Slovenian Spontaneous Speech Recognition and Acoustic Modeling of Filled Pauses and Onomatopoeas

ANDREJ ŽGANK, TOMAŽ ROTOVNIK, MIRJAM SEPESY MAUČEC
Laboratory for Digital Signal Processing
University of Maribor
Smetanova ul. 17, SI-2000 Maribor
SLOVENIA
andrej.zgank@uni-mb.si http://www.dsplab.uni-mb.si

Abstract: - This paper is focused on acoustic modeling for spontaneous speech recognition. This topic is still a very challenging task for speech technology research community. The attributes of spontaneous speech can heavily degrade speech recognizer's accuracy and performance. Filled pauses and onomatopoeias present one of such important attributes of spontaneous speech, which can give considerably worse accuracy. Although filled pauses don't carry any semantic information, they are still very important from the modeling perspective. A novel acoustic modeling approach is proposed in this paper, where the filled pauses are modeled using the phonetic broad classes, which corresponds with their acoustic-phonetic properties. The phonetic broad classes are language dependent, and can be defined by an expert or in a data-driven way. The new filled pauses modeling approach is compared with three other implicit filled pauses modeling methods. All experiments were carried out using a context-dependent Hidden Markov Models based speech recognition system. For training and evaluation, the Slovenian BNSI Broadcast News speech and text database was applied. The database contains manually transcribed recordings of TV news shows. The evaluation of the proposed acoustic modeling approach was done on a set of spontaneous speech. The overall best filled pauses acoustic modeling approach improved the speech recognizer's word accuracy for 5.70% relatively in comparison to the baseline system, without influencing the recognition time.

Key-Words: - speech recognition, acoustic modeling, filled pauses, onomatopoeas, Slovenian spontaneous speech, broadcast news, HMM

1 Introduction

In the today's world are speech recognition systems gathering on importance. There are a large number of different applications, which were transferred from laboratories into real-life environment. Many of them serve in different areas, where speech recognition systems are the only possible solution for providing content to users:

- Information Retrieval: public or commercial content [1, 2], healthcare, government, universities, media, public transport, directory enquires [3], Web browsing, ...
- Entertainment: gambling, dating, ring-tones, chat, games, ...
- Retail, Advertising: shopping, ticketing services [4], surveys, order status inquires, shipment status, ...
- Financial services: banking, stock exchange services, insurance, currency exchange inquires,...
- Messaging: Voice Mail, Unified Messaging, Unified Communication, ...
- Communication : communication portals, Auto-attendant, PBX provisioning, Call Centers, CRM,...

The quality of speech recognition hardly depends on the recognition task [5, 6] and on the availability of spoken and written language resources for particular language. The building of a new language resource is usually very expensive, due to the manual work, which is needed during the production process. The complexity of recognition task is mainly influenced by the speaking style and the type of vocabulary and language model used in system. According to these parameters, the following categories can be defined for a speech recognition system:

- Isolated speech recognition: each word is pronounced in isolated manner, a typical vocabulary has size of 10 – 100 words. Example: simple voice driven telecommunication applications.
- Connected speech recognition: words are pronounced in a connected manner, but the size of the vocabulary is still similar to the first

category. Example: recognition of connected digits (PIN, telephone number) [7] or simple directory enquire services.
- Large vocabulary continuous speech recognition: user can speak almost freely (e.g.: pronounce planned whole sentences), the vocabulary has size of several 10k words. Example: dictation application on a modern desktop computer.
- Spontaneous speech recognition: user can speak freely, without planning his/her speech. The vocabulary has size of several 10k words. Example: speech-to-speech translation system, "how can I help you?" telecommunication services.

As can be seen from the above categories, spontaneous speech presents the most challenging task for a speech recognition system, where significant degradation of speech recognition performance can occur. Additional problem can present the supported language, as some of them are more complex from the speech recognizer's point of view. The most frequently supported language in speech recognisers is English, where in the majority of cases already the standard approaches yield good results. On the other side are languages with characteristics, which require additional processing methods [8, 9]. Some of such languages are:

- Agglutinative languages: Turkish, Finnish, Hungarian …
- Tonal languages: Mandarine …
- Highly inflectional languages: Slovenian, Czech…

A vocabulary of a speech recogniser for highly inflectional language must be approx. 10 times larger than for an English speech recognition system to achieve the same out-of-vocabulary rate on the test set. Such large vocabulary decreases the accuracy and speed of a speech recognition system. Slovenian, which is being included in experiments in this paper, belongs to the group of highly inflectional languages with a relatively free word order. Free word order additionally decreases the performance of an N-gram based language model.

One of possible challenging applications for spontaneous speech recognition is real-time subtitling for live TV broadcasts. Such systems are of immense importance for deaf and partially deaf people, as they enable them to follow the current events in real-time. In a typical broadcast news show, approximately 50% to 75% of stories can be automatically subtitled using closed caption generated from the scripts (Figure 1).

SLAVKO
PRIJETI PREPRODAJALCI OROŽJA, POKI V MESTU POLICIJSKI ZVOČNI EFEKTI
T- LJUBLJANA, ATENE
X- Olimpijske igre
EDITA
ZA PROMOCIJO SLOVENIJE V ATENAH DESETKRAT MANJ DENARJA KOT V SYDNEYJU
02_Nap 0(Sprejem Goričanov K2 (Čurlič_B 3 ____ NL_0:23_6:59:50 38"_Stat_Aired ____SPREMENIL: haskaj_____KDAJ: 08/05/04 18:37:57_T- EDITA M. CETINSKI
X- SLAVKO BOBOVNIK
T- NOVA GORICA /zg.
K2-DVOPLAN (SLAVKO) Dober dan, cenjene gledalke in spoštovani gledalci.
K2-DVOPLAN (EDITA) Lepo pozdravljeni.
SLAVKO
Upajmo, da se bomo tako, kot so se danes veselili Novogoričani, BETA veselili tudi mi, ko se bodo naša dekleta in fantje vračali iz Aten. Na sprejemu slovenskih nogometnih prvakov, ki so včeraj s kar pet proti nič premagali danske prvake, se je danes zagotovo zbralo kakih 1000 ljudi.

03_izjava župana ____TONSKO_B 4 ____*_0:09_7:00:13__Stat_Aired ____SPREMENIL: golob_____KDAJ: 08/05/04 18:37:38_T- MIRKO BRULC
X- župan Mestne občine Nova Gorica
KONČA: ... Dragi Novogoričani. 30. aprila je *Evropa gledala* v Novo Gorico.
Danes pa po vaši zaslugi zopet gleda v Novo Gorico.
04_Nap 1 _____K2 ____ _ ____NL_0:23_7:00:22__Stat_Aired ____SPREMENIL: nakrst_____KDAJ: 08/05/04 18:42:28_K2, DVOPLAN
SLAVKO
O podrobnostih v športu, saj so prve minute Dnevnika namenjene precej manj prijetnim dogodkom.
EDITA
Na Slovenskem zunanjem ministrstvu so povedali, da so po njihovem mnenju navedbe v hrvaški diplomatski noti napačne in da so bili naši policisti ob nedavnih incidentih v Piranskem zalivu v vodah pod slovenskim nadzorom.
SLAVKO
Odgovora na noto pa na ministrstvu še niso napisali.

Figure 1: Fragment of a script from an evening TV news show, which is part of the Slovenian BNSI Broadcast News text database.

The remaining part isn't covered, as it contains live conversations (e.g. interviews, talk shows), where closed captions can't be generated from scripts or scenarios. Such part of script is shown on Figure 1, denoted as section *"03 izjava župana"*, where only the last few seconds are transcribed as guideline for the director. To cover these parts of a broadcast, applications using dedicated keyboards or speech recognition systems must be used.

There are two main approaches, how a speech recognition system can be used for an on-line subtitling system:

- Respeaking: the core of such subtitling system is a speaker dependent speech recogniser. Of-the-shelf dictation systems can be used for this task. Highly trained operator is constantly monitoring the TV show and respeaking each pronounced sentence into the system [10]. Usually some simplifications of the spoken content are necessary to achieve the required speed and accuracy. Such system can be found in several major broadcast companies all over the world.
- Automatic subtitling: the core is a fully automatic speech recognition system, which automatically monitors the input audio stream, and recognizes the pronounced sentences (Figure 2). Usually several processing passes are needed to achieve good performance. Such subtitling systems were proposed in [11, 12].

There are also other very challenging applications for spontaneous speech recognition systems, such as indexing of audio and video material, meeting transcriptions, speech-to-speech translation, etc.

Figure 2: Screenshot of a speech recognition based TV subtitling demo application for Slovenian language.

Speech recognition for such spontaneous conversations is a very complex task. There are three major types of disfluencies in spontaneous speech that influence the performance of any spontaneous speech recognition application:

- Filled pauses (FP): short words, which appear as interjection – e.g.: uh, aaa. They are language dependent.
- Word repetitions: disfluencies used by the speaker to gain time before continuing with the sentence.
- Sentence restarts: speaker pronounces the initial part of a sentence and then starts over again with a new initial part.

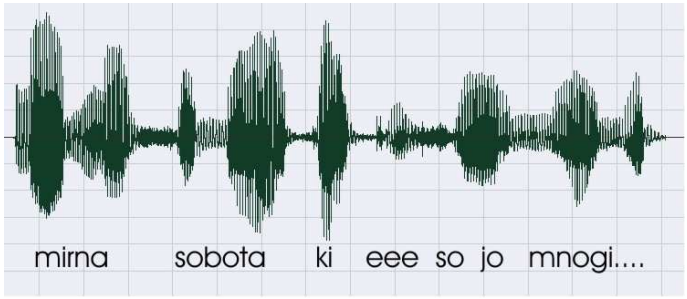

Figure 3: Spontaneous sentence from Slovenian BNSI Broadcast News speech database.

Figure 3 shows example of spontaneous sentence ("mirna sobota ki ee so jo mnogi") from Slovenian BNSI Broadcat News database. The shown sentence encompasses one filled pause – *"eee"*. The ratio of disfluencies in spontaneous speech hardly depends on the situation. In case when the speech is prepared in advance, there are far less disfluencies than in case of spontaneous speech in everyday situation.

This paper focuses on acoustic modeling of filled pauses and onomatopoeias for spontaneous speech recognition system. As the first ones are far more frequent in speech, but both categories are very similar from acoustic modeling point of view, we combined the two categories into one common category, called filled pauses. Although filled pauses and onomatopeas don't carry any true semantic information, it is still necessary to include them in modeling for speech recognition. Each filled pause disrupts the sequence of words, which is estimated with the acoustic and language model and so influences the overall accuracy of speech recognition system. In addition, disfluencies in spontaneous speech are often indicators of turn taking in a dialog, and can be as such used for dialog management in voice driven telecommunication services.

Several authors presented various methods for acoustic modeling of filled pauses. One major group of modeling approaches is based on some type of Gaussian Mixture Models (GMM) [13, 14, 15], whilst the other major group uses Hidden Markov Models (HMM) [16, 17, 18]. We propose a novel acoustic modeling approach for filled pauses in spontaneous speech recognition with HMM models based on phonetic broad classes. The proposed modeling approach will be compared with other methods for implicit acoustic modeling of filled pauses using the Slovenian speech database for the broadcast news domain.

The paper is organized as follows: the proposed method and other approaches for acoustic modeling of filled pauses are introduced in Section 2. The speech and text databases needed during the experiments are described in Section 3. The experimental design used for evaluation is presented in Section 4. Section 5 contains the results of the speech recognition experiments, while the conclusion and directives for future work are given in Section 6.

2 Spontaneous speech and modeling of filled pauses and onomatopoeas

There are two different types of filled pauses acoustic modeling from the speech recognizer's point of view. In the first case filled pauses are detected using an external module (e.g. GMM classification [13]), and speech recognizer than process only the part of speech without filled pauses (Figure 4).

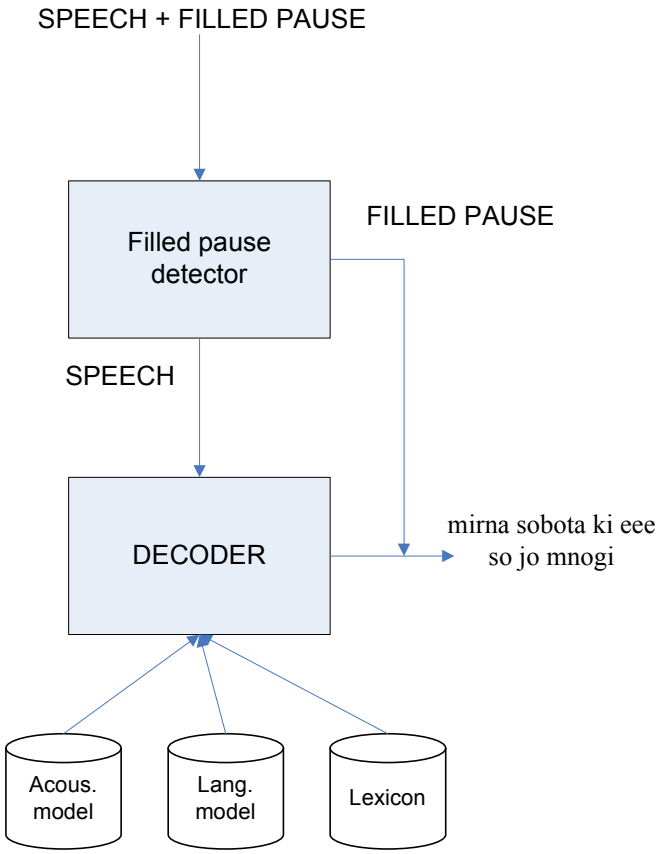

Figure 4: Explicit modeling of filled pauses in a speech recognition system.

In the second case acoustic models for filled pauses are part of the main speech recognition decoding process. This is called implicit modeling of filled pauses (Figure 5).

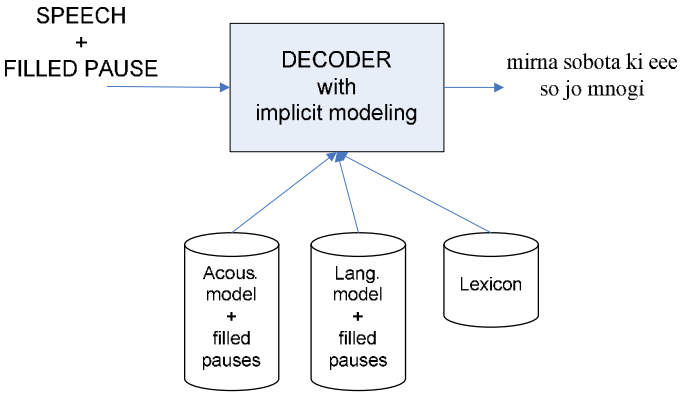

Figure 5: Implicit modeling of filled pauses in a speech recognition system.

2.1 Implicit modeling of filled pauses

In the basic acoustic modeling approach (AM1), all filled pauses use only one acoustic model. This results in combining all filled pauses, regarding their acoustic-phonetic properties, into one common model. In such a way, acoustic training material is grouped together, which is important in case of infrequent filled pauses (see Table 4). The drawback is that the modeling of acoustic diversities isn't taken into account. In our case, where the acoustic modeling was performed using the HMM, one three state left-right model was applied. The acoustic model for filled pauses was used as context-independent one and was as such also excluded from the phonetic decision tree based clustering of triphone acoustic models (see Section 4.1 for more details).

The second implicit acoustic modeling approach (AM2) uses a separate acoustic model for each type of filled pauses. Advantage is that such model covers all acoustic-phonetic properties of one type of filled pauses, but the problem can be with the amount of training material available for infrequent types of filled pauses. As for the first example, the HMM models are context independent.

The third kind of implicit modeling (AM3) is based on general acoustic models that are also used for speech modeling. Each filled pause is modeled with the speech acoustic models, according to its acoustic-phonetic properties. This solution usually assures enough training material for all types of filled pauses. The disadvantage lies in the fact that acoustic-phonetic properties of speech differ from those of filled pauses. The main difference is caused by duration of phonemes and levels of pitch. In case of this modeling approach, some of HMM models are context-dependent and therefore included in phonetic decision tree based clustering. The examples of all three implicit modeling approaches are presented in Table 1.

Table 1: Three different approaches of implicit acoustic modeling of filled pauses.

Filled pause	AM1	AM2	AM3
eee	filler	eee	e e
eem	filler	eem	e m
mhm	filler	mhm	m h m

There are three different filled pauses present in Table 1: *eee, eem,* and *mhm*. In case of AM1 acoustic models all three filled pauses are modeled with the common context-independent acoustic model *"filler"*. When AM2 acoustic models are applied, each filled pause has its own context-independent acoustic model for filled pauses (e.g. filled pause *eee* is modeled with *"eee"* acoustic model). In the last case, when AM3 acoustic models are applied each filled pause is modeled with context-dependent acoustic models for regular words – filled pause *mhm* is modeled with acoustic models *"m h m"* for regular words.

2.2 Implicit modeling of filled pauses based on phonetic broad classes

Considering all presented properties of described acoustic modeling approaches, a new method (AM4) how to model filled pauses is proposed. The basic idea is to use phonetic broad classes to model filled pauses. Phonetic broad classes are defined for each specific language, either by an expert phonetician or in a data-driven way. Phonemes with similar properties (e.g. open vowels) are grouped together in a particular phonetic broad class.

Class-01 *i i:*
Class-02 *m n v l b*
Class-03 *E i O u: E: e: ehr*
Class-04 *i: e:*
Class-05 *O u: o: W o w d-n ehr O:*
...

Figure 6: Slovenian phonetic broad class, defined in a data-driven way.

Example of Slovenian phonetic broad classes, defined in a data-driven way [19, 20] is shown on Figure 6. One of the smallest phonetic broad classes is *Class-01* with only two members *"i"* and *"i:"*. On the opposite side are phonetic broad classes, which have several members, as for example *Class-05* with 9 members.

Instead of using a separate acoustic model as in case of AM2, a group of acoustic models is used to model filled pauses. Groups should be defined in a way that they incorporate acoustically similar filled pauses with enough training material. The analysis of the training set showed (see Table 4) that 4 different categories should be defined: vowels, voiced consonants, unvoiced consonants, and mixed group. The last one is used for those filled pauses that can't be reliably categorized into the first three groups. The advantage of the proposed method is in the fact that are the acoustic models of filled pauses still separated from the acoustic models of speech. Therefore, they can better model peculiarities of filled pauses that strongly differ from speech. An example, how filled pauses are modeled with the proposed method is shown in Table 2.

Table 2: Modeling of filled pauses using the method based on phonetic broad classes.

Filled pause	AM4
eee	vowels
eem	mixed
mmm	voiced consonants
sss	unvoiced consonants

In proposed approach, each filled pause belongs to one of the possible phonetic broad class categories (vowels, mixed, voiced consonants, unvoiced consonants). The filled pause *eem*, which pronunciation is combination from vowels (*"e"*) and consonants (*"m"*) is member of category mixed. On the other side, the pronunciation of filled pause *eee* contains only vowels; therefore it is a member of the first category vowels.

3 Slovenian BNSI Broadcast News speech and text corpora

The primary language resource used during these experiments was the Slovenian BNSI Broadcast News database [21]. The BNSI database was designed in cooperation between University of Maribor, Slovenia and the Slovenian national broadcaster RTV Slovenia. The raw audio material was acquired from the archive of the broadcast company on DAT and DVD-R media. The captured audio signal was manually segmented, annotated and transcribed with tool Transcriber [22], according to recommendations on building Broadcast News spoken language resources.

The speech corpus comprehends two different types of TV-news shows. The first type is evening news where general overview of daily events is given. The second types of show are late night news where major events of the day are analyzed. In this type of news show are frequent longer interviews (up to 10 minutes), with high proportion of spontaneous speech.

The speech corpus consists of 42 news shows, which account for 36 hours of speech material. This material is further grouped into three sets: training, development

and evaluation, respectively. The size of the training set is 30 hours, whereas the size of the development and evaluation set is 3 hours each. Altogether 1565 different speakers are present in the BNSI database. The majority, 1069 of them, are male, while 477 are female. The gender of remaining 19 speakers was annotated as unknown.

It addition to the speech corpora, the text corpora (scenarios, transcriptions of speech corpus) was built. The text corpus is needed for developing the baseline set of language models. The Slovenian Vecer Newspaper text corpus was additionally incorporated in the language modeling. Properties of the BNSI Broadcast News database are given in Table 3.

Table 3: Slovenian BNSI Broadcast News speech and text database.

speech corpus:	
total length(h)	36
number of speakers	1565
number of words	268k
test corpus:	
number of words	11M
distinct words	175k

The evaluation set of the BNSI Broadcast News speech database is composed from 4 broadcasts in total length of approx. 3 hours. Typical broadcast news show comprises various types of speech: read or spontaneous, in studio or over telephone environment, with or without background [21, 23] (Figure 7).

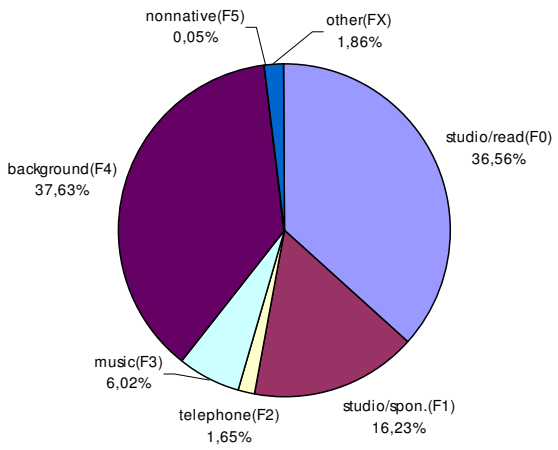

Figure 7: Ratio of various focus conditions in the BNSI speech database.

The goal in this experiment was to efficiently evaluate the acoustic modeling of filled pauses. Therefore only the utterances with spontaneous speech in clean studio environment (F1-focus condition [23]) were included in the evaluation set. There were 343 utterances with 3287 words in the evaluation set. The analysis showed that there were 155 different filled pauses in this evaluation set, which represent 4.72% of it. The training set comprises 24 broadcasts.

An analysis of all filled pauses that were found in the training set was also carried out. Those filled pauses with frequency higher than 5 are presented in Table 4.

Table 4: Statistics of filled pauses in the training set.

Filled pause	Frequency
eee	1833
sss	60
mmm	43
eem	40
zzz	21
uuu	16
ooo	14
vvv	12
ttt	12
aaa	12
nnn	10
iii	9
ppp	8
mhm	7
eeh	7

The most frequent filled pause in the training corpus is “eee”, with frequency 1833. The other filled pauses are far less frequent. The second one in Table 4 has frequency 60. There are altogether 15 filled pauses, which frequency is higher than 5. This distribution of frequencies between filled pauses support the idea of joining phonetically similar filled pauses in a same acoustic model, as the lack of appropriate training material for modeling of filled pauses can be foreseen.

4 Experimental design

The experimental design (Figure 8) is based on continuous density Hidden Markov Models for acoustic modeling and on n-gram statistical language models.

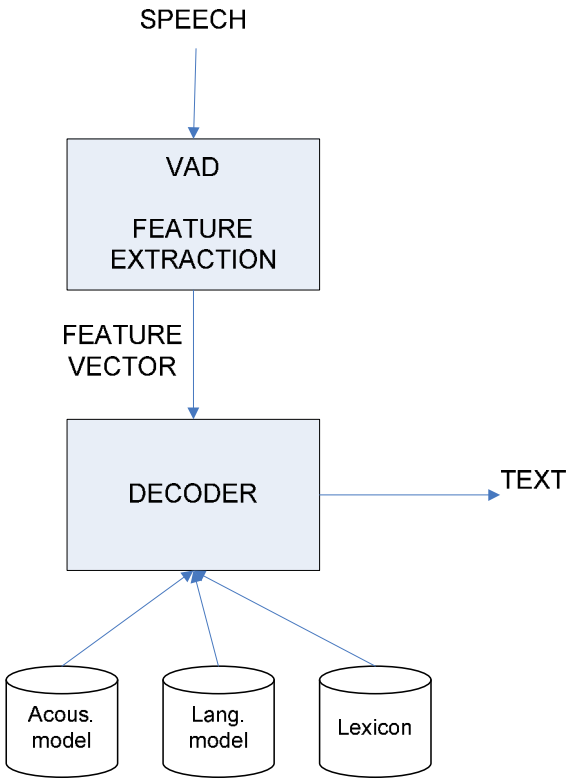

Figure 8: Block diagram of experimental speech recognition system.

The core module is a speech decoder, which needs three data sources for its operation: acoustic models, language model and lexicon.

4.1 Acoustic modeling

The frontend was based on mel-cepstral coefficients and energy (12 MFCC + 1 E, delta, delta-delta). The size of feature vector was 39. Also, the cepstral mean normalization was added to the feature extraction to improve the quality of speech recognition. The manually segmented speech material was used for training and speech recognition. 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 [24]. 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 [19, 20]. Three final sets of baseline triphone acoustic models with 4, 8 and 16 mixture Gaussian PDF per state were generated.

Our main task was the acoustic modeling of filled pauses. To exclude from the experiments influence of inter-speaker variations in pronouncing filled pauses, only the speaker independent acoustic models were applied. Consequently, this was reflected in lower accuracy of speech recognition system, than if also unsupervised speaker clustering and adaptation would be applied. The standard one-pass Viterbi decoder with pruning and limited number of active models was used for speech recognition experiments in the next section.

4.2 Language modeling and vocabulary

Language models were built using corpora of written language and transcribed speech. For LM training three different types of textual data were used: Vecer (corpus of newspaper articles in period 2000-2002), INews (TV show scripts in period 1998-2004) and BN-train (transcribed BNSI acoustic training set). The interpolation coefficients were estimated based on EM algorithm using a development set. The language model is based on bigrams. 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 [25, 26]. A possible reason for this is the usage of text corpora with speech transcriptions for language modeling.

Two types of language models were built. In the first model (LM1), all filled pauses and onomatopoeic words were mapped into unique symbol, which was considered as non-event, and can only occur in the context of a bigram and was given zero probability mass in model estimation. In the second model (LM2) filled pauses and onomatopoeic words were modeled as regular words.

Table 5: Statistics of language models used for modeling filled pauses.

	LM1	LM2
λ(BN-train)	0.2619	0.2665
λ(INews)	0.2921	0.2941
λ(Vecer)	0.4459	0.4392
perplexity	410	414

Language models built on the Vecer newspaper text corpus has the highest interpolation weight (0.4459 and 0.4392) for both types of language models. The interpolation weights for two other language models (INews and BN-train) are similar. The perplexities of language models, calculated on the evaluation set were 410 and 414, respectively. The higher value for LM2 is due to the unmapped filled pauses.

5 Results

The proposed method of acoustic modeling of filled pauses will be evaluated indirectly with word accuracy, using the speech recognition results. These speech recognition results will be also used to compare the modeling methods. The word accuracy is defined as:

$$Acc(\%) = \frac{H - I - D}{N} \cdot 100 \qquad (1)$$

where H denotes the number of correctly recognized words, I the number of inserted words, D the number of deleted words, and N the number of all words in the evaluation set. First, three different versions of the baseline system without modeling of filled pauses were evaluated, to check which system's topology performs best (Table 6).

Table 6: Speech recognition results without modeling of filled pauses for three different topologies of acoustic models.

	Acc(%)
Baseline 4 PDF	42.17
Baseline 8 PDF	49.82
Baseline 16 PDF	56.33

The simplest topology of acoustic models with 4 Gaussian PDF mixtures per state performed worst, with the 42.17% accuracy. When the number of mixtures was increased to 8 per state, the accuracy improved to 49.82%. The last baseline speech recognition configuration with 16 Gaussian mixtures achieved the best result with word accuracy of 56.33%. Thus the speech recognition performance was increased for 14.16% absolute. The relatively low performance of all three baseline systems is mainly due to the following facts: highly inflectional Slovenian language with high out-of-vocabulary rate, completely spontaneous type of conversations in the evaluation set and limitations of using speaker-independent acoustic models for this very complex speech recognition task. The disadvantage of the topology with 16 Gaussian mixtures per state, which yield the best result, is its complexity with high number of free parameters, which must be estimated. This results in increased computation time.

In the next step of evaluation four different filled pauses modeling techniques (AM1-AM4) were tested. Appropriate language models (LM1, LM2) were used in combination with the correct type of acoustic models. The results are presented in Table 7.

Table 7: Speech recognition results without and with acoustic modeling of filled pauses.

	Acc(%)
AM1+LM1	56.98
AM2+LM1	57.77
AM2+LM2	57.71
AM3+LM1	58.56
AM3+LM2	58.37
AM4+LM1	59.54

Already the basic modeling of filled pauses improved the speech recognition performance. The word accuracy increased to 56.98% when the AM1 and LM1 models were involved in test. The next version of acoustic models (AM2), where both types of language models (LM1, LM2) were tested, further improved the quality of speech recognition – the word accuracy increased to 57.77% and 57.71%. In this case each type of filled pause has its own acoustic model. There was almost no influence of the language model type on the speech recognition performance. The AM3 type of acoustic models, where the filled pauses were modeled with the same acoustic models as speech, achieved similar word accuracy (58.56% and 58.37%) as the AM3 type. There was again almost no influence of language model type on the word accuracy.

The AM4 acoustic models where the filled pauses were modeled with phonetic broad classes according to their acoustic-phonetic properties achieved the best overall result with word accuracy of 59.54%. In comparison with the baseline system the performance was increased by 5.70% relatively. The increase of word accuracy to the second best acoustic modeling approach was 1.67% relatively. There is practically no difference between all filled pauses modeling methods from the point of view of system's speed.

6 Conclusion

The analysis of speech recognition results showed the importance of acoustic modeling of filled pauses, where the best performance was achieved with the modeling technique, which balances between the diversity of acoustic-phonetic properties of filled pauses and the available spoken training data.

In the future the work will be oriented toward the definition of a data-driven approach of defining the

number and type of phonetic broad classes used for acoustic modeling of filled pauses. In such a way, further improvement of speech recognizer's accuracy can be achieved, which would also improve its usability for various speech recognition systems.

Acknowledgements: The work was partially funded by Slovenian Research Agency, under contract number: J2-9742-0796-06.

References:

[1] Billi, R., Castagneri, G., Danieli, M., 1997. Field trial evaluations of two different information inquiry systems. Speech Communication, Volume 23, Issues 1-2, October 1997, Pages 83-93.

[2] Žgank, A., M. Rojc, B. Kotnik, D. Vlaj, M. Sepesy Maučec, T. Rotovnik, Z. Kačič, A. Zögling Markuš, B. Horvat. Govorno voden informacijski portal LentInfo – predhodna analiza rezultatov. Jezikovne tehnologije 2002, Inštitut Jožef Stefan, Ljubljana.

[3] Gupta, V., Robillard, S., Pelletier, C., 2000. Automation of locality recognition in ADAS plus, Speech Communication, Volume 31, Issue 4, August 2000, Pages 321-328.

[4] Sket, G., B. Imperl (2002). M-vstopnica – uporaba avtomatskega razpoznavanaj govora v praksi. Jezikovne tehnologije 2002, Inštitut Jožef Stefan, Ljubljana.

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

[6] H. Marvi, "Speech Recognition Through Discriminative Feature Extraction". *WSEAS TRANSACTIONS on SIGNAL PROCESSING,* Issue 10, Volume 2, October 2006.

[7] S. A. R. Al-Haddad, Salina Abdul Samad, Aini Hussein, "Automatic Segmentation and Labeling for Continuous Number Recognition". *WSEAS TRANSACTIONS on SIGNAL PROCESSING,* Issue 9, Volume 2, September 2006.

[8] S. A. R. Al-Haddad, Salina Abdul Samad, Aini Hussein, M. K. A. Abdullah, "Automatic Segmentation and Labeling for Malay Speech Recognition". *WSEAS TRANSACTIONS on SIGNAL PROCESSING,* Issue 9, Volume 2, September 2006.

[9] R. Thangarajan, A.M. Natarajan, M. Selvam "Word and Triphone Based Approaches in Continuous Speech Recognition for Tamil Language". *WSEAS TRANSACTIONS on SIGNAL PROCESSING, Issue 3,* Volume 4, March 2008.

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

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

[12] Imai, T., Kobayashi, A., Sato, S., Tanaka, H., and Ando, A., "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, 2000.

[13] Wu, C. and Yan, G. "Acoustic Feature Analysis and Discriminative Modeling of Filled Pauses for Spontaneous Speech Recognition". *Journal of VLSI Signal Process*. Syst. 36, 2-3 (Feb. 2004), 91-104.

[14] Wu, Chung-Hsien, Yan, Gwo-Lang, "Discriminative disfluency modeling for spontaneous speech recognition", In: *EUROSPEECH-2001*, Aalborg, Denmark, pp. 1955-1958.

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

[16] S. Furui, M. Nakamura, T. Ichiba and K. Iwano "Analysis and recognition of spontaneous speech using Corpus of Spontaneous Japanese" *Speech Communication*, vol.47, pp.208-219 (2005-9).

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

[18] N. Seiichi, K. Satoshi, "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.

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

[20] Žgank, A., Kačič Z., Horvat, B., "Data driven generation of broad classes for decision tree construction in acoustic modeling", In: *EUROSPEECH 2003*, Geneva, Switzerland, 2505-2508, 2003.

[21] Žgank, Andrej, Verdonik, Darinka, Zögling Markuš, Aleksandra, Kačič, Zdravko. "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.

[22] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: Development and use of a tool for assisting speech corpora production". *Speech Communication,* Vol. 33, Issues 1-2, 5-22, 2001.

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

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

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

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