SERBO-CROATIAN LVCSR ON THE DICTATION AND BROADCAST NEWS DOMAIN

Peter Scheytt[1,2], Petra Geutner[2], Alex Waibel[1,2]

Interactive Systems Laboratories
[1]Carnegie Mellon University, Pittsburgh PA, USA
[2]University of Karlsruhe, Karlsruhe, Germany
scheytt@cs.cmu.edu, pgeutner@ira.uka.de, waibel@cs.cmu.edu

ABSTRACT

This paper describes the development of a Serbo-Croatian dictation and broadcast news speech recognizer. The intention is to generate an automatic text transcription of a news show, which will be submitted to a multilingual Informedia database. We outline the complete system development process using the JanusRTk, beginning with data collection, design and training of parameters, tuning and evaluation. We report on general recognition techniques like segmentation, adaptation and language model interpolation, as well as language specific problems, e.g. high OOV rate due to inflected word forms. We show that even a low amount of acoustic training data, combined with Web based interpolated language models, is sufficient to build up a fairly reliable automatic news transcription system, which yields a performance of 36.0% WE.

1. INTRODUCTION

The goal of our work was to develop a speech recognizer for Serbo-Croatian broadcast news. The automatic transcription generated by our system will be submitted to a multilingual Informedia database. Thus adding not only diversity of information, but also the possibility of trans-linguistic queries and multiple language document retrieval. First we show the development of a Serbo-Croatian dictation system for read news text. We outline the data collection, show how we determined the speech recognizer's base parameters for this new language: Phones and phone classes, pronunciation dictionary and Web based language models. Even with a rather low amount of acoustic data and in short time, we were able to train a quite stable system, which yielded a performance of 28.8% WE. The dictation recognizer served as the baseline for our broadcast news system, which we developed on the combined training data of both tasks. We report the effects of Vocal Tract Length Normalization (VTLN) and MLLR Adaptation, which were not as favorable as observed by other groups for English broadcast news. We present the gains obtained by language model interpolation and describe the language specific problems, e.g. high OOV rate due to rapid vocabulary growth and inflected word forms. Again, the few training utterances were no obstacle to build up a system, which surprised us positively with a recognition performance of 36.0% WE.

2. DATA COLLECTION

2.1 Speech Recording and Transcription

The audio data for the dictation system was collected in Croatia and Bosnia-Hercegovina. 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 DAT-recorder at a sampling rate of 48 kHz in stereo quality and further sampled down to 16 kHz with 16 bit resolution in mono quality. The read utterances were checked against the original text to eliminate major errors and mark spontaneous effects. This data was originally collected for the GlobalPhone project at Karlsruhe University.

Spkrs.	Articles	Rec. Length	Words	Vocab.
85	**131**	**18 h**	**89 K**	**17 K**

Table 1: Dictation System Database

The broadcast news data was collected at Karlsruhe University in Germany. A satellite dish and a dedicated PC, equipped with an MPEG encoder board, were installed to record the HRT evening news show, which is transmitted from Croatia via the Eutelsat satellite. 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. Three native speakers transcribed the news broadcasts, using a software tool developed specifically for this task. Similar to the HUB4 corpus for English broadcast news data, the Serbo-Croatian recordings were divided into segments, in which the acoustic conditions remained constant, and tagged according to the speaker, the channel quality and the background noises. The different tags in these three categories are shown in Table 2, where "Non-Serbo-Croatian" identifies a person speaking in another language than Serbo-Croatian, most often English. In addition to these acoustic tags, only the most frequent and clearly audible spontaneous effects were transcribed: Hesitation, breathing, some other human and non-human noises. The diacritical letters in Serbo-Croatian were transcribed applying the rules in Table 3.

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

Table 2: Acoustic Tags

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

Table 3: Serbo-Croatian Diacritics

Transcription time varied between 13 and 18 hours per news broadcast (approx. 40 min.). 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 in Realaudio format and converted them to 16 kHz, 16 bit Wave format for speech processing.

Source	Broadcasts	Rec. Length	Words	Vocab.
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: Broadcast News System Database

2.2 Text Acquisition and Preparation

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 almost daily. We retrieved text data from 20 different sources (television and radio stations, newspapers and news agencies). During text preparation, we encountered one major problem: Many sites simply map diacritics onto their corresponding non-diacritical letter; i.e. ć and č become c, đ becomes d, š is replaced with s and ž with z, which is no problem for native speakers. For language modeling, however, we have to insert the diacritics at the right position. To accomplish this we collected as many Serbo-Croatian texts with diacritics as were available, not necessarily only news texts. From these texts we generated a list, L_{C}, of correct words, which served as reference to convert the second list, L_{F}, of both correct and false word forms, which were extracted from the texts without special characters. First, all words in L_F that did not contain the letters c, d, s and z were marked as correct. Then all words, which occurred in L_F and L_{C}, were labeled correct. For some words this might be wrong in certain situations; depending on the context, 'grada' converts into 'grada' (grad, gen. sg., town) or 'građa' (građa, nom. sg., material). A trigram model should help to improve the converting accuracy in such cases. In the next step, all remaining words in L_F were assigned to the their nearest neighbor in L_C. When the Levenstein editing distance exceeded a certain threshold, this word pair was thought to be valid and the necessary conversions were performed (Table 5).

No Diacritics	Diacritics	Conversion
afirmisanog	**afirmisanoj**	**NO**
africki	**afrički**	**c → č**

Table 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. In the test text, 25% of the words were converted incorrectly using this mechanism, which is better than just leaving the words as they are, which produces an error rate of 70%. Thus the combined conversion error rate of the whole algorithm on the test text was 5%.

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

Table 6: Internet Text Database

3. DICTATION SYSTEM

3.1 System Overview

We used the JanusRTk speech recognizer to build up the Serbo-Croatian dictation system. We determined a phone set and generated a pronunciation dictionary almost automatically. We bootstrapped a first context independent system with an existing German recognizer. Our first language model was based on the few training utterances. Even with the low amount of available data given, we were able to develop a chain of improving context dependent systems, applying techniques like vocal tract length normalization (VTLN), adaptation (MLLR) and multiple corpora language model interpolation.

3.2 Bootstrapping and Initial Systems

One principle of Serbo-Croatian is "Write as you speak, and speak as you write". Thus the phone set corresponds almost exactly to the alphabet. The set of phones used in the JanusRTk speech recognizer is based upon 30 language phones, 4 noise phones and the silence phone (Table 7). Each phone was modeled by a left-to-right HMM with 16 diagonal Gaussians. A linear discriminant analysis (LDA) was used to reduce the 43 feature dimensions to a 32 dimensional input vector. The pronunciation dictionary was created by an automatic grapheme-to-phoneme tool, whose mapping rules were also derived from Table 7. Some manual adjustments were necessary for numbers, abbreviations, foreign words and names.

SCR	A	B	C	C1	C5	D	D1	DZ5
Letter	A	B	C	Ć	Č	D	Đ	DŽ
GER	A	B	TS	TSCH	TSCH	D	TSCH	TSCH

SCR	E	F	G	H	I	J	K	L	LJ	M	N
Letter	E	F	G	H	I	J	K	L	LJ	M	N
GER	E	F	G	X	I	J	K	L	L	M	N

SCR	NJ	O	P	R	S	S5	T	U	V	Z	Z5
Letter	NJ	O	P	R	S	Š	T	U	V	Z	Ž
GER	N	O	P	R	S	SCH	T	U	V	Z	SCH

Phone	+QK	+hGH	+hBR	+nGN	SIL
Descr.	Garb.	Hmn. Noise	Breath.	Oth.Noise	Silence

Table 7: Serbo-Croatian Phone Set

The first context independent Serbo-Croatian dictation system was trained upon the labels generated by JanusRTk for a German scheduling task using “label boosting”. The Serbo-Croatian phones were initialized by their closest German equivalent (Table 7).

System	Data	Labels	Vocab.	OOV	LM	WE
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 8: Results for Context Independent Systems

The labels written by this first recognizer turned out to be more accurate and were used to train the next context independent system. Both recognizers used a single trigram language model based on the training transcriptions (Table 8).

3.3 Advanced Systems

We introduced 54 acoustic and articulatory phone classes, which guided the training of phonetically tied Gaussian mixtures on the base of pentaphones.

System	Data	Labels	Vocab	OOV	LM	WE
D-CD-0	**5.5 h**	**DCI-1**	**10 K**	**13.6%**	**D-LM-0**	**39.4%**
D-CD-1	**12.5 h**	**DCD-0**	**18 K**	**8.5%**	**D-LM-1**	**36.6%**

Table 9: Results for Context Dependent Systems

While the first polyphone system was still based on a low number of training utterances, more audio data was available for the second context dependent system. Together with an increased vocabulary and speaker dependent VTLN during training, the WE rate went down significantly (Table 9).
Some more tests with the same vocabulary were performed on the D-CD-1 system. They showed the positive effect of VTLN and MLLR adaptation during testing, especially when the amount of training data is limited. Interpolated language models from different corpora also helped improve the system performance. The system particularly benefits from language model interpolation when the acoustic performance is weaker (Table 10). See Table 11 for details on language model interpolation.

System	VTLN/MLLR	LM	WE
D-CD-1	**NO**	**D-LM-0**	**36.6%**
D-CD-1	**NO**	**D-LM-1**	**34.2%**
D-CD-1	**SPK. DEP.**	**D-LM-0**	**28.4%**
D-CD-1	**SPK. DEP.**	**D-LM-1**	**28.2%**

Table 10: Results for System D-CD-1 with VTLN, Adaptation and Interpolated LM

LM	Crpra.	Words	Single PP	PP
D-LM-0	**1**	**42 K**	**480**	**N/A**
D-LM-1	**1**	**98 K**	**313**	**N/A**
D-LM-2	**2**	**98 K/1.6 M**	**313/322**	**260**
B-LM-0	**1**	**2.5 M**	**426**	**N/A**
B-LM-1	**3**	**2.5 M/80 K/8 M**	**426/1598/998**	**281**
B-LM-2	**2**	**2.5M/220 K Classes**	**426/344**	**268**

Table 11: LM Interpolation

4. Broadcast News System

4.1 System Overview

The Serbo-Croatian broadcast news recognizer was based on the above described dictation system. We started with a context dependent system that was trained on the broadcast news data alone; later we also included the dictation data. We studied the effects of segment clustering for MLLR adaptation during testing. Several word and class based language model interpolations were performed. We observed a very strong vocabulary growth (Figure 1) and a high OOV rate.

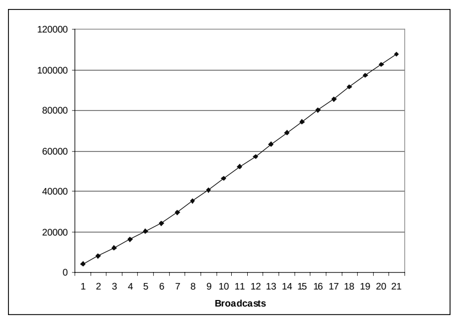

Figure 1: Vocabulary Growth

The reported results correspond with the PE (partitioned evaluation) test in the last HUB4 evaluation (December 1996), in which the segments and their constant acoustic properties were given for training and testing.

4.2 Bootstrapping and Initial System

In a baseline experiment we evaluated the performance of the dictation system D-CD-1 on the broadcast news test set. Due to the noisy conditions, even in the clean segments, the recognition performance dropped off. We used this system, however, to label our broadcast news data and trained a first recognizer B-CD-0 solely on 10 hours of transcribed recordings. This context dependent system was set up with 2 K codebook vectors over 24 input features. The vocabulary size was 29 K, the OOV rate 14.0% (Table 12).

System	Data	WE D-LM-1	WE D-LM-2	WE B-LM-0	WE B-LM-1
D-CD-1	**12.5 h (D)**	**75.9%**	**73.6%**	**N/A**	**N/A**
B-CD-0	**10.0 h (B)**	**N/A**	**N/A**	**45.2%**	**43.6%**

Table 12: Baseline Tests on the Broadcast News Data

We also observed the positive effect of language model interpolation. Its influence however reduced with the system becoming more stable and seemed to remain at an improvement level of about 1.5%, compared to a single language model. We applied three criteria do divide our text data into different corpora: Geographical origin (Serbia and Croatia), content source (television and radio stations on one side, newspapers and news agencies on the other) and language model perplexity. An interpolation of three different corpora yielded the best recognition results. We also noticed that the increased number of words we used to create the language models did not necessarily boost the recognition improvement.

4.3 Advanced System

The next system, B-CD-1, was developed on the combined dictation (12 h) and broadcast news data (13 h), with the latter occurring twice in the training set. We augmented the test vocabulary (31 K), which slightly reduced the OOV rate (13.5%), examined the effect of interpolated language models and also applied different adaptation techniques on the test data.

Classes	Portion of Data
male_clean_quiet	**28.0%**
male_clean_noisy	**20.0%**
male_dirty_quiet	**1.5%**
male_dirty_noisy	**7.5%**
female_clean_quiet	**20.0%**
female_clean_noisy	**12.5%**
female_dirty_quiet	**7.0%**
female_dirty_noisy	**0.5%**
nospeaker	**3.5%**

Table 13: Segment Clusters

For that reason the segments that shared the same acoustic conditions were assigned to one of the clusters in Table 13. The warping factor for VTLN and the adaptation were calculated on the hypothesis generated by the recognizer. The recognition gain by simply adapting over single segments was very low. Although observed for English broadcast news systems, adaptation on clustered segments, before adapting the single segments, did not improve recognition results.

System	MLLR	LM	WE
B-CD-1	**SNGL**	**B-LM-1**	**36.0%**
B-CD-1	**CLST+SNGL**	**B-LM-1**	**36.2%**
B-CD-1	**SNGL**	**B-LM-0**	**37.6%**
B-CD-1	**SNGL**	**B-LM-2**	**38.7%**

Table 14: Results for System B-CD-1 with Different Test Parameters

Serbo-Croatian is a strongly inflected language (Table 15), so the recognizer might be able to benefit from a class based language model, in which a class represents all inflections of one stem form. Although this approach reduced the language model perplexity, the overall system did not improve.

hotel **hotel**	**hotele, hotela, hotelu, hotelom, hotelu, hoteli, hotelima**
govoriti **to speak**	**govorim, govoriš, govori, govorimo, govorite, govore**

Table 15: Samples of Inflections

The number of inflections is one reason for the high OOV rate, the dominating factor in the WE rate. If we had a 64 K frequency based vocabulary, the Serbo-Croatian OOV would be 9.0%, for English broadcast an OOV rate of 0.6% was reported. This seems to be the most challenging problem to be solved in our future work.

5. SUMMARY

We described the development of a Serbo-Croatian dictation and broadcast news speech recognizer. We showed that the acoustic material of both domains can be combined to train a fairly reliable automatic transcription system, which generates output for a multilingual Informedia database. We showed the whole system development process, starting with data collection and design of the speech recognizer's base parameters for a completely new language. We examined the effects of segmentation, adaptation and language interpolation. Future work will concentrate on improvement of the system performance, by further attacking language specific problems, e.g. inflections and OOV rate.

Acknowledgements

This work was partly supported by the 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 authors and do not necessarily reflect the position or policy of the Government and no official endorsement should be inferred.

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

Thanks to Aleksandra Slavković, Sandra Yoon, Ljubomir Cvetković, Boris Tomaz and Manfred Weber.

Thanks to Michael Finke for his time, ideas and motivation. Without his work this paper would not have been possible.

6. REFERENCES

[1] Broadcast News Transcription Using HTK. *P.C. Woodland, M.J.F. Gales, D. Pye & S.J. Young.* Proc. of ARPA SR Workshop, Feb. 1997.

[2] The JanusRTk Switchboard/Callhome 1997 Evaluation System. *Michael Finke, Jürgen Fritsch, Petra Geutner, Klaus Ries and Torsten Zeppenfeld.* Proceedings of LVCSR Hub5-e Workshop, May 13-15, Baltimore, Maryland.