgh, PA, USA, 2006, pp. 117–120, ISSN 1990-9772.
- J. Kolorenč, "Automatic punctuation of automatically recognized speech," in *Studientexte zur Sprachkommunikation (ESSV 2005)*, vol. 36, 2005, pp. 291–297, ISSN 0940-6832.

[1] Only citations in which the citing and cited paper do not have any author in common are listed.