5

Modelling of Filled Pauses and Onomatopoeias for Spontaneous Speech Recognition

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

1. Introduction

With the growing availability of various content provided over state-of-the-art digital media is speech recognition becoming one of the main core technologies (Billi et al., 1997; Žgank et al., 2002; Gupta et al., 2000; Sket et al., 2002). Its task is to minimize the needed effort to access the particular part of content. The main content categories can be grouped in the following way:

- broadcasted media,
- public and governmental content,
- entertainment,
- education,
- meetings,
- personal communication,
- personal repositories,…

The common point of all items is that characteristics of such spoken content widely diverge from type of speech, which is commonly found in spoken language resources used for training automatic speech recognition systems (Maddi et al., 2006; Marvi, 2006; Al-Haddad et al., 2006a; Al-Haddad et al., 2006b; Thangarajan et al., 2008). The main issue, which influences the quality of speech recognition, is the presence of spontaneous speech with all its special requests and characteristics. A speaker in such scenario can speak freely, without planning his/her speech. The vocabulary has size of several 10k words, which hardly depends on the properties of language involved. For less inflectionally and morphologically complex languages (e.g.: English, Spanish, Italian,…), the size of 64k vocabulary words can cover more than 99% of words in the test set (out-of-vocabulary (OOV) rate). On the other side are complex highly inflectional and agglutinative languages (e.g.: Finnish, Hungarian, Slovenian, Czech …), where the same size of vocabulary produces the OOV of 10% or even more.

In such cases present all various effects of spontaneous speech an additional parameter, which reduces the quality and performance of speech recognition for several percents. The applications where such problems can occur are: speech-to-speech translation system, "how can I help you?" telecommunication services, TV subtitling services, spoken content indexing services...

Real time TV subtitling service as one of the emerging services (Lambourne et al., 2004; Brousseau et al., 2003; Imai et al., 2000) in current and future society with increased proportion of elderly people is gathering on importance. The proportion of broadcasted content, which can't be immediately subtitled from content scripts, hardly depends on the

Source: Advances in Speech Recognition, Book edited by: Noam R. Shabtai,
ISBN 978-953-307-097-1, pp. 164, September 2010, Sciyo, Croatia, downloaded from SCIYO.COM

show's type. In a typical broadcast news show, approximately 50% to 75% of stories can be automatically subtitled using closed caption generated from the scripts. Example of such Slovenian evening news show script is given on 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.

Fig. 1. Evening TV news show script, a part of the Slovenian BNSI Broadcast News database.

The remaining part of the show isn't covered, as it contains live conversations (e.g. interviews, talk shows), where closed captions can't be generated from scripts or scenarios. Example of such script part is shown on Figure 1, denoted as section "03 izjava župana", where only the last few seconds are transcribed as guideline for the director. These parts of shows must be covered with dedicated methods as is spontaneous speech recognition. Two methods can be used for producing closed captions: respeaking, where a highly trained operator respeakes all utterances in an of-the-shelf dictation system or a fully automated subtitling system, which must process the entirely show, usually in several steps.

Automatic recognition of spontaneous conversations is a very challenging task. There are three major groups of disfluencies in spontaneous speech that influence the quality and performance of any spontaneous speech recognition system:

- 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 2 shows example of spontaneous sentence ("mirna sobota ki ee so jo mnogi") from Slovenian BNSI Broadcast 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 when spontaneous speech is used in everyday situation.

The presented characteristics of spontaneous speech influence both types of models in a system – acoustic and language model. On the other side, the accents mainly influence the performance of acoustic models. Spontaneous conversation can involve a high degree of accented speech, depending of the discourse properties. In case of broadcast news language resources various groups of interviews include such discourse.

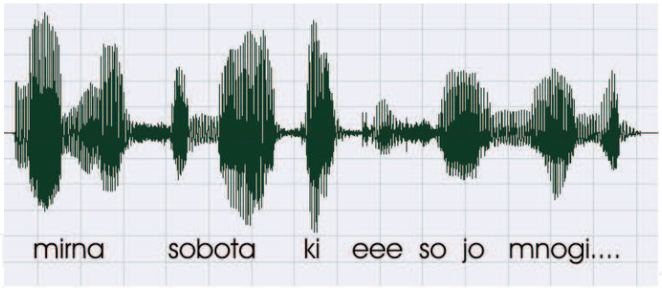

Fig. 2. Spontaneous sentence from Slovenian BNSI Broadcast News speech database.

Spontaneous speech is also a challenging task for language modelling. It is characterized by unconstrained speaking style, frequent grammatical errors, hesitations, starts-over, etc. Another problem is a limited amount of training data. The main source is audio transcription. Unfortunately, sources of written data do not exhibit characteristics of spoken language.

The research work presented in this chapter is oriented on modelling of filled pauses and onomatopoeias for spontaneous speech recognition system. A previously proposed filled pauses acoustic modelling approach will be further improved with an advanced training procedure. In addition to normally accented speech, also a heavily accented spontaneous speech of a non-native speaker will be included in the experiment. Filled pauses are one of the most frequent categories of spontaneous speech effects, which are present in real-life spoken language resources and will be as such included in our experiments. Onomatopoeias as another category are less frequent, but still very challenging for modelling. We have grouped both categories in one, called filled pauses. Although filled pauses and onomatopoeias don't carry any true semantic information, it is still necessary to include them in modelling 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. The methods proposed for modelling of filled pauses will be also evaluated on heavy accented speech, to show that modelling of filled pauses plays even more important role in such case of conversation.

The level of accented speech usually depends on the speaker and its role in the discourse. In addition to these properties, the language also plays an important role. There are some languages, where a large number of various accents can be found. Slovenian is one of such languages, with approximately 50 different accents. This makes any accent modelling an additionally challenging task.

The chapter is organized as follows: the current state-of-the-art is described in Section 2. Various filled pauses modelling approaches are presented in Section 3. The native and non-native spoken language resources are introduced in Section 4. The experimental design used for evaluation is described in Section 5. Section 6 contains the results of the speech recognition experiments, while the conclusion and directives for future work are given in Section 7.

2. Overview of current research work on topic of spontaneous speech recognition

In the last few years is the research area of spontaneous speech recognition gathering on importance. One of the prerequisites for this development was the increase in CPU power, as are the algorithms for spontaneous speech recognition very demanding on processing power.

In the area of acoustic modelling of filled pauses, several authors presented successful approaches, how to address this topic. The first group of methods is based on Gaussian Mixture Modelling (GMM) (Wu & Yan, 2004; Wu & Yan, 2001; Rangarajan & Narayanan, 2006). There are two main approaches possible. In the first approach, for each type of filled pauses a separate GMM model is build. The number of mixtures depends on the availability of spoken material per class. A separate class is used for modelling of normal spontaneous speech without any filled pauses. As the end results a system with multi GMM is being used for explicit (see Section 3) recognition of filled pauses in spontaneous speech. The second approach is based on only two GMM models. The first one represents filled pauses and the second one normal spontaneous speech. The main advantage of the second approach is that it is simpler to collect adequate amount of training material per class to train the GMM models. It also reduces the classification error between various types of onomatopoeias, as it can be sometimes extremely difficult to label separate sounds correctly. In general, the second approach yields better speech recognition results due to its higher modelling capability.

The second major group is based on modelling with Hidden Markov Models (HMM) (Furui et al., 2005; Stouten et al., 2006; Seiichi & Satoshi, 2007), usually in an implicit way (see Section 3 for details). The performance of this group of approaches depends on the quality of transcriptions of spoken language resources. Each filled pause must be correctly labelled and transcribed to be able to model it with an HMM model. There are several methods possible how can a filled pause be represented with an HMM. One approach is to use separate HMM models for filled pauses. Another approach uses the same HMM acoustic models for filled pauses and spontaneous speech. The second approach is more difficult and complex as acoustic-phonetic properties of both types usually differ. Therefore complex modelling approaches are needed to reduce this discrepancy. It is also possible to combine the above presented methods in one system.

The specifics of spontaneous speech presented above for acoustic modelling are also reflected on language modelling. Disfluencies (repetitions, hesitations, and sentence restarts) distinguish spontaneous from read speech to a great extant. N-grams base their word prediction on a local context of N-1 previous words. Early psycholinguistic experiments found that human subject asked to guess next word in the transcription a spontaneous speech required more guesses for words that had been proceeded by a hesitation (Goldman, 1968). The experiment indicates the difficulties of transition from modelling read speech to modelling spontaneous speech.

Disfluencies corrupt this context. First, the idea was to remove disfluencies from the context. Based on experiments it has been shown that simple clean-up is not the right way to recover the fluent order of meaningful words (Duchateau et al., 2004). If we eliminate disfluencies completely, we would lose some information.

In (Duchateau et al., 2004) the authors allow the system to pick the most probable option when both a context with and without disfluencies are available. In case of repetitions the results were improved significantly by offering the system the choice between removing or not removing the disfluency from the prediction context. For hesitations and restarts this method results in a small deterioration of the recognition rate. The research was later extended by developing a specialized preprocessor which operated independently of the search and which searches for filled pauses on the basis of acoustic and prosodic features that are not accessible to the recognizer (Stouten et al., 2006). A filled pauses detector was built. Two strategies for incorporating the posterior probabilities at the output of this detector into the search engine were proposed.

Filled pauses and onomatopoeias don't carry any true semantic information, but should be incorporated into the language model. The biggest difficulty is that statistical language models typically have very limited context, and by keeping filled pauses and onomatopoeias in context, information bearing word is lost. In (Stolcke et al., 1999) they are demarcated by events surrounding the words. They refer to them as Hidden Word-level Events (HWE). Models of HWE capture the specific prosodic characteristics of HWEs, such as intonation and duration patterns. The information from prosodic features was combined with statistical language models that describe the distribution of HWE in relation to words, part-of-speech, and other syntactical and lexical units.

Adaptation to speaker-dependent disfluencies was studied to adopt a system for disfluency removal. Disfluency removal makes sentences shorter, less ill-formed and thus facilitates the downstream processing by natural language understanding components such as machine translation or summarization (Honal & Schultz, 2005). The probability that a word is disfluent is composed of a weighted sum over the six models. The most prominent were the model of the length of the deletion region of a disfluency and the model of the position of a disfluency. Gradient descent method was used to automatically optimize the parameter weights.

Speaker-produced disfluencies were identified in a conditional random field-based approach (Fitzgerald et al., 2009). The authors emphasize false start regions, which are often missed in current disfluency identification approaches as they lack lexical or structural similarity to the speech immediately following. They find that combining lexical, syntactical, and language model-related features with the output of the state-of-the-art disfluency identification system improves overall word-level identification of these and other errors.

Although there has been significant work devoted to some spontaneous speech phenomena, we are still looking for an accurate and efficient language models for speech disfluencies.

3. Spontaneous speech and modelling of filled pauses and onomatopoeias

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 (Wu &Yan, 2004)), and speech recognizer than process only the part of speech without filled pauses (Figure 3).

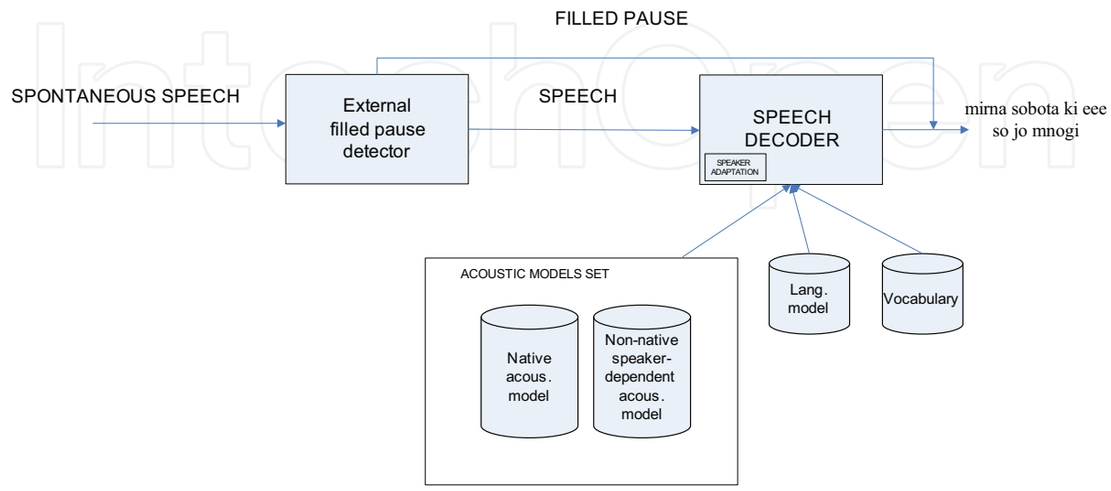

Fig. 3. Explicit modelling of filled pauses in a speech recognition system.

In the second case are acoustic models for filled pauses part of the main speech recognition decoding process. This is called implicit modelling of filled pauses (Figure 4).

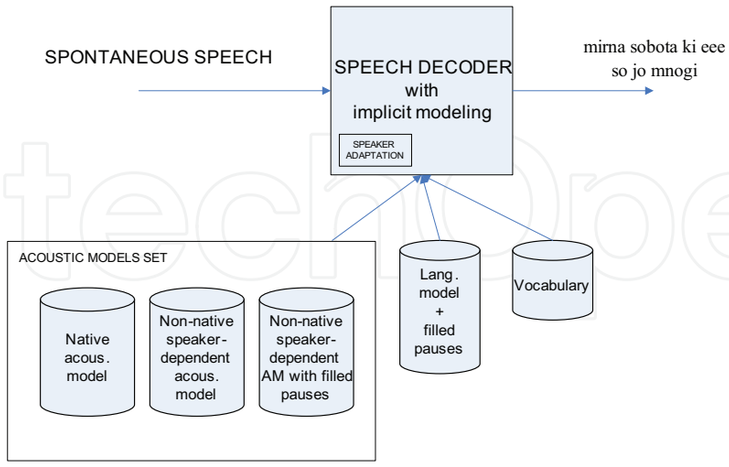

Fig. 4. Implicit modelling of filled pauses in a speech recognition system.

3.1 Implicit modelling of filled pauses

In the basic acoustic modelling 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 modelling of acoustic diversities isn't taken into account. In our case, where the acoustic modelling 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 5 for more details).

The second implicit acoustic modelling 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 modelling (AM3) is based on general acoustic models that are also used for speech modelling. Each filled pause is modelled 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 modelling approach, some of HMM models are context-dependent and therefore included in phonetic decision tree based clustering. The examples of all three implicit modelling approaches are presented in Table 1.

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 modelled 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 modelled with "eee" acoustic model). In the last case, when AM3 acoustic models are applied each

filled pause is modelled with context-dependent acoustic models for regular words – filled pause mhm is modelled with acoustic models "m h m" for regular words.

Filled pause	AM1	AM2	AM3
Eee	Filler	eee	e e
Eem	Filler	eem	e m
Mhm	Filler	mhm	m h m

Table 1. Three different approaches of implicit acoustic modelling of filled pauses.

3.2 Implicit modelling of filled pauses based on phonetic broad classes

Considering all presented properties of described acoustic modelling approaches, a new method (AM4) how to model filled pauses was proposed in (Žgank et al., 2008). 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:*
...

Fig. 5. Slovenian phonetic broad class, defined in a data-driven way.

Example of Slovenian phonetic broad classes, defined in a data-driven way (Žgank et al., 2005a; Žgank et al., 2003) is shown on Figure 5. 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 this 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 modelled with the AM4 method is shown in Table 2.

Filled pause	AM4
Eee	Vowels
Eem	Mixed
Mmm	voiced consonants
Sss	unvoiced consonants

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

In AM4 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. The AM4 method already proved promising results. The current focus is to evaluate the method with improved training procedure and on heavy accented speech.

4. Slovenian BNSI Broadcast News speech and text corpora

The primary language resource used during these experiments was the Slovenian BNSI Broadcast News database (Žgank et al., 2005b). 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 (Barras et al., 2001), 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. With usage of additional preprocessing steps on level of manual transcriptions the amount of training material which was prior excluded from the training set was reduced. Detailed analysis of speech recognition results showed statistically significant improved performance due to this additional step.

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 modelling. Properties of the BNSI Broadcast News database are given in Table 3.

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

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

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 (Žgank et al., 2005b; Schwartz et al., 1997) (Figure 6).

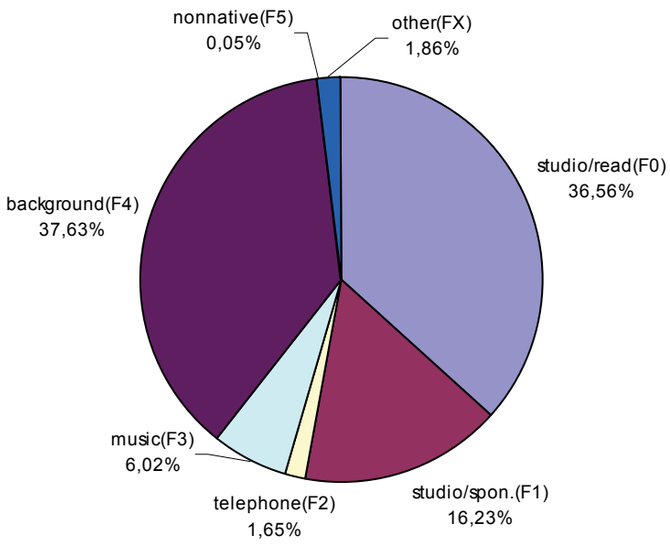

Fig. 6. Ratio of various focus conditions in the BNSI speech database.

The goal in this experiment was to efficiently evaluate the acoustic modelling of filled pauses. Therefore only the utterances with spontaneous speech in clean studio environment (F1-focus condition (Schwartz et al., 1997)) 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.

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

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

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 modelling of filled pauses can be foreseen.

The secondary spoken language resource was used for modelling and evaluating heavy accented speech. For this experiment, the Slovenian SINOD speech database (Žgank et al., 2006a) was used. The SINOD database was developed as a supplement to the BNSI Broadcast News database. It consists of two TV interviews, the first one with Russian non-native speaker (Table 5) and the other one with English non-native speaker of Slovenian. The same structure and transcription rules were applied as in the BNSI database. Here, only the part with the Russian non-native speaker of Slovenian was involved in the training and evaluation procedure. The secondary spoken language resource plays an important resource as it involves a high proportion of accented filled pauses, due to the non-native speaker involved. The presented spoken language resource has the drawback that only one speaker and its speaking style is involved in the heavy accented speech experiments. But the fact is that such spoken language material is extremely difficult to collect, especially for languages with smaller number of speakers. To reduce this characteristic of non-native spoken language resource, adaptation procedures presented in Section 5 were additionally incorporated.

speech corpus:	SINOD
total length(mm:ss)	28:20
number of sentences	642
distinct words	1010
test corpus:	
test set length (mm:ss)	8:36

Table 5. Slovenian non-native database SINOD (Russian speaker).

5. Experimental design

The experimental design (Figure 7) is based on continuous density Hidden Markov Models for acoustic modelling and on n-gram statistical language models.

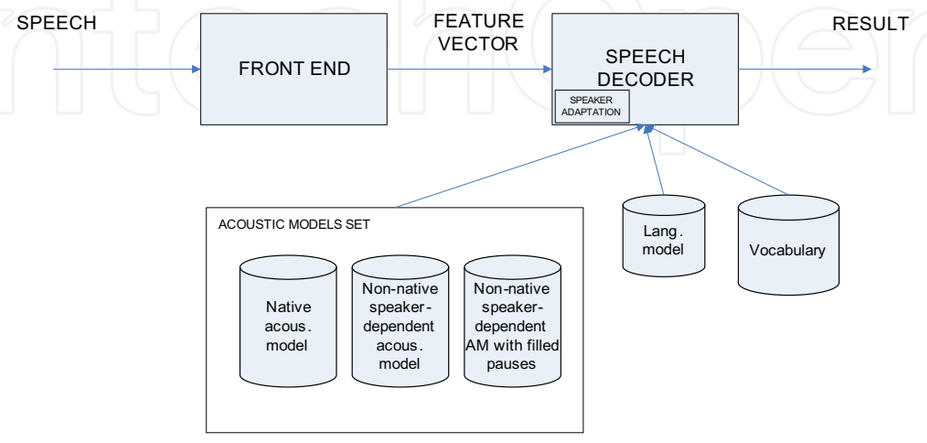

Fig. 7. Block diagram of experimental speech recognition system.

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

5.1 Acoustic modelling

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 (Žgank et al., 2006b). 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; Žgank et al., 2003). 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.

Our main task was the acoustic modelling 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 for native test set.

For the heavy accented speech with the non-native set using the SINOD speech database, the baseline BNSI acoustic models were first adapted to particular speaker using the Maximum Likelihood Linear Regression (MLLR) (Leggetter & Woodland, 1995) procedure. The MLLR was used in an iteratively way. During the first iteration, acoustic models were adapted on general transcription. Thereafter the forced realigning procedure was used to improve the general transcriptions for a particular speaker. During the second iteration, the improved transcriptions were used for MLLR speaker adaptation. In the last step of modelling heavy accented speech, all approaches for modelling filled pauses were applied to the set of speaker dependent acoustic models.

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.

5.2 Language modelling 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 (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.
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 modelled as regular words.

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

Table 6. Statistics of language models used for modelling filled pauses.

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.

6. Results

The proposed method of acoustic modelling 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 modelling methods for normal and heavy accented speech. The word accuracy is defined as:

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

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 modelling of filled pauses were evaluated on normal speech, to check which system's topology performs best (Table 7).

	Acc(%)
Baseline 4 PDF	50.90
Baseline 8 PDF	55.82
Baseline 16 PDF	62.15

Table 7. Speech recognition results without modelling of filled pauses for three different topologies of acoustic models recognizing normal speech.

The simplest topology of acoustic models with 4 Gaussian PDF mixtures per state performed worst, with the 50.90% accuracy. When the number of mixtures was increased to 8 per state, the accuracy improved to 55.82%. The last baseline speech recognition configuration with 16 Gaussian mixtures achieved the best result with word accuracy of 62.15%. Thus the speech recognition performance was increased for 11.25% 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. The increased complexity of training procedure, presented in Section 5, improved the performance for approximately 5% in overall if compared to system applied in (Žgank et al., 2008).

In the next step of evaluation four different filled pauses modelling 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 8.

	Acc(%)
AM1+LM1	62.73
AM2+LM1	62.96
AM2+LM2	62.98
AM3+LM1	63.60
AM3+LM2	64.37
AM4+LM1	64.95

Table 8. Speech recognition results without and with acoustic modelling of filled pauses.

Small improvement of recognition performance was already denoted for basic modelling of filled pauses on normal speech. The combination of AM1 and LM1 models increased the accuracy to 62.73%. Similar improvement of accuracy was achieved with the AM2 acoustic models, when LM1 and LM2 language models were used – the accuracy was 62.96% and 62.98% respectively. There was almost no influence of the language model type on the normally accented speech recognition performance. In case of AM3 acoustic models were filled pauses modelled in combined mode with normal speech. The evaluation of this approach showed word accuracy of 63.60% and 64.37% for each particular language model LM1 and LM2. In this case, the version of language model played an important role.

The last evaluation step for normally accented speech was focused on AM4 acoustic models where the filled pauses were modelled with phonetic broad classes according to their acoustic-phonetic properties. This approach achieved the best overall result with word accuracy of 64.95%. The baseline system performance was improved for 2.80% absolutely. Due to the improved training procedure, the improvement was smaller as in case of system described in (Žgank et al., 2008), although it was still statistically significant.

In the last step of evaluation, the heavy accented speech originating from the SINOD database was tested. The results for this case are presented in Table 9.

In case of SINOD database only the AM4 approach of modelling filled pauses was tested, as it already proved to be the most efficient one. The baseline SINOD system achieved the word accuracy of 65.74%. The improvement in comparison to the baseline system is result of

	Acc(%)
SINOD baseline	65.74
SINOD AM4	67.31

Table 9. Speech recognition results for heavy accented speech without and with filled pauses modelling.

applying MLLR procedure, although the increase of word accuracy is smaller than usual for speaker adaptation. The possible cause for this is the non-native origin of test speaker. In case, when the filled pauses were modelled using the proposed phonetic broad classes approach, the word accuracy increased to 67.31%. Thus the overall improvement for heavy accented non-native speech was 1.57%. The improvement is smaller as in case of native speech, but it still show, how important it is to model the filled pauses.

7. Conclusion

The new speech recognition system achieved statistically significant improvement of word accuracy in comparison with the previous version. The obtained speech recognition results clearly showed how important it is to adequately model filled pauses and onomatopoeias in spontaneous speech on level of acoustic and language models. The detailed analysis of speech recognition performance on filled pauses in non-native speech showed that there is still some room for improvements due to the complexity of this task.
The future work will be focused on various data-driven approaches, which will take into account the difference in pronouncing filled pauses and onomatopoeias in native and non-native speech. The detailed analysis of speech recognition results namely showed that this could further improve the performance of our system.

8. Acknowledgements

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

9. References

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

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

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

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.

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.

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.

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

Goldman-Eisler, F., (1968). *Psycholinguistics: Experiments in Spontaneous Speech*, New York: Academic Press.

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.

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.

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.

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.

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

Rotovnik, T., Sepesy Maučec, M., 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.

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

Stolcke, A., E. Shriberg, D. Hakkani- Tür, G. Tür, (1999). Modeling The Prosody Of Hidden Events For Improved word Recognition, *In Proc. EUROSPEECH*, vol. 1, pp. 307–310, Budapest.

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.

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

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

Wu, C. and Yan, G. (2004). "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.

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

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

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., Verdonik, D., Zögling Markuš, A., Kačič Z., (2006). SINOD - Slovenian non-native speech database. *Proc. LREC 2006*, Genova, Italy.

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.

Advances in Speech Recognition

Edited by Noam Shabtai

ISBN 978-953-307-097-1
Hard cover, 164 pages
Publisher Sciyo
Published online 16, August, 2010
Published in print edition August, 2010

In the last decade, further applications of speech processing were developed, such as speaker recognition, human-machine interaction, non-English speech recognition, and non-native English speech recognition. This book addresses a few of these applications. Furthermore, major challenges that were typically ignored in previous speech recognition research, such as noise and reverberation, appear repeatedly in recent papers. I would like to sincerely thank the contributing authors, for their effort to bring their insights and perspectives on current open questions in speech recognition research.

How to reference

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

Andrej Zgank and Mirjam Sepesy Maucec (2010). Modeling of Filled Pauses and Onomatopoeas for Spontaneous Speech Recognition, Advances in Speech Recognition, Noam Shabtai (Ed.), ISBN: 978-953-307-097-1, InTech, Available from: http://www.intechopen.com/books/advances-in-speech-recognition/modeling-of-filled-pauses-and-onomatopoeas-for-spontaneous-speech-recognition

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

IntechOpen