CAR2 - CZECH DATABASE OF CAR SPEECH

Petr POLLÁK, Josef VOPIČKA, Václav HANŽL,
Pavel SOVKA
Dept. of Circuit Theory
Czech Technical University in Prague
Technická 2, 16627 Praha 6
Czech Republic

Abstract

This paper presents 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. Tools for automated phoneme labelling based on Baum-Welch re-estimation were realised. The noise analysis of the car background environment was done.

Keywords

speech, speech enhancement, speech recognition, databases, labelling

1. Introduction

It is already possible to find working systems for automated speech processing in Czech republic. Since they start appearing in standard human life and since the requirement for their reliable performance in real life becomes to be the most important one, the databases of real life speech are the most important and necessary condition for a design of such systems.

The subject of this paper is to describe a database (DB) from the car environment (CAR2). With the development of car equipment and also mobile communications, one can find following typical tasks to be solved: automated speech recognition in car environment for applications in the car (voice dialing, voice control), enhancement of speech transmitted by mobile phone in the car, voice driven teleservices controlled from running car, etc. Presented DB is designed to be used for an evaluation and a design of these systems.

From this point of view the following requirements were defined for this new database:

- Real car-noisy speech is the most important part of the DB. It is necessary for the recognizer training under real conditions. This material shouldn't contain only real background environment in speech signals but also Lombard effect.
- The DB must contain separate speech signals and car noises. These signals are important for experiments with artificially mixed signals. It allows us to quantify exactly a different criteria of speech enhancement.
- The speech material in the DB should respect the most probable car applications and all Czech phonemes should be well covered.
- Longer natural sentences are also important for studying of noise adaptation ability of evaluated systems.
- Orthographic transcription must be well done with respect to irregularities of pronunciation. It must be manually checked.
- Labelling tools must be automated as much as possible.

2. Database description

The DB has three different parts: signals, signal labels and description, and tools.

2.1 Signals

According to the requirements mentioned above, following types of signals are in the DB:

1.	*speech*	speech signals recorded in a quiet car without background noise
2.	*noise*	background noise recorded in running car without speech
3.	*mix*	speech signals recorded in running car

Table 1. Types of signals in the database.

2.1.1 Speech corpus

Speech recorded in a quiet car without noise is important especially from the point of view of a evaluation and a design of final systems. It is suitable to have artificially mixed signals with the reference to clean speech during this period. Moreover, since the collection of the data in the real noisy car environment is very difficult, this part is currently the most important part of the DB.

Collected speech material was chosen with respect to final assumed car application, i.e. car information systems, simple digit and command recognition, etc. Also three phonetically rich sentences per each speaker are included in DB. It is necessary for having more phonetic material to train phoneme based speech recognizers and also for hav-

ing longer fluent speech sequence without pauses to study behavior of speech enhancement systems in real situation.

Final set of the items in *speech corpus* per each speaker[1] is summarized in the Table 2.

4	**Isolated digits** 0-9	sn01-sn04	xn01-xn04
2	**Connected digits** 0-9	sf01, sf02	xf01, xf02
3	**Natural numbers** (number of telephone, credit card, sheet No.)	st01, st02 si01	xt01, xt02 xi01
1	**Name and surname**	sj01	xj01
4	**City names**	sm01-sm04	xm01-xm04
5	**Commands**	sc01-sc05	xc01-xc05
2	**Commands in the sentence**	sp01-sp02	xp01-xp02
3	**Phonetically rich sentences**	ss01-ss03	xs01-xs03

Table 2. Speech material per each speaker for CAR2 version 3.0[2].

2.1.2 Noise corpus

The *noise corpus* contains signals with typical car background environment. They were collected in different cars and they are always divided into three groups according to their characteristics. It is summarised in Table 3.

Stationary noises (*n0-corpus*) same speed and situation, different gears and road surfaces	n0* n0001, n0002, ...
Non-stationary with slow changes (*n1-corpus*) relatively slow changes in speed	n1* n1001, n1002, ...
Non-stationary with fast changes (*n2-corpus*) changing of the gear, window opening, turn indicator lights, paved road, street noise, ...	n2* n2001, n2002, ...

Table 3. Type of noises in the database.

The last two groups are selected according to relation to changes in speech characteristics. Non-stationary noise with slow changes has slower changes in the characteristics comparing to the changes in speech characteristics. In the third group the speed of changes in the noise and speech characteristics are comparable.

2.1.3 Mix corpus

The *mix corpus* has the same structure *as speech corpus* described in table 2. The signals were recorded in real running car situations so all above mentioned types of background environment are included. The signals were sorted according to the type of the utterance, e.g. xi01 for isolated numerals. Type of the background environment is mentioned in the description file only. There is not any sorting from this point of view since the final structure of the DB would have been too complicated.

2.2 Labels and description

Necessary part of the DB is the description and labels of all signals. Since we want to use the final DB in different research activities from the pure speech recognition without noise (*speech corpus*) up to speech enhancement for communication purposes only, signals must be labelled from different aspects.

2.2.1 Orthographic transcription

The first important information is orthographic transcription of spoken utterance. It is then used for the creation of orthoepic transcription, i.e. a sequence of phonemes [5], using tool "*transc*" which was developed in our laboratory during previous research. Standard lower-case Czech alphabet is used for almost phonemes which is joined by a number of upper-case letters for some special phonemes, e.g. 'H' is for 'ch' etc.

Orthographic transcription is automatically generated from the prompt sheets of each recording session. It must be done exactly in words not in digits, symbols, etc. Moreover, it must be manually checked because really spoken utterance can differ from the prompted one and also exact form of the utterance cannot be sometimes predicted, typically for natural numbers. Also the irregularities in the pronunciation, so typical for Czech language, must be marked. From these points of view, following rules for orthographic transcription were defined:

- **Regular changes of the pronunciation** of written text do not have to be included into the orthographic transcription, for example voiced-unvoiced consonant changes on the boundaries of words, insertion or deletion of consonants, creation of diphthongs, etc.
- **Spoken forms** (colloquialisms) **or irregular pronunciations** are transcribed with the connection to the written form of the word, i.e. "(written form/spoken form)". The digits are special part of this case because some digits have several pronunciations.
- The transcription of **foreign words** is the same as in the previous case. It should be simply described "write what you hear" with the written form.
- **Non-speech sounds** like hesitations, load breath, filled pauses, laugh, etc. must be also marked.

Two examples of utterance annotation:

Orthographic transcription:

1. Měl panický strach [mlask].
2. Měl (panický/panycký) strach [mlask].

Orthoepic transcription:

1. mňel paňycký straH
2. mňel panycký straH

[1] Some items can be missing in final DB because of recording problems, too high background noise for clean *speech corpus*, etc.

[2] Speech corpus in older version slightly differs, but the most important items are the same.

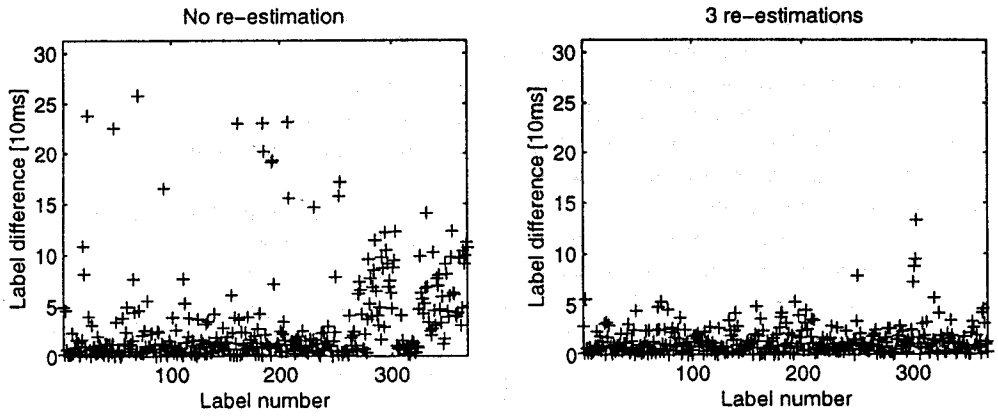

Figure 1. Manual and automated label comparison.

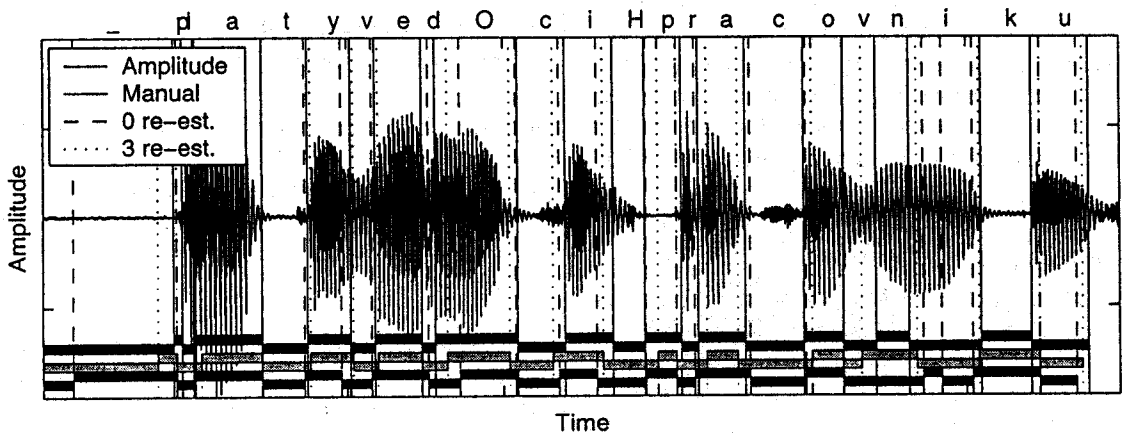

Figure 2. Illustrative example of automated labelling.

For the manual check of orthographic transcription (annotation) "*annotator*" tool is used. It is simple TCL/TK tool for editing prepared text connected to speech signal with the possibility of playing the signal. It runs on Linux platform.

2.2.2 Automated phoneme labeling

Correct orthographic transcription is used for phonetic transcription and then automated labelling is done using HTK toolkit [1]. Used phoneme models were trained on database of radio sport news and forecast (RADIO) [4]. The phonetic transcription is orthoepically corrected and transformed to the appropriate structure of HMM for each sentence. After that Viterbi recognition is performed.

Since this DB was collected in the environment which differed from the environment used for original models training, these models had to be adapted. The adaptation was done by Baum-Welch re-estimation algorithm [1]. Three re-estimations were done and new labels were evaluated at each step.

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.

The phoneme HMM adaptation was measured by the value of the average phoneme boundaries shift between label versions. The HMM adaptation was acceptable when this number fell below some threshold level. There were 3 adaptations for our case. Detail description can be found in [3].

The quality of the labelling can be measured only by the comparison with manually made labels. We had a set of manually labeled signals for which the difference between manually and automatically set labels were evaluated. The results are shown at the Figure 1 where these differences are shown for labelling done by Viterbi algorithm with models without any adaptation and for labelling with models after 3 re-estimations. Another illustrative example is on the Figure 2 where these three labels are evaluated for one signal.

The labeling was implemented also for noisy speech. The re-estimation was done in the same way as for the clean speech at the first time but the convergence of models re-estimation was slower. That was reason why the silence models were re-estimated in advance and after that they were added in to the group of the not re-estimated models. 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. Nevertheless, the results for noisy data were worse in comparison with the clean speech data.

2.2.3 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. The exceptions were foreign words and colloquialisms. These words had to be manually transcribed. This transcription can be found in orthographic section of the database.

2.2.4 Voice activity detection

Voice activity detector (VAD) is many times necessary part of speech enhancement systems. Since real VADs fail frequently in real noisy conditions, ideal voice activity detection (VAD) is very useful for algorithm evaluation.

VAD is automatically generated from final phoneme or word labels because our automated phoneme labelling gives satisfactory results. This conversion between phoneme labels and VAD was done by simple "perl" script.

2.2.5 SNR labels

Very important information in a noisy speech DB is the information about signal-to-noise ratio (SNR). However, the estimation of SNR from noisy speech is non-trivial problem. The following points must be taken into account:

- **noise power estimation** - We use two algorithms: averaging during speech pauses with VAD and Martins algorithm [2] using power minima tracking.
- **global, local, and segmental SNR definition** - Global SNR is affected by different pauses between speech activity in measured signal. Segmental SNR (SSNR) is defined as the average of local SNR, see equation (1), and when it is evaluated only during speech activity it corresponds to real noise level in the signal.

$$\mathrm{SNR} = \frac{1}{L} \sum_{i=0}^{L-1} 10 \log \frac{P_S(i)}{P_N(i)} \qquad (1)$$

The estimations of SSNR and (local SNR) are included in the DB for all speech signals, both in the corpora speech and mix. Its mean value for mix corpus is relatively very low, approximately -5dB.

2.3 Tools

The DB structure contains the directory "*tools*" were many different tools used for the data processing are joined. In the evaluation version there are:

- "*cz2cz*" - simple tool for the conversion of Czech texts with different coding (Linux),
- "*transc*" - tool used for the generation of phonetic transcription of the utterance from the orthographic transcription (Linux),
- "*annotator*" - TCL/TK application for the check of orthographic transcription (Linux),
- "*sox*" - shareware software for different sound format conversion (C-code, Linux),
- "*vplay*" - simple tool for raw sound file playing (Linux),
- "*m*" - directory with different MATLAB utilities (m-files, mex-files, dll-files) for the basic signal manipulations; e.g. *loadbin.m, loadwav.m, savebin.m, savewav.m*, etc.

Many of these utilities can be also found at our WEB *server "http://noel.feld.cvut.cz"* or as shareware programs somewhere else at the Internet.

3. Recording hardware description

All signals, inclusive *speech corpus,* were recorded in a car using DAT-recorder and then transferred directly to PC using sound card with digital input. In this case it is not necessary to model acoustic properties of car cab and it is possible to obtain the approximation of real noisy speech by artificial mixing of signals from *speech* and *noise corpora*, of course, without Lombard effect in this case.

3.1 Microphone placement

Optimal microphone placement was found at the top line in the car body. Exactly, the microphones were placed at sunshades according to Figures 3a. The speaker is assumed to be the driver. But it is very difficult to keep this requirement, especially for a collection in the running car. But the speaker is always close to the left channel microphone, i.e. the configuration according Figure 3b is used for speaking driver adjoined person.

This placement is called asymmetrical, i.e. from the point of view of speaker, and consequently speech signal in the second right channel is delayed. This strongly varies in the dependence on speaker head movement and it can cause the failure of algorithms based on coherence or correlation methods.

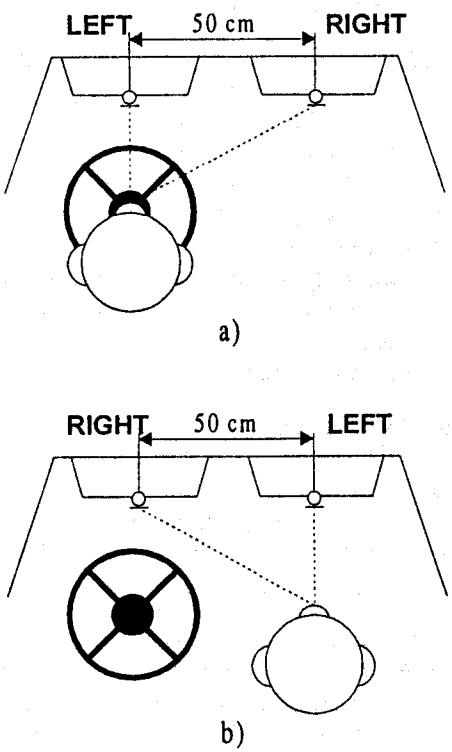

Figure 3. Microphone placements in a car.

From this point of view this placement is not the best one so some other possible placement were studied. But always other undesirable aspects appeared. E.g. 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, nevertheless, when the window of the car is opened, the signals are completely masked by air turbulence. So the placement used in our recordings is the result of the compromise from many different aspects.

3.2 Data digitalisation

The sound card SOUND BLASTER LIVE with the digital input/output was used for the data transfer into PC. The SPDIF interface was used and the down-sampling of the data to 16 kHz was provided using standard tools of mentioned sound card.

Final parameters of the signals are summarized in the following table 4.

Sampling frequency	fs = 16000 Hz
speech coding	PCM 16 bits
number of channels	2
microphone distance	50 cm
delay between speech signals	approx. 10 samples
file formats	WAVE, RAW (both)

Table 4. Parameters of signals in the DB.

These first versions of the DB contain also raw binary files of both left and right channels, "*.bil" and "*.bir" files. These files are widely used under UNIX platforms. Since we are interested in telephony speech applications, also band-limited and down-sampled (8 kHz) telephony data are included in the DB. The automated labelling was performed on these data because original HMMs were trained for telephony purposes.

4. CD-ROM

The database is currently on one CD-ROM with the structure described on Figure 4. There are currently two version of the DB: the version 3.0 and the older version 2.3 with data transfer using analogue output of DAT-recorder and slightly different *speech corpus* per speaker.

For the compatibility of the DB at different platforms we strictly use short names for all files (ISO 9660 format), all label files are in easily readable text formats, and all tools are designed to be easily executable at different platforms.

Character set for Czech texts in the label files is always ISO-8859-2 (iso-latin-2) which is used at UNIX platforms. Of course, the labels could be easily converted using e.g. "*cz2cz*" tool developed for Czech texts in our lab or many other tools available on the Internet.

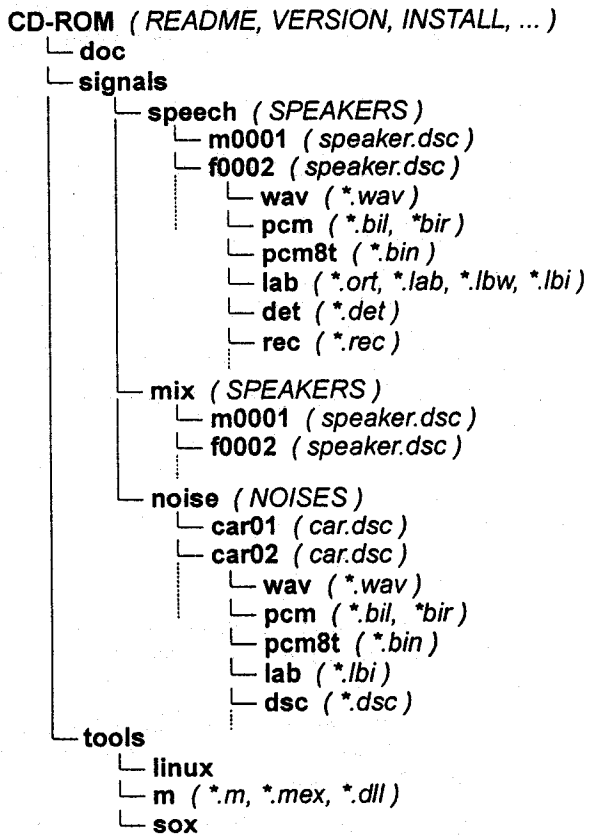

Figure 4. Structure of the data at the CD-ROM.

5. Conclusions

Current state:

- The dB contains currently the data recorded approximately by 100 speakers. It can be used for the evaluation and the design of systems in car noise environment, both for robust speech recognition and speech enhancement for transmission purposes.
- The *speech corpus* contains more than 70 speakers while the *mix corpus* approx. 20 speakers only because the recording in the running car is very difficult and expensive.
- Tools for automated labelling were designed and used for phoneme labelling of this DB with very optimistic results for the *speech corpus*.
- The labelling of *mix corpus* gave a little bit worse results, but some improvement was achieved by separate re-estimation of speech pause models only.
- The rules for the annotation of spoken utterances were defined with respect to current knowledge about phonetic inventory of Czech language and irregularities of Czech pronunciation.
- Since there is not any other large Czech speech DB with similar description, the *speech corpus* should be also used for the recognition in noisy free environment.

Future work:

- We assume the continuation of this work by the extension of speakers representation.
- The work will be focused especially to the *mix corpus* extension.
- Labelling techniques must be improved especially for noisy speech again.
- Final structure should contain also some table files and DB contents files to improve the orientation in the DB structure.

References

[1] YOUNG,S.: HTK Users Guide, 1993.

[2] MARTIN,R.: An Efficient Algorithm to Estimate of Instantaneous SNR of Speech Signals, In proc. of *Eurospeech'93*, Berlin 1993.

[3] POLLÁK,P.-VOPIČKA,J.-SOVKA,P.: Czech Language Database of Car Speech and Environmental Noise. In proc. of *Eurospeech'99*, Budapest 1999.

[4] VOPIČKA,J.: French-Czech Cross Language experiment. In *Poster 1998*, CTU FEE, Prague 1998.

[5] NOUZA,J.-PSUTKA,J.-UHLÍŘ,J.: Phonetic Alphabet for Speech Recognition of Czech. In Radioengineering, Vol.6, No.4, December 1997.

About authors...

Petr POLLÁK was born in 1966 in Ústí nad Orlicí (Czechoslovakia). He graduated in 1989 (Ing.) at Czech Technical University in Prague. After graduation he joined CTU and he has received CSc. degree in 1994 (Ph.D. equivalent). In 1996 he stayed 3 month at ENST Telecom-Paris. Currently, he works as teacher and researcher at Department of Circuit Theory of CTU FEE. His main research activities are in speech processing, especially speech enhancement, noise robust speech recognition, speech databases collection, etc. He is member of ISCA (International Speech Communication Association).

Josef VOPIČKA was born in Prague, Czechoslovakia, 12th December 1972. He received his M.S. degree at the Faculty of Electrical Engineering of the Czech Technical University in Prague (FEE CTU) in 1996. He is Ph.D. student at the same faculty. His main research interest is speech recognition, especially word spotting.

Václav HANŽL was born in Prague, Czechoslovakia, February 1st 1965. He received his M.S. and Ph.D. degrees at the Faculty of Electrical Engineering of the Czech Technical University in Prague (FEE CTU) in 1988 and 1994 respectively. After studies he joined FEE CTU in his current position of a researcher and teacher. He spent five months at Hewlett-Packard Laboratories, Bristol, UK in 1991 and six months at l'Ecole Nationale Superieure des Telecommunications (ENST), Paris, France in 1994-5. His main research interest is large vocabulary continuous speech recognition. He is member of ISCA (International Speech Communication Association).

Pavel SOVKA was born in Jihlava, Czechoslovakia, on February 4, 1957. He received the M.S. and Ph.D. degrees in electrical engineering from the Faculty of Electrical Engineering of the Czech Technical University in Prague (FEE CTU) in 1981 and 1986 respectively. From 1985 to 1991 he worked in Institute of Radioengineering and Electronics of Czech Academy of Sciences, Prague. In 1991 he joined the Departement of circuit theory CTU FEE. Since 1996 he has been an associated professor. Among his current research interests the application of adaptive systems to noise and echo cancellation, speech analysis, changepoint detection, and signal separation can be found. Mr. Sovka is member of the International Speech Communication Association (ISCA).