

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
Fakulta elektrotechnická

CZECH TECHNICAL UNIVERSITY IN PRAGUE
Faculty of Electrical Engineering

**Design and Creation of Speech Databases - Measurement of
Signal Quality**

Návrh a tvorba řečových databází - měření kvality signálů

Ing. Petr Pollák, CSc.

Habilitation lecture / Habilitační přednáška

11 June 2003

Summary

This contribution describes the problem of design and creation of databases for speech recognition and enhancement. It is an intersectorial research field covering the problems from electronics, through digital signal processing, up to phonetics and linguistics.

In the first part of this presentation, the brief overview of speech databases creation is mentioned. This work is focused on databases for speech recognition. That is the reason why the most important requirements are oriented to the coverage of the utterance variability with respect to environment, speaker, or utterance contents. Exact definitions of these requirements are usually summarised in the database specification, the first step of database creation. Next step is the recording. In this phase, different analysis algorithms should be integrated into recording procedure to measure signal quality or to detect bad items respectively. The annotation of collected data is the last step of the database creation.

The second part of this presentation deals with some methods for collected signal analysis, often integrated into database recording platform. More precise description is devoted to the definition and estimation of speech SNR and to the voice activity detection. Basic SNR criteria are defined: global SNR, segmental SNR, arithmetical segmental SNR. These criteria are compared especially from the point of view their estimation. The algorithm of voice activity detection based on cepstral analysis is described, commonly with its importance for the purposes of speech SNR measurement.

Finally, the overview of created databases is presented as the main conclusion of presented activity and further more general conclusions achieved in this research are also summarised.

Souhrn

Tato prezentace popisuje problematiku návrhu a tvorby databází pro účely rozpoznávání a zvýrazňování řeči. Jedná se o interdisciplinární problematiku zahrnující problémy od elektroniky, přes číslicové zpracování signálů, až po fonetiku a lingvistiku.

V první části této prezentace bude uveden stručný přehled problematiky tvorby řečových databází. Hlavní pozornost je věnována návrhu a tvorbě databází pro rozpoznávání řeči. Z toho pak vycházejí definované požadavky na maximální pokrytí variability řeči z hlediska prostředí, mluvčího i obsahu promluvy. Konkrétní definice těchto požadavků tvoří specifikaci databáze, první nezbytný krok při tvorbě nové databáze. Dalším zmíněným krokem je pak vlastní nahrávání. V této fázi je v rámci konstrukce nahrávací platformy vhodné maximálně integrovat algoritmy analýzy signálů, zejména měření kvality signálů resp. kritéria pro detekci nekvalitních položek. Posledním krokem tvorby databáze je pak její anotace.

Druhá část této prezentace je věnována vybraným metodám analýzy signálů, se zaměřením na definici a odhad SNR řečového signálu a detekci řečové aktivity. Jsou definována základní kritéria: globální SNR, segmentální SNR, aritmetické segmentální SNR. Uvedená kritéria jsou srovnána zvláště z hlediska rozdílných vlastností při jejich odhadu pro signál bez reference. Ilustrativně je popsán princip detekce řečové aktivity na bázi keprstrální analýzy a jeho vliv na zmiňované odhady SNR.

Závěrem je jako jeden z hlavních výsledků uveden přehled vytvořených databází a rovněž jsou shrnuty přínosy presentované práce pro další aktivity v dané oblasti výzkumu.

Keywords:

speech database, speech enhancement, speech recognition, signal-to-noise ratio, SNR, SNR estimation, voice activity detection, speech quality criteria, cepstral analysis

Klíčová slova:

řečové databáze, zvýrazňování řeči, rozpoznávání řeči, odstup signálu od šumu, SNR, odhady SNR, detekce řeči, kritéria kvality řeči, keprální analýza

Contents

1	Introduction	6
2	Speech databases overview	6
2.1	Specification of DBs for speech recognition	7
2.2	Collection of speech databases	8
3	Measurement of signal quality	10
3.1	Measurement of speech SNR	11
3.2	Methods for speech SNR estimation	12
3.3	Voice activity detection	15
4	Conclusions	17
	References	18
	Ing. Petr Pollák, CSc.	20

1 Introduction

Since 80-th years, the speech processing is one of the important research area at the Department of Circuit Theory, at Czech Technical University in Prague, Faculty of Electrical Engineering. During the years many different problems were solved: starting with basic speech analysis and the simplest isolated word recognisers, continuing in implementation of the system on signal processor board, speech enhancement, continuous speech recognition, robust speech recognition in real environment, building of Linux computer cluster for training huge recognisers, etc.

The speech databases were needed in all fields of our activities in pass: especially in speech enhancement and speech recognition. The needs could be summarised in following points:

- evaluation and testing of *speech enhancement* algorithms - the first very small database collected in real environment (car) was created for this purpose in 1993,
- *training of speech recognisers* - both recognisers based on HMM (Hidden Markov Models) or ANN (Artificial Neural Networks) must be trained on large databases, covering as much as possible variability of speech; trained models (neurons) are then basic elements of working recognition system,
- evaluation of *speech parametrisation* techniques - evaluation and testing of any parametrisation mean performance of some recognition task; especially the testing of parametrisation performance requires data collected in real environment.

Noise background simulation can be possible simplification or alternative way to collection of large amount of data in many different environment, situations, etc. Especially, when the background noise can be assumed as additive, the artificially mixed data (i.e. speech and noise collected separately) can represent some behaviour of noisy data very well. On the other hand, it cannot be omitted that speech production is influenced by the environment and situation (i.e. noise, stress, it is so called *Lombard effect*) and the difference from real data may be reasonable.

As the consequence it can be said that **the design and the collection of large speech databases are indivisible parts of the research in the field of speech recognition (enhancement).**

2 Speech databases overview

What is *speech database (DB)*? It could be formulated that it is a set of data containing the acoustic speech signals with required utterances with more or less detailed annotation, i.e. the description of speech signal contents.

Otherwise, we may meet also *text (lexical) databases*, i.e. large corpora of written text, different electronic lexicons, etc. For Czech should be mentioned *Czech national corpus* - <http://ucnk.ff.cuni.cz>, the large corpus of written Czech.

Concerning the speech DB, at our department we were interested in design and collection of:

- DB for speech enhancement - *not two large databases* of representative speech signals with real noise background, devoted to statistically significant testing of speech enhancement performance,
- DB for speech recognition - *very large databases* covering speech variability for purposes of system training.

Also other types of speech databases might be mentioned : *DB for speaker verification* or *DB speech synthesis* with little different requirements according to other final application.

Further text will be devoted mainly to the DBs for speech recognition which design and collection were the major part of the activity in this field at our department.

2.1 Specification of DB for speech recognition

Omitting technical details of data storage (speech data formats, storage media standards, required file structure, etc.), further principal requirements are given by following points. These specifications (requirements) originate from desired maximal coverage of speech variability in the database.

- **environment coverage**

However the target point should be robust recogniser reliable independently to the environment, the coverage of all possible environments with all their variabilities is real problem. That is the reason why the databases are collected usually in exactly defined (one or just several) environments.

- *Telephone* is one of the most often collected environment, mainly due to currently high development in telecommunications, including different services controlled by voice. The signal is collected in this case by close-talk microphone in telephone, the speech has then usually high quality, the distortion in telecommunication channel starts being less and less significant, so the recogniser in such systems can work with very low error rate.
- *Car* is also one of the most popular environment. The usage of voiced controlled systems is motivated by minimising of driver disturbance during car ride.
- Depending on environment and further signal processing, the data can be collected as *one-channel* \times *multi-channel*. Multi-channel data brings other information, but the complexity and consequently also costs are increasing. That is the reason why one-channel systems (databases) are still very popular.

- **speaker coverage**

The coverage of speaker variability should yield to the speaker independent recognition system as the result of the training on such balanced database. To cover the speaker variability as best as possible, the requirements with respect to *age*, *gender*, and *accent (dialect)* are typically defined.

- **database corpus**

Finally, the database corpus (i.e. the contents of the utterances) should be defined. Two basic groups of utterances are usually recorded: *phonetically rich material* and *application specific items*.

- The first group of utterances - *phonetically rich sentences* and *phonetically rich words* - is devoted to the well training of all basic acoustic elements of speech (phoneme, diphones, triphones). These sentences/words are typically collected from different texts as novels, newspapers, Internet, etc. Then the big collected corpus is processed to select suitable sentences of optimal length with sufficient representation of each phoneme/diphone/triphone according to the specifications for given corpus.
- The second groups of utterances contains mainly *application specific commands*, i.e. short utterances for control of different systems or devices.
- Other frequently used application specific utterances are *numerals* in different forms - basic digits, isolated digits, connected digits, natural numbers, etc.
- Other possible items more and more depending on exact application are *date, time, money amount, names, cities, streets, spelled letters, etc.*

Some short application specific utterances can be then used for the training of HMM based on word models, however, it can be supposed just fro the small vocabulary recognition systems.

2.2 Collection of speech databases

The design and collection procedure of speech DB can be summarised in following 3 points.

1. Definition of specifications

In this first step, all requirements (mentioned in previous section) should be exactly defined with respect to the final purpose of the DB.

2. Realisation of recordings

Within realisation phase, the recording platform must be designed and constructed.

- Different requirements will be defined for telephone, car, office, etc.
- Speech can be recorded on audio medium (usually tape) or directly digitised and stored in PC (notebook). Recording on audio medium can be provided by standard devices which is the main advantage allowing quick realisation of any recording in different environments. On the other hand, it requires more complex further processing with much manual work. Direct recording into the PC can save a lot of necessary manual work but the construction of recording platform is more difficult.
- Evaluation of good recording software may be very helpful, the interest should be focused to the maximal automation of recording procedure, on-line checks of recorded data, creation of final DB structure during the recordings.

Next step is the recordings, which is usually one of the most difficult part of the data collection. It requires precise organisation yielding to the efficient data collection. The recruitment of the speakers according to the specification may be also quite complicated.

3. Annotation

The last step is the creation of recorded data description, i.e. *annotation*. Several kinds of annotation information can be observed:

- orthographic transcription - contents of the utterance in words,
- phonetic transcription - real pronunciation of each utterance (i.e. information of spoken phonetic elements), this phonetic transcription can be omitted when pronunciation lexicon of words used in orthographic transcription is available,
- transcription with time marks for words or phonemes, it start being less required, current algorithms does not need this information,
- special marks for mispronunciation and other non-speech events (speaker or environmental noise).

Within our activities we worked mainly with orthographic and phonetic transcription with additional marks for non-speech events. Some experiments were done with automated generation of time marks.

It was a brief summary of general problems connected to speech databases collection. Other details could be found in habilitation thesis [15] or in other reports closely connected to exact speech database, e.g. [12], [22], [13], [23], [5], [4], [7], [6], [11].

3 Measurement of signal quality

The second part of this presentation is focused to analysis of collected signal. This field is again quite wide. Concerning creation of speech databases, just the analyses having some relation to this task will be mentioned. All of them are motivated by reasons summarised in following points:

- *criteria of signal quality*

These criteria are many times the part of database annotation, moreover, their evaluation during (after) the recording is used frequently for detection of badly recorded items. One of the most often used criterion is Signal-to-Noise-Ratio (SNR).

- *detection of speech activity*

The speech activity is required in many algorithms and also in above mentioned measurement of speech SNR. Moreover, the information about speech activity/in-activity may be used for automated stop of particular item recording.

- *detection of signal delay in multi-channel systems*

For multi-channel recordings, the microphone are placed at different positions so different delay may appear between the channel. Creating some multi-channel processing system (VAD, enhancement, parametrisation, ...) requires usually the synchronisation of input data.

- *general analyses, i.e. general description of signal characteristics*

These analyses give useful information about signals in database for purposes of further target processing. The results of such analyses can be also used for the optimisation of above mentioned algorithms because these algorithms are always optimised to speech and noise characteristics.

- *integrated recogniser*

One of the most sophisticated analysis may be simple recognition task. It may be used during the recording for automated identification of speaker, recording session, etc. Usually, it is used on the basis of simple isolated digit recognition for speaker code, prompt sheet number, etc.

In next part of this contribution, the measurement of speech SNR and speech activity detection will be discussed in details.

3.1 Measurement of speech SNR

Signal-to-Noise-Ratio (SNR) is well known criterion used for the quantification of noise level in the signal. Its basic definition is given by following formula

$$SNR = 10 \log_{10} \frac{\sigma_s^2}{\sigma_n^2}. \quad (1)$$

Nevertheless, the speech signal is characterised by many absolutely irregular pauses during the utterance. These pure criterion may be strongly influenced by different length of pauses in speech, see fig 1. Let us assume the case of stationary noise background in speech. It is clear, that signal variance σ_s^2 will be strongly influenced by different length of pauses within speech, although the noise variance σ_n^2 will be for stationary noise same.

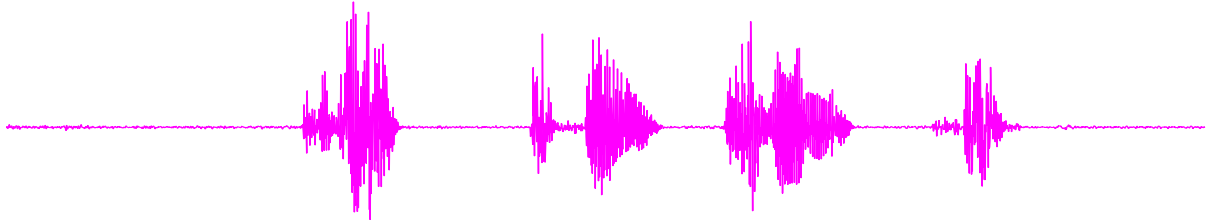


Figure 1: Clean speech signal

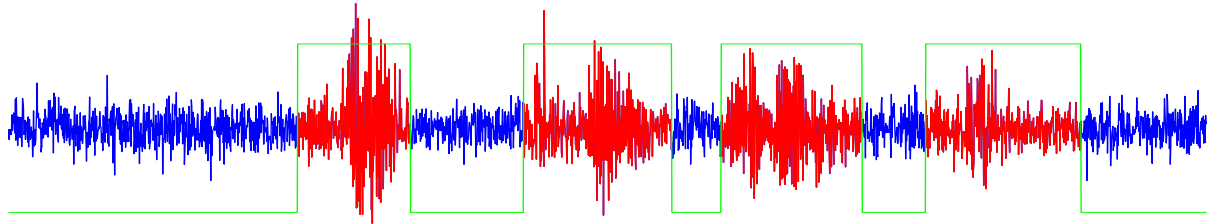


Figure 2: Speech signal with noise background and marked speech activity part

It seems to be clear that this disadvantage will be eliminated when the SNR is evaluated only over the signal part with speech activity, see fig. 2. The basic *Global SNR* is then for signals $s[n]$ and $n[n]$ with length l and VAD information $vad[n]$ ¹ defined as

$$SNR = 10 \log_{10} \frac{\sum_{n=0}^{l-1} s^2[n] vad[n]}{\sum_{n=0}^{l-1} n^2[n] vad[n]}. \quad (2)$$

The speech is typically processed applying the segmentation into quasi-stationary short-time segments (frames). Taking into account the used segmentation, *segmental SNR* is defined as average of *local SNR* evaluated over each processed segment. The formula for SSNR evaluation takes following form, where VAD_i is again voice activity information (on frame basis in this case)

$$SSNR = \frac{1}{K} \sum_{i=0}^{L-1} VAD_i \cdot 10 \log_{10} \frac{\sum_{n=0}^{M-1} s_i^2[n]}{\sum_{n=0}^{M-1} n_i^2[n]}. \quad (3)$$

¹ $vadn$ gives values 1 (speech) and 0 (non-speech) for each speech sample.

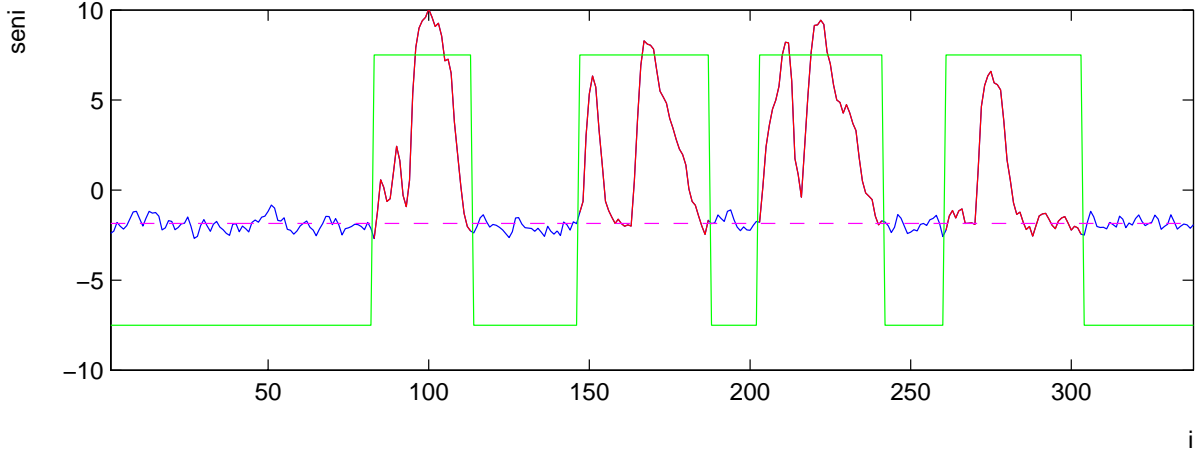


Figure 3: Speech signal log-energy with noise background and marked speech activity part

On fig. 3, the flow-graph of logarithmic signal energy is shown. For the case of stationary noise, estimated background energy is displayed by dashed line and the local SNR is in fact 10 times scaled difference of current frame logarithmic energy and mentioned estimation of background noise logarithmic energy.

For the estimation purposes, it may be useful to mention that standard segmental SNR can be rewritten as geometrical average of local linear signal-to-noise ratio followed by logarithmic conversion to decibels, i.e.²

$$SSNR = 10 \log_{10} \left(\prod_{j=0}^{K-1} VAD_j \cdot \frac{\sum_{n=0}^{M-1} s_j^2[n]}{\sum_{n=0}^{M-1} n_j^2[n]} \right)^{\frac{1}{K}}. \quad (4)$$

Finally mentioned criterion is based on modified method of averaging in formula (4), i.e. arithmetical averaging of linear ratios. We call it *arithmetical segmental SNR* and it is given as

$$SSNRA = 10 \log_{10} \left(\frac{1}{K} \sum_{i=0}^{L-1} VAD_i \cdot \frac{\sum_{n=0}^{M-1} s_i^2[n]}{\sum_{n=0}^{M-1} n_i^2[n]} \right). \quad (5)$$

All above described criteria give different values for same level of noise, however standard global SNR with VAD and arithmetical SSNR are very close. The most important evidence is that the criteria with VAD are independent on different pauses in utterance.

3.2 Methods for speech SNR estimation

Above described criteria can be easily evaluated when the speech and noise are known. Nevertheless, the evaluation may be difficult and may have limited precision when these criteria are evaluated from noisy signal only.

²Index j in formula (4) represents the signal frames with speech activity only and corresponding noise frames. Non-speech frames are excluded from this evaluation.

Let us assume that we have additive noise $n[n]$ in signal $s[n]$, i.e.

$$x[n] = s[n] + n[n]. \quad (6)$$

When the noise is uncorrelated with speech signal, additive relationship is fulfilled also in power-domain

$$\sigma_x^2 = \sigma_s^2 + \sigma_n^2. \quad (7)$$

Estimation with reference signal

The simplest estimation method originates from a priory knowledge about one signal in mixture. It is usually speech signal. When the reference to clean speech signal is available, the additive noise can be estimated as $n[n] = x[n] - s_{(ref)}[n]$. SNR can be then evaluated as

$$\widehat{SNR} = 10 \log_{10} \frac{\hat{\sigma}_s^2}{\hat{\sigma}_n^2} = 10 \log_{10} \frac{\hat{\sigma}_s^2}{\hat{\sigma}_{x-s}^2}. \quad (8)$$

In principle, using this estimation of noise signal, all criteria defined above can be evaluated using original formulae. This approach is typically advantageous when we working with simulated data processing by some noise suppression system. It must be mentioned that any distortion of speech (reference) signal is consequently represented as noise and increases measured level of noise. Similarly, also any delay has same consequences.

On the other hand, this approach is not suitable for estimation of SNR for data recorded in real noisy background. Due to the reasons mentioned above, it is not possible to use often recorded “close-talk” channel as a reference.

Estimation without reference

When any reference is not available SNR can be evaluated using following formula

$$\widehat{SNR} = 10 \log_{10} \frac{\sigma_x^2 - \hat{\sigma}_n^2}{\hat{\sigma}_n^2}. \quad (9)$$

It is based on assumption (7), so it can be used just for uncorrelated additive noise. On the other hand, when the noise background is not extreme, this method can be successfully used and in fact it is then reduced to the estimation of noise power (variance).

Several methods for **noise variance estimation** $\hat{\sigma}_n^2$ can be found. We are working mainly with one of the most frequently used, i.e. *averaging of σ_x^2 in speech pauses* [3], [8], [14]. Block or recursive averaging can be used with slightly different results. Due to implementation issues, recursive exponential averaging is one of the most frequently used approached.

$$\hat{\sigma}_{n,i}^2 = \begin{cases} p \cdot \hat{\sigma}_{n,i-1}^2 + (1-p) \cdot \sigma_{x,i}^2, & VAD_i = 0 \\ \hat{\sigma}_{n,i-1}^2, & VAD_i = 1 \end{cases}, \quad (10)$$

where

$$\sigma_{x,i}^2 = \frac{1}{M} \sum_{n=0}^{M-1} x_i^2[n]. \quad (11)$$

This algorithm requires VAD but it is required generally for the speech SNR. Next algorithm for noise variance estimation works without VAD. It is based on *tracking of σ_x^2 minimum*, see [10], [19].

$$\hat{\sigma}_{n,i}^2 = c \cdot \min \left(\sigma_{x,i}^2, \sigma_{x,i-1}^2, \sigma_{x,i-2}^2, \dots, \sigma_{x,i-L_o+1}^2 \right). \quad (12)$$

Properties of algorithms for speech SNR estimation

- It is assumed that speech and noise are additive and uncorrelated. Under these conditions speech SNR can be estimated, however, approximately min. 1 dB standard deviation should be taken into account.
- The estimation of global SNR is often corrupted by failure when $\sigma_x^2 < \hat{\sigma}_n^2$. Under this condition the linear ratio is negative and logarithm is not defined. In principle it could mean minimal signal level, i.e. very negative SNR. It is usually solved by thresholding because the value $-\infty$ is not suitable for further processing.
- Evaluating of $SSNR$ gives less frequent failures due to above mentioned logarithm evaluation. These failures appear at frame level and globally are compensated by further averaging.
- The failure of logarithm evaluation is minimised in $SSNRA$ estimation due to arithmetical averaging of local linear ratios.
- $SSNR$ and $SSNRA$ give different values, $SSNRA$ is more close to SNR .
- The $SSNRA$ seems to be the best criterion from the estimation points of view. Arithmetical average is less influenced by small local linear ratios (linear SNR_i) which are more often influenced by estimation errors, see fig. 4.

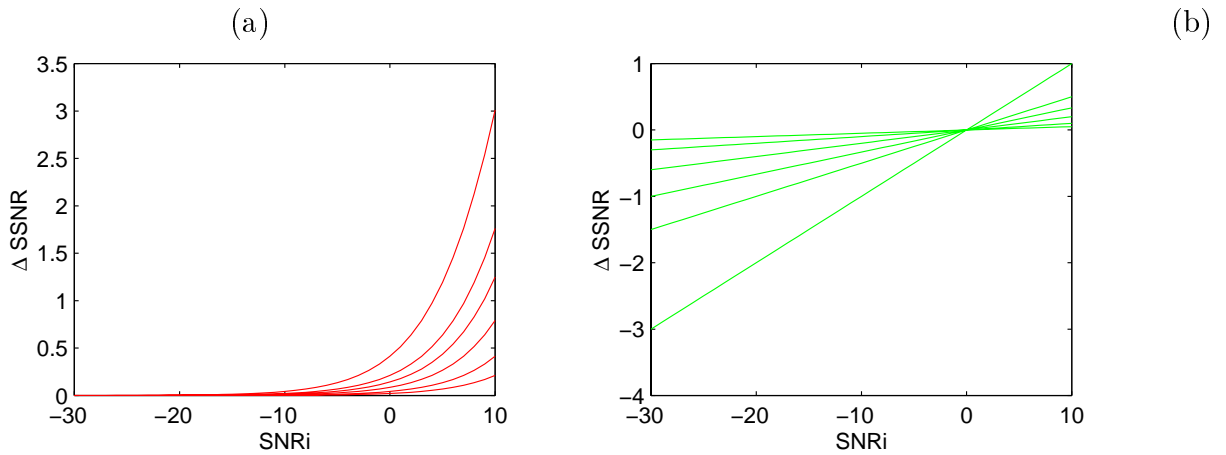


Figure 4: Influence of short-time SNR to SSNR and SSNRA (basic SSNR is 0 dB).

3.3 Voice activity detection

One of the most frequent requirement in many speech processing algorithms is the availability of voice activity information (speech detection). As mentioned above also the evaluation of SNR requires this information. Many types of VAD algorithms can be found: energy based, spectral, cepstral, coherence, etc. [2], [1], [9], [20], [16], [3], [17], [18], [21].

Further part is devoted to an illustrative example of cepstral voice activity detector, see the block scheme on fig. 5. It is based on evaluation of cepstral distance CD_i between current frame and averaged frame representing noise background, solid curve on fig. 6. Cepstral distance is correlated to spectral differences between two signals. This distance is then compared with adaptive threshold $CD_{p,i}$ and for values above the threshold the speech is detected. When pause is detected adaptive threshold and background cepstrum (\bar{c}) are updated.

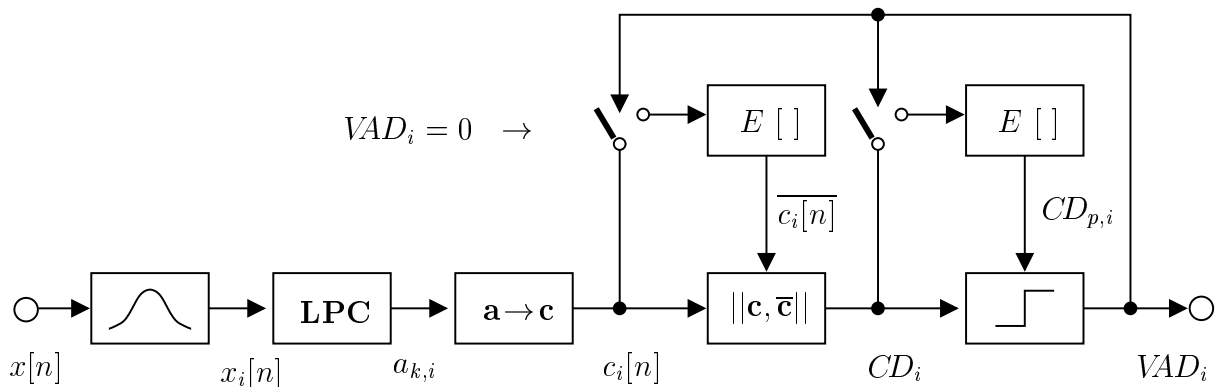


Figure 5: Integral cepstral voice activity detector

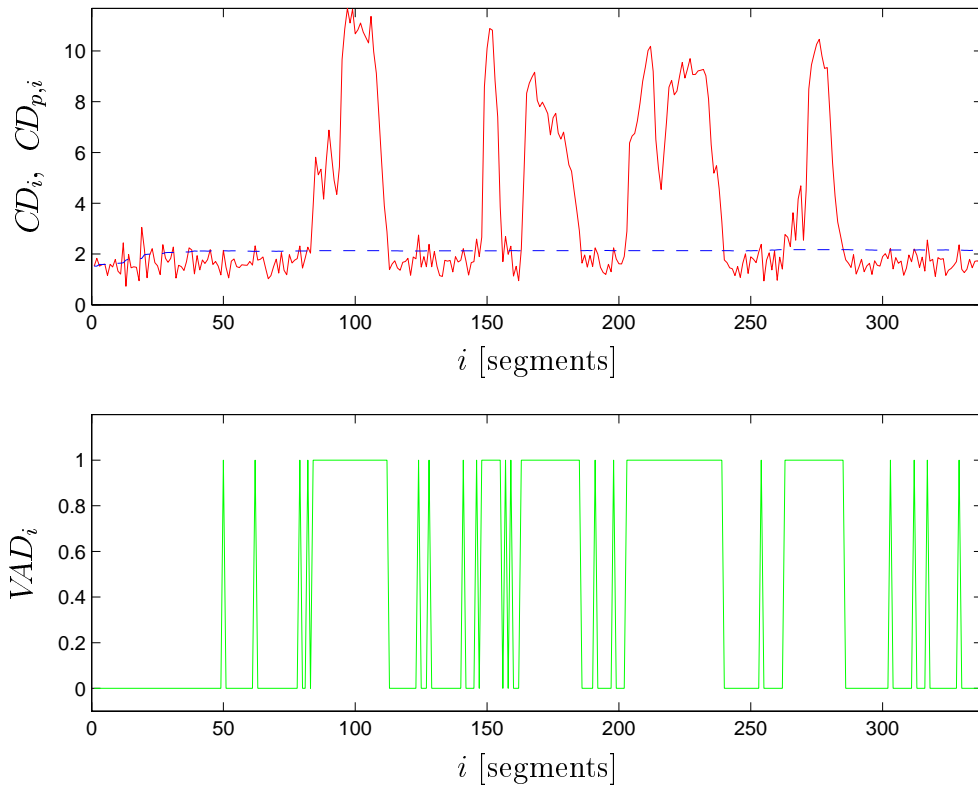


Figure 6: Illustrative output of cepstral voice activity detector

The problems of voice activity detection are, of course, more complex but for our purposes this informative description may be closed by following points:

- Instead of cepstral distance CD_i other characteristics may be used, e.g. energy, zero-crossing, coherence, etc. It means different complexity of the algorithm commonly with different reliability. Typically, energy detector is very simple and it may be well used in relative less intensive background, however it fails for high level background.
- On-line algorithms must work with adaptive threshold, off-line algorithms can work also with fixed threshold easily set from signal dynamics.
- Principle of the VAD is not too complex, the problems appear in situations when characteristics of speech and the background are simple, see the example from running car of fig. 7 and 8.
- During the collection of speech databases, VAD may be used as end-point detection. It was an example of SpeechDat collection when simple energy VAD was integrated into recording platform.

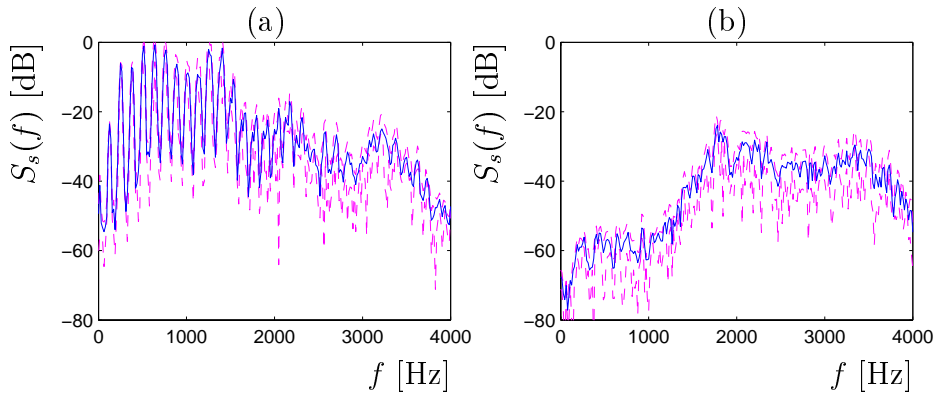


Figure 7: Speech spectrum: (a) voiced, (b) unvoiced

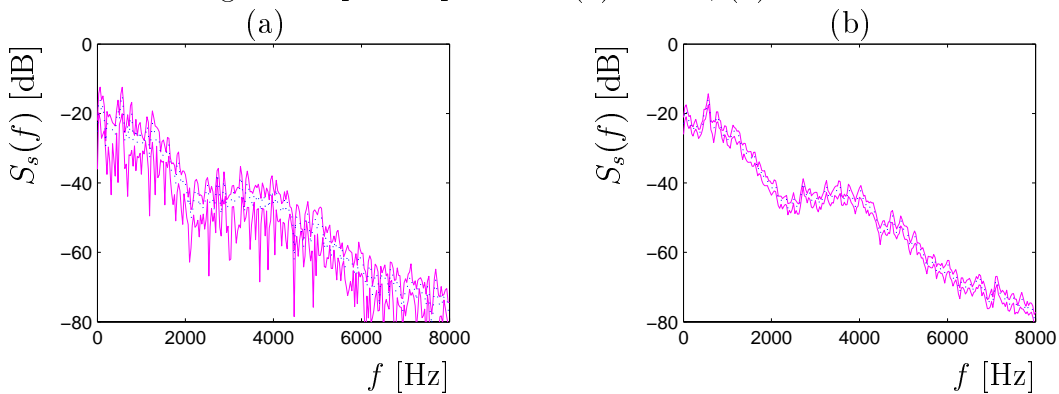




Figure 8: Car noise spectrum: (a) estimated from 64 ms, (b) estimated from 0.5 s

4 Conclusions

Looking for the most important conclusions of presented activities, collected speech databases should be mentioned in the first step. All collected databases are summarised in following points:

1. CAR2ECS - 54 | 62 speakers in staying | running car, 2 min. of speech per speaker,
2. ČÍSLOVKY - FIXED2CS - 1227 speaker over fixed telephone network, 5 min. of speech per speaker (isolated and connected digits),
3.  SpeechDat(E) - FIXED3CS - 1052 speaker over fixed telephone network collected within INCO-Copernicus project, 15 min. of speech per speaker (phonetically rich material - big corpus & application specific items),
4. CZKCC4 - 300 in car (staying | running), 20 min. per speakers (phonetically rich material - small corpus & application specific items),
5.  **Running European IST project** - SpeeCon: „Voice Driven Interfaces in Consumer Devices“ - 600 speakers in different environment (office, public, entertainment, car, children), 4 channels, 30 min. of speech (read and spontaneous).

Secondly, more general conclusions can be formulated in following points:

- Methodology for design and creation of speech databases were summarised, including design of corpus, collection methodology, creation of annotations, etc.).
- Methodology for analysis of collected data was defined. Several original criteria were tested and successfully applied during the collections of databases.
- Recording of the data in different environment was solve (mainly telephone, car) including the possibility of its automation.
- Czech SAMPA were standardised (April 2003) as the result of co-operative work of several Czech labs interested in this field, (this work was dealt by ing. Hanzl from our lab). This activity was also initialised during the collection of Czech SpeechDat database.

Finally, **student activities in this field** are typically provided within

- semester projects in the field of digital signal processing (subjects **31CZS**, **31ASI**); typical themes are analyses of signals in databases, analyses of particular algorithms (estimations of SNR, VAD), etc.,
- semester projects in hardware focused subjects (**31LBR**); it was realised the support for the recordings in the car,
- works during the collection and annotation; although it is just manual work, it can be done by students in the beginning of the study and it may be their first contact in further continuing activity in this field of research.

Reference

- [1] HAIGH, J. A. – MASON, J. S.: A voice activity detector based on cepstral analysis., In *Eurospeech'93 - Proceedings of the 3rd European Conference on Speech, Communication, and Technology*, Berlin, Sept. 1993, pp. 1103–1106.
- [2] HUANG, X. – ACERO, A. – HON, H.-W.: *Spoken Language Processing*, Prentice Hall, 2001.
- [3] JUNQUA, J.-C. – HATON, J.-P.: *Robustness in Automatic Speech Recognition.*, Kluwer Academic Publishers, 1996.
- [4] KIESSLING, A. – DIEHL, F. – FISCHER, V. – MARASEK, K.: Specification of databases - specification of corpus and vocabulary, Technical Report, SPEECON, Jul 2001. Deliverable D213, workpackage WP2.
- [5] KIESSLING, A. – DIEHL, F. – FISCHER, V. – MARASEK, K.: Specification of databases - specification of recording scenarios, Technical Report, SPEECON, Oct 2001. Deliverable D212, workpackage WP2.
- [6] KIESSLING, A. – DIEHL, F. – FISCHER, V. – MARASEK, K.: Specification of databases - specification of speakers, Technical Report, SPEECON, May 2001. Deliverable D215, workpackage WP2.
- [7] KIESSLING, A. – DIEHL, F. – FISCHER, V. – MARASEK, K.: Specification of databases - specification of annotation, Technical Report, SPEECON, Feb 2002. Deliverable D214, workpackage WP2.
- [8] KORTHAUER, A.: Robust estimation of the SNR of noisy speech signals for the quality evaluation of speech databases, In *Proc. of Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, May 1999.
- [9] LE FLOC'H, A. – SALAMI, R. – MOUY, B. – ADOUL, J.-P.: Evaluation of linear and non-linear subtraction methods for enhancing noisy speech., In *Speech Processing in Adverse Conditions*, Cannes-Mandelieu (France), Nov. 1992, pp. 131–134.
- [10] MARTIN, R.: An efficient algorithm to estimate the instantaneous SNR of speech signals., In *Eurospeech'93 - Proceedings of the 3rd European Conference on Speech, Communication, and Technology*, Berlin, Sep 1993, pp. 1093–1096.
- [11] POLLÁK, P.: Language specific peculiarities – Czech., Technical Report, SPEECON, May 2003. Deliverable D216CZE, workpackage WP2.
- [12] POLLÁK, P. – ČERNOCKÝ, J.: Specification of speech database interchange format., Technical Report, SpeechDat(E), Aug 1999. Deliverable ED1.3, workpackage WP1.
- [13] POLLÁK, P. – ČERNOCKÝ, J.: Final recruitment methodology and documentation of speakers typology for the final Czech database., Technical Report, SpeechDat(E), Dec 2000. Deliverable ED2.12.2.b, workpackage WP2.
- [14] POLLÁK, P.: Metody odhadu odstupu signálu od šumu v řečovém signálu, *Akustické listy*, vol. 7, no. 3, pp. 14–21, 2001. In Czech language.

- [15] POLLÁK, P.: *Desing and Creation of Speech Databases for Recognition and Enhancement*, Habilitation thesis, Czech Technical University in Prague, Prague, 2002. In Czech language.
- [16] POLLÁK, P. – SOVKA, P. – UHLÍŘ, J.: Cepstral speech/pause detectors, In *Proceedings of IEEE Workshop on Nonlinear Signal and Image Processing*, Neos Marmaras, Greece, June 1995.
- [17] PSUTKA, J.: *Komunikace s počítačem mluvenou řečí.*, Academia Praha, 1995.
- [18] RABINER, L. R. – JUANG, B.-H.: *Fundamentals of Speech Recognition.*, Murray Hill, New Jersey, USA, 1993.
- [19] RIS, C. – DUPONT, S.: Assessing local noise level estimation methods: Application to noise robust asr, *Speech Communication*, pp. 141–158, 2001.
- [20] SOVKA, P. – POLLÁK, P.: The study of speech/pause detectors for speech enhancements methods, In *EUROSPEECH'95 - Proceedings of the 4th European Conference on Speech Communication and Technology*, Madrid, Spain, September 1995, pp. 1575–1578.
- [21] VASEGHI, S. V.: *Advanced Digital Signal Processing and Noise Reduction*, 2-nd edition, John Wiley & Sons, Ltd., 2000.
- [22] ČERNOCKÝ, J. – POLLÁK, P. – HANŽL, V.: Definition of corpus, scripts, standards, and environmental and speaker specific coverage applied for speech databases, Technical Report, SpeechDat(E), Jul 1999. Deliverable ED1.12.2, workpackage WP1.
- [23] ČERNOCKÝ, J. – POLLÁK, P. – HANŽL, V.: Czech recordings and annotations on CD's - Documentation on the Czech database and database access., Technical Report, SpeechDat(E), Nov 2000. Deliverable ED2.3.2, workpackage WP2.

Ing. Petr Pollák, CSc.

Date and place of birth: 23 Mar 1966, Ústí nad Orlicí, Czechoslovakia

Nationality: Czech

Marital status: married, 3 children

Address: Czech Technical University, Faculty of Electrical Engineering - K331,
Technická 2, 166 27 Praha 6 - Dejvice, CZECH REPUBLIC

E-mail: pollak@feld.cvut.cz

Education:

1989 (Ing.) - graduated at Czech Technical University, Faculty of Electrical Engineering

1994 (CSc.) - defence of dissertation at CTU FEE

Languages: English, French, Russian, Czech

Brief chronology of employment:

1989-1992	postgraduate student	Department of Circuit Theory, CTU FEE
1992-1993	research fellow	-”-
1993-	assistant professor	-”-

Stages and study leaves:

Sep-Dec 1996	sabbatical stay	Telecom-Paris - ENST
--------------	-----------------	----------------------

Teaching experience:

Lectures: since 1994 - Basic circuit theory I & II

Seminars: since 1991 - Basic circuit theory I & II

since 1995 - Electronic circuits

since 1996 - Algorithms for signal processing

since 1998 - Digital signal processing

since 2000 - Electronic Circuits Laboratories

Supervising of graduate students:

15 successfully graduated students since 1993

Supervising of PhD students:

1 successful PhD student - 2003 - supervisor specialist

currently 4 PhD students - supervisor

currently 1 PhD student - supervisor specialist

Past grant support:

<i>institution</i>	<i>year</i>	<i>Project</i>
EC	1998-2000	INCO-Copernicus 977017 “Eastern European Speech Databases for Creation of voice Driven Teleservices - SpeechDat(E)”
EC	2003	IST-1999-10003 “SPEECON - Speech Driven Interfaces for Consumer Applications”
MŠMT	1999-2000	OK410/1999, OK410/2000

Research interest:

General research interest is in *digital signal processing* and its applications especially speech processing, processing of biological signals, etc. The most important fields of research interest are: *creation of very large speech databases, robust speech recognition, speech enhancement, speech identification in a noisy environment*. Programming and computer experiences: *Linux (Unix), MS Windows, C, MATLAB, Perl, LaTeX*.