We are IntechOpen, the world's leading publisher of Open Access books
Built by scientists, for scientists

5,400
Open access books available

133,000
International authors and editors

165M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS
BOOK CITATION INDEX
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

11

Speech Recognition System of Slovenian Broadcast News

Mirjam Sepesy Maučec and Andrej Žgank
University of Maribor, Laboratory for Digital Signal Processing
Slovenia

1. Introduction

Speech is the most natural form of expression which is why it accounts for the majority of communication and information around the world. Media monitoring is a crucial activity today. For the most part today's methods are manual, with human reading, listening and watching, annotating topics and selecting items of interest for the user. The huge amount of data we can access nowadays in different formats (audio, video, text) and through distinct channels revealed the necessity to build systems that can efficiently store this data and retrieve this data automatically. Unavoidable component of such systems is speech recognition engine. Different types of speech and speech environments pose different challenges and, therefore, require different engines to accurately process the speech.

Speech recognition of broadcast news (BN ASR) is designed for news-oriented content from either television or radio and it readily processes broadcasts that include news, multi-speaker roundtable discussions and debates and even open-air interviews outside of the studio. BN ASR is a challenging task for many years and different languages. This chapter summarizes our key efforts to build BN ASR system for Slovenian language.

BN ASR system open the possibility for many applications where the use of automatic transcriptions is a major attribute. One of applications is live subtitling (Brousseau et al., 2003; Imai et al., 2000; Lambourne et al., 2004), were BN ASR system processes audio input and creates closed captions (Figure 1). Another task is speaker tracking, which can be used to find parts of speech belonging to specific speaker (Leggetter et al., 1995) in an audio input (Figure 2). Speech content search and retrieval is also a very useful functionality, which can be applied based on speech recognition. Based on some key terms a user can index audio/video to create a searchable repository to find the exact clip they need and its transcript. Yet another challenging field is topic detection and topic tracking. The goal is to use the system for continuously monitoring a TV channel, and searching inside their news programs for the stories that match the profile of a given user.

The chapter describes in detail our Speech Recognition System of Slovenian Broadcast News (UMB BNSI system), which is still under development. The chapter is organized as follows. First in section 2 we overview research work on broadcast news speech recognition. Properties of the Slovenian language make transcribing Slovenian broadcast news a more challenging task than for example English language. In section 3 basic differences are outlined. Section 4 summarizes the speech and text corpora used for training and testing the

system. Section 5 introduces the baseline UMB BNSI system. Section 6 describes advances based on recent improvements on the system. The experimental results are given in section 7. Finally, section 8 states some conclusions.

Fig. 1. Example of live subtitling using the BN ASR system.

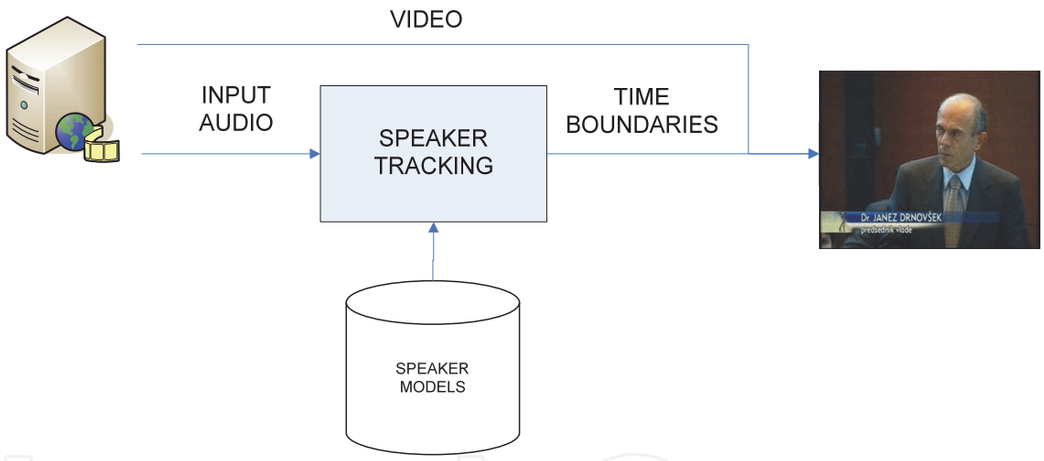

Fig. 2. Example of speaker tracking application designed with the BN ASR system.

2. Overview of research work on broadcast news speech recognition

Speech recognition has intrigued engineers and scientists for centuries. The problem of automatic speech recognition has been approached progressively. Based on major advances in statistical modeling of speech in the 1980s, automatic speech recognition systems have made considerable progress from then.

Broadcast News large vocabulary continuous speech recognition is one of the most challenging tasks today in the research field of language technologies. US agency DARPA was one of the key initiators in the area of Broadcast News system with the HUB campaigns. Several research groups took part in the HUB campaigns. The first experiments were performed for English language, thereafter also experiments for Spanish and Mandarin followed. Two main approaches to modeling BN ASR systems can be observed:

- increasing the complexity of system,
- increasing the amount of data for modeling.

The first approach resulted in increased processing times, therefore a dedicated faster subsystems (1xRT, 10xRT) were also developed. The main topic in increasing the complexity of the BN ASR system is how to model spontaneous speech, which is part of audio stream.

Analysis of disfluencies in spontaneous speech that shows their acoustic, prosodic and phonetic features influencing the speech recognition task were presented in (Batliner et al., 1995; Quimbo et al., 1998). First research work on spontaneous speech recognition was performed in (Godfrey et al., 1992; Stolcke et al., 1996). A set of various research works followed, focusing on improvements in different parts of spontaneous speech modeling (Siu et al., 1996; Peters et al., 2003; Stouten et al., 2003). Although permanent improvements were achieved modeling the spontaneous speech, the word error rate is still relatively high. Analyses of errors (Stouten et al., 2006; Rangarajan et al., 2006; Seiichi et al., 2007) that occur during the spontaneous speech recognition indicate the need for further development on this research topic.

The goal of increasing the amount of data available for training is particularly difficult for under-resourced languages. The majority of highly inflectional languages belong to this group. To overcome this problem, additional modeling approaches for highly inflectional languages must be integrated in the BN ASR system. The first research work on subword units for speech recognition in highly inflectional languages were performed for Serbian, Croatian and Czech language (Geutner et al., 1995; Byrne et al., 1999; Byrne et al., 2000). Different topologies of speech recognizer were implemented. Promising results were achieved during these tests. The first research work on subword units modeling for Slovenian language was presented in (Rotovnik et al., 2002). Further research work for Slovenian language followed in (Rotovnik et al., 2003). Achieved results showed that it is possible to achieve statistically significant improvements of results. It was indicated that an increase in the quality of acoustic models will be necessary to further improve the speech recognition results. Short subword units are very similar and consequently the confusability increases.

Another approach in modeling highly inflectional languages is based on increasing the number of words in the vocabulary (Nouza et al., 2004) or its adaptation (Geuntner et al., 1998). In the case when the first approach is used, the increased computational complexity is compensated with usage of simpler acoustic models, which may decrease the recognition results. When the second approach is used, very time consuming generation of new language models must be performed after each adaptation step.

The unsupervised and lightly supervised training of acoustic models was introduced in (Kemp et al., 1999; Lamel et al., 2002). The results confirm that such approach can be effectively used with automatically transcribed speech resources. Similarly effective results were observed, when discriminative training of acoustic models was incorporated (Woodland et al., 2000).

3. Speech recognition in highly inflected languages

Many techniques were first developed for English language and declared as language independent. Highly inflected languages make speech recognition a more difficult task in comparison to English due to their higher complexity (Maučec et al., 2004; Maučec et al., 2009). The concept of word formation is of great importance from the language modelling

point of view. The Slovenian language shares its characteristics to varying degrees with many other inflectional languages, especially the Slavic ones. In Slovenian, parts of speech (POS) are divided into two classes according to their inflectionality:

- the inflectional class: noun (substantive words), adjective (adjectival words), verb and adverb;
- the non-inflectional class: preposition, conjunction, particle and interjection.

Slovenian words often exhibit clearer morphological patterns in comparison with English words. A morpheme is the smallest part of a word with its own meaning. In order to form different morphological patterns (declinations, conjugations, gender and number inflections) two parts of a word are distinguished: a stem and an ending. There is one additional feature of the Slovenian language. Morphologically speaking some morphemes alternate in consonants, vowels and some in both simultaneously. Because of inflectionality, for Slovenian, approximately ten times larger recognition vocabulary is needed to assure the same text coverage as for English.

Word order in the Slovenian language does not play such an important role as in other languages (e.g. English language). The reason lies in the grammar of Slovenian language. There is a lot of grammatical information encoded in Slovenian words, which is in English language defined by the position in sentence. A simple sentence is presented as an example (Table 1). All six Slovenian word permutations form semantically logical sentences and are to be expected in spoken language. In contrast, English language does not support such freedom of word order choice. Therefore n-gram modeling, which is a standard in statistical language modeling, results in better language models for English language than for Slovenian language (Maučec et al., 2009).

Slovenian		English	
Maja študira angleščino.	✓	Maja studies English.	✓
Angleščino študira Maja.	✓	English studies Maja.	×
Študira Maja angleščino?	✓	Studies Maja English?	×
Študira angleščino, Maja?	✓	Studies English Maja.	×
Maja, angleščino študira.	✓	Maja English studies.	×
Angleščino Maja študira	✓	English Maja studies.	×

Table 1. Word permutations in Slovenian and English.

4. Speech and language resources

Speech and language resources are crucial in development of speech recognition systems. Speech databases are needed for acoustic modelling, and text databases for language modeling.

The main speech database used in our system was Slovenian BNSI Broadcast News (Žgank et al., 2005) speech database, which consists of two parts. The first part is speech corpus with transcriptions (BNSI-Speech) and the second part is text corpus (BNSI-text).

BNSI-Speech (Table 2) contains speech of news shows (evening news called TV Dnevnik, and late night news shows called Odmevi). It was captured in the archive of RTV Slovenia.

Speech signal has different acoustic properties (Figure 3) (Schwartz et al., 1997). Two most

frequent focus conditions in database are F0 (36.6%; read studio speech) and F4 (37.6%; read or spontaneous speech with background other than music). The high amount of F4 condition is caused by strict transcribers that very often assigned background, even if its level was very low. 16.2% of speech in the BNSI database is spontaneous (F1), while 6.0% is spoken in presence of background music (F3). Less than one hour of speech originates over the telephone channel (F2). Less than 0.1% of material was spoken by nonnative speakers (F5).

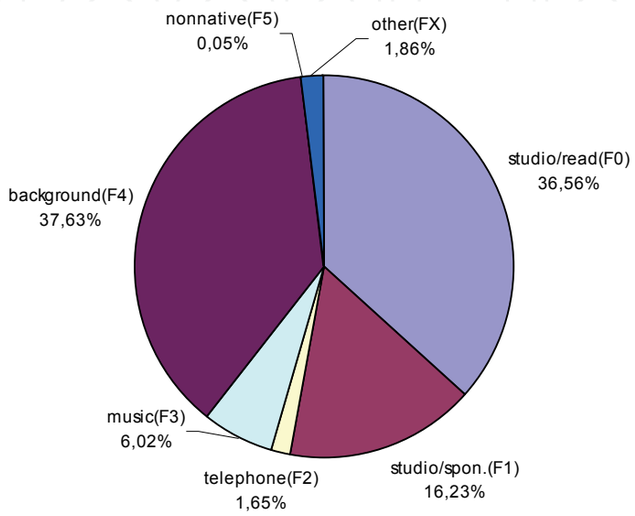

Fig. 3. Ratio of various focus conditions in the BNSI speech database.

The complete speech corpus consists of 36 hours of material (Table 2). The size of the training set is 30 hours. The next 3 hours are used for development set, which function is to fine tune the recogniser's parameters on it. The last 3 hours are used for evaluation set. The average length of a news show in the database is 51:22.

speech corpus:	BNSI-Speech
total length(h)	36
number of speakers	1565
number of words	268k

Table 2. Slovenian BNSI Broadcast News speech database

Table 3 shows some statistics of corpora used to train a language model. Transcriptions of BNSI-Speech corpus were used as the first database. This database was the smallest one. BNSI-text corpus is a collection of different TV scenarios. Some of scenarios were used by reporters and read from a teleprompter during a show. Both databases capture the characteristics of spoken language. Other two databases are collections of samples of written language. The Večer database is a collection of articles of newspaper Večer from 1998 till 2001. The largest database is FidaPLUS corpus (Arhar et al., 2007).

text corpus:	BNSI-Speech	BNSI-text	Večer	FidaPLUS
number of sentences	30k	614k	12M	46M
number of words	573k	11M	95M	621M
number of distinct words	51k	175k	736k	1.6M

Table 3. Slovenian text database

The material dates from the 1996 till 2006. The corpus is a composition of texts from different categories such as newspapers, magazines, books, the internet and other. Table 4 shows the proportion of different categories.

type:	percentage
internet	1.24%
books	8.74%
newspapers	65.26%
magazines	23.26%
other	1.5%

Table 4. Text variety in FidaPLUS corpus

FidaPLUS corpus is linguistically annotated and presented in the form of attributes of the element containing one corpus token. The information about all the possible lemmas and POS-tags is included in the corpus, together with the disambiguated single lemma and POS tag (see example in table 5). Although linguistic information is useful, it was not incorporated in language models discussed in this chapter.

excerpt from the corpus	translation to English
<w lemma="voditi" msd="Gppste--n-----n" lemmas="voditi voda vod" msds="Gppste--n-----n,Gpvsde--------n, Sozed,Sozem,Sozdi,Sozdt Sommi,Sommo" lemmass="voditi voda vod Voda" msdss="Gppste--n-----n,Gpvsde--------n, Sozed,Sozem,Sozdi,Sozdt Sommi,Sommo Slzed,Slzem"> vodi</w>	lead(V) lead(V), water(N), duct(N) lead(V), water(N), duct(N), Voda(NP)

Table 5. The verb …vodi… [to lead] taken from one sample sentence in the FidaPLUS corpus

We are modeling spoken language. There exist large amounts of written texts but we still lack adequate spoken language corpora. In our repository only two corpora are examples of spoken language (BNSI-Speech and BNSI-Text). Other two, Večer and FidaPLUS, are corpora of written language. It can be seen that the collection of texts is significantly diverse. Spoken sentences are short, and written sentences can be very large and complex. Word order in spoken language is much more relaxed than in written language (Duchateau et al., 2004 ; Fitzgerald et al., 2009 ; Honal et al., 2005). We discussed this phenomenon in previous section. Spoken sentences are often not grammatically correct. Written text is in most cases proof-read by professionals in a given language. Diversity of corpora should be taken into account when building a language model.

5. UMB BNSI baseline system

This section contains a description of the components in the UMB BNSI speech recognition system. The system is based on continuous density Hidden Markov Models for acoustic modelling and on n-gram statistical language models. It consists of three main modules, segmentation, features extraction, and decoding. The core module is a speech decoder, which needs three data sources for its operation: acoustic models, language model and lexicon. The block diagram of the baseline system is depicted in figure 4.

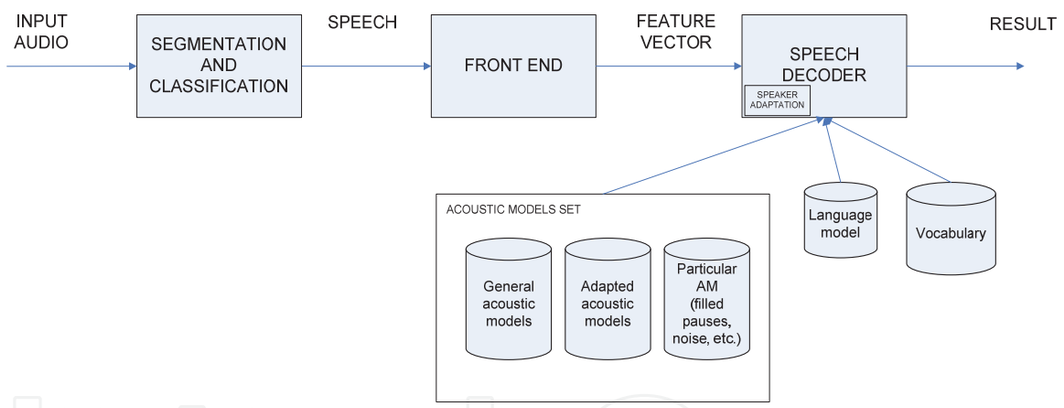

Fig. 4. Block diagram of experimental speech recognition system.

5.1 Segmentation

The main goal of the segmentation module is to produce homogeneous part of input audio stream. The Broadcast News topic can incorporate spoken material in adverse acoustic conditions. One of the most frequent cases is when is the journalist's voice mixed with background audio from the video segment. As a result of segmentation the homogeneous audio parts can be modeled with different acoustic models (wide-band vs. narrow-band), or even with complete separate speech recognition systems.

The three major segmentation criteria, which can be used in a Broadcast News speech recognition system, are:

- channel (narrow-band, wide-band),
- speech/silence/music/noise,
- gender (male, female, unknown).

Different methods can be used for acoustic segmentation: energy based, bandwidth based, Gaussian Mixture Models (GMM), Hidden Markov Models,… UMB BNSI system usually applies automatic acoustic segmentation based on multi-model GMM approach. In tests presented in this paper manual acoustic segmentation based on transcription files was used to exclude the influence of automatic acoustic segmentation on speech recognition results. Prior analysis showed that automatic acoustic segmentation decreases the speech recognition performance by approximately 2% absolute.

5.2 Feature extraction

Features are extracted from overlapping frames of homogeneous speech signal with duration of 32ms and frame shift of 10ms. Two different methods were used for frontend (i.e. feature extraction). The first one was based on mel-cepstral coefficients and energy (12 MFCC + 1 E, delta, delta-delta) and the second one was based on perceptual linear prediction (PLP). The size of baseline feature vector was 39 (Marvi, 2006). Also, the cepstral mean normalization was added to the MFCC feature extraction to reduce the influence of various acoustic channels (Maddi et al., 2006), which can be found in Broadcast News databases. This method significantly improved the speech recognition performance.

5.3 Acoustic modelling

The manually segmented speech material was used for training. This was necessary to exclude any influence of errors that could occur during an automatic segmentation procedure.

The developed baseline acoustic models were gender independent. The training of baseline acoustic models was performed using the BNSI Broadcast News speech database. The procedure was based on common solutions (Žgank et al., 2006). First the context independent acoustic models with mixture of Gaussian probability density function (PDF) were trained and used for force alignment of transcription files. In the second step, the context independent acoustic models were developed once again from scratch, using the refined transcriptions. The context-dependent acoustic models (triphones) were generated next.

The number of free parameters in the triphone acoustic models, which should be estimated during training, was controlled with the phonetic decision tree based clustering. The decision trees were grown from the Slovenian phonetic broad classes that were generated using the data-driven approach based on phoneme confusion matrix (Žgank et al., 2005a). Three final sets of baseline triphone acoustic models with 4, 8 and 16 mixture Gaussian PDF per state were generated. As some additional training data was won from the pool of outliers in comparison with the system described in (Žgank et al., 2008), additional training iterations were applied to context-dependent acoustic models. These transcriptions preprocessing steps showed significant improvement of log-likelihood rate per acoustic model according to an analysis.

5.4 Language modelling and vocabulary

The vocabulary contained the 64K most frequent words in all three corpora. The lowest count of a vocabulary word was 36. The out-of-vocabulary rate on the evaluation set was 4.22%, which is significantly lower than for some other speech recognition systems built for highly inflectional Slovenian language (Žgank et al., 2001; Rotovnik et al., 2007). A possible reason for this is the usage of text corpora with speech transcriptions for language modelling.

However in highly-inflected languages the number of possible word forms is very high. Many valid word forms are missing from the 64K vocabulary. If we enlarge the vocabulary, the complexity of a language model increases, which is demanding from a computational point of view. The vocabulary problem can be alleviated considerably by using sub-word units instead of words as basic vocabulary units. In our research this idea served as a starting point as well (Rotovnik et al., 2002), but did not bring any improvement in broadcast news domain.

Baseline language model was word-based bigram language model. All bigrams were included in the model. Katz back-off with Good-Turing discounting was used for smoothing. Language models were trained using SRI LM Toolkit (Stolcke, 2002).

A language model generated only from largest databases, Večer and FidaPLUS, would be too much adapted to the type of written language. When this language model is used in a UMB BNSI system it will not perform well. The sentences spoken in broadcast news do not match the style of the written sentences. A language model built only from Broadcast News transcriptions would probably be the most appropriate. The problem is that we do not have enough BN transcriptions to generate a satisfactory language model.

Baseline language model was built on first three text corpora. If we would merge all corpora into one big corpus, the influence of much smaller corpus of spoken language (BNSI-Speech) would be lost. Each text corpus was used for construction of one language model component. Individual components were then interpolated using BNSI-Devel set. The interpolation weights were: 0.26 (BNSI-Speech), 0.29 (BNSI-Text), and 0.45 (Večer). Final model contained 7.37M bigrams and resulted in perplexity of 410 on BNSI-Eval set.

5.5 Decoding

The standard one-pass Viterbi decoder with pruning and limited number of active models was used for speech recognition experiments in the next section. We applied additional fine tuning of decoder parameters on combined development set in comparison to the system described in (Žgank et al., 2008), to further improve the performance of speech recognition system.

The main characteristics of the baseline UMB BNSI system are summarized in table 6.

Baseline system	
Features extraction	MFCC, PLP
Features characteristics	window size: 32ms with 10ms frame shift
Acoustic model (AM)	inter-word context dependent trigraphemes
AM complexity	16 mixture Gaussian
Language model	interpolated bigram model
Vocabulary size	64000

Table 6. Characteristics of the baseline system

The baseline speech recognition system achieved 66.0% speech recognition accuracy when used with manual segmentation. This result is comparable to speech recognition system of similar complexity, which is used for highly inflected languages.

6. Improvements in the UMB BNSI system

This section describes recent improvements on the UMB BNSI system. The improvements in the area of acoustic modeling were mainly focused in the feature extraction module. MFCC and PLP feature vectors were used for all experiments, as they showed slightly different performance in various conditions. Beside the speech recognition accuracy also the decoding time can be significantly influenced by the feature vector type.
The influence of feature extraction characteristics on speech recognition performance was analyzed in the experiments. The characteristics observed were: frame length (32 ms versus 25 ms), size of filter bank (26 and 42) and number of MFCC coefficients (12 and 8). When acoustic models for the last two characteristics were developed, the clustering threshold for decision tree based clustering was modified to produce context dependent acoustic models of comparable complexity.
The main improvement in the language modeling procedure was introduction of FidaPLUS text corpus, which significantly increased the number of words in set. Having large text corpus makes transition from bigram to trigram reasonable.

7. Results of comparative experiments

Bigram and trigram language models were built. Independent language model components were constructed, using each database in separation for counting n-grams. If we will use all corpora together as one huge training corpus, the statistical dependencies typical for spoken language and represented by first two corpora, will be weaken by dependencies typical for written language and expressed by much larger training material. In each component Katz back-off with Good-Turing discounting was used for smoothing. Experiments with modified Kneser-Ney smoothing were also performed, but did not bring any improvements. Individual components were then interpolated using the BNSI-Devel corpus of 4 broadcast shows. Optimal interpolation weights for the corresponding 4 models were iteratively computed to minimize the perplexity of an interpolated model on BNSI-Devel corpus. Two interpolated models were build, bigram and trigram models. Table 7 contains interpolation weights for both of them.

component:	bigram	trigram
BNSI-Speech	0.20	0.18
BNSI-Text	0.28	0.24
Večer	0.15	0.12
FidaPLUS	0.37	0.46

Table 7. Interpolation weights for bigram and trigram models.

Perplexity on BNSI-Eval set of final bigram model was 359, and the perplexity of trigram model was 246. The number of bigrams redoubled in comparison to baseline system. As the result of adding the fourth language component the perplexity of bigram model improved by 12%. In trigram model 33.6M trigrams were added. Transition from bigram to trigram model brought 40% of improvement in perplexity. The transition was reasonable because of

the size of FidaPLUS corpus. At the same time the language model increased in size and slows down the decoding process.
Several experiments were performed to evaluate the improvements introduced in the UMB BNSI system. The first test was focused on evaluation of using MFCC or PLP feature extraction module in combination with the trigram language models (see Table 8). The results of bigram language models were used as a baseline value.

system:	Correct [%]	Accuracy [%]
bigram, MFCC	69.0	65.7
bigram, PLP	69.6	66.0
trigram, MFCC	70.7	67.5
trigram, PLP	71.4	68.0

Table 8. Recognition results obtained with bigram and trigram language models and MFCC and PLP features.

The more complex trigram language models improved the speech recognition performance by approximately 2%. The accuracy increased from 65.7% to 67.5% when MFCC feature extraction was used and from 66.0% to 68.0% when PLP feature extraction was applied. The disadvantage of using trigram language models is the increased complexity of speech recognition system, which results in increased decoding time.
The second evaluation step was focused on including the FidaPLUS text corpus to language modeling. The results are presented in table 9.

system:	Correct [%]	Accuracy [%]
bigram1, MFCC	70.0	67.4
trigram1, MFCC	73.6	71.0
bigram2, MFCC	60.9	57.7
trigram2, MFCC	73.4	71.1

Table 9. Speech recognition results with language models, improved with FidaPLUS text corpus.

The first type (bigram1, trigram1) of language models in table 9 was built in such a way that FidaPLUS text corpus was added to other baseline text corpora. In the second type (bigram2, trigram2), the FidaPLUS was added, but the Večer text corpus was deleted from the set as it is already included in the FidaPLUS corpus in great extent. The inclusion of FidaPLUS text corpus significantly improved the speech recognition results. The accuracy was increased by 3.6% absolute from 67.5%to 71.1%. In case of these experiments the speech decoder's vocabulary was identical for all four cases. This is the probable cause for the degraded speech recognition performance in case of bigram2 set. In this set the Večer text corpus was excluded from building the language models, but words from this corpus were still present in the lexicon. The frequencies of bigrams from Večer as subcorpus in the FidaPLUS text

corpus were not high enough to significantly influence the probabilities in the resulting bigram2 language model.

system:	Correct [%]	Accuracy [%]
bigram1, MFCC, 32ms	70.0	67.4
bigram1, MFCC, 25ms	70.4	67.8
bigram1, PLP, 32ms	70.9	68.0
bigram1, PLP, 25ms	70.5	67.7
trigram1, MFCC, 32ms	73.6	71.0
trigram1, MFCC, 25ms	73.5	71.0
trigram1, PLP, 32ms	73.9	70.9
trigram1, PLP, 25ms	73.6	70.7

Table 10. Speech recognition results using different types of feature extraction and various frame lengths.

The table 10 shows comparison between two different feature extraction frame lengths – baseline 32 ms and 25 ms. There is a small difference between comparable configurations (feature extraction type, language models) for two frame lengths, but it is statistical insignificant.

Various feature extraction configurations were used in combination with the bigram1 and trigram1 language models. The evaluation results are presented in table 11. The increased number of filters in filter bank decreased the speech recognition performance by 0.5% (bigrams) and 0.3% (trigrams). When only 8 mel-cepstral coefficients (8+1 case in table 11) were used, the accuracy decreased, as it was anticipated. The decrease was 4.0% with bigram language model and 3.6% with trigram language model. The advantage with using this configuration was the reduced decoding time, due to lower feature complexity. When bigram language models were applied the decoding time decreased by approximately 16%. The decrease with the trigram language models was approximately 19%. Such faster configuration with decreased accuracy can be successfully included in a speech recognition system with two iterations.

system:	Correct [%]	Accuracy [%]
bigram1, MFCCm	70.9	68.1
bigram1, MFCCm, FB42	70.3	67.6
bigram1, MFCCm, 8+1	66.0	64.1
trigram1, MFCCm	74.0	71.3
trigram1, MFCCm, FB42	73.7	71.0
trigram1, MFCCm, 8+1	69.7	67.7

Table 11. Speech recognition evaluation for various feature extraction configurations.

8. Conclusion

Speech recognition of Broadcast News is a very difficult and resource demanding task. The development of UMB BNSI system is a long-continued project.

The chapter described statistically significant improvements of UMB BNSI system. The analysis of speech recognition results showed the importance of acoustic and language models in speech recognition systems in broadcast news domain. A significant effort was devoted to reducing the complexity of the system. We succeeded to speed up the system by small loss of accuracy to prepare the system for the second pass with lattice rescoring. Our results suggest that these methodologies are well suited to the challenges presented by the Broadcast News domain.

The future work will be focused on implementing a second iteration of speech recognition with increased complexity. The analysis of results namely showed the possibility of overtraining in some evaluation steps, when only one speech recognition iteration was carried out.

We are still far from perfect recognition, the ultimate goal, nevertheless our current technology is able to drive a number of very useful applications, where perfect recognition is not needed, for example audio archive indexing.

9. Acknowledgements

The work was partially funded by Slovenian Research Agency, under contract number P2-0069, Research Programme "Advanced methods of interaction in telecommunication".

10. References

Arhar, Š., Gorjanc, V., (2007). Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. Jezik in slovstvo 52/2., 95--110.

Batliner, A., Kiessling, A., Burger, S., Nöth, E., (1995). Filled pauses in spontaneous speech. *In Proc. International Congress of Phonetic Sciences*, Stockholm, Sweden.

Brousseau, J., Beaumont, J.F., Boulianne, G., Cardinal, P., Chapdelaine, C., Comeau, M., Osterrath, F., Ouellet, P., (2003). Automatic Closed-Caption of Live TV Broadcast News in French, *Proc. Eurospeech 2003*, Geneva, Switzerland.

Byrne, W., Hajic, J., Ircing, P., Khudanpur, F., McDonough, J., Peterek, N., Psutka, J.(1999). Large vocabulary speech recognition for read and broadcast Czech, *Proc. Workshop on Text Speech and Dialog*, Plzen, Czech Republic, 1999, Lecture Notes in Artificial Intelligence, Vol. 1692

Byrne W., Hajič J., Ircing P., Krbec P. in Psutka J. (2000). Morpheme Based Language Models for Speech Recognition of Czech, *TSD 2000*, 2000.

Duchateau, J., T. Laureys, P. Wambacq, (2004). Adding Robustness to Language Models for Spontaneous Speech Recognition, *In Proc. ISCA Workshop on Robustness Issues in Conversational Interaction*, Norwich, UK.

Fitzgerald, E., K. Hall, F. Jelinek, (2009). Reconstructing False Start Errors In Spontaneous Speech Text. *In Proc. of the 12th Conference of the European Chapter of the ACL*, pp.255–263, Athens, Greece.

Geuntner P. (1995). Using Morphology Towards Better Large-Vocabulary Speech Recognition Systems, *ICASSP*, pp. 445-448, Detroit, 1995.

Geuntner P., Finke M., Scheytt P., Waibel A. in Wactlar H.(1998). Transcribing Multilingual Broadcast News Using Hypothesis Driven Lexical Adaptation, *DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, 1998.

Godfrey, J., Holliman, E., McDaniel, J., (1992). Switchboard: Telephone speech corpus for research and development. *In: Proc. International Conference on Acoustics, Speech and Signal Processing*, Vol. I, San Francisco, USA, pp. 517–520.

Honal, M., T. Schultz, (2005). Automatic Disfluency Removal On Recognized Spontaneous Speech - Rapid Adaptation To Speaker-Dependent Disfluencies. *Proc. of ICASSP, 2005*, vol 1, pp. 969-972.

Imai, T., Kobayashi, A., Sato, S., Tanaka, H., and Ando, A., (2000). Progressive 2-pass decoder for real-time broadcast news captioning, *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp.1937-1940, Istanbul, Turkey.

Kemp, T., Waibel, A., (1999). Unsupervised Training Of A Speech Recognizer: Recent Experiments, *In Proc. Eurospeech 1999*, Budapest, Hungary.

Lambourne, A., J. Hewitt, C. Lyon, S. Warren, (2004). Speech-Based Real-Time Subtitling Services, *International Journal of Speech Technology*, Vol. 7, Issue 4, pp. 269-279.

Lamel, L., Gauvain, J., and Adda, G., (2002). Lightly supervised and unsupervised acoustic model training, *Computer Speech and Language*, Volume 16, Issue 1, , January 2002, 115--129.

Leggetter, Woodland, (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Comput. Speech Lang*. v9 i2. 171-185.

Maddi, A., A. Guessoum, D. Berkani, (2006). Noisy Speech Modelling Using Recursive Extended Least Squares Method, *WSEAS TRANSACTIONS on SIGNAL PROCESSING*, Issue 9, Volume 2, September 2006.

Marvi, H., (2006). Speech Recognition Through Discriminative Feature Extraction, *WSEAS TRANSACTIONS on SIGNAL PROCESSING*, Issue 10, Volume 2, October 2006.

Maučec, M. S., Kačič, Z., Horvat, B. (2004). Modelling highly inflected languages. *Inf. sci.*, Oct. 2004, Issue 1/4, Volume 166, pp. 249-269

Maučec, M. S., Rotovnik, T., Kačič, Z., Brest, J.(2009). Using data-driven subword units in language model of highly inflective Slovenian language. *Int. j. pattern recogn. artif. intell.*, Mar. 2009, Volume 23, Issue 2, pp. 287-312.

Nouza, J., Nejedlova, D., Zdansky, J., Kolorenc, J.,(2004). Very Large Vocabulary Speech Recognition System for Automatic Transcription of Czech Broadcast Programs, *Proc. ICSLP 2004*, Jeju Island, Korea.

Peters, J., (May 2003). Lm studies on filled pauses in spontaneous medical dictation. *In: Proc. Human Language Technology conference/North American Chapter of the Association for Computational Linguistics Annual Meeting,* Edmonton, Canada, pp. 82–84.

Quimbo, F.C., Kawahara, T., Doshita, S., (1998). Prosodic analysis of fillers and self-repair in Japanese speech. *In: Proc. International Conference on Spoken Language Processing*, Sydney, Australia, pp. 3313–3316.

Rangarajan, V., S. Narayanan, (2006). "Analysis of disfluent repetitions in spontaneous speech recognition", *Proc. EUSIPCO 2006*, Florence, Italy.

Rotovnik T., Maučec M. S., Horvat B., Kačič Z., (2002). Large vocabulary speech recognition of Slovenian language using data-driven morphological models, TSD 2002.

Rotovnik T., Maučec M. S., Horvat B., Kačič Z., (2003). Slovenian large vocabulary speech recognition with data-driven models of inflectional morphology, *ASRU 2003*, U.S. Virgin Islands, 2003. pp. 83-88, 2003.

Rotovnik T., Maučec, M. S., Kačič Z., (2007). Large vocabulary continuous speech recognition of an inflected language using stems and endings, *Speech communication*, 2007, vol. 49, iss. 6, pp. 437-452.

Schwartz, R., H. Jin, F. Kubala, and S. Matsoukas, (1997). Modeling those F-Conditions - or not, in *Proc. DARPA Speech Recognition Workshop 1997*, pp 115-119, Chantilly, VA.

Seiichi, N., K. Satoshi, (2007). Phenomena and acoustic variation on interjections, pauses and repairs in spontaneous speech, *The Journal of the Acoustical Society of Japan*, Vol.51, No.3, pp. 202-210.

Siu, M., Ostendorf, M., (1996). Modeling disfluencies in conversational speech. *In: Proc. International Conference on Spoken Language Processing,* Vol. I, Atlanta, USA, pp. 386–389.

Stolcke, A., Shriberg, E., (1996). Statistical language modeling for speech disfluencies. In: *Proc. International Conference on Acoustics, Speech and Signal Processing,* Vol. I, Atlanta, USA, pp. 405–408.

Stolcke, A. (2002). *SRILM an Extensible Language Modeling Toolkit*. Proc. of the ICSLP, Denver, Colorado, September 2002.

Stouten, F., Martens, J.-P., (2003). A feature-based filled pause detection system for Dutch. *In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Virgin Islands, USA, pp. 309–314.

Stouten, F., J. Duchateau, J.P. Martens, P. Wambacq, (2006). "Coping with disfluencies in spontaneous speech recognition: Acoustic detection and linguistic context manipulation". *Speech Communication* 48(11): 1590-1606.

Woodland, P.C., Povey, D.,(2000). Large Scale Discriminative Training For Speech Recognition, *In ISCA ITRW Automatic Speech Recognition: Challenges for the Millennium,* pages 7-16, Paris, 2000.

Žgank, A., Kačič, Z., Horvat, B, (2001). Large vocabulary continuous speech recognizer for Slovenian language. *Lecture notes computer science, 2001*, pp. 242-248, Springer Verlag.

Žgank, A., Horvat, B., Kačič Z., (2005). Data-driven generation of phonetic broad classes, based on phoneme confusion matrix similarity. *Speech Communication* 47(3): 379-393.

Žgank, Andrej, Verdonik, Darinka, Zögling Markuš, Aleksandra, Kačič, Zdravko, (2005). BNSI Slovenian broadcast news database - speech and text corpus, *9th European conference on speech communication and technology*, September, 4-8, Lisbon, Portugal. Interspeech Lisboa 2005, Portugal.

Žgank, A., Rotovnik, T., Maučec Sepesy, M., Kačič Z., (2006). Basic Structure of the UMB Slovenian Broadcast News Transcription System, *Proc. IS-LTC Conference*, Ljubljana, Slovenia.

Žgank, A., Rotovnik, T., Sepesy Maučec, M., (2008). Slovenian Spontaneous Speech Recognition and Acoustic Modeling of Filled Pauses and Onomatopoeas, *WSEAS TRANSACTIONS on SIGNAL PROCESSING*, Issue 7, Volume 4, July 2008.

IntechOpen

IntechOpen

Speech Technologies

Edited by Prof. Ivo Ipsic

ISBN 978-953-307-996-7
Hard cover, 432 pages
Publisher InTech
Published online 23, June, 2011
Published in print edition June, 2011

This book addresses different aspects of the research field and a wide range of topics in speech signal processing, speech recognition and language processing. The chapters are divided in three different sections: Speech Signal Modeling, Speech Recognition and Applications. The chapters in the first section cover some essential topics in speech signal processing used for building speech recognition as well as for speech synthesis systems: speech feature enhancement, speech feature vector dimensionality reduction, segmentation of speech frames into phonetic segments. The chapters of the second part cover speech recognition methods and techniques used to read speech from various speech databases and broadcast news recognition for English and non-English languages. The third section of the book presents various speech technology applications used for body conducted speech recognition, hearing impairment, multimodal interfaces and facial expression recognition.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Mirjam Sepesy Maucec and Andrej Zgank (2011). Speech Recognition System of Slovenian Broadcast News, Speech Technologies, Prof. Ivo Ipsic (Ed.), ISBN: 978-953-307-996-7, InTech, Available from: http://www.intechopen.com/books/speech-technologies/speech-recognition-system-of-slovenian-broadcast-news

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2011 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License, which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.

IntechOpen

IntechOpen