Modeling Filled Pauses for Spontaneous Speech Recognition Applications

ANDREJ ŽGANK, TOMAŽ ROTOVNIK, MIRJAM SEPESY MAUČEC
Laboratory for Digital Signal Processing
University of Maribor
Smetanova ul. 17, SI-2000 Maribor
SLOVENIA

Abstract: - This paper is focused on acoustic modeling for spontaneous speech recognition applications. 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. Filled pauses and onomatopoeias present one of such important attributes. 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 new modeling approach is compared with three other 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 recording of TV news shows. The evaluation of the proposed acoustic modeling approach was done with a set of spontaneous speech. The overall best acoustic modeling of filled pauses improved the speech recognizer's word accuracy for 5.70% relatively in comparison to the baseline system.

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

1 Introduction

In the today's world are speech recognition applications gathering on importance. There are a large number of such applications, which were transferred from laboratories into real-life environment. The quality of speech recognition hardly depends on the recognition task. Applications which involve recognition of isolated or connected words can achieve almost perfect accuracy in normal environment conditions. In case of planned speech, the speech recognition applications (e.g. computer dictation) still perform very well. The most challenging task for speech recognition applications presents the spontaneous speech, where significant degradation of speech recognition performance can occur. Additional problem can present the supported language, as are some of them more complex from the speech recognizer's point of view (e.g. highly inflectional languages). Slovenian also belongs to such group of languages.

One of possible challenging applications for spontaneous speech recognition is real-time subtitling of 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. 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. To cover these parts of a broadcast, applications using dedicated keyboards or speech recognition systems (Figure 1) must be used [1, 2, 3]. 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.

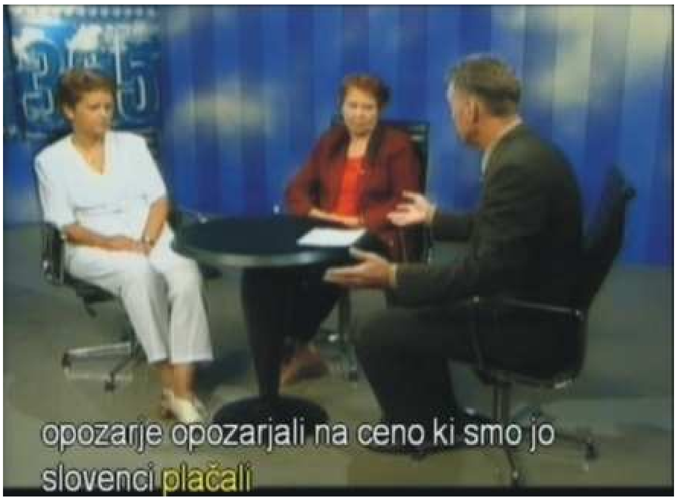
opozarje opozarjali na ceno ki smo jo
slovenci plačali

Figure 1: Screenshot of a speech recognition based TV subtitling demo application.

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), word repetitions and sentence restarts.

This paper focuses on acoustic modeling of filled pauses and onomatopoeias for spontaneous speech recognition applications. 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. 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) [4, 5, 6], whilst the other major group uses Hidden Markov Models (HMM) [7, 8, 9]. 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 Modeling filled pauses for spontaneous speech recognition

There are two different types of filled pauses acoustic modeling from the speech recognizer's point of view. In the first case are acoustic models for filled pauses part of the main speech recognition decoding process. This is called implicit modeling of filled pauses. In the second case are filled pauses detected using an external module (e.g. GMM classification [4]), and speech recognizer than process only the part of speech without filled pauses.

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

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. 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 are filled pauses 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

3 Speech and text corpora

The primary language resource used during these experiments was the Slovenian BNSI Broadcast News database [10]. It consists of speech and text corpora (scenarios, transcriptions of speech corpus), which were needed for developing the baseline set of acoustic and 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 [10, 11]. 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 [11]) 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. 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 is based on continuous density Hidden Markov Models for acoustic modeling and on n-gram statistical language models.

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 [7]. The procedure was based on common solutions [12]. 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 [13, 14]. The final set of baseline triphone acoustic models consisted of 16 mixture Gaussian PDF per state.

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, was this 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: 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 [15, 16]. 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

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 both spoken text sources (INews and BN-train) are similar. The perplexity of language models, calculated on the evaluation set was 410 and 414. 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:

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

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. The baseline system, without modeling of filled pauses and four different filled pauses modeling techniques (AM1-AM4) were tested. Appropriate language models (LM1, LM2) were used in combination with the acoustic models. The results are presented in Table 6.

Table 6: Speech recognition results without and with acoustic modeling of filled pauses.

	Acc(%)
Baseline	56.33
AM1+LM1	56.98
AM2+LM1	57.77
AM2+LM2	57.71
AM3+LM1	58.56
AM3+LM2	58.37
AM4+LM1	59.54

The baseline speech recognition system achieved the word accuracy of 56.33%. The relatively low performance is mainly due to the following facts: highly inflectional Slovenian language, completely spontaneous type of conversations in the evaluation set and limitations of using speaker-independent acoustic models for this very complex speech recognition task.

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 has each type of filled pause 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.

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 applications.

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

References:

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

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

[7] 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).

[8] 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).

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.