Long Recording Segmentation Based on Simple Power Voice Activity Detection with Adaptive Threshold and Post-Processing

Petr Pollák, Josef Rajnoha

Czech Technical University in Prague, Faculty of Electrical Engineering
ČVUT FEL K13131, Technická 2, 166 27 Praha, Czech Republic
pollak@fel.cvut.cz, rajnoj1@fel.cvut.cz

Abstract

This paper describes the method of long recording segmentation based on Voice Activity Detection (VAD). Power based detection using an adaptive threshold derived from power dynamics is the core of presented approach. Simple post-processing based on long time sub-segmentation is used for smoothing of primary VAD output to obtain target start-point and end-point detection of particular utterances within long recordings. Because the algorithm is based on simple power VAD it can be much more easily implemented in comparison to approaches based on speech recognition. Though presented approach is so simple it gives quite robust and satisfactory results for pure segmentation task. The tests with two different data types proved satisfactory results same as practical usage during the creation of new speech corpora.

1. Introduction

The segmentation of long recordings is a procedure which is frequently needed by speech technology applications and it is used also for different purposes within general activities in this research field. Typically, it can be required for the purposes of the segmentation of long recordings into particular utterances during the speech database creation. It is often used procedure to capture long recording, especially, when spontaneous speech data are collected. As such speech databases are used mainly for training purposes, stored signals should be split into shorter utterances and precisely annotated. Also for the annotations itself, it is very reasonable to have such pre-segmentation of long recordings available to listen and transcribe selected part of the record only. This application was one important motivation for the work described in this paper.

Currently, the systems for large vocabulary continuous speech recognition (LVCSR) are often used for full automated transcription of the contents of different audio records and they start giving very satisfactory results. The segmentation into subparts can be then by-product of this procedure. On the other hand, LVCSR represents usually very complex system with high computational costs, with hard requirements to available proper large language model, and with necessity of very efficient implementation of decoding procedure. The creation of such efficient recognition engine means usually several years of very special hard work. Moreover, the accuracy of this transcription is topic dependent and it is much more difficult for non-fluent or spontaneous speech. Consequently, it is still suitable to look for more simple approach which is not based on automated speech recognition.

The approach based on voice activity detection (VAD) can be such simpler choice. These techniques are not topic or language dependent because pure general acoustic signal analysis is the core of these algorithms [1]. We can meet different approaches of VAD from the simplest power (energy) based ones [2], through spectral or cepstral ones [1], [3], up to sophisticated approaches combining several characteristic carrying different information [4], [5], [6]. Though the complexity and requirements of particular algorithms can differ, their implementations are much more easy, the computational costs are also very small, so these systems can be easily implemented in real time without the necessity of difficult special optimizations.

In this paper, we would like to present one simple approach of utterance segmentation which is based on power VAD. The basic decision is done using continuously updated adaptive threshold tracking the condition of analyzed signal and final decision is obtained on the basis of simple post-processing over slightly higher long-time period. Optimized setting of this approach can give very good results and it can be used in several applications.

2. Utterance segmentation algorithm

Our segmentation algorithm is based on power VAD. The basic motivation is, of course, to have an approach without high computational costs and with simple implementation. We can also assume that we have relatively high quality speech signal because it is captured by head-set microphone. Even sometimes we must work with slightly disturbed signals, power approach seems to be sufficiently robust for given conditions.

The segmentation algorithm has two layer structure. Within the first layer, basic short-time speech activity detection is performed, in the second layer, this primary detection is smoothed by post-processing within long-time frames.

2.1. Primary short-time VAD detection

Primary detection is based on short-time power (energy) analysis computed on the frame basis. Instantaneous signal power on the sample level represents too detailed information which is not reasonable in this case. Moreover, frame approach offers further extension in using more complex characteristics, e.g. cepstral analysis.

Standard setup of short-time analysis is used for power computation, i.e. the frame of fixed length is moving over the signal with 50% overlapping. However, the overlapping is not necessary for such power computation, we are using it for the compatibility with spectral based algorithms where it must be used due to frame weighting. Frame length must be sufficiently long not to track power changes within one fundamental period of speech (approx 10 ms), but it should not exceed typical length of quasi-stationary part of speech signal (approx 30 ms). In our

approach we use 20 ms frame length.
Frame power is then easily computed as

$$P(t) = \frac{1}{N} \sum_{n=0}^{N-1} s[t \cdot M + n] \,, \tag{1}$$

where $P(t)$ represents power in t-th frame, $s[n]$ signal sample, N frame length, and M frame step.

Short-time power $P(t)$ is then compared with the threshold $P_{thr}(t)$ to obtain short-time speech activity detection in the sense

$$vad(t) = \begin{cases} 1, & \text{if } P(t) \geq P_{thr}(t), \\ \\ 0, & \text{if } P(t) < P_{thr}(t). \end{cases} \tag{2}$$

The definition of this threshold is the basic problem in such way of classification. Fixed threshold is not suitable and the adaptive one can be defined different way. We have proposed the definition based on adaptively updated power dynamics, i.e.

$$P_{thr}(t) = P_{min}(t) + \frac{p}{100} \cdot (P_{max}(t) - P_{min}(t)) \,, \tag{3}$$

where p represents the percentage of dynamics range which is added to current minimal power for threshold definition. Working with long records, minimal and maximal powers are updated. It can be done on the basis of exponential averaging according to following formulae

$$P_{max}(t) = \begin{cases} q_{max1}\, P_{max}(t-1) + (1-q_{max1})\, P(t), \\ \qquad\qquad \text{if } P(t) \geq P_{max}(t-1), \\ \\ q_{max2}\, P_{max}(t-1) + (1-q_{max2})\, P(t), \\ \qquad\qquad \text{if } P(t) < P_{max}(t-1), \end{cases} \tag{4}$$

$$P_{min}(t) = \begin{cases} q_{min1}\, P_{min}(t-1) + (1-q_{min1})\, P(t), \\ \qquad\qquad \text{if } P(t) \leq P_{min}(t-1), \\ \\ q_{min2}\, P_{min}(t-1) + (1-q_{min2})\, P(t), \\ \qquad\qquad \text{if } P(t) > P_{min}(t-1). \end{cases} \tag{5}$$

Constants q_{max1}, q_{max2}, q_{min1}, and q_{min2} control the speed of given updates. Generally, the speed of the increase of maximal power and decrease speed of minimal power should be higher than for the forgetting of these values. Especially, the increase of power minimum must be extremely slow. Related values of time constants used in our setup were following

$$\begin{array}{lcl} \tau_{max1} & \approx & 0.2 \text{ s} \\ \tau_{max2} & \approx & 2 \text{ s} \\ \tau_{min1} & \approx & 0.1 \text{ s} \\ \tau_{min2} & > & 1 \text{ min} \end{array}$$

Finally, when longer speech pause is present in the record, short-time dynamics is continuously decreasing as far as it can be very small. Consequently, the speech can be badly detected in such frames. To avoid this phenomena, minimal dynamics P_{dmin} of speech segment is defined. When the dynamics falls bellow this threshold, the noise is detected independently of anything else, i.e.

$$vad(t) = 0, \quad \text{if } P_{max}(t) - P_{min}(t) < P_{dmin}. \tag{6}$$

This threshold definition can efficiently track middle long-time speech dynamics and it can be used for the comparison with short-time power. Illustrative example is in the fig. 1.

2.2. VAD post-processing

Above described basic VAD detection contains typically a lot of short-time errors, very short parts of speech or pauses up to single frame ones can appear in this output. For the purposes of utterance segmentation we need to realize so called start-point and end-point detection. This can be performed by following post-processing which takes into account longer time frames for the decision about speech presence. The algorithm can be described by following parts.

2.2.1. Long-time frame processing

Firstly, the output is split into relatively long-time frames (i.e. buffers of short-time outputs) without overlapping. We are using buffer of the length $T = 0.5$ s, which is related to the typical length from the perception point of view. The detection should not vary within such time so the final outputs are unified for all short-time frames inside of this long-time buffer.

$$\mathit{buff}_i(t) = vad(i \cdot K + t) \quad \text{for } t = 0, 1, .., K-1, \tag{7}$$

$$\mathit{vadbuff}_i = \begin{cases} 1, & \text{if } \dfrac{1}{K} \displaystyle\sum_{t=0}^{K-1} \mathit{buff}_i(t) \geq 0.2 \,, \\ \\ 0, & \text{if } \dfrac{1}{K} \displaystyle\sum_{t=0}^{K-1} \mathit{buff}_i(t) < 0.2 \,. \end{cases} \tag{8}$$

The constant $K = T/M$ represents the number of short-time frames within given long-time buffer.

2.2.2. Start-point detection

As the processing of records with spontaneous and conversational speech is supposed, such records can contain e.g. one word speech parts, i.e. the minimal duration of one utterance can be very short. It means, that once the voice activity is detected within one long-time buffer, the start-point is given.

$$vad_i = \mathit{vadbuff}_i \qquad \text{if } \mathit{vadbuff}_i = 1 \,, \tag{9}$$

Generally, the minimal duration of the utterance can be defined, then the start-point of voice activity is placed at the beginning of the first buffer when given number of speech buffers is reached.

2.2.3. End-point detection

End-point detection is limited by following constraint. It is supposed that reasonable pause is much longer than short inter-word pause. On the basis of this idea, similarly as it was mentioned above we can define minimal pause length to stop the utterance. When speech activity has started, next long-time buffer is always supposed to be the speech one, until given number P of long-time buffers (related to minimal pause length) is classified as non-speech, i.e.

$$vad_i = 1, \quad \text{if } vad_{i-1} = 1 \,, \tag{10}$$

$$vad_{i-j} = 0, \quad \text{for } j = 0, .., P-1, \quad \text{if } \mathit{vadbuff}_{i-j} = 0 \,. \tag{11}$$

The value of constant P is given by supposed minimal pause which is required to be detected. Typically, it should be the value related to the minimal speech pause between 0.5 - 2 s.

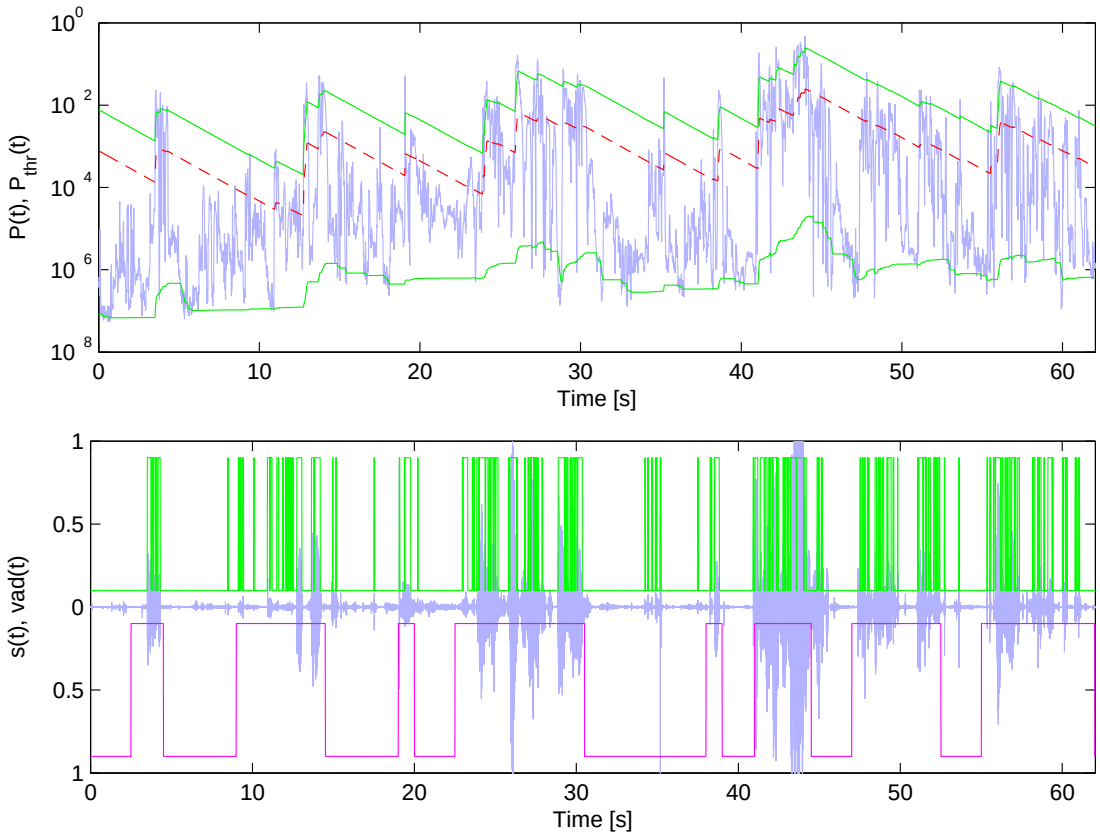

Figure 1: Illustrative example of detection process with threshold computation details

2.2.4. Final smoothed detection

Obtained detection within long-time buffers represents in the principle the final smoothed detection. Formally it can be written as

$$VAD(i \cdot K + t) = vad_i, \qquad \text{for } i = 0, 1, .., L-1\,, \quad \text{for } t = 0, 1, .., K-1. \tag{12}$$

Of course, above described algorithm gives the output where the boundaries between speech parts and pauses are placed discretely with the time interval T (0.5 s in our case). Though such step is quite rough it is acceptable for the purposes of start-point and end-point detection in following two principle applications: utterance selection followed by automated speech recognition or utterance extraction for storing within speech database. When speech is not truncated, slight extension at the beginning or at the end of particular utterance can be acceptable for both mentioned applications. Moreover, sometimes this utterance extension can be required, e.g. for the background adaptations during feature extraction, etc.

3. MATLAB implementation for long recordings

Concerning the tuning of current system setting and concerning further development, MATLAB environment was chosen for the implementation. It offers full support for computation of many different characteristics and immediate perfect graphical output to observe the influence of particular parameters.

On the other hand, MATLAB is usually used for off-line processing when whole and more frequently shorter signals are processed. When long recordings need to be processed (e.g. 1.5 hour of speech in our case), such approach starts reaching memory limits of the system and special implementation for this purpose must be used. It can be very similar to standard off-line processing with several minor changes only. Some of them are summarized in following items.

- Firstly, signal must be continuously loaded on frame basis and similarly, output should be written immediately into defined external output. MATLAB offers standard low-level file I/O functions for these purposes using equivalent conventions well known from C/C++. Finally, this solution is quite easy and it has the structure more close to possible real-time implementation within some target application.
- Secondly, when some short-time parameters are kept in internal memory for further usage, the vector or matrix size cannot be changed within each step of the processing. The re-allocation of the memory for rather complex structure of MATLAB variable is very slow and consequently it can mean critical increase of computational time.
- Similarly, it is necessary to maximize the usage of special vector or matrix operation respectively instead of standard for-cycle because its performance in MATLAB interpreter is again very slow.

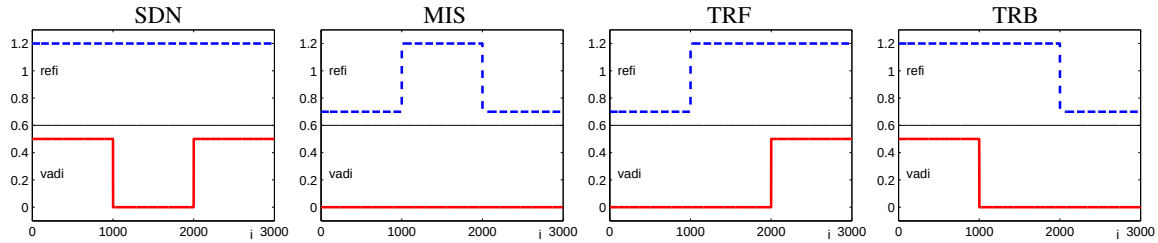

Figure 2: Speech detection error categories used for the classification.

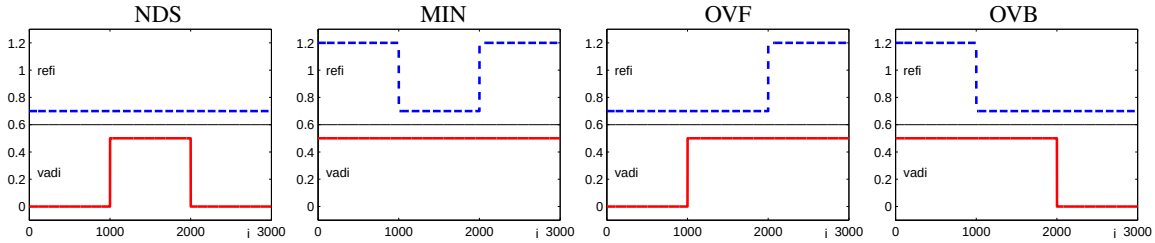

Figure 3: Non-speech detection error categories used for the classification.

4. Applications and experimental results

Proposed algorithm with very simple implementation was designed for the segmentation of long recordings and for the detection of speech activity at the input of voice controlled systems. Algorithm gives reliable results in the real environment.

4.1. Segmentation of long recordings

The basic motivation for the usage of this algorithm described briefly in the introduction is to use it for the segmentation of long recordings. It represents frequent application needed during the collection of different speech corpora. It is standard practice to capture long part of speech data at once and to store such session later in separate files corresponding to particular utterances. If long speech data are saved at once, at least time boundaries of particular utterances are marked for further extraction. We have collected several such corpora where boundaries of particular utterances were marked manually [7], [8], [9].

To overcome this hard manual work above mentioned algorithm was used for automated pre-segmentation. Of course, the accuracy of described approach is not 100%, so boundaries must be still manually adjusted. On the other hand, it can be done already more easily within further step representing annotation of utterance content. For this procedure, LDC tool *Transcriber* is used [10], so our MATLAB implementation generates already required empty XML formatted *.trs file with preset utterance boundaries.

4.2. Detection of voice controlled system input

The implementation in testing voice-controlled page in demo section at WEB page of our lab [11] is our second application of given algorithm. Start-point and end-point detection based on this approach is used here for the selection of control command from continuously listened input. Although the performance can depend strongly on the conditions of captured speech signal, we have observed satisfactory performance in our tests.

4.3. Results on segmentation of long recordings

Proposed algorithm is now used mainly during the collection of two spontaneous speech databases, containing technical lectures [7] on DSP topic and free conversation between three speakers [8]. We have the first results of the precision of proposed automated segmentation.

4.3.1. Classification criteria

For objective classification, the following criteria are used.

- NDS& SDN : Non-speech Detected as Speech
  Speech Detected as Non-speech
- MIS & MIN : MIssed Speech, MIssed Non-speech
- OVF & OVB : OVerlap at Front, OVerlap at Back
- TRF & TRB : TRuncation at Front, TRuncation at Back

They are derived from the criteria published in [13] and they are presented illustratively in fig. 2 and 3. Then also following summarizing criteria are used.

- ERror in Speech

$$\mathrm{ERS} = \mathrm{SDN} + \mathrm{MIS} + \mathrm{TRF} + \mathrm{TRB} \qquad (13)$$

- ERror in Non-speech

$$\mathrm{ERN} = \mathrm{NDS} + \mathrm{MIN} + \mathrm{OVF} + \mathrm{OVB} \qquad (14)$$

- global ERRor in detection

$$\mathrm{ERR} = \mathrm{ERS} + \mathrm{ERN} \qquad (15)$$

All these criteria are computed usually in percents related to the total number of short-time frames, i.e. for global ERR criterion it means mean value from absolute value of difference between reference and classified information about speech activity

$$\mathrm{ERR} = E\Big[|VAD_{ref}(t) - VAD(t)|\Big] \cdot 100\,. \qquad (16)$$

Particular errors are computed as number of badly classified frames in given category divided by total number of frames.

Secondly, the average values of time duration of particular errors are evaluated. These values are marked as $asdn$, $amin$, $atrf$, $atrn$, etc. The tool *vadcrit* implementing these criteria is available at WEB page [11].

	ERR	ERS	ERP
sig 1	15.12	2.90	12.22
sig 2	11.83	1.81	10.02
sig 3	15.03	2.37	12.67
sig 4	18.97	2.83	16.14
sig 5	16.49	3.49	13.00
sig 6	11.85	3.27	8.58
sig 7	11.47	2.00	9.48
Mean	14.40	2.67	11.73
Std	*2.82*	*0.63*	*2.58*

Table 1: Error rates for conversational speech [%]

	SDN	MIS	TRF	TRB	NDS	MIP	OVF	OVB
sig 1	1.19	0.00	0.00	1.72	2.44	3.20	5.16	1.37
sig 2	0.60	0.00	0.00	1.21	0.67	2.80	4.80	1.75
sig 3	0.45	0.00	0.00	1.91	3.72	3.46	4.91	0.57
sig 4	0.61	0.00	0.00	2.20	5.15	4.09	5.86	1.04
sig 5	0.98	0.00	0.00	2.51	2.53	4.00	5.37	1.08
sig 6	1.03	0.00	0.00	2.24	1.42	2.61	4.16	0.40
sig 7	0.67	0.00	0.00	1.32	2.75	2.78	3.52	0.32
Mean	0.79	0.00	0.00	1.87	2.67	3.28	4.82	0.93
Std	*0.27*	*0.00*	*0.00*	*0.48*	*1.46*	*0.59*	*0.77*	*0.52*

Table 2: Relative particular errors for conversational speech [%]

	ERR	ERS	ERP
sig 1	5.65	1.20	4.45
sig 2	8.19	2.87	5.32
sig 3	5.68	1.60	4.08
sig 4	6.68	2.42	4.26
Mean	6.55	2.02	4.53
Std	*1.19*	*0.75*	*0.54*

Table 3: Error rates for DSP lectures [%]

	SDN	MIS	TRF	TRB	NDS	MIN	OVF	OVB
sig 1	0.83	0.00	0.01	0.37	0.00	1.44	1.43	1.58
sig 2	2.03	0.00	0.01	0.82	0.00	0.53	2.24	2.55
sig 3	1.33	0.00	0.01	0.27	0.00	0.73	1.57	1.77
sig 4	1.71	0.00	0.00	0.70	0.00	0.66	1.78	1.82
Mean	1.48	0.00	0.01	0.54	0.00	0.84	1.76	1.93
Std	*0.52*	*0.00*	*0.01*	*0.26*	*0.00*	*0.41*	*0.35*	*0.43*

Table 4: Relative particular errors for DSP lectures [%]

Database	asdn	amis	atrf	atrb	ands	amin	aovf	aovb
Conversational speech	450	1	9	419	2233	1440	879	346
DSP lectures	182	0	10	88	0	565	155	155

Table 5: Average durations of particular errors for both databases [ms]

4.3.2. Results on used databases

Presented results were achieved for two different types of data. Concerning the duration and signal quality, both two types of speech recordings were similar. In both cases, we processed the signal from relatively high quality head-set close-talk microphone. The differences could be found in speaking style and in background disturbance.

The data from lectures [7] contain more fluent speech, with medium speed, with minimal amount of longer pauses, with rather precise articulation. Consequently it means better situation from the point of view of signal quality. But when speaker made only very short pauses between particular sentences, it could yield sometimes to longer block of speech without detected shorter pause using proposed method.

On the other hand, conversational speech data [8] contain very natural and informal conversation between 3 speakers which was spoken very fast, with changing intensity of speech, and with a lot of non-speech events. Also though head-set microphone is used for all 3 speakers, soft cross-talk appears in recordings from all of them. But as the records contain the conversation, longer non-speech parts are in the signals so it is easier to resolve particular utterances.

Consequently, we have used the same setup for the first primary detection based on acoustic analysis but a little bit different settings of post-processing procedure such as different duration of non-speech part at the end of utterances was supposed in these two databases.

Following tables 1 and 3 give general overview about achieved results. Though the results are presented for several recordings only, the durations of all of them were very long, so these results should be statistically relevant. Typical duration of one recording in DSP lecture database is about 30 minutes and typical duration of one session in conversational speech database is about 90 minutes. The results have also similar trend for all analyzed sessions. We can see very small error in the detection of speech activity which means that the particular utterance should not be missed, just boundaries are usually moved.

Tables 2 and 4 show in more details how much particular error appears in obtained results. We can see that the error in non-speech parts detection is mainly in the overlapping at both boundaries. Especially for rather high quality speech, this is dominant error. For conversational speech we can observe also higher number for other two categories of non-speech errors. We have minimal error in undetected speech parts and only very small rate of badly detected speech pauses. On the other hand, a little bit higher error is in missed pauses. For conversational speech we can see more frequent classification of pauses as speech due to cross-talks appearing in analyzed speech.

At the end of this discussion it must be said that manual adjustment of boundaries can be done in different precision. For given purposes, annotators were asked for the placement of boundaries with a reserve to guarantee that speech is not truncated on these boundaries. The real error in noise can be consequently a little bit less than in presented results due to this fact.

Average lengths of particular errors presented in the table 5 illustrate typical duration of particular errors. We can see that truncation appearing at the end of speech segments is rather

small or that the extension of speech activity parts is also about 150 ms only for lecture database. The highest values are for missed pauses due to frequent placement of single boundary between two sentences instead of marking a short pause. The values are worse for conversational speech because primary short-time detection fails more frequently due to discussed more disturbing conditions.

5. Conclusions

We have presented simple and reliable algorithm for automated detection of speech activity within very long signals. The most important results can be summarized in following points.

- Simple algorithm of VAD based on power analysis with the classification on the basis of continuously updated adaptive threshold was completed by the post-processing in long-time sense. Proposed procedure has very low computational cost and it can be easily implemented, e.g. in MATLAB as it was done in our case.
- This utterance detection was used mainly for automated pre-segmentation of long recordings during the creation of large speech corpora. It represents reasonable help for further manual annotations when the annotator can just precise automatically set boundaries commonly with the annotation of utterance content.
- Analysis of achieved precision after manual correction of set boundaries has showed very low error in the detection of speech. It was 2.02% for lecture database and 2.67% for conversational speech.
- A little higher error was observed in the detection of the non-speech parts, i.e. 4.53% for lecture database and 11.73% for conversational speech. The increase of this error for conversational data was given mainly by incorrect detection of cross-talks as speech activity of main speaker in processed channel and also by missing of short pause detection between particular utterances. Also the extension of speech activity duration was a typical error.
- The application of proposed algorithm at the input of voice controlled demo WEB-page gave also satisfactory results.
- As minor result, extended criteria for the classification of voice-activity detection have been defined. They enable precise analysis of voice-activity detection classification. A simple tool implementing this analysis is available.
- The improvement of whole system is supposed to be done by analyzing additional characteristics in short-time analysis to overcame failing due to higher level of environmental noise, cross-talks, etc. More robust performance in the real environment should be the main target of this further work.

6. Acknowledgements

The research was supported by grants GAČR 102/08/0707 "Speech Recognition under Real-World Conditions", GAČR 102/08/H008 "Analysis and modelling biomedical and speech signals", and by research activity MSM 6840770014 "Perspective Informative and Communications Technicalities Research".

7. References

[1] P. Sovka and P. Pollák, "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.

[2] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Transactions on Speech and Audio Processing*, vol. SAP-10, no. 2, pp. 109–118, FEB 2002, iSSN 1063-6676.

[3] J. A. Haigh and J. S. Mason, "A voice activity detector based on cepstral analysis." in *Eurospeech'93 - Proceedings of the 3rd European Conference on Speech, Communication, and Technology*, Berlin, Sep. 1993, pp. 1103–1106.

[4] ITU, "International Telecommunication Union Recommendation G.729, annex b - A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70," 1996.

[5] ETSI, "European Standard EN 300 965 - digital cellular telecommunications system (Phase 2+); Full rate speech; Voice Activity Detector (VAD) for full rate speech traffic channels," 2000.

[6] Y. Kida and T. Kawahara, "Evaluation of voice activity detection by combining multiple features with weight adaptation," in *Proc. of Interspeech 2006, 9-th International conference on Spoken Language Processing*, Pitssburgh, Sep 2006.

[7] J. Rajnoha and P. Pollák, "The database of technical lectures," 2009, *http://noel.feld.cvut.cz/speechlab*.

[8] L. Kočková-Amortová, P. Pollák, and M. Ernestus, "Nijmeghen corpus of causal speech," 2009, (Corpus is currently under creation).

[9] M. Ernestus, *Voice assimilation and segment reduction in casual Dutch, a corpus-based study of the phonology-phonetics interface.* Utrecht: LOT, 2000.

[10] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: A free tool for segmenting, labeling and transcribing speech," in *Proc. of the First international conference on language resources & evaluation (LREC)*, Granada, Spain, 1998, pp. 1373–1376.

[11] P. Pollák, "Web-pages of speech processing group," *http://noel.feld.cvut.cz/speechlab*.

[12] J. Rajnoha and P. Pollák, "Czech spontaneous speech collection and annotation: The database of technical lectures," in *2nd International Conference on Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions, & 18th Czech-German Workshop on Speech Processing*, Prague (Czech Republic), 2008, under publishing in Lecture Notes of Computer Science.

[13] J. Rosca, R. Balan, N. P. Fan, C. Beaugeant, and V. Gilg, "Multichannel voice detection in adverse environments," in *Proceedings of EUSIPCO 2002*, Toulouse, France, Sep 2002.