# CZECH LANGUAGE DATABASE OF CAR SPEECH AND ENVIRONMENTAL NOISE

*Petr Pollák, Josef Vopička, and Pavel Sovka*

Czech Technical University in Prague,
ČVUT FEL K331, Technická 2, Praha 6, Czech Republic
E-mail: pollak@feld.cvut.cz, vopickaj@feld.cvut.cz, sovka@feld.cvut.cz

## ABSTRACT

This paper will present new Czech language two-channel (stereo) speech database recorded in car environment. The created database was designed for experiments with speech enhancement for communication purposes and for the study and the design of a robust speech recognition systems. It respects car noise environment which is currently at the top of the interest. Tools for automated phoneme labelling based on Baum-Welch re-estimation were designed. The noise analysis of the car background environment was done.

## 1. INTRODUCTION

Speech processing systems start playing more important role in the human life; voice driven control, information, or security systems, noisy speech preprocessing for communication purposes, etc. For reliable performance in real life.

These systems must be robust to background environmental noise. The environment in the running car is at the top of the interest. Typical application could be the pre-processing of noisy speech for telephony purposes from the car or automated speech recognition in car environment.

For the design of such system, the database of car speech and noise is the first and the most important requirement. We describe the creation of two-channel car speech database in this paper. The following problems were solved:

- database structure definition,
- data collection and digitalisation,
- tools for automated labelling,
- noise analysis of collected data.

## 2. DATABASE STRUCTURE

The contents of the database was extended from the database of isolated and sequentially spoken digits with respect of wider area of database using. The database should be used for the speech enhancement experiments and for the design of car-oriented speech recognition systems. From this point of view the following requirements were defined :

- The database must contain separate speech signals and car noise. These signals are important for experiments with artificially mixed signals, for evaluation of speech enhancement algorithms, because it allows us to quantify exactly different criteria of speech enhancement.
- Real car-noisy speech must be included for the confirmation of the simulated experiments. The evaluation of different criteria is harder because they must be estimated. The study of Lombard effect studies in robust speech recognisers is possible in comparison of experiment results with artificially mixed signals and real noisy speech.
- Isolated numerals, city names, word commands are included to respect the typical car applications.
- Three sentences per each speaker are included for involving longer fluent speech without pauses. It is important for studying of noise adaptation ability of speech enhancement systems. It is also necessary for having more phonetic material to train speech recogniser.

### 2.1. Types of signals

According to requirements mentioned above, three basic groups of signals are in the database.

| | | |
|---|---|---|
| 1. | *clean speech* | signals recorded in a quiet car |
| 2. | *background noise* | recorded in a running car without speech |
| 3. | *noisy speech* | signals recorded in a running car |

#### *2.1.1. Speech material per speaker*

Different types of utterances were recorded from each speaker. It is summarized in the following table.

| |
|---|
| Isolated digits $0 \div 9$ |
| Connected digits $0 \div 9$ |
| Natural numbers (telephone numbers) |
| City names |
| One word commands |
| Commands in the sentence |
| Phoneticaly rich sentences |

*2.1.2. Background noise signals*

The signals of this group contain the typical car background environment. They map almost all possibilities of background noise at different modes of the ride. Also the characteristics of different cars were recorded.

## 3. RECORDING AND HARDWARE

All signals were recorded in the car into DAT-recorder and transferred to PC directly by using sound card with digital input.

### 3.1. Microphone placement

Since the two-channel signals (stereo) were recorded, the microphone placement in the car had to be solved. Finally, the optimal placement was found at the top line in the car body. Exactly, the microphones were placed at driver sunshade and at the sunshade in the front of right side place in a car. That means asymmetrical placement from the speaker (which should be driver) point of view.

This placement is not optimal for algorithms using coherence or correlation methods. The results are strongly influenced by different delay between two signals in the channels which strongly varies in the dependence on speaker head movement. The placement on the left side jamb of car body seemed to be better because the delay between the signals was not changed by the head movement. Unfortunately, when the window of the car was opened, the signals were completely masked by the air turbulence. It means that this placement is unusable.

### 3.2. Data digitalisation

The sound card SOUND BLASTER LIVE with the digital input/output was used for the data transfer into PC. Finally, the parameters of the signals are summarised in the following table.

| | |
|---|---|
| *sampling frequency* | $f_s = 16000$ *kHz* |
| *speech coding* | PCM 16 *bits* |
| *microphone distance* | 50 *cm* |
| *delay between signals* | cca 10 *samples* |

## 4. DATA LABELLING

Since the assumed usage of the database is very wide, all signals in the database must be well labelled from many different aspects as SNR, speech/pause sequences, and typically the most important labels for words and for phonemes.

### 4.1. Orthographical transcription

The variety of pronunciation of Czech words brings the need of the orthographical transcription. This problem can be divided into three parts.

- Regular changes of the pronunciation of written text. These changes does not have to be included in to the orthographical transcription. They are for example voiced/unvoiced consonant changes on the boundaries of words, insertion or deletion of consonants, creation of diphthongs, etc.
- Spoken forms of words (colloquialisms). It is important to well transcribe this kind of changes and also to safe connection to the written form of this word. The numerals are special part of this case where one number can have several pronunciations.
- Foreign words. The transcription rules are the same like in the previous case. It could be simply described "write what you hear".

### 4.2. Phoneme labelling

Semiautomatic phoneme based labelling was done using HTK toolkit [1], [5]. Models were trained on database of radio sport news and forecast [3]. The text transcription was orthoepically corrected and transformed to the appropriate structure of HMMs for each sentence. After that Viterby recognition was performed. Since this database was collected in special environment which differed from the origin of the HMMs, models had to be adapted. The adaptation was done by Baum-Welch re-estimation algorithm [5]. Three re-estimations were done and new labels were evaluated at each step. The phoneme HMMs adaptation was measured by the value of the average phoneme boundaries shift between label versions. The HMMs adaptation was acceptable when this number fell below some threshold level.

The labelling was implemented also for noisy speech. The re-estimation was done by the same way like for the clean speech at the first time, but the boundaries shift decreased slowly. One simple approach brought an improvement. The approach was to re-estimate the silence models in advance. It means that silence models were re-trained and after that they were added in to the group of the non re-estimated models. Finally this group of models was re-trained in three rounds of the re-estimation and results (labels) were evaluated. The average boundaries shift decreased more rapidly and reached lower level in this case. The results for the noisy data are worse in comparison with the clean speech data. This results will be shown later in the results section.

Technical data of the HMM

- down-sampled binary data (8kHz),
- 13 mel-cepstral coefficients, 10ms frame rate,
- 40 phoneme and 3 silence models,
- each phoneme model has 5 states, the first and the last state are non emitting, each state has 3 mixtures and each mixture has 3 streams,
- the delta and delta-delta coefficients are generated on the fly.

,

### 4.3. Word and sentence labels

The database was labelled also in words. The models of words were composed from phoneme models. This composition was driven by fixed orthoepical rules for

the Czech language. Exceptions were especially foreign words and colloquialisms. These words had to be manually transcribed. This transcription can be found in orthographical section of the database.

### 4.4. Voice activity detection

The final labels were used for detection of the voice activity. The conversion between labels and speech detection was done by simple perl script.

## 5. NOISE ANALYSIS

The data analysis from the noise point of view was also done. Typical characteristics as spectra, coherences, SNRs which are important for parameter setting of studied algorithms were found.

From this point of view three basic categories of the noise were recorded.

| |
|---|
| *Stationary noise*<br>ride with constant speed at different gear on the different surface of the road |
| *Non-stationary with slow changes*<br>increasing and decreasing speed without change of the gear, the changes in noise characteristics are slower than the changes in speech characteristics |
| *Non-stationary with quick changes*<br>gear change, window opening, paved road, street noise, the rates of the characteristic changes are same for the speech and noise |

Segmental SNR were estimated and statistically analysed for all speech signals. Its mean value is relatively very low approximately -5dB.

## 6. RESULTS OF EXPERIMENTS

### 6.1. HMM adaptation

| *iteration* | *speech* | *mix* | *noise advance* |
|---|---|---|---|
| *0 x 1* | 0.04994 | 0.48731 | 0.58859 |
| *1 x 2* | 0.01261 | 0.31992 | 0.14543 |
| *2 x 3* | 0.00890 | 0.12833 | 0.06445 |

Table 1: Average boundaries shift in seconds after models re-estimation.

In table 1 it is shown how the HMMs adapt during the re-estimation. Each of the numbers in this table the average phoneme boundaries shift between label versions. It means that the labels are made for every version of the prototypes. After that the average distance between two following versions is computed. For example the expression *0 x 1* on the first row means that the numbers in this column are the results of the comparison of labels made without any re-estimation (*0*) and labels after the first (*1*) re-estimation.

There are three columns of results.

- The *speech* column shows results on clean speech part of the database,

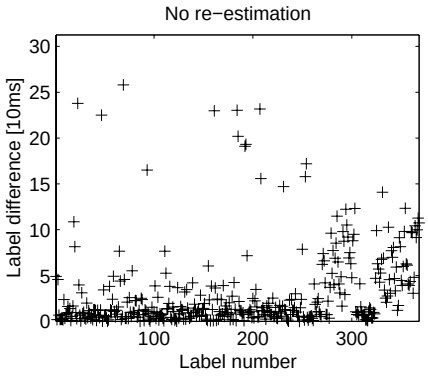


Figure 1: Automatic and manually made labels comparison, no re-estimation.

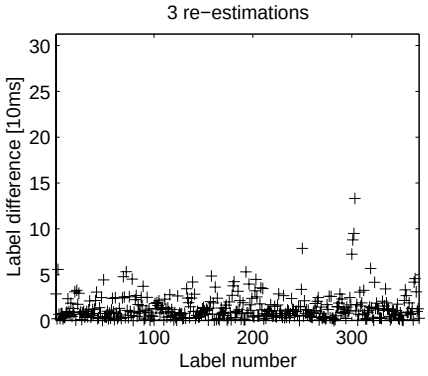


Figure 2: Automatic and manually made labels comparison after the 3rd re-estimation.

- *mix* results on the noisy part,
- The *noise advance* column shows results on the same data, but the silence models are re-estimated in advance.

As it can be seen the average boundary shift decreases with the number of re-estimations. The final shift for the clean speech is lower then 10ms. The same approach for the noisy data brought 12 times worse result. The shift decreased faster when we re-estimated the silence models in advance. The final result was better but still 7 times worse then for the clean speech.

### 6.2. Comparison with the manually made labels.

| *iteration* | *0* | *1* | *2* | *3* |
|---|---|---|---|---|
| *mean [ms]* | 31.8 | 14.4 | 14.5 | 13.8 |
| *std [ms]* | 44.5 | 15.4 | 18.3 | 14.9 |

Table 2: Automatic and manually made labels comparison.

The numbers in table 1 gave us information about the quality of the HMMs adaptation but we needed some objective criterion about the quality of the labels. There is only one way how to get this information. It is the comparison with the manually labelled representative sample of the data. The results of labelling using this method are reliable and we will demonstrate that the new phoneme boundaries are more

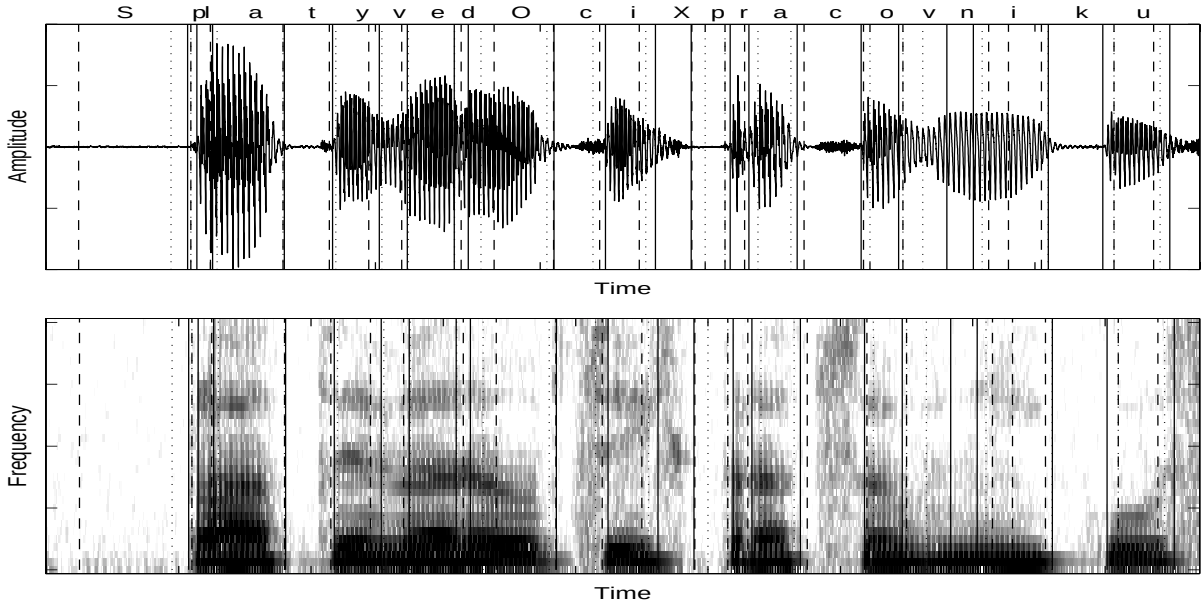


Figure 3: Labels comparison in amplitude and spectral view.

precise after above mentioned model re-estimations.

Representative sample of the data (16 sentences of clean speech with the 368 phoneme boundaries) were manually labelled by 3 men. The average values of the manually made labels were computed. This average labels were compared with the automatically generated labels. Results were statistically evaluated.

Table 2 shows this comparison. The numbers in the first row show the order of re-estimations of the phoneme HMMs. The average difference between automatic and manually made labels (*mean*) is in the second row. The standard deviation of this difference (*std*) is in the last row. This table gives us information about the importance of the re-estimation. Improvement can be seen after the first re-estimation. Higher orders do not bring next changes. This can be caused by small number of compared data.

Figure 1 and figure 2 show distances of all compared labels and graphically demonstrate numbers in table 2.

Figure 3 demonstrates labels positions on the one of manually labelled signals. There can be seen manual (solid vertical line), before models re-estimation (dashed line) and after 3rd re-estimation (dotted line) labels.

## 7. CONCLUSIONS

- The clean speech signals were collected from more than 100 speakers.
- Since the collection of real noisy speech is more complicated, this part of database contains at this time the data from 20 speakers only.
- Noise signals of car background environment were collected in 5 different cars. Approximately $20 \div 30$ min per car of typical background environment were recorded.
- Tools for automated labelling of the collected data were designed. The experiments showed that they give satisfactory results.
- The database can be used for speech enhancement evaluation or for training of robust speech recognition system.
- Since there isn't any publicly available Czech database with similar description, the clean part of the database should be used also for the study and the design of different automated recognisers of Czech speech without environmental noise.

### Acknowledgements

This study was supported within the project COST 249 "Speech recognition over telephone line" and the grant GACR 102/96/k087 "The theory and application of speech communication in Czech".

## REFERENCES

[1] V. Hanžl and J. Uhlíř. Different approaches to definition of elements used in speech recognition. COST 249 meeting., 1996.

[2] O. Mella and D. Fohr. Semi-automatic phonetic labelling of large corpora. In *Eurospeech'97 - Proceedings of the 5th European Conference on Speech, Communication, and Technology*, Rhodos, September 1997.

[3] P. Pollák and P. Sovka. Database of car speech, analysis of collected data, tools for automated labeling. In *8-th Czech-German Workshop on Speech Processing*, Prague, Czech Republic, September 1998.

[4] J. Vopička. French-Czech cross-language experiment. In *Poster 1998*, Prague, 1998. CTU FEE.

[5] S. Young. *HTK - User's Guide*. Cambridge, 1993.