Speaker Non-speech Event Recognition with Standard Speech Datasets

J. Rajnoha

A non-speech event modelling approach to speech recognition is presented in this paper. A speaker independent spoken Czech digit recogniser is used for this purpose, and speaker generated non-speech events are modelled. Because it is important for the recogniser to be trained on suitable data, the paper shows some factors that influence the occurrence of the modeled non-speech events in the training database. Some results achieved on the analysed training database are then shown. In the experiments on forced alignment the recogniser eliminates almost all the insertion error, which is a promising property for subsequent training. However, experiments with a different basis for the non-speech event models provide almost the same results, so the difference seems to be not so significant for recognition.

Keywords: speech recognition, digit recognition, non-speech events, training database, forced alignment.

1 Introduction

Automatic speech recognition has become a very popular field of research, and the results come into our lives in many forms, e.g. in voice controlled machines, like PCs and mobile phones.

These command controlled systems often work with real, spontaneous speech, which is different from the read speech used in the clean laboratory conditions of a research centre. This noisy kind of speech is recorded in a real environment, which causes some noise addition to the speech signal. The speech is also created "on the fly" and the speaker has to think of the next word while speaking. All this causes many non-speech events to be present in such speech. Therefore it is important to make the recogniser robust against these events, and to take the events into account and ensure that they are not incorrectly recognised.

The long-term goal of our work is to create a hidden Markov model (HMM) based digit recogniser, which will suppress the non-speech event influence. One solution is to model the non-speech events that affect speech. For this purpose it is important to have a large training database with many occurrences of each modelled item, e.g. phoneme, to get a general description of the item. Therefore in the first part of this work, two Czech speech databases are analysed for the presence and quality of the speaker non-speech events. In this very beginning of our work, only speaker-generated non-speech events are taken in account because of their position between words of recognised speech. This enables us to consider the events to be another word and to model them easily. For this purpose, the databases were inspected for the presence of the events in different situations.

The second part of the work is concerned with speaker non-speech event modelling. A robust Czech digit sequence recogniser based on HMMs of Czech phonemes is trained on the analysed database, and the results gained with the non-speech event recognition feature are presented. An HMM Toolkit [2] is used for this purpose.

Based on these results, forced-alignment experiments are shown in the third part of the work. The recogniser is used to re-recognise the training data. This will remove unsuitable event marks in the transcription and it also enables to remark the event, which should improve the recognition results.

2 The database for the non-speech event recognition task

As noted above, it is important for automatic speech recogniser training to have a speech database that has enough occurrences of each type of modelled unit, e.g. phonemes. In the non-speech event modelling task this calls for a sufficient number of non-speech events present in the database, which is the way to get a general description of the modelled event.

This work deals with speaker non-speech events that appear in between particular words, so the database is analysed for the presence of these events.

In most cases, speech recognition systems are trained on a database of mainly read speech. This helps to complete the database, because it is known approximately what was said. However, read speech is different from the spontaneous speech used in voice communication with a machine. Read speech is low in speaker non-speech events. Unlike a spontaneous speaker, a reader is not forced to think while speaking and so the occurrence of hesitation is low. The same problem occurs with other possible non-speech events, like lip-smacking, because the reader has better control over his mouth.

The recognizer presented here was trained on two different sets of speech. The first database (SPEE) is a collection of Czech speaker records in a different environment, which was inspected to contain only the items in a silent environment for training purposes. The SPEE set includes speaker non-speech events divided into two classes: filled pauses (FIL), which are pauses in speech filled with some sound (common in hesitation), and other speaker non-speech events (SPK).

To increase the number of non-speech events present in the training set of speech data, one other dataset was used. It is a collection of records from a car (TEM), which was also inspected to contain only silent items from a standing car. This database divides the events into several classes, which helps to create more different models of non-speech events and to describe these events more accurately.

Table 1 shows the number of events marked in the transcription of the whole dataset and of the selected clean training subset for both the SPEE and the TEM dataset.

It can be seen that the training selections (clean) contain notably fewer speaker non-speech events marked in the tran-

© Czech Technical University Publishing House

Table 1: Number of non-speech events

	utterances	SPK-**event**	FIL-**event**
SPEE-**all**	**180213**	**108474**	**7382**
SPEE-**train**	**63024**	**33138**	**1856**
TEM-**all**	**221318**	**46742**	**691**
TEM-**train**	**38391**	**11532**	**153**

scription. The analysis in 2.2 shows that higher SNR (lower noise level) in speech leads to lower occurrence of the events. Some compromise is therefore needed in the choice between clean speech (more understandable for the recogniser) and the number of non-speech events for training the non-speech event robust recogniser.

The reason for the low number of non-speech events in the TEM-training subset is that the training subset is only a small fragment of the whole TEM dataset. Only the standing car items were taken for training purposes.

The datasets also include some records that are not suitable for standard phoneme training, e.g. web-page addresses or spelled utterances. While the average occurrence of a non-speech event is 0.64 events per one utterance in the whole SPEE dataset and 0.56 events per utterance in the training subset, the average rate for web-page utterances is 0.89. However, these records were not used in the training phase, which decreases the number of non-speech events in the training subset.

The analyses follow. These try to find groups of records with some property that is important in the speaker non-speech event distribution in the training subset. This can help to discover some inefficiency in the non-speech event training process.

2.1 Event distribution in speaker's age

The datasets include some basic information about the speaker. This can help us to find whether some group of people can influence the training dataset in some way. One such kind of information is the age-class distribution and the number of speaker non-speech events for different age-classes.

Figure 2 shows that the presence of speaker non-speech events is not much influenced by the speakers age till about 65 years, and only for higher age the amount rises. There are not many speakers from the 65+ age-class in the datasets (Fig. 1), and this increase is not high enough to cause harmful effects, like training the recogniser for a specific kind of event only.

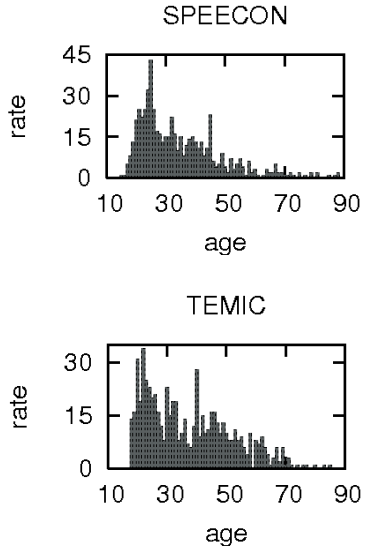

Fig. 1: Age-class distribution in the datasets

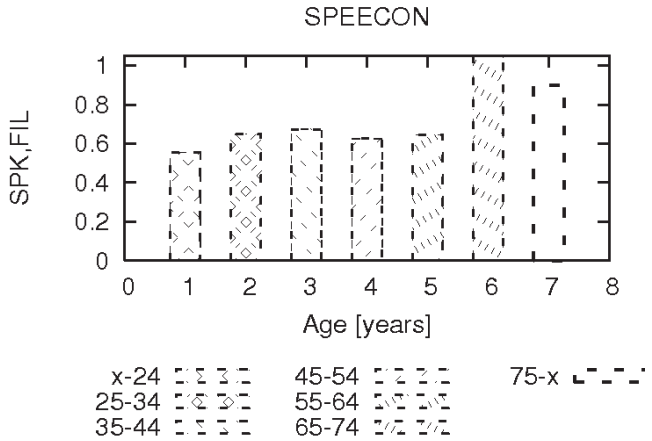

Fig. 2: Non-speech events distribution in age for the speech datasets

2.2 Event distribution in different noise levels

The SPEE database includes an SNR estimation for the records. It enables the user to analyse the influence of a noisy environment on the occurrence of a non-speech event in speech. There is no SNR estimate for the TEM database, but the information on car type and engine state can also partially describe the environment.

The graph in Fig. 3 shows that the distribution of the average non-speech event rate has its maximum in the SNR group between 10 and 15dB. For lower SNR, the rate falls mainly because of the noise that covers the event. A similar effect can be seen in TEM, where the items without a running engine in the background contain more non-speech event marks. For higher SNR ranges the rate also falls. In this environment the

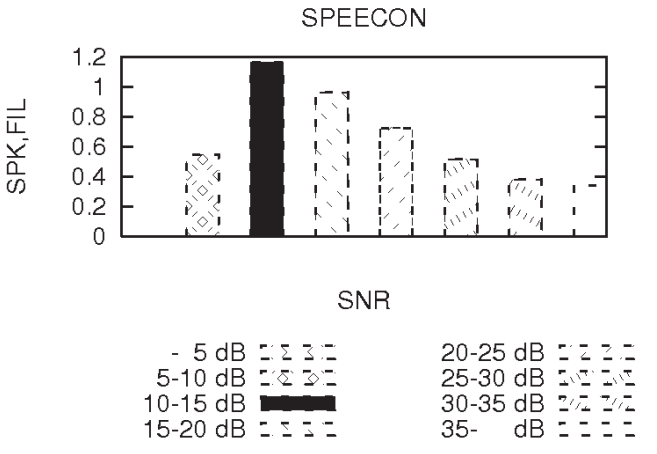

Fig. 3: Non-speech events in the training SPEE dataset in different noisy environment

© Czech Technical University Publishing House

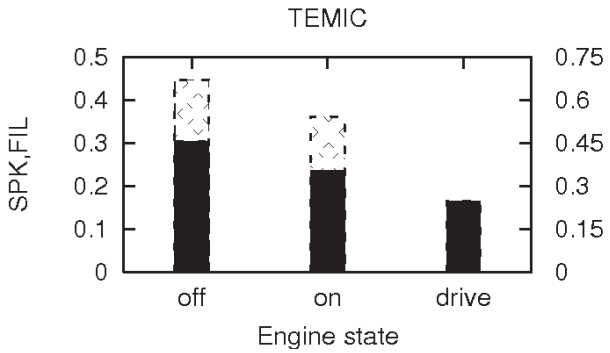

Fig. 4: Non-speech events in the whole (full filled bar) and training (lined bar) TEM dataset in different noisy environment

speaker is not disturbed by the environmental noise so he pronounces properly and this leads to a lower non-speech event rate.

2.3 Event distribution for different workgroups

The realization team for creating the two databases was divided into two parts with different supervisors and workplaces within the Czech Republic. This led to a different dialect distribution within the groups, as the records from the second group were spoken mainly by the speakers with a Moravian dialect. As shown in graph 5, this separation led to a difference in the annotations. Unlike the second group, the workers in the first group had to use fewer non-speech event marks in the transcription, and these items have a notably lower rate of non-speech event occurrence.

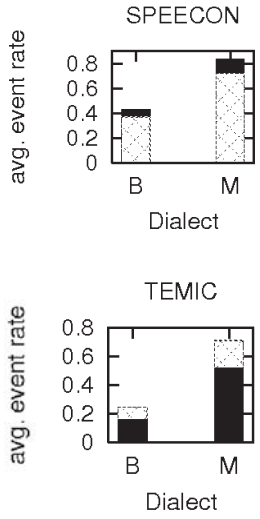

Fig. 5: Non-speech events in the whole (full filled bar) and training (dashed bar) datasets for different dialect/workgroup

This may have caused some of the marked non-speech events from the second group to be less loud or less prominent. Because such events do not decrease the recognition score, it is quite undesirable to train the recogniser to separate them from common background noise. If these unimportant event marks were not there, the recogniser would take only significant events in account and would be able to model the significant events more properly. This is one reason for using forced alignment on the training dataset see below.

3 Previous tests

The analysed databases were used to create a Czech digit sequence recogniser that models speaker non-speech events [1]. In the first step, only two classes of speaker non-speech events were modelled and both datasets could be used. In subsequent processing, the SPK event was divided into separate plosive and fricative events. Therefore only TEM was used for subsequent training, because there was no information on these properties in the SPEE dataset. The recogniser was tested on the selection of both datasets (because of the different environment).

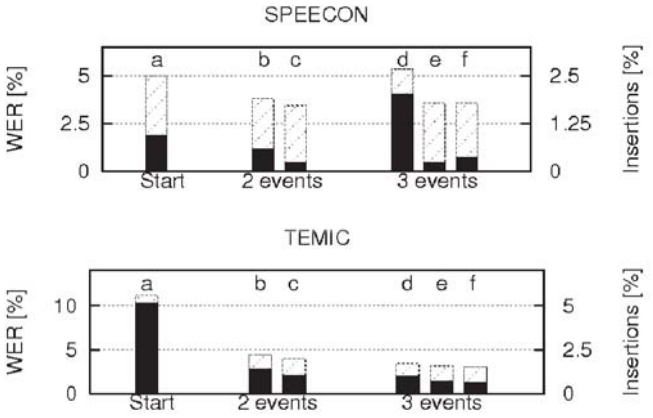

Fig. 6: Word error rate (dashed bar) and insertion error rate (full filled bar) on the SPEE-testdat (top) and on the TEM--testdat (bottom)

Figure 6 shows the recognition results for the testing data derived from SPEE and TEM in terms of word error rate

$$WER = \left(\frac{D + S + I}{N} \right) \cdot 100\ \% \qquad (1)$$

where N means number of recognised words, and D stands for deleted words, S for substituted words and I for incorrectly inserted words.

At the beginning of non-speech event modelling there was a recogniser trained on SPEE only (col. a). In the first step TEM was added to the training set and also two basic non--speech event marks were added to the set of models. The first retraining decreased the error rate even without using non--speech events (col. b), but the recogniser that takes the events into account has better results (col. c). Additional retraining brought no improvement, so only TEM was used for subsequent phases and the general SPK mark was divided into two classes: fricative (BRE) and plosive (PL) event. After two retraining steps the results show that, unlike in the case of non-speech event modelling (col. e, f), the word error rate and insertion error rate increase rapidly without using the additional models (col. d). So the experiment therefore shows that the non-speech event modelling approach helps recognition improvement.

4 Experiments on forced-alignment

As noted above, the SPEE database divides speaker non--speech events into two classes only, but it is better for the recogniser to have more classes of these events. The more

© Czech Technical University Publishing House

classes are used, the more accurately the events can be described and modelled. Therefore three models of non-speech events were used in the recogniser above, which meant using only TEM for subsequent retraining and the training dataset size decreased.

Using the SPEE database for subsequent training again needs a decision on, whether the SPK mark in the transcription closer to a BRE or to a PL event. The HTK system enables a feature called forced alignment, which tries to recognise the training database and puts the most fitting (probable) form of the given word into the record. This is used for deciding between pronunciation variants of one word, and in this work it was used in a similar way for speaker non-speech events.

In this experiment the SPEE dataset was re-recognised using the recogniser above. Because the recogniser is able to classify the SPK event, the result of this recognition is the SPEE database, which has 3 classes of speaker non-speech events, like in the TEM database.

4.1 Breath-like noise

The recogniser above uses two types of BRE model. One is based on the phonetically similar phoneme "f", while the second takes advantage of the length of silence model. So it was necessary to decide which should be used for subsequent processing. The recognition results (Fig. 6, col. e, f) show that in the two cases the recognition score is almost the same.

It seems that both models have the same ability to describe the event, but using forced alignment on the SPEE dataset discovered the quality of these models. This was done by comparing of the results of the re-aligned SPEE when using the BRE models with a different basemodel.

The recogniser was trained in several retraining steps. In the last step (*phase 2*) and in one preceding step (*phase 1*) a forced alignment of the SPEE database was performed Table 2 shows the difference between the 'f'-based and silence-based forced alignment in those two succeeding retraining phases. This led to 4 comparisons. We analysed whether one recogniser had marked the non-speech event with the same mark (BRE vs PL) as another recogniser (substitution). And we noted as a deletion the situation, when one of them had not placed a mark where the second one marked the event.

In the case of 'f'-based BRE models, the event was marked the same way as by other recognisers in 89.5% cases at maximum. This shows that the recogniser does not have stable models and these models continue to change their properties notably while being trained. As a result one event will serve as a BRE event for a while, but then it will be used to train the PL model, because of the non-stability of the forced-alignment results.

Table 2: Comparison of re-aligned BRE non-speech events in different retraining phases

		silence, phase 1	'f', phase 2
'f'-based, phase 1	**Subst.**	**24.18 %**	**10.44 %**
	Deletion	**5.34 %**	**4.14 %**
silence-based phase 2	**Subst.**	**4.44 %**	**18.77 %**
	Deletion	**3.86 %**	**7.19 %**

On the other hand, the comparison of silence-based models in both succeeding phases shows that these models act in a rather stable way. Therefore the event will be used to train only one type of non-speech event model, and seems to be more suitable for subsequent processing.

4.2 Aligned data-based recogniser

The first analysis (listening test) of the re-aligned data discovered faults in the ability of the recogniser to decide whether the SPK event is of BRE or PL type. In some cases the event was too close to the beginning word, sometimes the event was too silent, so high accuracy could be expected. However, the results expanded the training database, which could help the recognition.

Based on the results above retraining was performed with the use of the re-aligned SPEE training dataset. Both kinds of BRE models were used to check whether the comparison experiment has any effect on recognition accuracy.

Table 3: Recognition results after re-alignment with a different BRE basis

	Acc[%]	Insertions
'f'-based, original	96.44	3
sil-based, original	96.44	2
'f'-based, phase 1	95.96	3
sil-based, phase 1	95.96	2
'f'-based, phase 2	96.20	1
sil-based, phase 2	96.32	1

Table 3 shows that using a different base model for a BRE event has no significant influence on recognition accuracy. After two steps of retraining, the recogniser was able to eliminate all insertions except for the one that remained. But even when the accuracy does not achieve the value of the original recogniser, the re-alignment seems to bring an improvement for subsequent training.

4.3 Alignment against listening

As noted above the TEM dataset divides non-speech events into several classes. These events were marked by the human annotator, so the original transcription can be used as a good basis for re-alignment quality comparison. This transcription can be considered as a good estimate of the non-speech event class, and if the re-alignment phase marks some plosive event as BRE in many cases (or vice versa), the recogniser is unable to describe the difference between these events and it needs to be trained better.

Section 2.3 shows that not all the non-speech events marked in the original transcription of the speech datasets can be considered as a suitable pattern for training the event model. So the re-alignment can help to reduce the difference between the marked events by removing silent events, which can be modelled by the silence model.

© Czech Technical University Publishing House

Table 4 shows the number of non-speech event marks BRE and PL that were deleted from the original training subset (annotated by a human being) in different training phases and which were substituted for each other. For the SPEE dataset there was no human-aligned transcription for these non-speech event classes, so there is no information about substituted marks.

Table 4: Comparison of non-speech events marked in the original and re-aligned training subsets

phase	Deleted	Substituted
SPEECON		
phase 1 before re-align	1116	–
phase 2 before re-align	1723	–
phase 1 after re-align	2169	–
phase 2 after re-align	2282	–
TEMIC		
phase 1 before re-align	1381	424
phase 2 before re-align	1381	566
phase 1 after re-align	1296	458
phase 2 after re-align	1286	448

The number of deleted marks of non-speech events rises in the case of SPEE subset, but the difference between the number of deleted marks in particular phases decreases. The recogniser therefore tends to some final form of non-speech event models that consider about 2300 marks in the original transcription as too silent or inappropriate in some other way. For the TEM subset this number does not change notably. This may be because the recogniser was trained on the TEM subset in all 4 phases, while the SPEE was used only in the last two training phases.

The substitutions in Table 4 show that training leads to a decreasing number of substituted marks. This effect means that the recogniser classifies the non-speech events similarly to the human annotator, and so the recogniser can re-align the data more precisely. The substitutions are more often caused by marking the PL event as a BRE event. A simple listening test showed that plosive non-speech events are followed by or mixed with breath in some cases and only the PL event is marked, so the substitution does not necessarily indicate a bad model of a non-speech event.

5 Conclusion

This paper describes some analyses and tests in a non-speech event modeling task. The analyses of the training datasets show some properties that can influence recognition accuracy. Then some recognition tests were performed to find the best way to model speaker non-speech events. A spoken Czech digit sequence recogniser based on phoneme HMMs was used for this purpose.

The speech databases used for the experiments were analysed, and it was found, that a part of both sets contains a notably different non-speech event rate. This was caused by the different supervision in the annotation phase of database creation. The distribution in different noise backgrounds supports the intuitive conclusion that a high noise level covers non-speech events, and so the occurrence rate decreases. For highly silent environments the rate is also lower, so not only the cleanest items are best for the non-speech event recognition task.

The analysed datasets were used for training the recogniser, and although they were not checked to ensure that they contained only suitable non-speech items, using non-speech event modeling feature brought a notable improvement. This recogniser was used to re-align one of the training datasets to get a more accurate description of the non-speech events. This reduced the insertion error. The choice of the model that stands as a basis for non-speech events seems to be of less importance, because after the retraining difference slowly disappears.

Acknowledgments

The work presented here was supported by GAČR 102/05/0278 “New trends in voice technologies research and usage”, AVČR 1ET201210402 “Voice technologies in information systems”, IGA MZ ČR NR8287-3/2005 and research activity MSM 6840770014 ”Research in the Area of the Prospective Information and Navigation Technologies”.

References

[1] Rajnoha, J.: Modeling of Speaker Non-Speech Events in Robust Speech Recognition. *Proceedings of the 16th Czech-German Workshop on Speech Processing*, Prague: Academy of Sciences of the Czech Republic, Institute of Radioengineering and Electronics, 2006, p. 149–155.

[2] Young, S. et al.: *The HTK Book (for HTK Version 3.2.1)* Cambridge University Engineering Department, 2002.

[3] Gajic, B., Markhus, V., Pettersen, S. G., Johnsen, M. H.: Automatic Recognition of Spontaneously Dictated Medical Records for Norwegian. *COST278 and ISCA Tutorial and Research Workshop – ROBUST 2004*, 2004.

[4] Shriberg, E. E.: Phonetic Consequences of Speech Disfluency. *Proceedings of the International Congress of Phonetic Sciences*, San Francisco, 1999, p. 619–622.

[5] *SPEECON project webpage*. http://www.speechdat.org/speecon.

Josef Rajnoha
e-mail: rajnoj1@fel.cvut.cz

Dept. of Circuit Theory

Czech Technical University in Prague
Faculty of Electrical Engineering
Technická 2
166 27 Praha, Czech Republic