Serbo-Croatian Speech Recognition of Broadcast News within a Multilingual Informedia Project

Diploma Thesis

Peter Scheytt

Interactive Systems Laboratories
Carnegie Mellon University, Pittsburgh, USA
University of Karlsruhe, Germany

Supervisors:
Prof. Dr. Alex Waibel
Dipl. Inf. Petra Geutner

Pittsburgh, PA - 31 December 1997

Keywords: Serbo-Croatian, Large Vocabulary Speech Recognition, Broadcast News Recognizer, Dictation Recognizer, Informedia, Multilingual, Out-Of-Vocabulary Rate, Language Normalization.

Ich erkläre hiermit, daß ich vorliegende Arbeit selbständig verfaßt und keine anderen als die angegebenen Hilfsmittel verwendet habe.

Peter Scheytt - Pittsburgh, 31 December 1997.

ABSTRACT

We outline the complete system development process of a Serbo-Croatian speech recognizer for the dictation and broadcast news domain. We describe the data collection process, give an introduction to the Serbo-Croatian language and explain the basic concepts of the Informedia project. Our speech recognition system will be part of the multilingual extension to the existing English Digital Video Library.

We build a dictation system, which serves as a baseline and prototype for the training of the actual broadcast news recognizer. We describe the effects of different normalization and adaptation strategies, examine the effects of multi-corpora language model interpolation, and attack language specific problems due to rapid vocabulary growth and high out-of-vocabulary (OOV) rate. The dictation and broadcast news recognizer are finally unified, and we are able to develop a unique system for both tasks based on a fairly low amount of acoustic training data. We achieve results that are comparable to the 1996 Hub-4 evaluation for English broadcast news: 29.5% WER for broadcast news and 20.9% WER for dictation data.

Future work will concentrate on further language normalization techniques and a better combination of morpheme- and word-based recognition. With more acoustic data becoming available the system performance is also expected to improve. The training of an online recognition system, which can be plugged into the Informedia system to enable spoken queries, is another task for the near future.

ACKNOWLEDGEMENTS

This work was partly supported by the Defense Advanced Research Projects Agency under contract No. N66001-97-D-8502, Delivery Order 0001. The views and findings contained in this material are those of the author and do not necessarily reflect the position or policy of the Government of the United States of America and no official endorsement should be inferred.

Thanks to Radio B92 (Belgrade, Serbia), HRT (Zagreb, Croatia), Radio Free Europe/Radio Liberty (South Slavic Service), RTS (Belgrade, Serbia) and all other organizations that provided data for this project.

My thanks go to Aleksandra Slavković, Ljubomir Cvetković and Boris Tomaz, who did an excellent job transcribing the Serbo-Croatian broadcast news and helped me with language related questions. I also want to thank Manfred Weber and Sandra Yoon, who worked hard to keep the data collection running. We really had a good "bi-continental" group with people working in Karlsruhe and Pittsburgh.

I want to express very special thanks to Michael Finke, without his help and support this thesis would not have been possible. I appreciate his technical advice as well as his driving motivation.

Petra, who spontaneously agreed to be my direct supervisor, did a great job and I am really happy about the results of our common work.

The team with Petra and Michael was a unique experience, a very special cooperation - we had fun as well as success.

Thanks to Alex Waibel for giving me the opportunity to come to Pittsburgh and providing the necessary infrastructure. It was also his idea to work on the Serbo-Croatian speech recognizer, which turned out to be a lucky choice.

Ana had to suffer from my extended work schedule and some very stressful moments. I want to thank her for her love, patience and support.

Last but not least I want to thank my parents. Their support of all kinds helped me to reach the point where I am now.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF PICTURES

TABLE OF TABLES

DEUTSCHE ZUSAMMENFASSUNG

In der folgenden Arbeit beschreiben wir die Entwicklung eines serbokroatischen Spracherkenners für gelesene Zeitungstexte und Nachrichtensendungen. Unsere Bemühungen sind Teil des **Multilingualen Informedia Projektes** an der Carnegie Mellon University, welches die Arbeiten der Informedia-Gruppe (digitale Video-Bibliothek), des Language Technologies Institute (Übersetzung und mehrsprachige Anfragen) und der Interactive Systems Laboratories (Spracherkennung) miteinander verbindet. Das Ziel dieses Projektes ist es, die vorhandene digitale Informedia-Bibliothek, die bisher nur englische Dokumente beinhaltet, auch auf andere Sprachen auszudehnen. Das Informedia-System ist eine Multimedia-Datenbank, die Bilder, Video-, Ton- und Textinformationen speichert; dabei werden Technologien aus den Bereichen der Spracherkennung, des Sprach- und Bildverstehens eingesetzt, um Ton- und Videodokumente zu transkribieren, zu segmentieren und zu indizieren. Dieselben Techniken werden dann verwendet, um intelligente Anfragen durchzuführen, was einen schnellen Zugriff auf bisher unstrukturierte, heterogene Datenquellen erlaubt. Das serbokroatische Modul ist ein erster Schritt hin zu einem multilingualen Informedia-System, das die Bearbeitung von Dokumenten in mehreren Sprachen ermöglicht.

Wie wir gesehen haben, werden bei der Forschung im Informedia-Bereich Methoden der Spracherkennung, des Bildverstehens und der Verarbeitung natürlich-sprachlicher Äußerungen eingesetzt. Wir werden die beiden letzten Punkte in einer kurzen Einführung behandeln, jedoch liegt unser Hauptaugenmerk auf der Entwicklung des serbokroatischen Erkenners unter Verwendung von **JanusRTk** (Janus Recognition Toolkit). Wir werden die einzelnen Schritte vorstellen, die notwendig sind, um einen Spracherkenner mit großem Wortschatz zu bauen: Wir werden dabei sowohl die Datensammlung und -aufbereitung vorstellen als auch die Grundbausteine eines Erkenners bestimmen. Dazu gehören Phonemsatz und -klassen, Aussprachewörterbuch und Sprachmodelle. Schließlich beschreiben wir ausführlich das Training des Erkenners für gelesene Zeitungstexte. Dieses Diktiersystem verwenden wir als Basis für die Entwicklung des eigentlichen Nachrichtenerkenners.

Im weiteren werden wir die Auswirkungen unterschiedlicher Normalisierungs-, Segmentierungs- und Adaptionstechniken vorstellen. Den Nutzen, der von der Verwendung interpolierter Sprachmodelle herrührt, werden wir ebenso untersuchen wie die besonderen Eigenschaften der serbokroatischen Sprache. Sie zeichnet sich durch eine hohe Anzahl von Inflektionen aus, geographisch unterschiedliche, dialektische Varianten spielen ebenfalls eine große Rolle. Das starke Anwachsen des Vokabulars und die damit verbundene hohe Rate von Worten, die vom Erkenner nicht erkannt werden können, sind Themen, die wir zusätzlich untersuchen werden.

Die Kapitel sind wie folgt strukturiert:

Das Multilinguale Informedia Projekt

Wir geben eine kurze Einführung zum Thema "Informedia - Digitale Video-Bibliothek" und "Multilinguales Informedia Projekt". Die Grundkonzepte werden beschrieben und mit möglichen Anwendungs-Szenarien illustriert.

Die serbokroatische Sprache

Dieses Kapitel vermittelt die Grundlagen der serbokroatischen Sprache, seine Geschichte, Aussprache und Morphologie. Wir zeigen Eigenschaften und Zusammenhänge auf, die zum Verständnis der später behandelten Methoden und Techniken sehr nützlich sind.

Die Datensammlung

Wir werden kurz auf die Sammlung und Nachbearbeitung von Audio-Daten für die beiden Bereiche Diktier- und Nachrichtenerkennung eingehen. Außerdem schildern wir die Beschaffung und Aufbereitung großer Mengen von Textdokumenten. Insbesondere werden wir unseren Algorithmus zum Konvertieren von serbokroatischen Sonderzeichen vorstellen: Der größte Teil der auf dem Internet gefundenen Texte enhielt keine diakritischen Zeichen, die wir automatisch einfügten, um nicht existierende Worte in gültige umzuwandeln.

Der Diktier-Erkenner

Wir beginnen mit der Entwicklung eines Systems zur Erkennung gelesener Zeitungstexte. Die akustisch reinen Diktier-Aufnahmen unterscheiden sich in ihrer Qualität recht stark von den Nachrichtensendungen, die neben spontanen Äußerungen auch Hintergrundgeräusche oder Telefongespräche enthalten. Trotz dieser Unterschiede, nutzen wir das Diktier-System, um mit den Eigenheiten der serbokroatischen Sprache vertraut zu werden und die Grundbausteine eines Erkenners zu entwerfen; dazu gehören Phonemsatz, Aussprachewörterbuch und Sprachmodelle.

Der Nachrichten-Erkenner

Auf der Grundlage des Diktier-Erkenners, trainieren wir ein serbokroatisches Spracherkennungssystem für Nachrichtensendungen. Wir werden die Entwicklung eines ersten Erkenners und darauf aufbauend eine Kette sich verbessernder Systeme vorstellen, die unter Verwendung verschiedener Normalisierungs- und Adaptionstechnologien mehr und mehr an die besonderen Probleme des Serbokroatischen angepaßt werden. Wir zeigen, daß unser System trotz einer sehr begrenzten Trainingsmenge von weniger als 30 Stunden Tonaufnahmen in der Lage ist, eine Leistung von 29.5% Wortfehler-Rate auf Nachrichtensendungen und 20.9% Wortfehler-Rate auf gelesenen Zeitungstexten zu erzielen.

1

INTRODUCTION

In this work we describe the development of a large vocabulary Serbo-Croatian speech recognizer for dictation and broadcast news data. Our efforts are part of the **Multilingual Informedia Project** at Carnegie Mellon University, which combines the work of the Informedia group (digital video library), the Language Technologies Institute (translation, cross language retrieval) and the Interactive Systems Laboratories (speech recognition). The goal of this challenging project is to extend the existing Informedia Digital Video Library, which is based on English documents, to other languages. The Informedia system is a multimedia database consisting of images, video, audio and text data. It integrates speech, language and image understanding technology to transcribe, segment and index the linear audio and video documents. The same tools will be applied to accomplish intelligent search and selective retrieval, thus enabling a fast user access to huge amounts of formerly unstructured, heterogeneous data sources. The Serbo-Croatian module is the first step towards a multilingual Informedia system, which enables submission and retrieval of documents in various languages.

As we have seen, research in the Informedia domain has to deal with speech recognition, image understanding and natural language processing. Although we will give a quick overview of the latter two issues, our focus lies on the development of the Serbo-Croatian recognition system using **JanusRTk** (Janus Recognition Toolkit). We will outline in detail all steps necessary to build a large vocabulary speech recognizer: This will cover data collection and preparation, design and identification of the basic parameters, including phone set and classes. We will show the building of a pronunciation dictionary and language models, and finally demonstrate the training of a dictation system, which served as baseline for the actual broadcast news system.

We will report on the effects of different normalization, segmentation and adaptation techniques, and also the benefits we can gain from interpolated language models. The characteristics of Serbo-Croatian are another issue we will talk about: Inflected word forms, geographical language variants, rapid vocabulary growth and thus a very high OOV rate. We will present different techniques to attack these problems.

The chapters in this work are structured as follows:

The Multilingual Informedia Project

We present a short introduction to the Informedia Digital Video Library and the Multilingual Informedia Project. Basic concepts and ideas are described. We also sketch possible application scenarios.

The Serbo-Croatian Language

The information given in this chapter provides the basic concepts of the Serbo-Croatian language, its history, dialects, pronunciation and morphology. It is useful to understand techniques used later in this work.

Data Collection

We will briefly outline the process of collecting acoustic data for both the dictation and broadcast news task. The acquisition and preparation of large amounts of text data is another issue discussed in this chapter. In particular we will present our diacritic conversion algorithm, which was used to modify web-retrieved text documents without diacritical characters so that they contained the correct letters at the right position.

Dictation System

We start by building a dictation speech recognizer, which is trained on a database of read Serbo-Croatian newspaper texts. The acoustically clean dictation data is quite different from the broadcast news recordings, which contain also spontaneous utterances, background noises or telephone speech. Despite these differences, the development of the dictation system helped us to become familiar with the characteristics of the Serbo-Croatian language and determine the basic parameters for the speech recognizer, such as phoneme set, pronunciation dictionary and language models.

Broadcast News System

On the basis of the dictation recognition system, we build a Serbo-Croatian speech recognizer for the broadcast news domain. We describe the training of the baseline recognizer and a chain of improving systems, which are developed subsequently to deal with the specific problems for the Serbo-Croatian broadcast news task. We show that despite the very limited amount of training data a performance of 29.5% WER on broadcast news and 20.9% WER on read newspaper articles can be achieved.

2

THE MULTILINGUAL INFORMEDIA PROJECT

The objective of the Multilingual Informedia Project is to extend the existing English Informedia System to a greater variety of languages, including Serbo-Croatian. The Informedia Digital Video Library [5] is a research initiative at Carnegie Mellon University, which started in 1993. The information presented in this chapter was mainly obtained from Informedia's web site at http://www.informedia.cs.cmu.edu. Refer to this URL for additional documentation.

2.1 The Informedia Digital Video Library

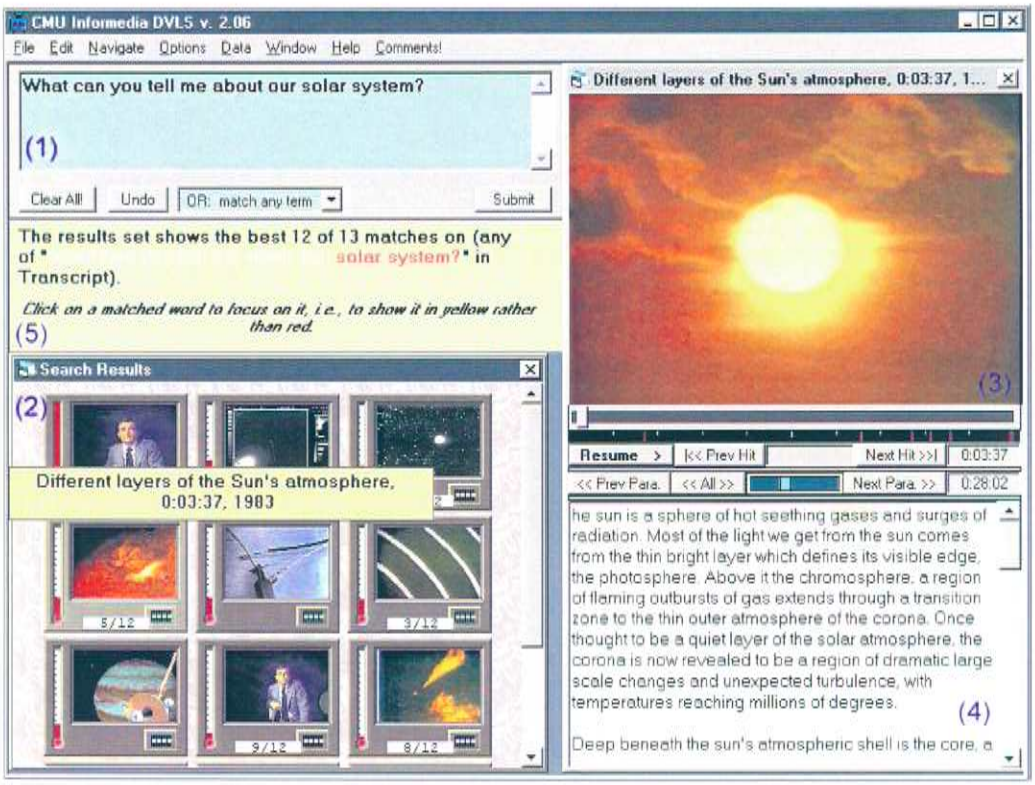
The Informedia Digital Video Library is a multimedia database consisting of digital video, audio, images, text and other related materials. Informedia automatically encodes, segments and indexes the data to enable a full-content and knowledge-based search and retrieval [16]. The automatic preparation of the data requires research in the fields of speech recognition, image understanding and natural language processing.

The growth of the worldwide data superhighway has increased the amount and availability of information. The difficulty lies in the search and retrieval of particular documents. Informedia's goal is to close this gap by providing techniques that enable a more efficient access to many different information sources and types. This might have a strong impact on the way information is delivered in areas such as education, training, sport or entertainment.

One area of particular interest is the news domain. Once a radio or television broadcast has been transmitted, it is very hard to find a specific information. Scanning through all the audio and video data manually would be too much effort, so other solutions have to be found. The Informedia project uses automatic transcription (and also closed caption if available) to extract the content of news transmissions and generate a time alignment between the audio and the text data. Additional video processing helps to further structure the content: Story boundaries are determined by analyzing changes in the acoustic conditions as well as in the video signal. Other algorithms make use of word frequency and statistical analysis to find text pieces, which belong to the same topic. All these different techniques for extracting, structuring and indexing the news data are being applied to overcome the linear access restrictions, to ease and speed up the retrieval of relevant information.

The screenshot in Picture 2-1 shows some features of the Informedia Digital Library: The upper left window (1) contains the query, which either can be spoken into a microphone or typed in directly. The window on the lower left side (2) displays several icons, which refer to different news stories matching the query ordered by their relevancy. By clicking on one of these icons the accompanying video can be viewed (3). It is also possible to read the transcript (4); the actual text position is marked by a sliding gray cursor. The different keywords, which were extracted from the query, can be seen in the window located under the query (5).

The following sections provide a quick overview of the building blocks in the Informedia Digital Library: Image understanding, speech recognition and information retrieval. We further describe the characteristics when adding multilingual features to the existing system. The last paragraph focuses on possible real world applications of a multilingual Digital Library.



Picture 2-1: Screenshot of the Informedia Digital Video Library Demo Application

2.2 Image Understanding

The analysis of the video portion in the Informedia database is primarily used to identify scene breaks and to select static frame icons that are representative of a scene [39].

To answer a user query by showing a half-hour long news show is rarely a reasonable response. Therefore a broadcast has to be segmented into stories with certain topics. Different properties of the images in the video stream are examined to find the segment boundaries. Acoustic (e.g. energy-based silence detection or similarity methods) and content-oriented text analysis techniques [3] complete this visual analysis.

The algorithms used during image processing are based on primitive image features such as color histograms; another approach deals with the interpretation of camera motions. Using **color histogram analysis**, video is segmented into scenes through comparative difference measures. Images with little histogram disparity are considered to be equivalent. By detecting significant changes in the weighted color histogram of successive frames, image sequences can be separated into individual scenes. **Optical flow analysis** is an important method of visual segmentation and description based on interpreting camera motion. By measuring the velocity that individual regions show over time, a motion representation of the scene is created. Drastic changes in this flow indicate random motion, and therefore, new scenes [18].

2.3 Speech Recognition

The Informedia Digital Library uses speech recognition technology for both Library Creation as well as Library Exploration [17].

During **library creation** a speech recognizer is used to generate an automatic transcript of a news show. Accompanied by timing information these text documents form the basis for all further processing. The CMU Sphinx recognition engine has been used to process English news.

Concerning **library exploration** an Informedia query can either be typed in or spoken into a microphone. A spoken query is recognized and then submitted to the database. In general that might not reach the accuracy of using the keyboard, but offers the convenience of a natural interface to the Digital Library. So far only English queries can be handled.

2.4 Information Retrieval

To identify and rank the contents of one news segment, the well-known technique of TF/IDF (term frequency/inverse document frequency) [32] is used to determine critical keywords and their relative importance for the video document.

To calculate the overall frequency of words, it is therefore necessary to collect a large number of texts in a specific language. Considering the overall frequency of the words in a particular document, the relevance of an expression can be computed. This method is applied to the documents in the Digital Video Library as well as to the query. Additionally, very frequent words, also called "stopwords", e.g. articles, pronouns etc., are excluded from the frequency analysis.

2.5 Multilinguality

The extension to multilinguality adds not only diversity of information, but also enables queries that retrieve documents in several languages. By gathering news from various countries in many different languages we can gain a broader view of national and international events. Thus an essential requirement for a multilingual Informedia system is to provide some basic translation functions for both the phrasing of the query and the display of the retrieved documents.

Cross-language queries operate on documents of different languages; the four most common techniques to accomplish this are [37]:

- Dictionary-based
  - Manually constructed dictionaries and thesauri
  - Machine-readable dictionaries
- Corpus-based
  - Aligned parallel corpora
  - Latent Semantic Indexing (LSI) [7]
  - Similarity thesauri
- Machine translation
- Combinations
  - Dictionary and corpora
  - Corpora and machine translation

After entering a query in one language and extracting the relevant keywords, we can apply the strategies described above to search through multiple documents in different languages at the same time without the need to reformulate the original query.

One can imagine three different usage scenarios:

1. The cross-language or translingual feature is useful for people that speak various languages, because they do not have to reenter the query to obtain multilingual documents.
2. Users that have a passive knowledge of a language also will have a huge benefit from the enhanced query functionality. A person, who understands a language, but is not able to form a query himself, now can accomplish this by asking a question in his mother tongue and have the Informedia system take care of the translation.
3. For people who want to obtain documents in a language they are not familiar with at all, the Multilingual Digital Library additionally provides translated subtitles that accompany the transcript of the original language.

The graph in Figure 2-1 illustrates the concepts described above and displays the basic units of a Multilingual Informedia System and the relationships between them.

Note: In this work we mainly concentrate on the development of a Serbo-Croatian speech recognizer which creates an automatic transcription and time alignment of broadcast news. We trained, however, a first prototype of an online speech recognizer, which can handle voice queries to the Informedia system.

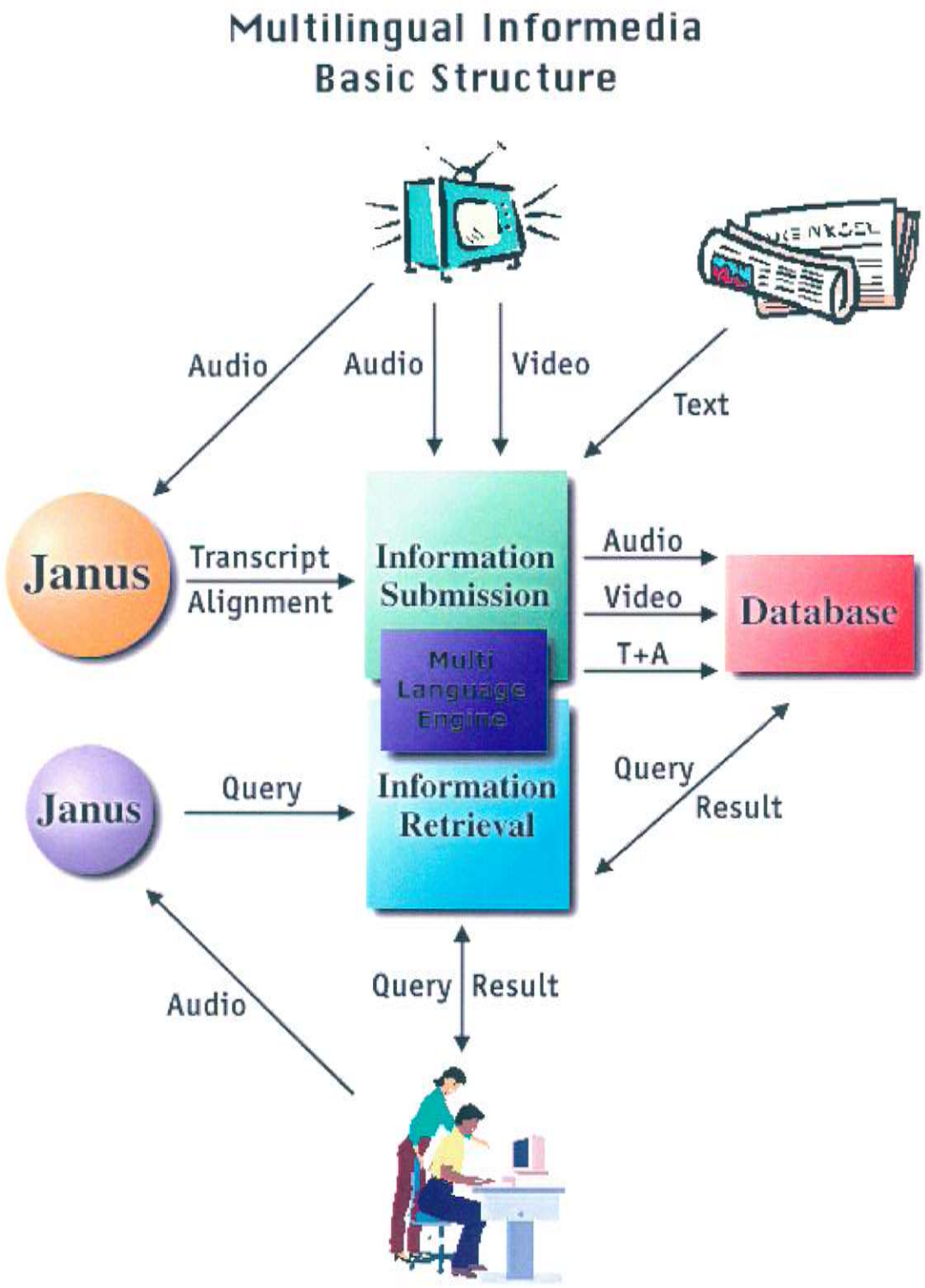


Figure 2-1: Basic Structure of the Multilingual Informedia Project

2.6 Application Scenarios for Multilingual Informedia

Although the Informedia Digital Video Library Project and its multilingual extension are still undergoing research, we would like to take a look at the various possibilities of its deployment in education, teaching, entertainment or news delivery. The following scenarios outline different possible applications of an Informedia database in general with the positive effects of multilingual features being obvious.

Online Applications

Using data from the Internet, radio or television stations, this type of application provides access to the latest available information. A single user might monitor web broadcasts (a market with an immense potential in the future benefiting from increasing net capacities and improving streaming technology), digitally record radio and television shows, which he receives terrestrially or by satellite. The Informedia software could then be used to process the data on his personal computer and then store it in his private archive. Daily news briefings, weekly summaries or customized programs are possible retrieval scenarios.

The data collection necessary to build a Digital Video Library requires financial and other resources that probably suggest another approach. Instead of having many isolated private archives the idea of **information centers** could be an alternative: **News kiosks** record data from many different sources and countries. Users access these central data management systems over dedicated connections, the Internet or telephone lines. Queries could be text- or speech-based. In some cases only the data necessary for search operations needs to be stored in a central database, which would contain a pointer to the place where the actual information is located. The deployment of such information providers depends heavily on the further development of fast data connections.

Offline Applications

Less time critical data might be stored on CD-ROMs or DVDs and sent out to possible users. Foreign language classes are another application that could benefit from multilingual databases. High School and University courses would be available to a much larger number of students in many different places and various languages. Multilingual Informedia could provide a much easier and sophisticated access to information retrieved from discussions in the United Nations or the European Parliament, which host speakers of many different languages.

3

THE SERBO-CROATIAN LANGUAGE

This chapter will present a short introduction to the Serbo-Croatian language: We will look at its historical foundation, dialects and pronunciation. The alphabet, writing system and the spelling are other issues discussed over the next few pages. We will end with a paragraph on Serbo-Croatian morphology. Please refer to [6][27][22] for further reading; the information in the following sections was mainly compiled from these monographs.

3.1 Historical Background

The line, which divided Europe into east and west, Orthodox and Catholic, runs right through the part of the Balkans where Serbo-Croatian is spoken. Various states have prospered at different times in this region.

The ancestors of the South Slavs arrived in the Balkans during the sixth and seventh century and within the next two centuries the first Slav states of the area were founded. There were two main sets of dialects: East South Slavonic would later develop into Bulgarian and Macedonian, while West South Slavonic was the basis for Slovene and Serbo-Croatian.

Christianity was accepted in the ninth century, with certain political repercussions. The tenth-century Croatian kingdom looked to Rome in matters of religion. Serbia's adoption of Orthodoxy meant that it looked first to Constantinople and later to Moscow for support. The picture was complicated by the invasion of the Turks who occupied the Kosovo, Bosnia and Herzegovina. At the end of the nineteenth century Croatia was part of the Austro-Hungarian Empire, which also took over Bosnia and Herzegovina. It was not until 1918 that the different groups were united into one state.

3.2 The Language: Dialects and Varieties

Three main dialect groups had emerged, which take their names from the interrogative pronoun 'what?': **Čakavian** ('ča?'), **Kajkavian** ('kaj?') and **Štokavian** ('što?'). Kajkavian was spoken in the north, Čakavian in the west and Štokavian in the east, center and southwest. The dialectal fragmentation strengthened by the Turkish influence impeded the development of a common literary language. It was **Vuk Karadžić** (1787-1864) who proposed a literary language based on a single dialect, the Štokavian dialect of East Herzegovina. This led to the Literary Accord which was signed between Serbs and Croats (Vienna, 1850). It justified the use of Štokavian as the literary language and gave rules for writing it.

There are however three dialectal varieties of the Štokavian – **Ekavian** in Serbia, **Ijekavian** in Montenegro, Bosnia-Herzegovina and parts of Croatian and **Ikavian** in parts of Croatia and Herzegovina. The official Serbo-Croatian language is therefore spoken and written in the Ekavian or Ijekavian variants. The Ijekavian and Ekavian differ in the usage of the sound combination '**ije**' or '**je**' and the sound '**i**' in Ijekavian, which is opposed by the sound '**e**' in Ekavian; see Table 3-1 for examples.

Ijekavian	Ekavian	English
rijeka, rjeka	reka	river
lijepo, ljepo	lepo	beautiful
riječnik, rječnik	rečnik	dictionary
vijetar, vjetar	vetar	wind
htio	hteo	he wanted to

Table 3-1: Ijekavian and Ekavian Variants of Standard Serbo-Croatian

3.3 The Writing System

The original alphabet was Glagolitic. In the eastern, Orthodox area, Cyrillic replaced it from the twelfth century on. In the west, the Latin alphabet was introduced in the fourteenth century, under Catholic influence. From the sixteenth century until the Second World War, some Moslem writers in Bosnia used the Arabic script.

In 1818, Vuk Karadžić justified and used a new version of Cyrillic. This was a major reform involving simplifying the alphabet, using a single letter per sound and adopting a phonemically based orthography ("**Speak as you write, and write as you speak.**"). The equivalent reform for the Latin alphabet was carried out a little later, using diacritic symbols like in the Czech model. The two modern alphabets are given in Table 3-2.

Latin	Cyrillic	Latin	Cyrillic
A a	А а	L l	Л л
B b	Б б	Lj lj	Љ љ
C c	Ц ц	M m	М м
Ć ć	Ћ ћ	N n	Н н
Č č	Ч ч	Nj nj	Њ њ
D d	Д д	O o	О о
Dž dž	Џ џ	P p	П п
Đ đ	Ђ ђ	R r	Р р
E e	Е е	S s	С с
F f	Ф ф	Š š	Ш ш
G g	Г г	T t	Т т
H h	Х х	U u	У у
I I	И и	V v	В в
J j	Ј ј	Z z	З з
K k	К к	Ž ž	Ж ж

Table 3-2: Modern Serbo-Croatian Latin and Cyrillic Alphabets

3.4 Phonology

Serbo-Croatian has one of the smallest phoneme inventories in the Slavonic family. It does not have the range of palatalized consonants found e.g. in Russian. We can identify 25 consonants

(of these r, which is trilled, can be syllabic) and five vowels. Please see also [15] for more information on this topic.

Consonants

The following table of the Serbo-Croatian consonants is taken from [15][22]. In section 5.4 we will use a larger set of acoustic and articulatory features to determine our set of phone classes for the development of a context dependent speech recognizer. For a fairly complete discussion of Serbo-Croatian phonology see [15][22].

	bilabial	labio dental	dental	alveolar	palato alveolar	palatal	velar
Plain Stop voiceless voiced	 p b		 t d				 k g
Affricate voiceless voiced			 c		 č dž	 ć đ	
Fricative voiceless voiced		 f v		 s z	 š ž		 x
Nasal	m		n			nj	
Lateral			l			lj	
Trill				r			
Semi-Vowel						j	

Table 3-3: Articulation of Serbo-Croatian Consonants

Vowels

The following synopsis of Serbo-Croatian vowels is also taken from [15][22].

Tongue	front	center	back
high	I		u
middle	e		o
low		a	

Table 3-4: Articulation of Serbo-Croatian Vowels

Accentuation

Serbo-Croatian is a very melodic language and therefore clearly identifies different accents based on tone, length and pitch. The representation of stress markers in Table 3-5 is taken from [15].

	long	short
falling tone stressed syllables rising tone	^ ‘	`` `
unstressed syllables	-	

Table 3-5: Serbo-Croatian Accentuation

Pronunciation Examples

The synopsis of pronunciation examples, given in Table 3-6, is thought to create a general feeling for the sound of the Serbo-Croatian language [1][15][22].

Phone	Croatian	German	English	Spanish
A	ona, rad	hatte, natürlich	father	mano
B	dobar, Zagreb	Bein, heben	bag	
C	centimetar, utakmica	zahm, sitzen	rats	
Ć	ćevapčići, kući	Mädchen, Flittchen	children (towards: tune)	
Č	četiri, uči	Matsch, rutschen	church	
D	dan, sedam	dumm, Adel	dog	
Đ	đuvec, mlađa		jar (towards: duke)	
DŽ	džezvica, udžbenik		jar	
E	gleda, evo	fett, Männer	bed	
F	fabrika, kafa	Vater, Affe	fun	
G	govori, digao	Gift, Segen	get	
H	Ahmet, novih	ach	Loch Ness	jugar, mujer
I	pita, bilo	sieben, nie	police	si
J	jedan, rijeka	Katzenjammer, Jahrestag	yes, boy	
K	ko, istok	Kamm, Socke	ski	caro
L	lampa, bilo	Lampe	like, let	la
LJ	ljudi, nedjelja		million	llegar, amarillo
M	malo, znam	Murmel, Dame	meet	
N	narodni, stanica	nackt, wenig	note	
NJ	Njemačka, pičinje		onion	mañana, España
O	korzo, odgovara	Post, offen	port	bola
P	petak, opet	pumpen, Post	spy	palma
R	zar, odgovor, prvi			regular, perro
S	sada, silazi	reißen, Riß, Slalom	six	
Š	šeta, sta	schief, pfuschen	ship	
T	ti, utorak	toll, Platte	stop	tu
U	uči, u fabriku	klug, Schuh	boot	tu
V	voda, provesti	Vase, wo	very	
Z	zove, jezik	sagen, Eisen	zero	
Ž	možda, muž	Journalist, Genie	measure	

Table 3-6: Serbo-Croatian Pronunciation with some German, English and Spanish Examples

3.5 Morphology

This paragraph gives a quick overview on Serbo-Croatian morphology [1][6][21]. It is intended to create a general understanding for the techniques that will be used later in this work: Morpheme-oriented recognition, class-based language models and research on stemming-related OOV rate reduction that is based on this work.

Serbo-Croatian has been generally conservative, maintaining most of the categories of Common Slavonic forms. Seven cases have been preserved, together with three genders, which are distinguished in the plural as well as the singular.

Nouns

The chart given here shows the main types of noun declension.

Feminine a-stem	Singular	Plural	English
Nominative	žena	žene	woman
Vocative	ženo	žene	woman(!)
Accusative	ženu	žene	(I saw the) woman
Genitive	žene	žena	(the) woman's (book)
Dative	ženi	ženama	(I gave it to the) woman
Instrumentalis	ženom	ženama	(with the) woman
Locative	ženi	ženama	(I went to the) woman

Masculine o-stem	Singular	Plural	English
Nominative	zakon	zakoni	law
Vocative	zakone	zakoni	
Accusative	zakon	zakone	
Genitive	zakona	zakona	
Dative	zakonu	zakonima	
Instrumentalis	zakonom	zakonima	
Locative	zakonu	zakonima	

Neuter o-stem	Singular	Plural	English
Nominative	selo	sela	village
Vocative	selo	sela	
Accusative	selo	sela	
Genitive	sela	sela	
Dative	selu	selima	
Instrumentalis	selom	selima	
Locative	selu	selima	

Neuter i-stem	Singular	Plural	English
Nominative	stvar	stvari	thing
Vocative	stvari	stvari	
Accusative	stvar	stvari	
Genitive	stvari	stvari	
Dative	stvari	stvarima	
Instrumentalis	stvarju/stvari	stvarima	
Locative	stvari	stvarima	

Table 3-7: Feminine, Masculine and Neuter Declensions of Serbo-Croatian Nouns

Adjectives

There are two different forms of adjectives:

- Definite: dobri čovek ('the good man')
- Indefinite: dobar čovek ('a good man')

They distinguish by inflection only in the masculine singular form.

If used pronominally, the adjective corresponds in case, number and gender with the noun. For animate nouns the accusative form of the masculine singular adjective is identical to the genitive, for inanimate nouns, however, nominative and accusative are the same. The examples in Table 3-8 refer to the definite variant, with optional forms given in brackets.

Masculine	Singular	Plural	English
Nominative	mladi	mladi	young
Vocative	mladi	mladi	
Accusative	as nom. or gen.	mlade	
Genitive	mladog(a)	mladih	
Dative	mladom(e)	mladim(a)	
Instrumentalis	mladim	mladim(a)	
Locative	mladom(e)	mladim(a)	

Neuter	Singular	Plural	English
Nominative	mlado	mlada	young
Vocative	mlado	mlada	
Accusative	mlado	mlada	
Genitive	mladog(a)	mladih	
Dative	mladom(e)	mladim(a)	
Instrumentalis	mladim	mladim(a)	
Locative	mladom(e)	mladim(a)	

Feminine	Singular	Plural	English
Nominative	mlada	mlade	young
Vocative	mlada	mlade	
Accusative	mladu	mlade	
Genitive	mlade	mladih	
Dative	mladoj	mladim(a)	
Instrumentalis	mladom	mladim(a)	
Locative	mladoj	mladim(a)	

Table 3-8: Masculine, Neuter and Feminine Declension of Serbo-Croatian Definite Adjective

Verbs

We will close this paragraph with a chart showing the three main conjugations of Serbo-Croatian verbs.

I Conjugation		II Conjugation		III Conjugation	
imati	to have	govoriti	to speak	tresti	to shake
imam	I have	govorim	I speak	tresem	I shake
imaš	you have	govoriš	you speak	treseš	you shake
ima	he has	govori	he speaks	trese	he shakes
imamo	we have	govorimo	we speak	tresemo	we shake
imate	you have	govorite	you speak	tresete	you shake
imaju	they have	govore	they speak	tresu	they shake

Table 3-9: Main Conjugations of Serbo-Croatian Verbs (Simple Present)

4

DATA COLLECTION

The first step in developing a speech recognizer is to collect the necessary data for acoustic training and language modeling: Speech samples have to be recorded and transcribed, a large amount of text documents must be acquired and prepared. We will shortly outline the data collection for the dictation system, in the remainder of this chapter we describe in detail the creation of the broadcast news database of audio and text documents.

4.1 Dictation Data

Speech Recording

The audio data for the dictation system was collected in Croatia and Bosnia-Herzegovina. Native speakers were asked to read 20 minutes of news texts, extracted from the HRT (Croatian Radio and Television) web site and Obzor Nacional, a Croatian newspaper. The speech was digitally recorded using a portable Sony DAT-recorder TDC-8 and a close-speaking Sennheiser microphone HD-440-6 at a sampling rate of 48 kHz in stereo quality and further sampled down to 16 kHz with 16-bit resolution in mono quality [35].

Transcription

The read utterances were checked against the original text to eliminate major errors and mark spontaneous effects, like hesitation, breathing and some other human and non-human noises. This data was originally collected for the GlobalPhone project at Karlsruhe University [35]. Refer to http://werner.ira.uka.de for further information on the GlobalPhone project and its goals.

Database

The final dictation database is shown in Table 4-1. At the beginning of the system development process, there was less training material available, as data preparation still was going on. The data was further partitioned into a training set of 77 speakers, a development test set and an evaluation test set, both consisting of 4 speakers. There were 50 female and 35 male speakers in our acoustic database for the dictation task.

Speakers	Articles	Recording Length	Words	Vocabulary
85	131	18 h	89 K	17 K

Table 4-1: Dictation Database of Acoustic Recordings

4.2 Broadcast News Data

Data Recording

The broadcast news data was collected at Karlsruhe University in Germany. A satellite dish and a dedicated PC, equipped with an Optibase MPEG encoder board, were installed to record the HRT evening news show, which is transmitted from Croatia via the Eutelsat satellite. Later we also recorded the evening news delivered by the Serbian channel RTS (Radio and Television of Serbia), which is located in Belgrade, Serbia. The television signal was digitally recorded in MPEG format (target bit rate: 1.008 Mbit/s, audio bit rate: 0.192 Mbit/s, sampling rate: 44.1 kHz). For speech recognition the audio signal was uncompressed and sampled down to 16 kHz with 16-bit resolution. The graph in Figure 4-1 illustrates the hardware setup for data collection.

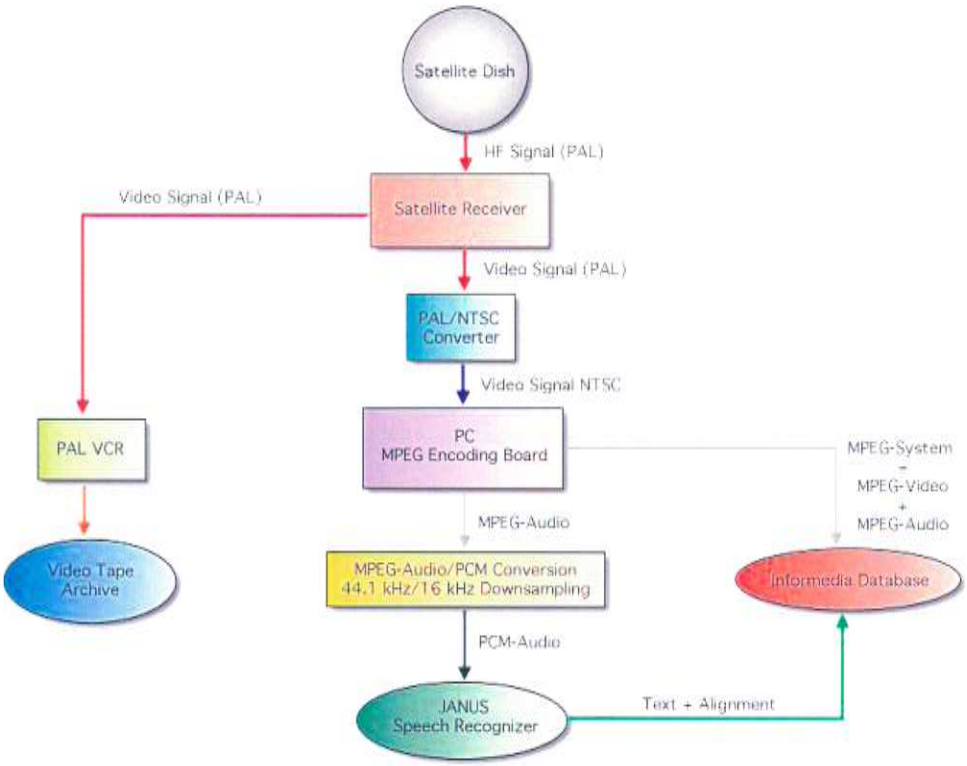


Figure 4-1: Hardware Setup for Broadcast News Data Collection

Transcription

Three native speakers transcribed the news broadcasts, using a software tool developed specifically for this task. Similar to the **focus conditions** in the HUB4 corpus for English broadcast news data [11], the Serbo-Croatian recordings were divided into segments, in which

the acoustic conditions remained constant. These segments were tagged according to the speaker's characteristics, channel quality and background noises.

The different tags in these three categories are shown in Table 4-2, where "Non-Serbo-Croatian" identifies a person speaking in another language than Serbo-Croatian, most often English.

Speaker	Channel	Noise
Male	Clean	Music
Female	Telephone	Second Speaker
Non-Serbo-Croatian	Distorted	Conference
Unknown	Unknown	Street
None		Static Noise
		Other
		None

Table 4-2: Acoustic Tags

In a later stage of the system development we gave up the time consuming tagging of the segments. Experiments showed that the benefit from tagged segments in terms of system performance did not justify the costly efforts employed during data preparation (see paragraph 6.3). In addition to acoustic tags, only the most frequent and clearly audible spontaneous effects were transcribed:

- Hesitation
- Breathing
- Generic human noise
- Generic non-human noise
- Garbage (false starts, completely unintelligible utterances, etc.)

The diacritical letters in Serbo-Croatian were transcribed applying the rules in Table 4-3

Diacritic	Ć ć	Č č	Đ đ	Š š	Ž ž
ASCII	C1 c1	C5 c5	D1 d1	S5 s5	Z5 z5

Table 4-3: Serbo-Croatian Diacritics

Transcription time varied between 13 and 18 hours per news broadcast, which lasts approximately 40 minutes. This is fairly long and there are several reasons for that:

- No close caption or teletext was available
- Speakers speak very fast
- Many noisy segments, which take longer to transcribe
- Acoustic labeling of the segments consumes a lot of time

In addition to the television broadcasts, we downloaded some radio news from the Radio Free Europe/Radio Liberty web site (http://www.rferl.org) in Realaudio format and converted them to 16 kHz, 16 bit Wave format for speech processing.

Database

Table 4-4 contains the information on the final broadcast news database of audio recordings. Starting with few data at the beginning of the system development, the completed transcriptions were added subsequently to the training set, thus enabling a more accurate training of the acoustic models.

Source	Broadcasts	Recording Length	Words	Vocabulary
HRT (MPEG)	23	15 h	118K	24 K
RFE/RL (RA)	7	0.5 h	7 K	2.5 K
Total	30	15.5 h	125 K	25 K

Table 4-4: Broadcast News System Database

4.3 Text Acquisition and Preparation

Text Acquisition

For language modeling we searched the Internet for news texts in Serbo-Croatian. There were few sites at the beginning of the project, but the amount of available data later increased significantly. We retrieved text data from 20 different sources in Serbia, Croatia and Bosnia-Herzegovina:

- Television stations
- Radio stations
- Newspapers
- News agencies

Text Preparation

During text preparation, we encountered one major problem: Many sites simply map diacritics onto their corresponding non-diacritical letter:

- ć and č become c
- đ becomes d
- š is replaced with s
- ž is replaced with z

This fact is no problem for native speakers, but for the purpose of language modeling, however, we have to convert the non-diacritical characters into diacritics at the right position.

Diacritic Conversion Algorithm

This task was accomplished by collecting as many Serbo-Croatian texts with diacritics as were available, and using them to create a wordlist L_{D2}. The union of wordlist L_{D1}, extracted from news texts with diacritics, and L_{D2} resulted in a list, L_C, of correct words. This wordlist was used to convert the second list, L_F, of both correct and incorrect word forms, which were extracted from the texts without special characters. The goal of the conversion algorithm was to subsequently move all words from L_F to L_C. Applying different rules, we either verified the correctness of the words in L_F or converted them, so that in both cases they could be inserted into L_C.

First, all words that occurred in L_F and L_C were labeled correct and moved from L_F to L_C. For some words this might be wrong in certain cases; depending on the context, 'grada' converts to 'grada' (grad, gen. sg., town) or 'građa' (građa, nom. sg., material). Therefore a trigram model was used to improve the conversion accuracy. As next step, all words in L_F that did not contain the letters c, d, s and z were moved to L_C. Then all remaining words in L_F were assigned to the their nearest neighbor in L_C. When the Levenstein editing distance between the two candidates did not exceed a certain threshold, the word pair was thought to be valid and a conversion, if necessary, was performed (Table 4-5). The thus converted and verified word forms were moved from L_F to L_C.

No Diacritics	Diacritics	Conversion
afirmisanog	afirmisanoj	None
africki	afrički	c → č

Table 4-5: Conversion Pairs

When applying this operation to a separate test text, 2% of the words among the qualifying pairs were not converted correctly. In the last step of the text conversion algorithm, we generated a letter trigram model. This model was used to score the likelihood of the different possible character sequences (switching the potential diacritic candidates c, d, s and z) for the remaining words in L_F. The sequence with the highest score was chosen and inserted into L_C. In the test text, 25% of the words were converted incorrectly using this mechanism, which is still better than just leaving the words as they are, which would lead to an error rate of 70%. Thus the combined conversion error rate of the whole algorithm on the test text was 5%. Refer to Figure 4-2 for a visualization of the algorithm described in the preceding paragraph.

Database

By converting the texts without diacritics we doubled the amount of text data available for language modeling. Refer to Table 4-6 for details.

Character Set	Web Sites	Words	Vocabulary
Diacritics	7	5 M	236 K
No Diacritics	13	6 M	216 K
Total	**20**	**11 M**	**353 K**

Table 4-6: Internet Text Database

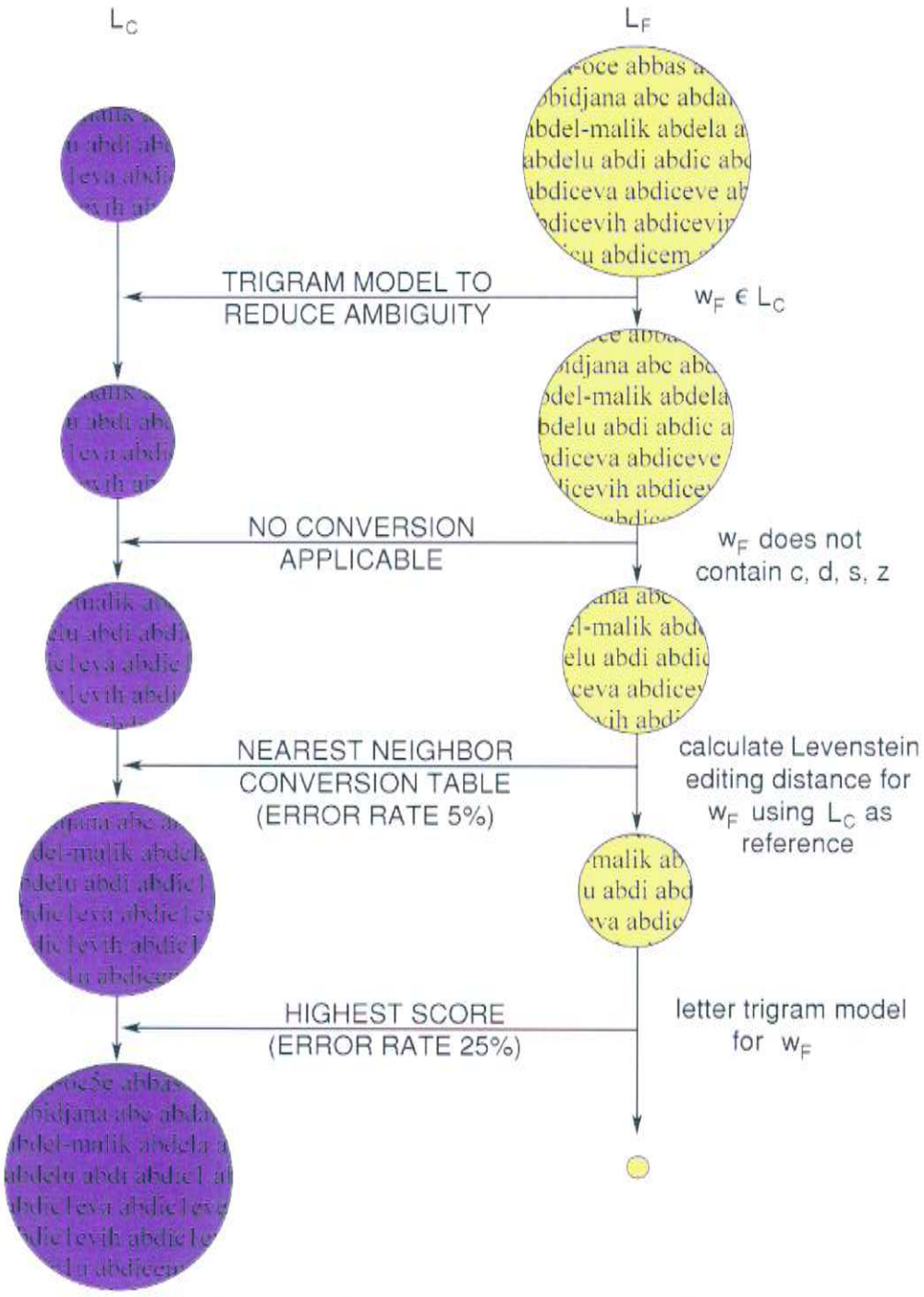


Figure 4-2: Diacritics Conversion Algorithm

5

THE DICTATION SYSTEM

In the following chapter, we will describe the development of the Serbo-Croatian dictation speech recognizer using JanusRTk [9]. We will focus on the design of the fundamental units, phone set and pronunciation dictionary, followed by an outline of the initialization procedure based on a German recognizer for conversational speech. After the description of the first context independent system, we will concentrate on developing a context dependent system, including identification of a set of phone classes and polyphonic modeling. We will report on the effects of different normalization and adaptation techniques as well as the benefit we can obtain from using multiple corpora language model interpolation during system evaluation.

The work on the dictation system helped us to become familiar with the Serbo-Croatian language and its particular difficulties when building a speech recognizer. We were able to apply this knowledge during the development of the broadcast news system, which was trained in a rather short time.

5.1 Initialization

Phone Set

As Serbo-Croatian has a phonemically oriented writing system (see section 3.4), the identification of the phone set is straightforward: Every letter corresponds to a phone. This relation can be observed in Table 5-1. The speech recognizer therefore is based on 30 language phones. For the representation of other acoustic events, we augmented our phone set by four noise models and one for silence (Table 5-2).

Letter	A	B	C	Ć	Č	D	Đ	DŽ	E	F	G	H	I
Phone	A	B	C	C1	C5	D	D1	DZ5	E	F	G	H	I

Letter	J	K	L	LJ	M	N	NJ	O	P	R	S	Š	T	U	V	Z	Ž
Phone	J	K	L	LJ	M	N	NJ	O	P	R	S	S5	T	U	V	Z	Z5

Table 5-1: Serbo-Croatian Letters and Corresponding Phones

Event	Garbage	Human Noise	Breathing	Other Noise	Silence
Phone	+QK	+hGH	+hBR	+nGN	SIL

Table 5-2: Noise and Silence Events with Corresponding Phones

Each phone is represented by a 6-state, left-to-right Hidden Markov Model (HMM), with a set of two states having the same parameters. A state is modeled by a mixture of 16 Gaussians with a diagonal covariance matrix. By doubling the number of states, we impose a duration

constraint without having to estimate a larger number of parameters from the same amount of data, which would be the case, if we had six different states. We preserved this basic structure throughout all phases of the dictation system. For the broadcast news recognizer, however, we reduced the number of states to three. The reason was that the speakers in the news shows spoke faster, and the duration constraint therefore seemed inappropriate.

Pronunciation Dictionary

For most Serbo-Croatian words the pronunciation can be derived from the spelling. And even the digraphs **lj, nj,** and **dž** cause little problem: The combination **l + j** does not occur, while **n + j** and **d + ž** are rare; an example is nadživeti 'to outlive', where d + ž represent a separate sound. A rather simple grapheme-to-phoneme tool was sufficient to create the pronunciation dictionary. Manual adjustments were, however, necessary for names, abbreviations, foreign words, numbers and pronunciation variants.

5.2 Bootstrapping

Preprocessing

The preprocessing for the dictation system starts with the extraction of a MFCC-based (Mel-scale Frequency Cepstral Coefficients) feature vector for every 10 ms. The window size is 20 ms. To obtain the final feature vector we apply a truncated LDA (Linear Discriminant Analyses) transformation to the concatenation of MFCCs, their first and second order derivatives, the short-term power coefficient including its first and second order derivative, and the zero crossing value. The LDA thus reduces the dimensionality of the feature vector from 43 to 32. Cepstral mean subtraction was performed to attenuate dissimilarities between different speakers and various channel conditions.

First Labels

When building a recognizer for a new task and/or language, we could start with randomly initialized parameters, i.e. codebook vectors and distribution weights. An alternative approach is to take an existing recognizer, if available. The parameters of that system, although trained possibly for another language and/or task, can then be used to replace the random values, mentioned above, which usually helps to speed up the system development process.

With JanusRTk we perform this type of initialization by writing a set of labels. Labels are the alignment between the audio recording and its transcription, and therefore provide information about corresponding symbols in the textual representation and acoustic events in an utterance. These labels are essential for the training of the new models, using the Kmeans and the Viterbi algorithm.

The labels for our first context independent Serbo-Croatian dictation system were generated by a recognizer for a German scheduling task [10]. Therefore we initialized each Serbo-Croatian phone with its closest German equivalent (Table 5-3). Additionally, we applied the technique of **'label boosting'**, which - prior to writing labels - performs an MLLR (Maximum Likelihood Linear Regression) adaptation to adjust the parameters to a particular speaker or acoustic condition.

Phone	A	B	C	C1	C5	D	D1	DZ5	E	F	G	H	I
GER	A	B	TS	TSCH	TSCH	D	TSCH	TSCH	E	F	G	X	I

Phone	J	K	L	LJ	M	N	NJ	O	P	R	S	S5	T	U	V	Z	Z5
GER	J	K	L	L	M	N	N	O	P	R	S	SCH	T	U	V	Z	SCH

Table 5-3: Serbo-Croatian and Corresponding German Phones

When recalling the pronunciation examples presented earlier (see paragraph 3.4), we might conclude that there are better choices to initialize the Serbo-Croatian phones. A Spanish recognizer could provide more accurate models for 'LJ', 'NJ' and R; an English one probably would do better for the different 'SCH' (S5 and Z5) and 'TSCH' (C1, C5, D1, DZ5) sounds. Although we were aware of these drawbacks, we thought the German recognizer to be a good overall choice. We should also keep in mind that the first labels lose their impact with the following systems being based on the alignment that was generated by the preceding one. New labels written by an improved recognizer allow the training of better models in the next iteration.

5.3 Context Independent Recognizer

Starting with the labels written by the German system, we developed a series of two context independent Serbo-Croatian recognizers. After the first iteration we calculated a new alignment between audio and transcription, and trained the second system upon these new labels. The available training data was limited to 5.5 hours. Both systems were evaluated on the same set of test utterances, which consisted of equal parts from four different speakers. Refer to Table 5-4 for further details on the test database. A summary of the results is shown in Table 5-5.

Speakers	Utterances	Length	Words	Vocabulary
4	98	21 min	2895	1291

Table 5-4: Test Database for Dictation Systems

System	Data	Labels	Vocabulary	OOV	LM	WER
D-CI-0	5.5 h	GSST	10 K	13.6%	D-LM-0	52.7%
D-CI-1	5.5 h	D-CI-0	10 K	13.6%	D-LM-0	50.4%

Table 5-5: Results for Context Independent Systems

The vocabulary for training and testing was identical, as there were no Serbo-Croatian texts for frequency analyses available at that time. Both recognizers (D-CI-0 and D-CI-1) used a single trigram language model based on the training transcriptions. The aspects of language modeling will be discussed in detail throughout paragraph 5.5.

5.4 Context Dependent Recognizer

The next goal was to build a context dependent system, which is based on phonetically tied models. This approach differentiates between the contexts in which a central phone occurs. Depending on the surrounding phones a different model is computed. Normally a context width of 1 (**Triphones**) or 2 (**Quinphones**) to the left and right is chosen. The wider the context, the more phone combinations are possible. For some of them, however, there might

not be enough training data to estimate the model parameters accurately as some combinations were not at all seen during training. The solution therefore is to have some allophonic[1] models with similar acoustic properties share the same parameters. This improves not only the robustness of our models, but also generalizes the system, for it is now possible to map unseen phone combinations to their corresponding class-based models. The phone classes presented in the next paragraph represent the sets of similar phones and were used to train a context dependent system with JanusRTk.

Phone Classes

The set of acoustic and articulatory phone classes shown in Table 5-6 guided the training of phonetically tied Gaussian mixtures on the base of quinphones. The categories were compiled from previous German and English recognizers and particularly from information found in [1] [6] [15] [22]. (There are some redundant classes, e.g. 'VIBRANT' and 'SYLLABIC', which are listed for reasons of completeness.) For more information see also paragraph 3.4 on phonology.

NOISES	+QK +hGH +hBR +nGN
HUMAN-NOISES	+hGH +hBR
SILENCES	SIL
CONSONANT	B C C1 C5 D D1 DZ5 F G H J K L LJ M N NJ P R S S5 T V Z Z5
OBSTRUENT	B D G H K P T F S S5 V Z Z5 C C1 C5 D1 DZ5
SONORANT	M N NJ L LJ J R V
ACUTE	T D S Z C1 D1 S5 Z5 NJ LJ J R C C5 DZ5
CONTINUOUS	F S Z S5 Z5 H M N NJ L LJ V J
VOWEL	A E I O U
VOICED	B D D1 DZ5 G J L LJ M N NJ R V Z Z5
UNVOICED	C C1 C5 F K P S S5 T H
COMPACT	C1 D1 S5 Z5 K G H J
DIFFUSE	P B F M V
LATERAL	L LJ
NASAL	M N NJ
STOP	B D G K P T
FRICATIVE	F S S5 V H Z Z5
AFFRICATE	C C1 C5 D1 DZ5
LABIAL	B M P F V
BILABIAL	B M P
LABIO-DENTAL	F V
APICO-DENTAL	T D C S Z N L R
DENTAL	D T C N L
ALVEOLAR	S Z R
PALATO-ALVEOLAR	C5 DZ5 S5 Z5
PALATAL	C1 D1 J LJ NJ
VELAR	G K H
STRIDENT-STRONG	C C1 C5 T

[1] Multiple phone sequences

STRIDENT-WEAK	D D1 DZ5
STOP-VOICED	B D G
STOP-UNVOICED	P T K
STOP-BILABIAL	P B
STOP-DENTAL	T D
STOP-VELAR	G K
FRICATIVE-VOICED	V Z Z5
FRICATIVE-UNVOICED	F S S5 H
FRICATIVE-LABIO-DENTAL	F V
FRICATIVE-ALVEOLAR	S Z
FRICATIVE-PAL-ALV	S5 Z5
FRICATIVE-VELAR	H
AFFRICATE-VOICED	D1 DZ5
AFFRICATE-UNVOICED	C C1 C5
AFFRICATE-DENTAL	C
AFFRICATE-PAL-ALV	C5 DZ5
AFFRICATE-PALATAL	C1 D1
SEMI-VOWEL	J
VIBRANT	R
SYLLABIC	R
VOWEL-FRONT	E I
VOWEL-CENTER	A
VOWEL-BACK	O U
VOWEL-TONGUE-HIGH	I U
VOWEL-TONGUE-MIDDLE	E O
VOWEL-TONGUE-LOW	A

Table 5-6: Acoustic and Articulatory Phone Classes for Serbo-Croatian Phones

Note: An earlier designed set of phone classes, consisting of seven categories less than the preceding one, performed only slightly worse (0.1% absolute). JanusRTk does not seem to be too sensitive to small changes within the set of phone classes.

Performance and Results

The introduction of context dependent models resulted in a significant improvement of recognition performance (Table 5-7). The test set was the same as in Table 5-5. Again, for details on language modeling please refer to paragraph 5.5.

System	Data	Labels	Vocabulary	OOV	LM	WER
D-CD-0	5.5 h	D-CI-1	10 K	13.6%	D-LM-0	39.4%
D-CD-1	12.5 h	D-CD-0	18 K	8.5%	D-LM-1	36.6%

Table 5-7: Results for Context Dependent Systems

System D-CD-0 yielded 39.0% WE, which is an absolute 11.0% better than the best context independent recognizer D-CI-1. The increase in word accuracy encountered between the first (D-CD-0) and the second (D-CD-1) context dependent system has several reasons:

1. **More training data was available.** By that time, the preparation of the second half of the dictation data was finished and we were able to incorporate these utterances into our training set leading to models with a higher robustness.
2. **The number of codebooks was increased from 2500 to 3000.**
3. **A larger vocabulary and reduced OOV rate.** The new words found in the additional training data were inserted also into the test vocabulary. Although the OOV rate went down significantly, we did not benefit from that as much as could have been expected. This indicates that our models are not very stable yet. A rule of thumb states that one OOV word produces between 1 and 2 errors during recognition, which we found to be true for Serbo-Croatian. The reduction of the OOV by a certain percentage should then lead to an almost identical reduction in the word error rate. We did not encounter that behavior here, but observed it, however, for some experiments in the broadcast news domain (see section 6.5).
4. **The language model was based on the larger training corpus.**

Normalization and Adaptation

As already mentioned (see paragraph 5.2) label boosting was used to write labels with a better accuracy. This method performs a linear transformation on the existing codebooks to better represent the properties of a particular speaker or utterance.

Our strategy was now to apply the **MLLR adaptation (Maximum Likelihood Linear Regression)** [26] during testing. In general, a transformation for the component mixture means is derived by linear regression using a maximum likelihood optimization criterion similar to the standard ML training algorithm for HMMs (Hidden Markov Models). The best use is made of the available adaptation data by defining equivalence classes of regression transformations and tying one regression matrix to a number of component matrixes. The regression equivalence classes for tying transformations were defined by using a between mixture component distance measure to place similar component into the same regression class. The original means are mapped to the new, unknown means by a linear transform estimated from adaptation data: $\mu = W \nu$, where W is the $n \times (n + 1)$ transformation matrix and ν is the extended mean vector: $\nu = [1, \mu_{1,\ldots}, \mu_n]^t$. In our case a linear transformation was calculated for each of the four test speakers on the results obtained by a first recognition pass.

In the same way we computed a **VTLN (Vocal Tract Length Normalization)** [25][40], which is done during preprocessing. This normalization technique extenuates differences between various speakers and channels, and has the effect that the data is distributed more uniformly over the feature space. If we assume a uniform tube with length L for the model of the vocal the frequency axis during the signal-processing step, to make speech from all speakers appear as if it was produced by a vocal tract of a single standard length. The frequency warping is done using a piecewise linear transformation of the frequency axis that has fixed point at 0 kHz and at the Nyquist frequency (8 kHz in the case of 16 kHz dictation or broadcast news data). We choose a point A below the Nyquist frequency. The map from 0 kHz to A is a line through the fixed point at the origin with a slope chosen from the range 0.88 to 1.2 (discrete steps). The map from A to the Nyquist frequency is a line that intersects the previous line at A and ends at the fixed point at the Nyquist frequency. We refer to such a map as warp scales.

For female speakers the slope is commonly less than 1.0, which compresses the frequency axis. Male speakers usually select a warp scale greater than 1.0, which expands the frequency axis (Figure 5-1). The frequency rescaling is applied after the FFT has been calculated: Let X and Y denote the original and transformed frequency axes, and let f: X → Y be a warp scale. Given y in Y, there is a unique x in X with y = f(x). We then set FFT(y) = FFT(x). This transformed FFT is the basis for all further processing (extraction of spectral and cepstral features).

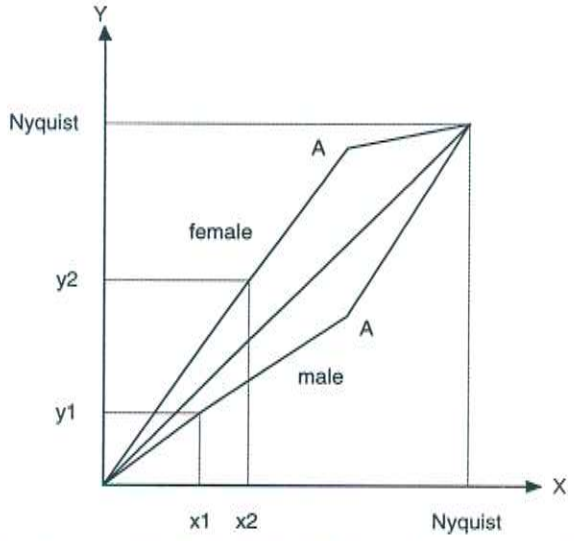


Figure 5-1: Warp Scales for Vocal Tract Length Normalization

A speaker dependent VTLN was already used during the training of system D-CD-1. Similar to adaptation, we computed the VTLN coefficients for each speaker in the test set on the basis of a first recognition pass.

Both normalization methods, VTLN and MLLR adaptation, were then applied to the test data during a second recognition pass. The results are shown in Table 5-8.

System	VTLN/MLLR	LM	WER
D-CD-1	NO	D-LM-1	36.6%
D-CD-1	SPK. DEP.	D-LM-1	28.4%

Table 5-8: Results for System D-CD-1 with VTLN and Adaptation

The normalization/adaptation test was run with the same vocabulary and language model and on the same set of utterances. The impressive improvement in performance indicates that the models trained so far were not very robust yet. This is no surprise at all, as the amount of training data was very low.

5.5 Language Modeling

This paragraph will give an overview on the different language models used for testing the dictation system. In particular we will present the usage of interpolated language models on the basis of web-retrieved text documents and the results obtained by applying them to the dictation test data. We used standard trigram, bigram and unigram backoff models.

Corpus Selection

We had gathered text material from different sources (television and radio stations, newspapers and news agencies) and geographical origin (Croatia and Serbia) which now had to be categorized into separate corpora. Table 5-9 provides more information on the data retrieved from different sites on the Internet. (Note: The figures in Table 5-9 reflect the final state of our text database, so the numbers might differ from Table 5-10.)

Corpus	Source	Type	Origin	Words	Diacr.
B92	www.siicom.com/odrazb	Radio	Serbia	3.50 M	Yes
Fonet	www.beocity.com/fonet	Radio	Serbia	0.55 M	Yes
Sezam	www.sezampro.yu	Radio	Serbia	0.35 M	Yes
Glas Koncila	jeronim.hbk.hr/GK/gk.htm	News[2]	Croatia	1.20 M	Yes
Hina	www.hina.hr	Agency[3]	Croatia	0.10 M	Yes
Varazdinske	www.hrvatska.com/varazdinske.vijesti	News	Croatia	0.10 M	Yes
Medjimurje	www.medjimurje.com	News	Croatia	0.30 M	Yes
Duga	www.suc.org/news/duga	News	Serbia	0.20 M	No
HRT	www.hrt.hr	TV	Croatia	1.00 M	No
Naša Borba	www.nasa-borba.co.yu	News	Serbia	2.00 M	No
Nin	www.suc.org/news/nin/archive	News	Serbia	0.50 M	No
Novine	serbia.net/novine	News	Ser/Cro	0.05 M	No
Republika	www.yurope.com/zines/republika	News	Ser/Cro	0.50 M	No
VOA	www.voa.gov/miscl/croatia	Radio	Croatia	0.10 M	No
RFE/RL	www.rferl.org/BD/ss	Radio	Ser/Cro	0.20 M	No
Vreme	www.vreme.com	News	Serbia	0.10 M	No
Tanjug	www.suc.org/news/tanjug	Agency	Serbia	0.30 M	No
SRNA	www.suc.org/news/srna	News	Serbia	0.35 M	No
Svetlost	www.svetlost.co.yu	News	Serbia	0.35 M	No
Tjednik	vukovar.unm.edu/tjednik	News	Croatia	0.30 M	No

Table 5-9: Text Corpora for Language Modeling

At this point, we were considering only documents that contained diacritical characters, as our text conversion algorithm (see section 4.3) was not developed fully yet.

[2] "News" refers to newspapers and all other categories of news, which do not fit elsewhere
[3] "Agency" refers to news agencies and similar sites

We defined a number of criteria for partitioning the data into categories:

1. **Geographical origin.** This is important due to the differences between Serbian and Croatian vocabulary (please refer to chapter 3 on Serbo-Croatian language for further information).
2. **Type of data.** Different types of text data, e.g. radio news stories and newspaper articles, might show different characteristics, which should be reflected in the categorization of the news texts.
3. **Corpus size.** To estimate robust tri-, bi- and unigram probabilities, the language model should be based on a sufficient amount of data.
4. **Perplexity of corpus language model.** Although language model perplexity does not directly relate to word accuracy, it is an indicator for the "difficulty" of a task or corpus, and therefore might give hints for a reasonable partition.
5. **Number of corpora.** The more language models are used for the interpolation process, the longer it takes to decode an utterance. The trade-off between possible performance and actual cost of computation must be handled accordingly.

Given the low amount of available data at this point, we based our decision mainly on geographical origin and corpus size, thus obtaining a corpus consisting of Croatian documents (D-CRO-0), and another one containing all Serbian texts (D-SER-0).

Synopsis of Language Models

The table below lists important numbers for the different language models used during the evaluation of the dictation system. The first language model (D-LM-0) is based on the initial set of training data (D-DIC-0), which consisted of 5.5 hours acoustic recordings and 42 K words respectively; we used it for system D-CD-0. The second model (D-LM-1) is based on 12.5 hours of audio data and 98 K words (D-DIC-1), which led to a significant reduction in perplexity. This model was used for the baseline test of our dictation system D-CD-1. We also calculated a standard trigram model for the two corpora selected in the paragraph above.

LM	Corpus	Words	PP
D-LM-0	D-DIC-0	42 K	480
D-LM-1	D-DIC-1	98 K	304
D-LM-2	D-CRO-0	1.6 M	613
D-LM-3	D-SER-0	2.5 M	1055
D-LM-4	D-LM-1/D-LM-2	98 K/1.6 M	260

Table 5-10: Language Modeling for Dictation System

Interpolation

We will outline the aspects of interpolating the three chosen corpora (D-DIC-1, D-CRO-0, and D-SER-0) in the following paragraph. We are using the hierarchical interpolation module provided by JanusRTk. Given two individual language models, a weighted sum of ngram probabilities is calculated to obtain the new score. By subsequently combining pairs of language models we can interpolate any number of models (hierarchy).

Interpolation Experiments

The following tables will show the results for the interpolation of different corpora.

γ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
PP	1055	864	792	733	697	667	645	630	619	**612**	613

Table 5-11: Interpolation with γ * D-CRO-0 + (1-γ) * D-SER-0

γ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
PP	304	262	**260**	266	277	290	308	332	369	431	613

Table 5-12: γ * D-CRO-0 + (1-γ) * D-DIC-1

γ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
PP	304	**263**	268	280	298	319	351	395	460	576	1055

Table 5-13: γ * D-SER-0 + (1-γ) * D-DIC-1

The results of the first interpolation served as basis for the second pass. We chose the best weights according to the lowest perplexity. Although the so determined value for γ does not necessarily lead to the best performance in terms of word error rate, we considered the perplexity to be a reasonable indicator for the difficulty of a language model.

γ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
PP	1055	598	482	413	370	338	312	295	278	267	**260**

Table 5-14: γ * (0.8 * D-DIC-1 + 0.2 * D-CRO-0) + (1-γ) * D-SER-0

γ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
PP	613	442	378	342	318	297	284	272	266	263	**263**

Table 5-15: γ * (0.9 * D-DIC-1 + 0.1 * D-SER-0) + (1-γ) * D-CRO-0

γ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
PP	304	262	**261**	266	277	291	309	333	373	431	612

Table 5-16: γ * (0.9 * D-CRO-1 + 0.1 * D-SER-0) + (1-γ) * D-DIC-0

According to perplexity, we selected the optimal coefficients from the tables above:

1.0 * (0.8 * D-DIC-1 + 0.2 * D-CRO-0) + 0.0 * D-SER-0.

This means that we only had to interpolate the Croatian language model and the one based on the training texts.

Results with Interpolation

We applied this newly generated language model (D-LM-4) during test with speaker dependent normalization and without. The performance we gained is shown in Table 5-17.

System	VTLN/MLLR	LM	WER
D-CD-1	NO	D-LM-1	36.6%
D-CD-1	NO	D-LM-4	34.2%
D-CD-1	SPK. DEP.	D-LM-1	28.4%
D-CD-1	SPK. DEP.	D-LM-4	28.2%

Table 5-17: Results for System D-CD-1 with Normalization and Interpolation

It can be observed that the benefit from interpolated language models is particularly high (+2.4% absolute), when the acoustic performance is still weaker. With the phone models becoming more robust, the gain from interpolation decreases to + 0.2% absolute.

5.6 Summary and Conclusion

We have given a complete overview on the system development process of a context dependent Serbo-Croatian dictation speech recognizer. We discussed the design of the phone set and classes, the pronunciation dictionary, the effects of different normalization techniques and the gain we can obtain from interpolated language models. We showed that a rather low amount of acoustic data is sufficient to build a fairly robust system.

At this point we switched over to the training of the broadcast news recognizer, which is based on the dictation system and uses many features and techniques described above.

6

THE BROADCAST NEWS SYSTEM

Based on the experience gained during the development of the dictation system and the availability of a working recognizer for that task, we focused all our efforts on the training and tuning of a recognition system for the broadcast news domain. This chapter presents the details on bootstrapping the system, reports the results for the recognizer based on context dependent models, and shows the effects of different normalization and adaptation strategies based on data segmentation. Another paragraph will give an overview of the language modeling for this task. We will also explain the various approaches to overcome one of the major obstacles towards a better system performance, namely the high OOV rate. Serbo-Croatian is not a homogenous language, so we introduced some techniques to unify Serbian and Croatian variants; this work is outlined in the last section on language normalization. Information on English broadcast news recognition can be found in [30][4][2][41][31][20][19][36][38][12][33].

6.1 Bootstrapping

Baseline Experiment

The dictation system, which was trained on 12.5 hours of read news texts and yielded 28.2% word error rate on the dictation test set, was now used for a first recognition run on the broadcast news data[4]. The result obtained through this experiment was the baseline for all further systems.

System	Data	WER D-LM-1	WER D-LM-4
D-CD-1	12.5 h (D)	75.9%	73.6%

Table 6-1: Results for Baseline Experiment on Broadcast News Data

We had expected a drop-off in recognition performance, but not in such a drastic manner. When testing a system that was trained with clean speech (e.g. read news articles recorded in studio quality) on distorted data (e.g. conversational telephone speech or broadcast recordings) other groups reported an increase of word errors in the range between 15 and 20% relative in respect to an evaluation on clean data [8]. We concluded that our models yet were not very robust (an effect of the little training data), particularly for the television recordings which were extremely noisy. Even the studio segments showed many distortions, probably resulting from bad recording conditions such as snow or rain.

[4] All tests reported in this chapter are evaluated under the same conditions as the HUB4 PE (partitioned evaluation) experiments, where segments and their acoustic conditions are known prior to the recognition run.

First Labels

Although the dictation system did not perform outstandingly well on the broadcast test data, the first labels in a forced alignment run on the new training data, consisting of 10 hours transcribed Croatian news broadcasts, were created with it. This time the label boosting technique was applied to the single segments, which were identified during manual transcription. The warping factors for VTLN were calculated in the same way.

6.2 Context Dependent Recognizer

System Properties

We used the same preprocessing as for the dictation system and each phone was modeled by a mixture of 16 Gaussians, we changed however some other parameters. The dimensionality of the feature vector was reduced to 24; this speeds up the system turn-around time during development without decreasing the performance too much (roughly 1-2% absolute) [8]. The number of states in the HMM for each phone was 3. The reduction from 6 to 3 states attenuates the duration constraint; this is appropriate, as the speech in the broadcast news data was much faster than encountered for the read newspaper articles.

First System

The first context dependent system (B-CD-0) was built on the basis of quinphone models with an overall number of 2000 codebooks, we used the same phoneme classes as for the dictation system. The test vocabulary consisted of 29 K word entries, which led to an OOV rate of 14.0%.

System	Data	WER B-LM-0	WER B-LM-4
B-CD-0	10.0 h (B)	45.2%	43.6%

Table 6-2: System Performance for First Broadcast News System

After only one training iteration on acoustically similar data the recognition performance was much better than we could obtain with the dictation system for the same test data. We observed that the effect of interpolated language models was comparable to the one we encountered for the earlier dictation systems. We gained an improvement of 1.6% absolute. This result was a little disappointing, as B-LM-4 was calculated on far more text data and thus was expected to do better than the language models for the dictation system (supposing that the unigram, bigram and trigram models were more accurate). More experiments and results with different language models will be discussed in paragraph 6.4.

6.3 Advanced Systems

For our next system, B-CD-1, we combined the dictation data (12.5 h) and the available broadcast news recordings (13 h) in the relation of 1:2, thus a total of 38.5 h of acoustic data was presented to the recognizer during training. We doubled the number of codebooks, because the amount of training data was sufficient to estimate 4000 codebooks reliably. We also added another 2000 words to the test vocabulary, yielding a total of 31 K words, which slightly reduced the OOV rate to 13.5%.

The Realaudio data, which had been decompressed and sampled up to from 11.025 kHz to 16 kHz, was included in our training set. Later we excluded it, as we were not sure about its effects on the system performance. The Realaudio format (http://www.realaudio.com) uses a lossy compression algorithm to minimize the bandwidth for audio and video broadcasts over the Internet. The high data reduction and poorer audio quality could introduce undesirable acoustic effects and therefore lead to inadequate models, although their robustness might benefit from the telephone-like sound quality of Realaudio recordings. Future work might deal with the usage of this type of data.

System	LM	WER
B-CD-1	B-LM-0	38.3%

Table 6-3: Results for B-CD-1

The application of the above described techniques led to a significant reduction in error rate (Table 6-3): 7.6% absolute, with the largest improvement probably resulting from the increased number of codebooks trained on a larger amount of acoustic data.

Segmentation

In the next series of tests we tried to benefit from the knowledge of segments and their acoustic conditions in our test set. We identified the 9 different classes listed in Table 6-4, which also shows the percentage of data they cover in the training set. The distribution in the test set was quite similar.

Classes	Portion of Data
male_clean_quiet (mcq)	28.0%
male_clean_noisy (mcn)	20.0%
male_dirty_quiet (mdq)	1.5%
male_dirty_noisy (mdn)	7.5%
female_clean_quiet (fcq)	20.0%
female_clean_noisy (fcn)	12.5%
female_dirty_quiet (fdq)	7.0%
female_dirty_noisy (fdn)	0.5%
nospeaker	3.0%

Table 6-4: Acoustic Classes/Segment Clusters

Adaptation

We expected an improvement in performance by using adaptation on the test set. While the linear transformation for the dictation data was computed speaker dependently, we had to choose another approach for the broadcast news task, because no information about the identity of the speakers was available.

In the first experiment we calculated a transformation for every single segment. We only used their transcription-based boundaries; the information about their acoustic characteristics was discarded. For the second test we used the acoustic classes shown in Table 6-4, assuming there was a way to classify and cluster the different segments (this again conforms to the HUB4 partitioned evaluation). We reduced the number of classes, however, and computed an adaptation for man_clean (mcq), man_dirty (mcn, mdq, mdn), wom_clean (fcq), and wom_dirty (fcn, fdq, fdn). By merging clusters, an almost equal part of data (man_clean

28.0%, man_dirty 29.0%, wom_clean 20% and wom_dirty 20%) was available for each adaptation, which should make the resulting models more robust. By computing a linear transformation for these newly defined classes, we obtained four condition-specific sets of acoustic models. We used them as basis for another adaptation, which was calculated for every single segment. Figure 6-1 illustrates this two-stage adaptation.

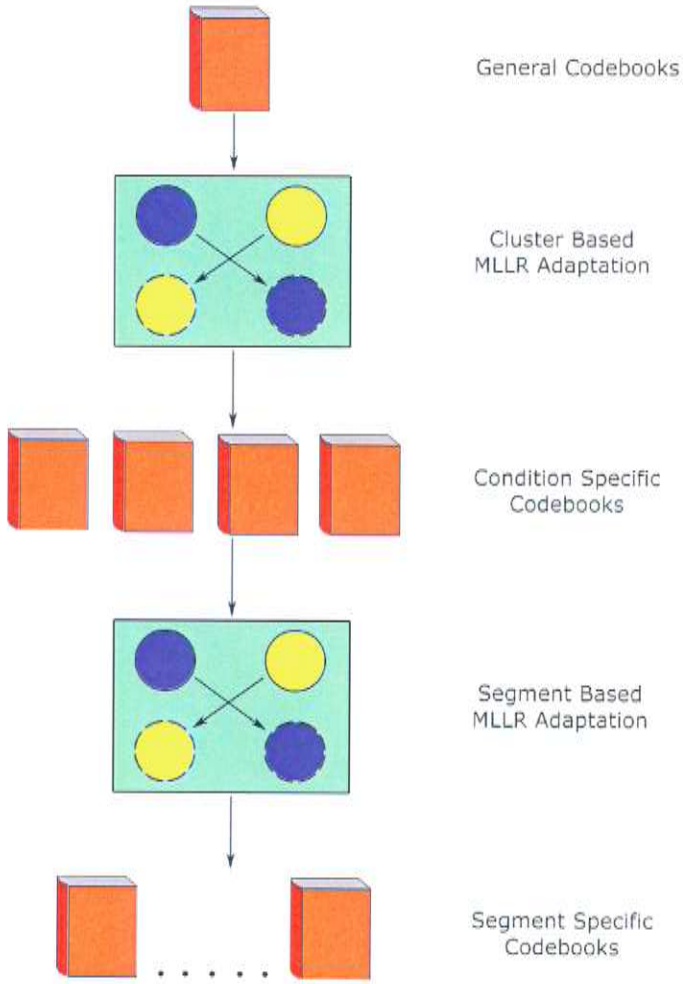


Figure 6-1: Two-Stage Adaptation

Normalization

Similar to the experiments for the dictation domain, we applied VTLN also on the broadcast news data. This time, however, there was no speaker information available. We used the transcript-based segments to determine the warp factors for the training data. The corresponding values for the test utterances were computed in the same way, based on the results of a first recognition run.

Results

The results for different adaptation strategies and language models are shown in Table 6-5. The best word error rate was gained with a one-stage adaptation on single segments, which outperformed the two-stage linear transformation by 0.2% absolute. This contradicts the results reported for English broadcast systems [41], where similar adaptation approaches yielded a significant improvement in performance (9.4% absolute). In addition our adaptation was guided by a confidence measure that might be too restrictive. During the calculation of the adaptation we only considered those words, which had been recognized with a confidence value of 1.0 in the first decoding run. This means that the entry for a particular time frame in

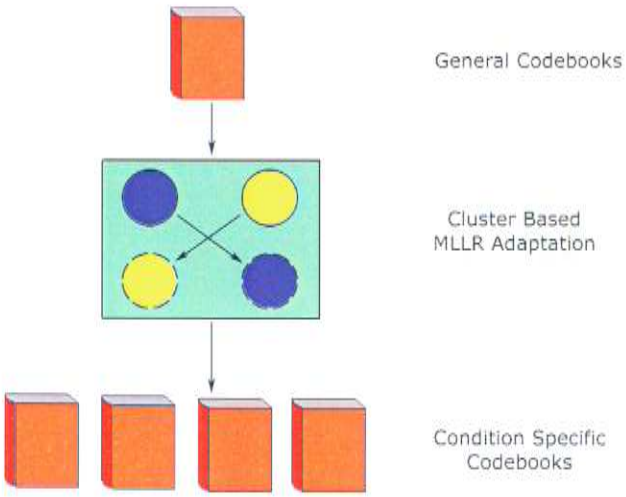


Figure 6-2: Condition Specific Adaptation

all lattice paths had to match the first best hypothesis. Given the observation that the Serbo-Croatian system recognizes the sequence of phonemes quite accurately, it might be a good idea to experiment with confidence values smaller than 1.0. We would therefore hope that the phoneme sequence is almost correct, although the word sequence might be wrong. It is also possible that the automatic clustering of acoustically similar segments might lead to better adaptation. Both issues (tolerant confidence measure and automatic segmentation) are, however, subject to future research and will not be covered in this work. We also observed that the interpolation with a class based language model did not improve the recognition performance, the best result was gained with an interpolation of three word-based language models. See section 6.4 for more details on language modeling.

System	Adaptation	LM	WER
B-CD-1	One-Stage	B-LM-3	36.0%
B-CD-1	Two-Stage	B-LM-3	36.2%
B-CD-1	One-Stage	B-LM-1	37.6%
B-CD-1	One-Stage	B-LM-5	38.7%

Table 6-5: Results for Different Adaptation Techniques and Language Models

In an earlier stage of development we had experimented with condition specific recognizers (Figure 6-2). On the basis of the general models obtained after training, we calculated an MLLR adaptation for each of the four male- and female-based clusters given in Table 6-4

using the same training data. During testing we used the appropriate recognizer for a single segment (assuming we knew its acoustic characteristics). Unfortunately recognition performance dropped off by 1.4% relative, which again differs from results reported for English broadcast news [36].

6.4 Language Modeling

For the different language models presented in Table 6-6 we used the broadcast news texts of the acoustic training (B-BRN-0), the Croatian (B-CRO-0) and the Serbian (B-SER-0) news documents. The classes for B-LM-4 were generated in the following way: We compiled a list of the 60 most common grammatical suffixes (noun, adjective and verb endings). We chopped off these particles from the back of all fitting words and gained a second list of word stems. When applying this algorithm to all available text documents (more than 10.5 million words), we obtained 220,000 syntax-oriented classes. In B-LM-5 the class-based language model was combined with a normal word-based one.

LM	Corpus	Words	PP
B-LM-0	B-BRN-0	80 K	1621
B-LM-1	B-CRO-0	2.5 M	432
B-LM-2	B-SER-0	8 M	1014
B-LM-3	B-LM-0/1/2	2.5 M/80 K/8 M	282
B-LM-4	B-BRN/CRO/SER-0	10.5 M/220 K Classes	344
B-LM-5	B-LM-1/B-LM-4	2.5M/220 K Classes	268

Table 6-6: LM Interpolation

6.5 OOV Rate Reduction

One of the characteristics of the broadcast news domain is a rapid vocabulary growth (se Figure 6-3and Figure 6-4). But while for an English 64k-system an OOV rate between 0.6% and 0.7% was reported [12], we would have to deal with 9% for a Serbo-Croatian recognizer of the same size. There are several reasons for that:

1. Serbo-Croatian is a strongly inflected language (see paragraph 3.5). One effect, which we observed, was that in many cases the word stem was recognized correctly, but the wrong suffix was chosen. Either because the acoustic models were still too weak or the language model triggered the incorrect form, or – this was far more often the case – the right word inflection was simply not in the vocabulary. The Hypothesis Driven Lexicon Adaptation (**HDLA**), shortly outlined below and further discussed in [13] [14], is one way to compensate this, another technique – **morpheme-based recognition** – is presented in the next section.
2. The various dialectic variations of Serbian ('reka') and Croatian ('rijeka', 'rjeka'), which exist for many words, are another source of increased vocabulary growth and OOV rate. We will attack this problem in the paragraph on **language normalization**.
3. Divergent tendencies after the war in former Yugoslavia. The new countries try to differentiate themselves by introducing new or archaic words, e.g. Serbian: 'ekonomija', Croatian: 'gospodarstvo' (archaic), which both mean 'economy'. We increased the vocabulary size to account for these effects.

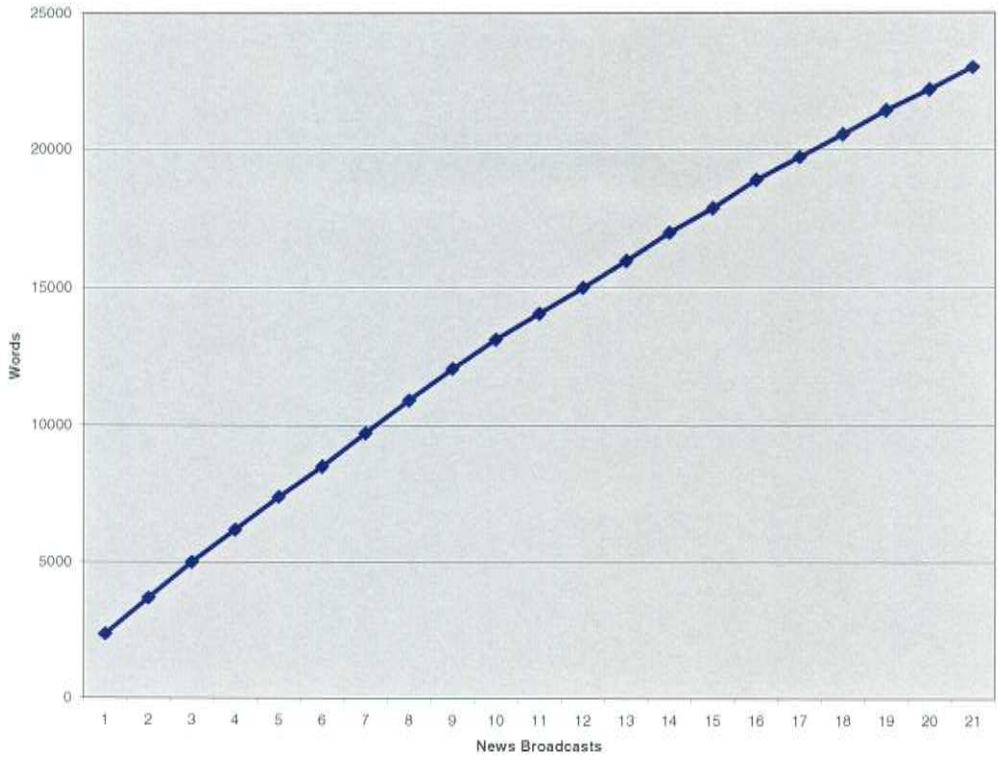


Figure 6-3: Vocabulary Growth Per Broadcast

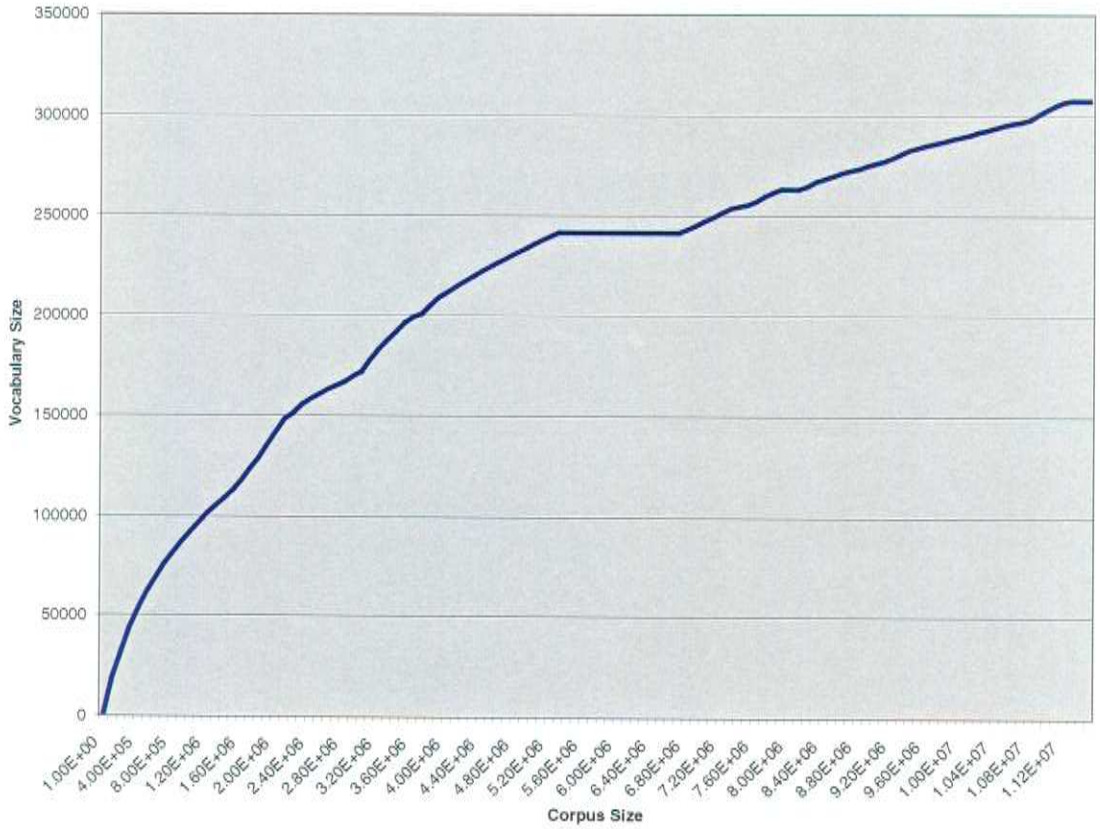


Figure 6-4: Vocabulary Growth / Corpus Size

The negative effect of the high OOV rate on system performance becomes even more evident, when we consider that almost every recognition error, which resulted from an OOV word, was followed by another misrecognition: The biggest part of errors for our 31 k Serbo-Croatian recognizer was OOV related.

Morpheme-based Recognition

We examined two different approaches to generate a morpheme vocabulary as the lexicon of our recognizer.

1. Suffix-based Stemming

The first idea was to use a list of grammatical suffixes, similar to the one used for class-based language modeling, which was described in paragraph 6.4. After chopping off the endings (Table 6-7) we obtained a list of stems. Suffixes and stems were the new units of recognition (morphemes).

Suffix Length	Suffixes
6	itijim ijskoga ijskome stvenu stvima ijskih ijskim ijskog ijskoj ijskom ijinim ijinih ijinog ijinoj ijinom ijiske
5	enima evima ijama ijima itije itiji ljama ljima nijeg nijem nijih nijoj nijom nikom njega njici njicu nogih nogim nosti novih novim novog novoj novom ovima ovske skije skoga stava stvai stvim stvom ijska ijske ijski ijsko ijsku nijim ijinu ijino ijina ijine ijini ijska
4	able ajte anje anom enom evih evog ijom itih itim itog itoj itom jema jemo jete ljim nija nije niji niju nika nike niku njeg njem njih njim njoj njom noga nogi nome nost nova nove novo novu osmo oste ovih ovim ovog ovoj ovom skih skim skog skoj skom stva stvi stvo stvo stvu tima anih anog anoj ijih ijim ijin anov
3	aju ale ama ana ani ano anu ega emo emu ena enu ete eva eve evi hom hov ija ije iji ijo iju ima ime imo imu ita ite iti ito itu jeg jem jih jim jmo joj jom jte lja lje lji ljo lju nih nik nim nja nje nji njo nju nog noj nom nov oga oga oma ome omi omu ova ove ovi ovo ovu ska ske ski sko sku smo ste tom ane os5e jes5
2	aj an eg em en ga ha hu ia ih ij im ja je ji ju la le li lo ma me mo na ne ni no nu og oh oj om ov ta te ti tu us es5 s5e
1	a e g h i j m o u s5

Table 6-7: Suffixes of Different Length Based on Grammatical Forms

2. Similarity-based Stemming

For the second method, we clustered all words in our 12M-word corpus into similarity classes, which was done by calculating the Levenstein editing distance between them (similar to the technique used for diacritic text conversion). For each word or cluster, respectively, we computed the closest neighbor and expanded the classes recursively (bottom up clustering). The algorithm terminated, when a distance threshold was exceeded. We determined the longest common prefix for each of the resulting classes, thus gaining a list of one prefix and several suffixes. The global prefix and suffix list was used to form the new morpheme dictionary for recognition.

Table 6-8 gives some examples out of the resulting word lists, using different stemming methods. (**Note**: At this point we used a shorter suffix list than the one shown in Table 6-7.)

Suffix-Based	Similarity-Based
ak	ak
akadem	aka
akademic	akade
akademij	akadem
akademijin	akademi
akademik	akademij
akademnij	akademsk
akademsk	

Table 6-8: Word List for Different Stemming Methods

Results

We did not retrain our acoustic models, which presumably would have improved recognition performance, as the context dependent polyphone models become more accurate, particularly when they contain word boundary tags. We used a single trigram language model, which leads to a loss of context information for a system that is based upon morphemes as recognition units. A word-based recognizer utilizes a group of three words for trigram modeling, but in a morpheme-based system the context might reduce to a sequence of only two words. The integration of 4-gram or even 5-gram language models could compensate this drawback. Another idea is to use a morpheme-based lexicon for acoustic scoring, but a word-based language model [8]. A synopsis of our experiments is presented in Table 6-9.

	Vocabulary	OOV Rate (Word-based)	Morpheme Error	Word Error
Word-based Baseline	31 K	13.6%	34.9%	44.9%
Morpheme-based (Suffix List)	17 K	7.5%	39.5%	53.3%
Morpheme-based (Suffix List)	31 K	5.5%	36.6%	51.4%
Morpheme-based (Similarity)	17 K	3.7%	51.5%	59.0%

Table 6-9: Morpheme-Based Recognition Results

We observed that the morpheme recognition based on a suffix list worked better than the similarity-based approach. While the gain in terms of OOV reduction was far bigger for the cluster-based morphemes, it seems that the grammar-derived dividing rules result in a more 'natural' way of splitting the words: The acoustic and/or language models fit better for this syntax-driven word chopping. In general, however, the word-based approach yielded better results, although the OOV rate decreased significantly for the morpheme-based systems. Some of the reasons (no acoustic retraining, loss of context) were mentioned in the section above, another one could be found in the simplicity of the morpheme-merging algorithm. To convert the morpheme hypothesis into words, we checked whether a morpheme existed in a large word vocabulary list. If it was found, no action was undertaken, otherwise it was merged with the following word. The idea behind this technique was, that word splitting resulted in a

maximum number of two morphemes per word. However, the algorithm might work incorrectly for morphemes that exist as a whole word and prefix for other words as well. Future work might additionally concentrate on the development of better merging strategies to improve the morpheme-based word error rate.

Adaptive Vocabulary

For the Hypothesis Driven Lexicon Adaptation (**HDLA**) word lattices for all test utterances are created in a first recognition run. The lattice is then used to determine which words are most likely uttered in the segment, namely the words in the lattice themselves and possible derivations that can be found in a global word list. Using the words in the lattice we obtain a list of stems and suffixes. All words in the global word list which have the same stem as the lattice words are added to the new lexicon, replacing the least frequent words that did not show up in the lattice. This leads to an utterance-specific vocabulary of the same size as the original one. The basic idea behind this algorithm is that a large number of words in the hypothesis are decoded incorrectly because only the inflection ending is wrong whereas the stem was recognized correctly. In many cases this was due to the fact that the right word was not even in the lexicon, thus constituting an OOV word. Additionally an utterance-specific language model was calculated and used in the second recognition run based on the adapted vocabulary. Hypothesis Driven Lexicon Adaptation helped to reduce the OOV rate and also led to improved performance. Please refer to [13] [14] for a detailed discussion of this technique and the recognition results that we obtained.

Language Normalization

In this paragraph we will present our approach to make the Serbo-Croatian language more uniform, with positive effects on acoustic and language modeling.

Unifying Serbian and Croatian Variants

We tried to eliminate the differences between Serbian and Croatian variants of the Serbo-Croatian language. As we have seen, there might exist up to three forms of the same word (not counting inflections), e.g. 'reka', 'rjeka' and 'rijeka'. We transformed them into pronunciation variants of the same word, thus reducing the overall vocabulary size. We also introduced the new vocabulary entry <number>; different numbers are modeled as a pronunciation variant of this generic symbol.

Dictionary Entry Before Normalization	Dictionary Entry After Normalization
reka R E K A	reka R E K A
rijeka R I J E K A	reka(2) R I J E K A
rjeka R J E K A	reka(3) R J E K A
1.-e P R V E	<number>(769) P R V E
1.-og P R V O G	<number>(770) P R V O G
1.-oga P R V O G A	<number>(771) P R V O G A

Table 6-10: Dictionary Entries Before and After Language Normalization

Acoustic Models

When we calculated the forced alignment for this new set of vocabulary and pronunciation variants, we allowed the system to determine which variant fits best. Sometimes the difference between, say 'reka' and 'rjeka', is very hard to hear during transcription, so it might be better to have the recognizer decide which variant is the best one.

Language Models

The unification of different Serbian and Croatian variants reduced the number of tokens in the language model. By representing the dialectic variants with only one single form, we expected the language models to better reflect the true word distribution. Using the generic <number> concept for cardinal and ordinal numbers should emphasize this effect, too. Table 6-11 shows the effects of our language normalization efforts on a selected piece of transcribed training audio data. We did not only unify the dialectic variants, but also corrected transcription errors and typos.

We normalized all broadcast news training texts manually and applied the generated mapping rules to all text material, including the dictation training data, and the Serbian and Croatian corpora used for language modeling; for the latter some more data had become available in the meantime. We tried to automate the normalization process - with poor results, however: There are two many ambiguous cases that can not be handled by a simple replacement algorithm and therefore produce a lot of errors by introducing non-existing word forms. Future work will include the development of tools that enable a semi-automatic conversion procedure and the generation of mapping rules for the most frequent words in all text documents.

Transcription Before Normalization	Transcription After Normalization
evo jedne informacije koju smo upravo **pimili** u vladi je odrz5an sastanak izmed1u predstavnika hrvatske vlade na c5elu s **podprijedsjednikom** doktorom ivicom kosovic1em i izaslanstva **untasa** na c5elu sa gerhardom fis5erom voditeljem **zivilnih** poslova **untasa** na ovom **posljednjem** pripremnom sastanku razgovaralo se o pripremi **prijedstojec1ih** izbora u hrvatskom podunavlju predstavnici hrvatske vlade i **nntasa usuglas5avali** su nac5in i vremenski okvir priprema za provedbu izbora koji c1e se odrz5ati s5esnaestog oz5ujka posebno su raspravljani nac5in i **mijesta** izdavanja dokumenata i registracija birac5a koji z5ive na tom podruc5ju hrvatska vlada je dala jamstvo da je osigurala sva potrebna sredstva za provod1enje registracije birac5a koji su u podruc5je hrvatsko podunavlje dos5li nakon **tisuc1udevetsto** i prve godine takod1er je bilo **rijec5i** i o **glasovanju** hrvatskih prognanika **podprijedsjednik** kostojic1 predstavnicima **untasa** dao je **prijedlog mijesta** za **glasovanje** koji je sastavljen u dogovoru s uredom za prognanike zajednicom prognanika te z5upanima gradonac5elnicima i nac5elnicima prognanih gradova i **opc1ina** koordinacija javnih poduzec1a za pitanja hrvatskog podunavlja na kojem je danas5njem sastanku	evo jedne informacije koju smo upravo **primili** u vladi je odrz5an sastanak izmed1u predstavnika hrvatske vlade na c5elu s **potpredsednikom** doktorom ivicom kosovic1em i izaslanstva **untaes-a** na c5elu sa gerhardom fis5erom voditeljem **civilnih** poslova **untaes-a** na ovom **poslednjem** pripremnom sastanku razgovaralo se o pripremi **predstojec1ih** izbora u hrvatskom podunavlju predstavnici hrvatske vlade i **untaes-a usaglas5avali** su nac5in i vremenski okvir priprema za provedbu izbora koji c1e se odrz5ati s5esnaestog oz5ujka posebno su raspravljani nac5in i **mesta** izdavanja dokumenata i registracija birac5a koji z5ive na tom podruc5ju hrvatska vlada je dala jamstvo da je osigurala sva potrebna sredstva za provod1enje registracije birac5a koji su u podruc5je hrvatsko podunavlje dos5li nakon **tisuc1u devetsto** i prve godine takod1er je bilo **rec5i** i o **glasanju** hrvatskih prognanika **potpredsednik** kostojic1 predstavnicima **untaes-a** dao je **predlog mesta** za **glasanje** koji je sastavljen u dogovoru s uredom za prognanike zajednicom prognanika te z5upanima gradonac5elnicima i nac5elnicima prognanih gradova i **ops5tina** koordinacija javnih preduzec1a za pitanja hrvatskog podunavlja na kojem je danas5njem sastanku

Table 6-11: Transcription Before and After Language Normalization

By applying the manually created conversions we were able to reduce the total number of words from 312 k to 307 k. Obviously this number would further decrease, when using the above described enhancements for the normalization method. With a cutoff value of 14 (or more) occurrences, which leads to a self-coverage of 93% (Figure 6-6), we determined the new test lexicon of 51 k entries, including 1.2 k normalized and 0.8 k number pronunciation variants. Given this new test vocabulary, we calculated language models for the three base corpora: Acoustic training texts (dictation and broadcast news data), Croatian and Serbian news.

LM	Corpus	Words	PP
B-N-LM-0	B-BRN-1	220 K	610
B-N-LM-1	B-CRO-1	3.5 M	500
B-N-LM-2	B-SER-1	8 M	1035
B-N-LM-3	B-LM-0/1/2	220 k/3.5 M/8.0 M	347

Table 6-12: LM Interpolation for Normalized Texts

With the increased number of words in the lexicon the OOV rate reduced to 7.8%, whereas a 49 k vocabulary would have led to 10.1% on texts without normalization (see Figure 6-5). This is due to the fact that some OOV words now occur implicitly as pronunciation variants. We also observed that the vocabulary based on the acoustic training data produces a smaller OOV rate than the lexicon derived by frequency analyses on all available text data. This is something we expected, because the training data is closer related to the test set than a collection of more general text corpora, although they belong to the same topic. We chose the conventional approach of frequency-based vocabulary selection, which simplifies future work and also the HDLA algorithm for adaptive vocabulary.

Results

System B-CD-2 is trained on 45 h combined dictation and broadcast news data and consists of 4000 codebooks with 32-dimensional feature vectors. The system was tested on both, the dictation and the broadcast news domain.

System	Test Set	WER B-N-LM-0	WER B-N-LM-3
B-CD-2	Broadcast News	33.1%	29.5%
B-CD-2	Dictation	24.5%	20.9%

Table 6-13: Results for System B-CD-2

The performance for both domains improved significantly. We observed that the effect of language model interpolation also increased. The benefit from this technique using normalized corpora was more than twice the improvement we got earlier for the original texts. We think this is due to the fact that the probability mass, which was split on up to three dialectic variants, now is unified on one canonical form. This leads to more accurate models and better performance through language normalization.

6.6 Conclusion

Although the dictation baseline system did not work well on broadcast news data, we were able to apply many of the techniques we had used during its development. Additionally we experimented with different adaptation and normalization techniques, and introduced some

methods to deal with language specific problems, such as an extremely high OOV rate due to a large number of inflected word forms. Again we used multi-corpora language model interpolation with positive effects on the system performance in particular after the application of our language normalization procedure. Despite the very limited training data, we trained a chain of improving broadcast news systems on the combined dictation and broadcast acoustic data. It turned out that this "hybrid" system worked also better for the dictation domain. We assume that the more robust models account for this effect.

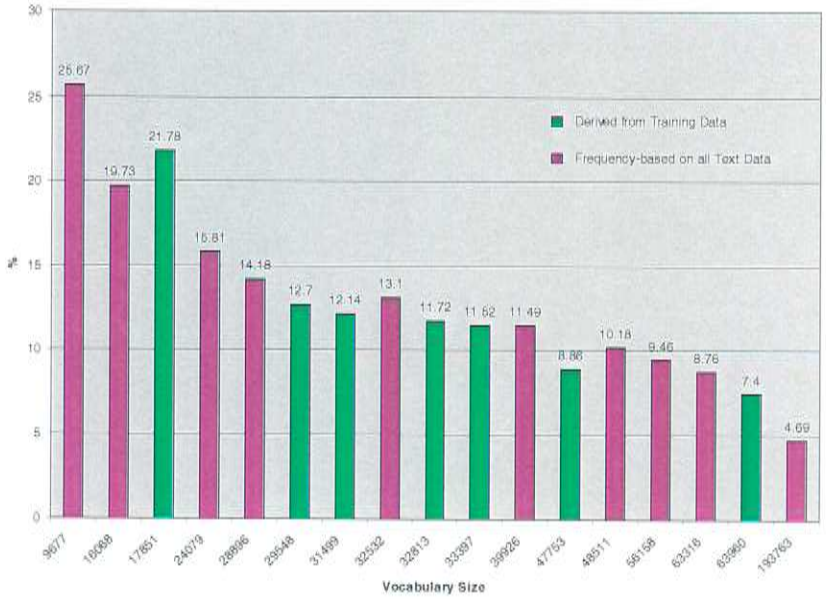


Figure 6-5: OOV Rates for Different Vocabulary Sizes

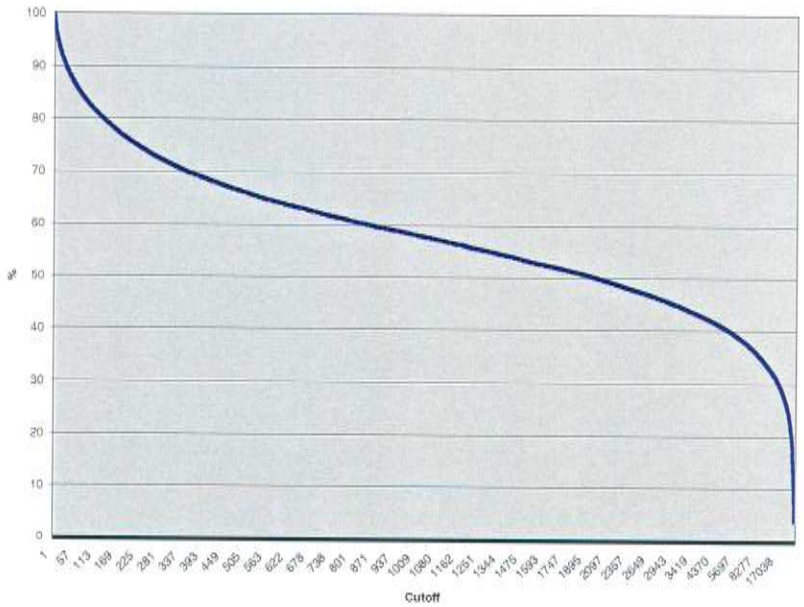


Figure 6-6: Coverage for Different Cutoff Values

7

SUMMARY AND CONCLUSION

We outlined the complete system development process of a Serbo-Croatian speech recognizer for the dictation and broadcast news domain. We described the data collection process, gave an introduction to the Serbo-Croatian language and explained the basic concepts of the Informedia project. Our speech recognition system will be a part of the multilingual extension to the existing English Digital Video Library.

We built a dictation system, which served as a baseline and prototype for the training of the actual broadcast news recognizer. We described the effects of different normalization and adaptation strategies, examined the effects of multi-corpora language model interpolation, and presented techniques to attack language specific problems such as rapid vocabulary growth and high OOV rate due to a large number of inflected word forms. The two distinct development lines for dictation and broadcast news were finally merged into one system for both tasks. Despite the fact that our system was based on a fairly low amount of acoustic training data, we yielded results that were comparable to the 1996 Hub-4 evaluation for English broadcast news: 29.5% WER for broadcast news, and 20.9% WER for dictation data.

Apart from applying standard techniques such as vocal tract length normalization, adaptation and multi-corpora language model interpolation, we had to deal with some particular problems for Serbo-Croatian: The available acoustic training data was very limited. The usage of alternative data sources, e.g. Realaudio recordings from the Internet, might relieve the situation. Concerning text data collection this approach has already been applied. Large amounts of Serbo-Croatian documents were obtained from the Internet, which was our only source for language model data, as no other corpora, e.g. on CD-ROM, were available. The fact that Serbo-Croatian is a highly inflected language makes it also different from English. The techniques, which we presented to account for that, could also be useful for other languages, e.g. Russian or Spanish.

8

FUTURE WORK

Although the obtained results for both the dictation and broadcast news recognizer were quite satisfying, there are still some open issues for the future:

- **More acoustic training data.** Retraining of the complete system with more acoustic data would certainly help to improve the existing recognizer.
- **Alternative data sources.** Considering the use of alternative data sources and studying its effects on system performance are upcoming tasks. Realaudio data, sometimes accompanied by some kind of transcription, can be obtained quite easily over the Internet. Its usage might speed up data collection efforts, which usually consume a lot of time. Examining the effects of compressed audio data might also provide new information for acoustic modeling.
- **Language normalization.** Further work on the language normalization techniques to conclude the process of unifying the different dialectic variants will lead to a homogeneous corpus and presumably better results.
- **Morpheme-based recognition.** Better combination of morpheme- and word-based recognition might help to increase the recognition accuracy for highly inflected languages such as Serbo-Croatian.
- **Segmentation and classification.** Automatic segmentation and classification of the audio stream might help to increase system performance in general, and the positive effects we can obtain from adaptation, in particular.
- **Confidence measure.** Experiments with a less restrictive confidence measure could also lead to results that favor the benefit from vocal tract normalization and adaptation as well.
- **Online recognizer.** The training of an online recognition system, which can be plugged into the Informedia system to enable spoken queries, is a desirable extension to the existing keyboard-based user interface for Serbo-Croatian. We already developed a first prototype, which has to be integrated into the system yet.

BIBLIOGRAPHY

[1] *Awde, N.,* Hippocrene Practical Dictionary: Serbo-Croatian-English/English-Serbo-Croatian, 1996.

[2] *Bakis, R., Chen, S., Gopalakrishnan, P., Gopinath, R., Maes, S., and Polymenakos, L.,* Transcription of Broadcast News Shows with the IBM Large Vocabulary Speech Recognition System. Proceedings of the 1997 DARPA Speech Recognition Workshop, Westfields International Conference Center, February 1997.

[3] *Beeferman, D., Berger, A., and Lafferty, D.* Text Segmentation Using Exponential Models. Proceedings of the Second Conference On Empirical Methods in NLP, Providence, RI, 1997.

[4] *Che, D. Yuk, S. Chennoukh, and J. Flanagan,* Development of the RU Hub-4 System. Proceedings of the 1997 DARPA Speech Recognition Workshop, Westfields International Conference Center, February 1997.

[5] *Christel, M., Kanade, T., Mauldin, M., Reddy, R., Sirbu, M., Stevens, S., and Wactlar, H.,* Informedia Digital Video Library. Communications of ACM, 38(4): 57-58 (1995).

[6] *Comrie, B. (Editor),* The World's Major Languages, Oxford University Press, August 1990, p.p. 391-409: *Greville Corbett,* Serbo-Croat.

[7] *Dumais, S. T., Furnasa, G. W., Landauer, T. K. and Deerwester, S.* "Using Latent Semantic Analysis To Improve Information Retrieval." In Proceedings of CHI: Conference on Human Factors in Computing, New York: ACM, 281-285, 1988.

[8] *Finke, M.,* Personal Communication.

[9] *Finke, M., Rogina, I., Woszczyna, M., Westphal, M., and Sloboda, T.,* The JanusRTk Tutorial, Interactive Systems Laboratories, Pittsburgh, PA, USA and Karlsruhe, Germany, 1993-1997.

[10] *Finke, M., Geutner, P., Hild, H., Kemp, T., Ries, K., Westphal M.,* The Karlsruhe-Verbmobil Speech Recognition Engine, IEEE International Conference on Acoustics, Speech and Signal Processing, Munich, 1997

[11] *Garofolo, John S., Fiscus, Jonathan G., Fisher, William M.,* Design and Preparation of the 1996 HUB-4 Broadcast News Benchmark Test Corpora, Proceedings of the 1997 DARPA Speech Recognition Workshop, Westfields International Conference Center, February 1997.

[12] *Gauvain, J.L., Adda, G., Lamel, L., Adda-Decker, M.,* Transcribing Broadcast News – The LIMSI Nov 1996 Hub-4 System. Proceedings of the 1997 DARPA Speech Recognition Workshop, Westfields International Conference Center, February 1997.

[13] *Geutner, P., Finke, M., and Scheytt, P.,* Adaptive Vocabularies for Transcribing Multilingual Broadcast News. To appear in Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP-98), Seattle, WA, USA, April 1998.

[14] *Geutner, P., Finke, M., Scheytt, P., Waibel, A., and Wactlar, H.,* Transcribing Multilingual Broadcast News Using Hypothesis Driven Lexicon Adaptation. To appear in Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne, February 1998.

[15] *Gvozdanović, J.,* Tone and Accent in Standard Serbo-Croatian (With a Synopsis of Serbo-Croatian Phonology), Österreichische Akademie der Wissenschaften, Vienna, 1980.

[16] *Hauptmann, A., and Wactlar, H.,* Indexing and Search of Multimodal Information. Submitted to International Conference on Acoustics, Speech and Signal Processing (ICASSP-97), Munich, Germany, April 1997.

[17] *Hauptmann, A., and Witbrock, M.,* Informedia News-On-Demand: Using Speech Recognition to Create a Digital Video Library, May 1997.

[18] *Hauptmann, A., Witbrock, M., and Christel, M.,* News-On-Demand – An Application of Informedia Technology. D-LIB Magazine, September 1995.

[19] *Jin, H., Kubala, F., Schwartz, R.,* Automatic Speaker Clustering. Proceedings of the 1997 DARPA Speech Recognition Workshop, Westfields International Conference Center, February 1997.

[20] *Kubala, R., Jin, H., Matsoukas, S., Nguyen, L., Schwartz, R., Makhoul, J.,* The 1996 BBN Byblos Hub-4 Transcription System. Proceedings of the 1997 DARPA Speech Recognition Workshop, Westfields International Conference Center, February 1997.

[21] *Kunzmann-Müller, B.,* Grammatikhandbuch des Kroatischen und Serbischen, P. Lang, Frankfurt am Main, New York, 1994.

[22] *Langenscheidt's Sprachführer:* Kroatisch und Serbisch, ISBN 3-468-22312-9.

[23] *Langenscheidt's Universal Dictionary:* English-Croatian/Croatian-English, 1988

[24] *Langenscheidt's Universal-Wörterbuch:* Kroatisch-Deutsch/Deutsch-Kroatisch, ISBN 3-468-18311-9.

[25] *Lee, L., and Rose, R. C.,* Speaker Normalization using Efficient Frequency Warping Procedures, International Conference on Acoustics, Speech and Signal Processing (ICASSP-96), Atlanta, GA, USA, May 1996, p.p. 353-356.

[26] *Leggetter, C.J, and Woodland, P.C.,* Speaker Adaptation of Continuous Density HMMs Using Multivariate Linear Regression, International Conference on Acoustics, Speech and Signal Processing (ICASSP-96), Atlanta, GA, USA, May 1996.

[27] *Leskien, A.,* Grammatik der serbokroatischen Sprache, I: Lautlehre, Stammbildung, Formenlehre, Carl Winter Verlag, Heidelberg, Germany, 1914.

[28] *Morton Benson,* English-Serbo-Croatian Dictionary, University of Pennsylvania Press, 1979.

[29] *Morton Benson,* Serbo-Croatian-English Dictionary, University of Pennsylvania Press, 1971.

[30] *Pallett, D. S. and Fiscus, J. G.,* 1996 Preliminary Broadcast News Benchmark Test. Proceedings of the 1997 DARPA Speech Recognition Workshop, Westfields International Conference Center, February 1997.

[31] *Placeway, P., Chen, S., Eskenazi, M., Jain, U., Parikh, V., Raj, B., Ravishankar, M., Rosenfeld, R., Seymore, K., Siegler, M., Stern, R., and Thayer, E.,* The 1996 Hub-4 Sphinx-3 System. Proceedings of the 1997 DARPA Speech Recognition Workshop, Westfields International Conference Center, February 1997.

[32] *Salton, G. Automatic Text Processing.* Addison-Wesley Publishing Company, Inc., Reading, MA, 1989.

[33] *Sankar, A. Stolcke, L. Heck, F. Weng,* SRI H4-PE System Overview. Proceedings of the 1997 DARPA Speech Recognition Workshop, Westfields International Conference Center, February 1997.

[34] *Scheytt, P., Geutner, P., and Waibel, A.,* Serbo-Croatian LVCSR on the Dictation and Broadcast News Domain. To appear in Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP-98), Seattle, WA, USA, April 1998.

[35] *Schultz, T., Westphal, M., and Waibel, A.,* The GlobalPhone Project Multilingual LVCSR with Janus-3. Multilingual Information Retrieval Dialogs: 2nd SQEL Workshop, Plzen, Czech Republic, April 1997.

[36] *Schwartz, R., Jin, H., Kubala, F., Matsoukas, S.,* Modeling Those F-Conditions – Or Not. Proceedings of the 1997 DARPA Speech Recognition Workshop, Westfields International Conference Center, February 1997.

[37] *Sheridan, P. and Ballerini, J.P.,* Experiments in Multilingual Information Retrieval Using the SPIDER System, Institute for Information Systems, ETH Zürich. SIGIR 1996.

[38] *Siegler, M. A., Jain, U., Raj, B., Stern, R. M.,* Automatic Segmentation, Classification and Clustering of Broadcast News Audio. Proceedings of the 1997 DARPA Speech Recognition Workshop, Westfields International Conference Center, February 1997.

[39] *Smith, M., and Kanade, T.,* Video Skimming and Characterization through Techniques in Language and Image Understanding. CMU School of Computer Science Technical Report CMU-CS-95-186R (revised 12/96).

[40] *Wegmann, S., McAllaster, D., Orloff, J. and Peskin, B.,* Speaker Normalization on Conversational Telephone Speech. International Conference on Acoustics, Speech and Signal Processing (ICASSP-96), Atlanta, GA, USA, May 1996, p.p. 339-341.

[41] *Woodland, P.C., Gales, M.J.F., Pye, D., and Young, S.J.,* Broadcast News Transcription Using HTK. Proceedings of the 1997 DARPA Speech Recognition Workshop, Westfields International Conference Center, February 1997.