ASR systems in Noisy Environment: Analysis and Solutions for Increasing Noise Robustness

Josef RAJNOHA, Petr POLLÁK

Dept. of Circuit Theory, Czech Technical University, Technická 2, 166 27 Prague, Czech Republic

rajnojos@fel.cvut.cz, pollak@fel.cvut.cz

Abstract. *This paper deals with the analysis of Automatic Speech Recognition (ASR) suitable for usage within noisy environment and suggests optimum configuration under various noisy conditions. The behavior of standard parameterization techniques was analyzed from the viewpoint of robustness against background noise. It was done for Mel-frequency cepstral coefficients (MFCC), Perceptual linear predictive (PLP) coefficients, and their modified forms combining main blocks of PLP and MFCC. The second part is devoted to the analysis and contribution of modified techniques containing frequency-domain noise suppression and voice activity detection. The above-mentioned techniques were tested with signals in real noisy environment within Czech digit recognition task and AURORA databases. Finally, the contribution of special VAD selective training and MLLR adaptation of acoustic models were studied for various signal features.*

Keywords

Robust speech recognition, robust ASR, front-end, parameterization, feature extraction, noisy speech, spectral subtraction, voice activity detection.

1. Introduction

Automatic speech recognition (ASR) systems are currently used in many applications in our everyday life. Due to the rapid development in this field all over the world we can see many systems and devices with voice input and output, e.g. automated information systems, personal dictation systems converting speech to text, systems for automated transcription of audio/video recordings or radio or TV on-line inputs, devices in cars controlled by voice, etc. Such a wide application area brings frequent usage of such systems also in noisy environment, so the issue of noise robustness represents the main topic of many research activities.

While current ASR systems working in noiseless environment can usually achieve very high accuracy, they may fail notably in an environment with background noise [1-5]. The solutions which overcome such failure are based firstly on using noise-robust features for the representation of speech signal and secondly on special modeling which takes into account degradation of analyzed signal. The solutions based on proper feature extraction originate usually from auditory-based features, i.e. Mel-Frequency Cepstral Coefficients (MFCC) [6] and Perceptual Linear Prediction (PLP) coefficients [7], which are most often used in the current ASR systems. Their performance in noisy environment can be improved by noise suppression algorithms such as Spectral Subtraction (SS) [8-10], Wiener filtering, or Minimum mean square error short time spectral amplitude estimator [11]. These methods are based on heuristic approaches so their performance under real conditions with highly non-stationary or unpredictable noise may be limited.

Secondly, the noise robustness of ASR can be increased by noise compensating methods applied in the classification phase of speech recognition. Standardly used methods such as multi-condition training [12], HMM composition and decomposition [13], parallel model combination (PMC) [4], or simple retraining of acoustic models to target environment are based on a-priori knowledge of training data or particular assumptions on the noise reduction algorithm. These approaches can provide reasonable solution, but they need a large amount of matching or almost matching [14] training data to obtain proper target acoustic models. This extensive coverage of target environment in training material increases the data collection expenses, and, moreover, full coverage of all conditions in real environment is not possible anyway. To overcome this problem, clean speech data can be mixed with independently collected noise recordings for off-line training to improve modeling of real noisy signal. But some real speech data from noisy environment need to be available in any case. The second group of these back-end techniques uses the adaptation of acoustic models to particular environment on real noisy signals. Adaptation techniques are based on Maximum-Likelihood Linear Regression (MLLR) [15] or Maximum A Posteriori (MAP) adaptation [16] and they are usually applied to speaker adaptation, but the usage for the transformation of acoustic models to match environmental conditions is possible as well.

This paper describes comprehensive experimental analysis of speech processing algorithm that combines the above-mentioned techniques to obtain efficient scheme for robust ASR. Attention is paid to the robustness of the pro-

posed system against additive background noise while using standard methods. Firstly, the performance of standard auditory-based features in noisy environment with several minor modifications is analyzed. These techniques are then supplemented with noise suppression algorithm and voice activity detection to increase their noise robustness. The advantage of using model adaptation based on MLLR algorithm together with robust front-end is also demonstrated. Such adaptation can provide satisfactory results especially in cases where only small portion of cost-intensive noisy data is available [17]. Particular methods are tested in various conditions with main focus on car noise.

2. Auditory Based Feature Extraction

Currently, the most often used speech features for ASR are based on the short-term spectral amplitude carrying principal information on speech on the basis of human perception. Due to the continuous character of speech and co-articulation, TRAP features based on longer temporal context are also used [5], [18]. Auditory-based processing of speech signal is typically applied within these techniques, simulating human perception and smoothing the influence of speech variability for particular realizations (intra-speaker) and also different speakers (inter-speaker). Techniques lacking this auditory modeling are currently rather rare.

2.1 MFCC and PLP Features

As it is important for further discussion about noise robustness of studied features, the basic description of *MFCC* [6] and *PLP* [7] is presented, along with their principal block schemes in Fig. 1 and 2.

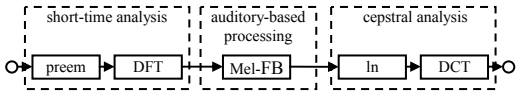

Fig. 1. Block scheme of MFCC feature extraction.

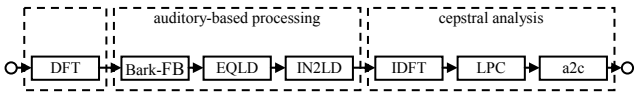

Fig. 2. Block scheme of PLP feature extraction.

Generally, both methods are based on three similar processing blocks: firstly, basic short-time Fourier analysis which is the same for both methods, secondly, auditory-based filter bank (FB), and, thirdly, cepstral coefficients computation. Both methods use principally similar auditory modeling based on Mel- or Bark-scale with non-linear frequency warping bringing similar contribution to recognition results [19]. While the shapes of filters in FBs are slightly different, the widths and the number of filters are similar (22 bands for Mel-FB and 19 bands of Bark-FB for the sampling frequency of 16 kHz). PLP uses also `EQLD` block (Equal LouDness) modifying the spectrum on the basis of frequency sensitivity of human hearing [7] and `IN2LD` block (INtensity-TO-LouDness) changing spectral dynamics according to the power-law of hearing [20]. MFCC changes frequency sensitivity only on the basis of standard pre-emphasis before short-time Fourier transform. The most significant difference between these two techniques lies in the final computation of cepstral coefficients. Autoregressive (AR) modeling is used in the case of PLP while MFCCs are computed directly using Discrete Cosine Transform (DCT) of the logarithmic auditory-based spectrum.

Our earlier experiments as well as other published results have proved the advantage of using PLP in clean conditions while MFCC technique gives better results for increasing noise level. It can be explained generally by the lower robustness of used AR modeling in the computation of PLP cepstral coefficients against noise. A more detailed study of the noise robustness of these known features under different conditions is one of the goals of this study.

2.2 Modified Auditory Based Features

The above-mentioned similarity of principal blocks and their different particular properties lead to the idea of exchanging principal blocks of MFCC and PLP, e. g. in [20], [21], or [22] and analyze more precisely the contribution of each particular block, especially from the viewpoint of noise robustness. These modifications are discussed in the following section.

The first modified method called *RPLP (Revised PLP)*, was described in [20]. The computation algorithm follows MFCC computation where the DCT-based transformation is replaced by AR modeling with additional decreasing of spectral dynamics using `IN2LD` block. The authors of this modification have presented the improvement in the accuracy of ASR recognition based on this technique against using standard methods. The most important contribution of this method lies in double suppression of spectral dynamics before LPC, however, FB band count, shape, and non-linearity scaling have rather minor effect on achieved accuracy as was shown in [19], [20], [23].

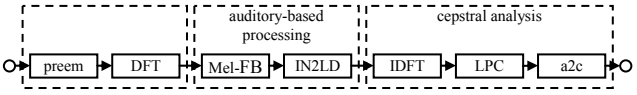

Fig. 3. Block scheme of RPLP feature extraction.

MFLP (Mel-Frequency Linear Prediction cepstral coefficients) is a technique similar to RPLP, except for the fact that it does not contain the `IN2LD` block. Experimenting with these features should analyze the contribution of spectral dynamics decreasing before LPC computation as well as the effect of minor differences between PLP and MFCC-based FB setting.

The last method that we call *BFCC (Bark Frequency cepstral coefficients)* uses filter bank from standard PLP,

but the cepstrum is then computed directly via DCT. This process combines the advantage of direct computing of the cepstrum with auditory-based processing of speech spectrum. The motivation for using DCT is to overcome frequent failing of AR modeling of speech signal in the presence of noise. The PLP-like auditory based analysis applies both equal loudness correction and the application of power law.

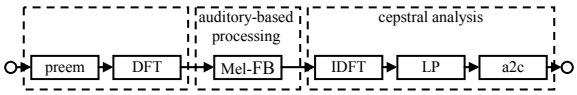

Fig. 4. Block scheme of MFLP feature extraction.

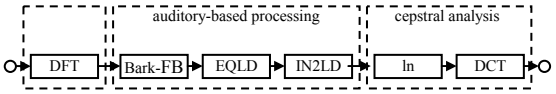

Fig. 5. Block scheme of BFCC feature extraction.

2.3 Summary of Analyzed Basic Features

Many other feature variants have been studied and published, e.g. adjusting the nonlinear frequency scaling [19], using other filter-bank shapes [23], [24] or different number of bands in filter-bank [20], [25], etc. However, these approaches are not examined in this work, as it would bring several other optional parameters and exceed the scope of this work. Moreover, many of these particular changes often brought only minor changes in recognition accuracy. Finally, we are comparing the performance of 5 above mentioned features with the following initial assumptions and knowledge:

- *MFCC* is the most frequently used method with acceptable performance under various conditions.
- *PLP* is a standard method giving very good results for the recognition under clean conditions as LPC analysis is less noise robust than DFT (DCT).
- *RPLP* is a technique similar to PLP. Experiments in [20] demonstrated the improvement of recognition accuracy against PLP under noisy conditions.
- *MFLP* is a technique that, similarly to RPLP, applies Mel-FB processing with LP-based analysis, but it preserves higher spectral dynamics of the signal.
- *BFCC* features are complementary to MFLP and they should complete the analysis of the differences between DCT and LPC-based cepstral coefficients.

2.4 Noise Robust Features

The above-described feature extraction techniques were designed without special attention to the robustness against background noise. If speech is corrupted by background noise, achieved accuracy of an ASR using these features is usually much worse, so the elimination of this degradation must be solved by additional noise suppression technique. The inclusion of frequency-domain noise suppression techniques into the above mentioned basic feature extraction procedures is suggested and described in the following section.

Spectral subtraction (SS), which is one of the most commonly used noise suppression techniques, is frequently applied in many noise robust front-end processing schemes [26], [27]. It eliminates additive background noise, as it is based on the subtraction of an additive component in the spectral domain. The principle of this technique is simple and it usually provides satisfactory results. On the other hand, the level of noise suppression as well as possible distortion of cleaned speech depends strongly on proper estimation of magnitude spectrum of the background noise.

Particular algorithms of spectral subtraction differ mainly in above-mentioned estimation procedure. Typically, the estimation is made from non-speech parts of the utterance detected by voice activity detector (VAD) [11] or by minimum statistics [28]. Iterative algorithm based on modified adaptive Wiener filtering [29], which we call extended spectral subtraction (ESS), is used in our feature extraction framework. This algorithm works without VAD and it can eliminate stationary or non-stationary noise with rather slow changes in characteristics. Further details describing these techniques can be found in [29].

All techniques mentioned above are based on spectral subtraction and they can be easily included into the feature extraction procedure. The subtraction in the frequency domain can be applied directly to the output of short-time DFT, which is typical usage (Fig. 6), or it can be placed after auditory-based FB, where the signal still has frequency domain representation. Both variants of SS placement give similar results [30], however, the position before FB seems to be generally more robust as the application of FB smoothes possible distortions after the application of SS within smoothing of short-time magnitude spectrum.

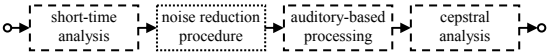

Fig. 6. Inclusion of spectral subtraction into feature extraction.

The paper presents a study of the influence of spectral subtraction for all above mentioned feature extraction techniques, as the effect of SS in particular features can differ due to different types of FB and cepstral analysis approaches.

2.5 Front-end Robustness Analysis

The main purpose of this article is to analyze the performance of ASR from the viewpoint of noise robustness in various environmental conditions. All tests were performed with recordings from real-world environment with various levels of noisy background. Firstly, experiments on small vocabulary speech recognition (Czech digits) were

conducted and secondly, the tests on AURORA3 databases were performed. Experimental setups for Czech digits recognition and for AURORA3 are described in the following section along with the results of experiments on noise robust features.

2.5.1 Experimental Setup

A speaker-independent connected ***Czech digit recognizer*** was created using HTK Toolkit [31] with the following parameters: phoneme-based context independent acoustic modeling, 44 Czech monophone HMMs, modeling of short pause and silence, standard left-to-right 3-state structure without state skips (excepting short-pause model), 32 mixtures, 3 streams for static, dynamic and acceleration features, simple loop grammar for particular digits. Context independent models of monophones were used in this part due to smaller amount of data in particular subsets for training more complex models.

The recognizer uses loop grammar with 10 digits with the same probability of occurrence and with several pronunciation variants. Despite performing rather simple recognition task with small vocabulary, the recognition of digits with possible repetitions and without any further restriction is a task where proper feature extraction can affect the target accuracy quite strongly.

No additional word insertion penalty tuning was used within WER minimization and one common setup of the recognizer based on previous results was used within the experiments.

General setup for feature extraction uses: 12 cepstral coefficients complemented by the energy in the form of static, dynamic and acceleration coefficients, 25 ms length of short-time frame, 10 ms frame step, the AR model used in LPC-based techniques of order 13.

The studied features were computed by ***CtuCopy*** [32], created in our lab as an extension of HTK tool *HCopy*. This tool makes it possible to apply different parameterization settings in combination with noise suppression techniques. The tool was used and firstly described in [26]. The current, updated version is now available for public use [32].

The recognition results were analyzed standardly in terms of Word Error Rate (*WER*) and relative *WER* reduction (*WERR*)

$$WER = \frac{S + D + I}{N} \cdot 100\,\% \,, \tag{1}$$

$$WERR = \frac{WER_{base} - WER}{WER_{base}} \cdot 100\,\% \tag{2}$$

where N means the total number of words in the test dataset and *S, D, I* represent substituted, deleted and inserted words respectively. WERR is computed against defined baseline word error rate WER_{base}.

2.5.2 Train and Test Datasets

The performance of digit sequence recognizer was tested on the selections of Czech SPEECON database [33]. Same as other databases from this family, it contains 16 kHz data recorded in various kinds of environment using more types of microphones, more than 550 speakers and 300 utterances per speaker with various content, i.e. phonetically rich sentences, digits, commands, application phrases, or spontaneous speech. The whole database was precisely revised and utterances containing mispronunciations or possible transcription inaccuracies were removed. This cleared corpus has been called ALL and all utterances in cleared corpus could be used for training of monophone HMMs.

To test the recognition under various conditions, further particular subsets have been created. The ALL set has been divided on the basis of different noise level according to the recording environment. While the CLEAN subset contains sessions recorded in offices and living rooms with relatively quiet background, the NOISY subset contains sessions collected in car and public places such as hall or open area. The OFFICE subset represents clean environment with very low level of background noise which is very typical for the usage of ASR applications. This subset contains approximately half of all sessions of the whole SPEECON database with usually very high SNR (> 20 dB). Generally, the real level of the background noise in particular subsets can vary significantly. The SNRs of utterances in particular subsets estimated during the database recording were analyzed, see Fig. 7 and it was observed that especially NOISY subset contained highly disturbed utterances from different environments (car and public places) which caused bi-modality of SNR distribution for NOISY and ALL subsets. It means finally highly mis-matched conditions for the recognition in these subsets. Training and testing data were then chosen within each subset. The amounts of speech material for the experiments in particular subsets are summarized in Tab. 1.

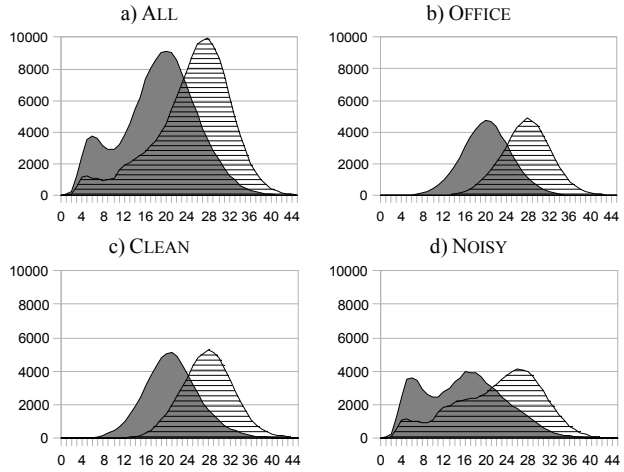

Fig. 7. SNRs in SPEECON subsets. (Hatched graph: head-set microphone, grey graph: hands-free microphone).

Name	Train		Test	
	[ses]	[hours]	[ses]	[hours]
ALL	531	141.7	59	0.63
OFFICE	190	51.6	21	0.23
CLEAN	220	59.7	25	0.26
NOISY	273	71.7	30	0.32

Tab. 1. Train and test subsets.

The data from two channels with different noise level were used for our experiments, i.e. channel 0 (CS0) recorded by head-set microphone and channel 1 (CS1) recorded by the microphone from NOKIA mobile phone hands-free set [33]. Signals from channel CS1 contain higher level of noise, which is also demonstrated by the distribution of SNR for the training signals in Fig. 7a-d.

2.5.3 Results of Czech Digits Recognition with Standard Features

Particular techniques were compared within small vocabulary ASR performed on the above mentioned data with various levels of background distortion. The models were trained on training data that correspond to the test subset.

This experiment shows rather small robustness of features without spectral subtraction in real environment. The comparison of particular methods under various conditions presented in Tab. 2 shows slight advantage of PLP against MFCC in the case of cleaner speech signal (silent environment, higher microphone quality). On the other hand, the recognition with all training data leads to better results for MFCC, which makes it more suitable for general conditions, where the noise level is not known or in conditions with changing noisy background. The parameterization method which gives the best recognition performance is highlighted in Tab. 2.

(a)	MFCC	PLP	MFLP	BFCC	RPLP
OFFICE	**5.74**	6.14	9.35	6.14	6.14
CLEAN	7.19	6.74	9.47	8.90	**6.62**
NOISY	13.98	18.97	16.47	19.95	**11.84**
AVG	8.97	10.62	11.76	11.66	**8.20**
ALL	14.53	13.94	13.57	13.48	**10.37**

(b)	MFCC	PLP	MFLP	BFCC	RPLP
OFFICE	**9.48**	10.68	14.02	12.15	10.41
CLEAN	10.73	9.93	13.36	12.21	**9.47**
NOISY	**15.14**	24.67	20.04	31.43	16.74
AVG	**11.78**	15.09	15.81	18.60	12.21
ALL	**15.04**	18.65	18.56	20.02	16.00

Tab. 2. Results of Czech digit recognition using head-set microphone (a) and hands-free set (b) for standard features.

Compared to the standard methods, RPLP technique brings decrease in recognition error for noisy conditions. As the bold numbers in Tab. 2 demonstrate, this technique reaches the best recognition performance in all conditions which are supposed to contain higher level of background noise. MFLP and BFCC achieve higher error rates against standard methods in all conditions. In the case of a more distorted channel, RPLP method gives results similar to the best MFCC, while MFLP overcomes PLP in noisy environment and BFCC gives the worst performance.

2.5.4 Results of Czech Digits Recognition with Noise Robust Features

To decrease the influence of noisy conditions, ESS was applied within the parameterization procedure. The resulting performance of digit recognition in various backgrounds is shown in Tab. 3.

The results show that the proposed front-end processing method can bring an improvement especially for NOISY subset with higher level of additive noise. In the case of PLP it gives 26.7% of WERR for NOISY subset and 18.8% of WERR in average. We can observe that PLP can outperform MFCC after the application of ESS. On the other hand, rather small reduction of WER can be caused by the type of signals in this database, where not all of the recordings contain short starting non-speech part needed for proper initialization of ESS. Moreover, high improvements cannot be expected by using ESS because the character of the distortion is not only additive, but the signal from real environment also contains some non-stationary noises and reverberation. Therefore, the contribution of the method varies highly in particular conditions, as ESS can suppress the noise just with slow non-stationary character.

(a)	MFCC	PLP	MFLP	BFCC	RPLP
OFFICE	6.94	**5.47**	9.88	7.08	6.01
CLEAN	7.88	**6.51**	11.30	8.45	7.76
NOISY	**11.40**	13.09	17.36	16.30	11.67
AVG	8.74	8.62	12.85	10.61	**8.48**
ALL	11.70	11.75	14.63	14.58	**11.52**

(b)	MFCC	PLP	MFLP	BFCC	RPLP
OFFICE	**10.41**	11.48	12.15	14.29	11.21
CLEAN	10.96	11.19	12.10	13.58	**10.84**
NOISY	15.85	24.13	16.30	34.55	**12.91**
AVG	12.41	15.60	13.52	20.81	**11.65**
ALL	13.21	18.28	14.72	30.30	**12.43**

Tab. 3. Results of Czech digit recognition using head-set microphone (a) and hands-free set (b) for noise robust features using ESS.

2.5.5 AURORA3 Experiments using Noise Robust Features

Standardized experiments for the comparison of the proposed methods with other published ASR results were also performed using AURORA3 recognition test [34]. These tests use different recognition setup. Unlike the previous SPEECON tests, the database of spoken digits is used here for training and the recognizer is based on the models of whole words.

Tab. 4 summarizes the results for 3 languages and various levels of matching in training and testing conditions, i.e. well matched (WM), medium mismatch (MM), and high mismatch (HM). The average value (AVG) gives

overall recognition performance for the comparison against baseline results, and the bold number refers to the overcoming of baseline results. Baseline results were achieved for standard MFCC coefficients without noise reduction and additional reference results were achieved also for parameterization based on ETSI ES 202 050 standard [36]. It uses two-stage speech enhancement algorithm and additional voice activity detection. As presented results analyze the contribution of ESS only, they are better comparable with standard MFCC. ETSI standard is therefore used as a reference in further experiments with noise reduction framework presented in the next section.

	Spanish						
	baseline	MFCC	PLP	MFLP	BFCC	RPLP	*ETSI*
WM	*13.15*	15.18	**10.04**	**12.14**	**10.74**	**10.72**	*6.58*
MM	*26.26*	33.68	28.59	33.96	29.83	**24.32**	*13.27*
HM	*57.77*	**56.78**	60.30	**56.24**	62.65	**56.63**	*15.79*
AVG	*32.39*	35.21	32.98	34.11	34.41	**30.56**	*11.88*

	Finnish						
	baseline	MFCC	PLP	MFLP	BFCC	RPLP	*ETSI*
WM	*9.61*	**8.09**	**6.56**	**5.16**	9.76	**6.44**	*2.52*
MM	*27.63*	30.16	60.19	**22.02**	33.31	46.31	*12.72*
HM	*68.94*	**51.84**	**64.66**	**62.01**	77.42	**59.72**	*18.83*
AVG	*35.39*	**30.03**	43.80	**29.73**	40.16	37.49	*11.36*

	Danish						
	baseline	MFCC	PLP	MFLP	BFCC	RPLP	*ETSI*
WM	*22.20*	**17.04**	**18.30**	**18.37**	**18.52**	**17.50**	*15.87*
MM	*53.60*	55.93	**49.72**	**48.87**	**51.55**	**50.00**	*38.98*
HM	*68.10*	**59.21**	69.24	**58.66**	79.27	**65.20**	*37.81*
AVG	*47.97*	**44.06**	**45.75**	**41.97**	49.78	**44.23**	*30.89*

Tab. 4. WER on Aurora3 for noise robust features using ESS.

As these data mainly contain speech with additive car noise with small level of reverberation, the contribution of ESS noise reduction is evident. The results are good especially for WM conditions, where LPC based techniques with ESS have achieved significant improvement in comparison to the given baseline (24% WERR for PLP and Spanish, 46% WERR for MFLP and Finnish, 21% WERR for RPLP and Danish).

3. Noise Robust Acoustic Modeling

The techniques that solve noise robustness on the basis of acoustic modeling are discussed within this section. Generally, the case of matching the training data for target environment for particular SPEECON subsets as it was used in the previous section can also be taken for noise specific modeling. Nevertheless, we will focus on another technique in the following section: selective matched training using VAD.

3.1 VAD in Feature Extraction

When speech is disturbed by an environmental background, non-speech parts of such signal can be influenced much more strongly. Such distortion can be the source of many faults in the result of ASR systems though different noise suppression methods can improve the accuracy of target ASR. Typically, disturbed non-speech segments can often be recognized as speech and it yields to increasing WER during the recognition or bad tuning of acoustic models during the training phase. Concerning this fact, VAD algorithm is therefore used as a frame dropping technique to remove potentially bad non-speech segments from the processed signal.

As the recognizer works with the cepstral representation of the signal, we use these features also for the VAD algorithm. It is based on smoothed differential cepstrum computation followed by cumulative distance computation, thresholding and final smoothing of binary output. Acceptable accuracy and possible usage of pre-computed differential cepstrum coefficients (used in ASR) represent big advantages of this approach, which is described in more detail in [35].

Depending on the conditions, the VAD algorithm can remove the non-speech parts of the signal, but occasionally also some frames with speech activity. This serious disadvantage can strongly influence the accuracy of target recognition. On the other hand, this problem of removing some speech frames seems to be acceptable during the training phase and further presented experimental results proved its utility.

Regarding this behavior, so called VAD selective training was used within the following experiments i.e. VAD-based frame dropping algorithm was used only in the training phase of ASR procedure. The motivation to apply this algorithm is in the removing of pauses and possibly strongly distorted parts of speech, which contributes to more accurate phone modeling. Despite removing silence frames within the parameterization of training subset, some parts of data with non-speech character are not removed by the detector due to used smoothing algorithm in VAD post-processing. These frames are used for training the models of non-speech elements.

3.2 Experiments with Noise Robust Acoustic Modeling

This section presents the results of experiments with improved acoustic modeling for increasing the robustness of ASR. BFCC features were excluded from these experiments, as their presence led to strongly unsatisfactory results in the previous experiments. The first tests use the same data as the previously presented feature-oriented experiments while the contribution of adaptation techniques is analyzed on another database recorded in car environment.

3.2.1 VAD Based Selective Training

Tab. 5 shows the obtained results of Czech digit recognition after proposed VAD selective training, which

significantly improved the recognition performance. VAD was used also for testing. Compared to the recognition without VAD, the results give even more than 50% WERR, mainly for highly disturbed conditions.

(a)	MFCC	PLP	MFLP	RPLP
OFFICE	**3.74**	4.14	4.81	3.87
CLEAN	4.22	4.79	5.59	**4.00**
NOISY	**9.53**	11.22	12.73	11.49
AVG	**5.83**	6.72	7.71	6.45
ALL	**7.27**	**7.27**	9.83	7.54

(b)	MFCC	PLP	MFLP	RPLP
OFFICE	**5.08**	6.01	5.74	5.34
CLEAN	5.83	6.16	6.51	**5.48**
NOISY	10.77	12.06	10.95	**10.33**
AVG	7.23	8.08	7.73	**7.05**
ALL	8.64	9.14	9.00	**7.91**

Tab. 5. Results of Czech digit recognition using head-set microphone (a) and hands-free set (b) for noise robust features using ESS and VAD selective training and testing.

As it was mentioned before, the result of VAD was smoothed to avoid false short detections. Despite such correction, the resulting speech could be affected by the detection and some part of speech could be removed. Therefore VAD detection was used only for training, not for testing in the following experiment.

The results in Tab. 6 show further improvement of recognition score and WER reaches up to 2% for clean environment.

(a)	MFCC	PLP	MFLP	RPLP
OFFICE	3.47	**2.00**	4.41	3.60
CLEAN	3.54	**2.85**	4.91	3.65
NOISY	**8.64**	10.15	10.77	8.73
AVG	5.22	**5.00**	6.70	5.33
ALL	6.22	6.31	9.10	**6.12**

(b)	MFCC	PLP	MFLP	RPLP
OFFICE	**4.81**	5.07	6.54	5.07
CLEAN	**4.57**	4.79	6.96	5.37
NOISY	7.66	10.95	9.88	**7.03**
AVG	**5.68**	6.94	7.79	5.82
ALL	6.99	7.72	8.18	**6.44**

Tab. 6. Results of Czech digit recognition using head-set microphone (a) and hands-free set (b) for noise robust features using ESS and VAD selective training and only ESS for testing.

3.2.2 VAD Selective Training in AURORA3 Test

The above-mentioned results with VAD selective training were confirmed again by AURORA3 test. It can be observed in Tab. 7 that the achieved results outperform in strong majority not only the AURORA baseline, but also the ETSI standard [36] (bold numbers).

The comparison of proposed methods shows their behavior within rather simple noise reduction scheme. The results of proposed techniques outperform ETSI standard in many cases, though ETSI standard is based on multiple noise reduction algorithms. VAD detection was used also in the testing phase, as the smoothing algorithm together with modeling of longer speech parts (words) in AURORA tests gives more precise results than in case of the phoneme modeling (previous experiments). The resulting error rate decreased by almost 40% for well matched Danish and by 20% for highly mismatched Danish against ETSI. The contribution was significant especially for LPC based features as PLP or RPLP.

	Spanish					
	baseline	MFCC	PLP	MFLP	RPLP	ETSI
WM	*13.15*	**4.77**	**5.10**	6.76	**4.45**	*6.58*
MM	*26.26*	**11.35**	**13.08**	14.51	**11.25**	*13.27*
HM	*57.77*	18.34	17.49	20.83	18.52	*15.79*
AVG	*32.39*	**11.49**	11.89	14.03	**11.41**	*11.88*

	Finnish					
	baseline	MFCC	PLP	MFLP	RPLP	ETSI
WM	*9.61*	4.60	3.42	3.46	3.51	*2.52*
MM	*27.63*	21.00	21.75	18.74	19.08	*12.72*
HM	*68.94*	23.82	31.55	**18.27**	29.82	*18.83*
AVG	*35.39*	16.47	18.91	13.49	17.47	*11.36*

	Danish					
	baseline	MFCC	PLP	MFLP	RPLP	ETSI
WM	*22.20*	**9.83**	**10.26**	**10.96**	**9.81**	*15.87*
MM	*53.60*	**29.59**	**30.82**	**30.40**	**28.84**	*38.98*
HM	*68.10*	**33.74**	**31.93**	**30.50**	**30.55**	*37.81*
AVG	*47.97*	24.39	24.34	23.95	23.07	*30.89*

Tab. 7. WER on Aurora3 with ESS and with selective VAD training.

4. Acoustic Model Adaptation in a Car Environment

While the robustness in terms of sensitivity to various conditions was analyzed in previous experiments, this section describes the contribution of analyzed framework within environmental adaptation. For this purpose, data from more specific conditions of car environment were used commonly with different recognition setup.

4.1 MLLR and Regression Classes

The MLLR technique estimates and applies affine transform of the HMM parameters in terms of likelihood maximization [15]. This algorithm is advantageous especially in situations where small amount of adaptation data is available, as it can share the transforms among several models. This is exactly the case of the continuously changing environmental characteristics, e.g. in automotive applications. Based on our previous experiments and other published work we use the simplest case, where only Gaussian means in HMMs are transformed while other parameters stay unchanged.

Global adaptation is the simplest and basic approach in MLLR which uses only one transform for all HMMs. It makes it possible to collect the adaptation database without special demands, e.g. on a sufficient number of occurrences of each modeled element. When more data is available for the adaptation, more transforms can be estimated independently to characterize more precisely the influence of noisy background on different groups of speech elements divided into particular regression classes. We use the division into speech and non-speech class, as it is the simplest form of categorization representing different nature of the signals within these groups and preserving low requirements to the adaptation data.

If adaptation data are available before the recognition, static adaptation can be performed and the data are processed in one block. This is very useful when general speaker independent system is to be adapted to certain background conditions without the adaptation to any speaker. Within this study, the general system is adapted to noisy car environment and the results are compared with standard retraining based on Baum-Welch re-estimation.

4.2 Experimental Setup

This section compares proposed parameterization techniques in robust speech recognition task in a car with environmental adaptation. As there is more available data for training for given conditions, the ASR system could be extended in comparison to previous parts. The following section describes the different setup.

4.2.1 Triphone Based Czech Digit Recognizer

The experiments on MLLR were performed on small vocabulary speaker independent ASR task with the following specification: Czech digit sequence recognizer based on tools from the HTK toolkit [31] (*HDecode*), HMMs of cross-word triphones, Gaussian functions with 32 mixtures, 25 ms segmentation with 10 ms step, and simple unigram language model with uniform probability and several pronunciation variants. This setting was selected with the intention to run simple system, that provides adaptation procedure and that can be simply expanded to more complex speech recognition system in further work.

4.2.2 Car Speech Database

The database involves recordings of 700 speakers captured by two channels with original sampling frequency 44.1 kHz, which were for our purposes down- sampled to 16 kHz for compatibility with other experiments. Finally, we have 2 sets of data from 700 speakers in FAR-TALK channel collected by AKG far-talk microphone which is strongly favorable for in-car application and 329 speakers in CLOSE-TALK channel collected by Sennheiser head-set microphone (a source of high quality speech signal)[1]. Data from both channels (far-talk/close-talk) were divided into basic retraining subset (500/242 speakers), adaptation subset (100/42) and test subset (100/42). This division preserved similar distribution of car classes in particular subsets as it significantly influenced the noise level.

All sessions were recorded under three different conditions: standing car with engine off (acronym OFF), standing car with engine on (ON) and running car (DRV). Our experiments were realized with the purpose of analyzing the behavior of ASR within these three significantly different environmental conditions.

4.2.3 Adaptation Procedure

The overall adaptation procedure started with baseline acoustic models trained on head-set microphone channel from CLEAN subset of SPEECON database. As these general models would be very inefficient when they are used in car environment, they were retrained on car specific data to fit in-car environment. It avoids the need of training the models from scratch; only one single-pass retraining step is performed. It was performed on OFF part of train subset by standard Baum-Welch re-estimation. The models are then adjusted from the viewpoint of environmental conditions, not for channel characteristics within the following adaptation procedure.

The adaptation was in all instances performed on signals from particular adaptation subsets (OFF, ON, DRV). All available data of each subset were used, but the adaptation with smaller amount of data is also possible. The recognition performance with adapted HMM models was analyzed for all the above-mentioned features with and without applied ESS noise suppression.

4.2.4 Results of MLLR Adaptation

In almost all cases in Tab. 8 and Tab. 9, using MLLR outperforms the baseline system and also the system with retrained models by single-pass retraining. The contribution of this method can be observed mainly in highly distorted background (DRV). The results correspond with the experiments mentioned above. Similarly to clean conditions, applying MLLR leads to higher performance of PLP-like parameterization and the RPLP technique gives the best overall accuracy, even though it loses its performance in clean environment (OFF) compared to standard methods. Though the results for MFLP are comparable to standard methods in clean conditions, it achieves higher error rate in noisy environment similarly to the above-noted experiments.

Achieved improvement is significant especially for DRV conditions where we can observe 87.1% WERR with respect to the baseline (WER from 13.78 to 1.77) or 73.9% WERR related to results after single pass retraining (WER

[1] The remaining part of CZKCC database involves the Peiker far-talk microphone recordings.

no ESS	OFF				ON				DRV				AVG			
	MFCC	PLP	MFLP	RPLP	MFCC	PLP	MFLP	RPLP	MFCC	PLP	MFLP	RPLP	MFCC	PLP	MFLP	RPLP
baseline	2.26	2.00	1.83	1.22	1.22	4.62	1.83	0.88	23.41	10.50	17.28	13.78	9.00	5.70	7.00	**5.30**
SP	1.48	1.30	1.48	1.74	0.81	2.78	1.15	0.54	28.27	13.04	29.83	6.78	10.00	5.70	11.00	**3.00**
MLLR	1.65	1.13	1.48	1.22	0.95	1.97	0.88	0.41	2.89	2.89	2.77	1.77	1.80	2.00	1.70	**1.10**

ESS	OFF				ON				DRV				AVG			
	MFCC	PLP	MFLP	RPLP	MFCC	PLP	MFLP	RPLP	MFCC	PLP	MFLP	RPLP	MFCC	PLP	MFLP	RPLP
baseline	1.39	1.91	1.22	1.39	0.68	1.63	1.09	0.81	17.16	8.51	14.14	10.26	6.40	**4.00**	5.50	4.20
SP	1.65	1.04	1.22	1.74	1.43	1.22	0.95	0.68	13.01	5.59	19.61	5.79	5.40	**2.60**	7.30	2.70
MLLR	1.30	0.70	1.04	1.57	1.09	1.02	0.81	0.68	1.90	2.06	2.64	1.56	1.40	**1.30**	1.50	**1.30**

Tab. 8. Czech digit ASR WER on car data recorded by head-set microphone.

no ESS	OFF				ON				DRV				AVG			
	MFCC	PLP	MFLP	RPLP	MFCC	PLP	MFLP	RPLP	MFCC	PLP	MFLP	RPLP	MFCC	PLP	MFLP	RPLP
baseline	20.48	19.65	17.79	16.39	26.87	25.87	22.66	22.12	53.52	48.52	43.44	41.83	34.00	31.00	28.00	**27.00**
SP	2.14	1.98	2.66	2.24	2.90	2.06	4.61	1.62	16.25	18.91	27.52	12.38	7.10	7.70	12.00	**5.40**
MLLR	2.72	2.11	2.69	2.94	1.84	1.34	3.02	1.75	7.04	6.32	8.86	5.91	3.90	**3.30**	4.90	3.50

ESS	OFF				ON				DRV				AVG			
	MFCC	PLP	MFLP	RPLP	MFCC	PLP	MFLP	RPLP	MFCC	PLP	MFLP	RPLP	MFCC	PLP	MFLP	RPLP
baseline	15.14	15.52	14.18	12.83	18.58	20.29	15.40	16.15	44.09	42.27	38.63	41.38	26.00	26.00	**23.00**	**23.00**
SP	3.36	2.24	2.34	2.08	2.21	1.71	2.56	1.78	13.83	15.10	19.01	14.54	6.50	6.30	8.00	**6.10**
MLLR	3.36	2.46	2.40	2.27	2.52	1.47	2.21	1.78	6.23	6.97	7.75	5.52	4.00	3.60	4.10	**3.20**

Tab. 9. Czech digit ASR WER on car data recorded by far-talk microphone.

from 6.78 to 1.77) for the case of RPLP without ESS. For RPLP with used ESS, 84.8% WERR with respect to the baseline (WER from 10.26 to 1.56) and 73.1% WERR related to results after single pass retraining (WER from 5.79 to 1.56) were achieved.

5. Conclusions

This paper presents a comprehensive analysis of several approaches of robust speech recognition for different feature extraction techniques and procedures of HMM training or adaptation respectively. Large amount of experiments were realized and the main focus was put on the performance of studied techniques in strongly disturbed real environment. The following points summarize the main results of this study.

- *Modified feature extraction* algorithms were studied to analyze the effect of particular computation steps of signal processing. Especially the influence of speech dynamics suppression within LPC and human-like signal processing served as the inspiration for the proposed techniques.
- The comparison of *standard methods* showed the supposed higher robustness of MFCC against PLP in noisy environment (14% vs. 19% WER). The PLP technique performed better in clean conditions without background distortion (7.2% vs. 6.7%) or cleaned by additional noise suppression techniques (6.9% vs. 5.5%). However, the accuracy was still worse for PLP under highly disturbed conditions.
- Concerning the proposed *modified methods*, the best overall score comparable to standard methods was observed for the RPLP technique. Lower performance of MFLP- and BFCC-based recognition shows positive effect of the suppression of speech dynamics for LP-based techniques.
- Additional noise reduction schema concerning extended spectral subtraction and VAD selective training increased recognition performance significantly and the results outperformed also 2-stage ETSI standard in many cases. Error rate reduction by 20 to 40% was achieved within AURORA3 tests.
- *The MLLR technique* significantly contributes to high improvement of recognition accuracy even in strongly distorted environment of a driven car. LP-based techniques supplemented by noise reduction method outperform even MFCC (e.g. 1.8% against 1.1% of avg. WER for recognition without ESS).

The study showed the advantage of using LP-based feature extraction algorithms within ASR with less disturbed background, both due to clean environmental conditions and due to using noise suppression methods.

Acknowledgements

The presented work has been supported by GAČR 102/08/0707 “Speech Recognition under Real-World Conditions”, GAČR 102/08/H008 “Analysis and modeling biomedical and speech signals”, and research activity MSM 6840770014 “Perspective Informative and Communications Technicalities Research”.

References

[1] OPENSHAW, J. P., MASON, J. S. On the limitations of Cepstral features in noise. In *Proc. ICASSP,* 1994, vol. 2, p. 49-52.

[2] WET, F. de, CRANEN, B., VETH, J. de, BOVES, L. A comparison of LPC and FFT-based acoustic features for noise robust ASR. In *Eurospeech 2001*, p. 865-868.

[3] CHOI, E. Noise robust front-end for ASR using spectral subtraction, spectral flooring and cumulative distribution mapping. In *Proc. 10th Australian Int. Conf. on Speech Science and Technology*, SST'04. Sydney (Australia), Dec. 2004, p. 451-456.

[4] GALES, M. J. F., YOUNG, S. J. *Parallel Model Combination for Speech Recognition in Noise.* Technical report CUED/FINFENG/ TR 135, Cambridge, England, 1993.

[5] HERMANSKY, H., SHARMA, S. Temporal patterns (TRAPs) in ASR of noisy speech. In *ICASSP '99: Proc. of IEEE Int. Conf. on the Acoustics, Speech, and Signal Processing*. Washington DC (USA), IEEE Computer Society, 1999, p. 289-292.

[6] DAVIS, S., MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, Aug 1980, vol. 28, p. 357-366.

[7] HERMANSKY, H. Perceptual linear predictive (PLP) analysis of speech. In *Proc. JASA*, April 1990, vol. 87, no. 4.

[8] KANG, G. S., FRANSEN, L. J. Quality improvement of LPC-processed noisy speech by using spectral subtraction. *IEEE Trans. on ASSP*, June 1989, vol. 37, no. 6, p. 939-942.

[9] BABA ALI, B., SAMETI, H., SAFAYANI, M. Likelihood-maximizing-based multi-band spectral subtraction for robust speech recognition. *EURASIP Journal on Advances in Signal Processing*, 2009. Article ID 878105, 15 p.

[10] XU, C., LIU, Y., YANG, Y. S., et al. A system for Mandarin short phrase recognition on portable devices. In *Proc. of Int. Symp. on Chinese Spoken Language Processing*, 2004.

[11] EPHRAIM, Y., MALAH, D. Speech enhancement using a minimum mean square error short time spectral amplitude estimator. *IEEE Trans. on ASSP*, Dec. 1984, vol. 32, no. 6, p. 1109-1121.

[12] MING, J., JANCOVIC, P., HANNA, P., STEWART, D. Modeling the mixtures of known noise and unknown unexpected noise for robust speech recognition. In *Proc. of Eurospeech'2001*. Aalborg (Denmark), 2001, p. 579-582.

[13] VARGA, A. P., MOORE, R. E. Hidden Markov model decomposition of speech and noise. In *Proc. ICASSP*, 1990, p. 845-848.

[14] LIAO, Y. F., FANG, H. H., HSU, C. H. Eigen-MLLR environment/speaker compensation for robust speech recognition. In *Proc. Interspeech'08*. Brisbane (Australia), September 2008, p. 1249-1252.

[15] LEGGETTER, C. J., WOODLAND, P. C. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language*, April 1995, vol. 9, no. 2, p. 171-185.

[16] GAUVAIN, J. L., LEE, C. H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans. on SAP*, 1994, vol. 2, no. 2, p. 291-298.

[17] FISHER, A., STAHL, V. Database and online adaptation for improved speech recognition in car environments. In *Proc. ICASSP'99*, p. 445-448.

[18] HERMANSKY, H. TRAP-TANDEM: Data-driven extraction of temporal features from speech. In *Proc. of ASRU'03*. Martigny (Switzerland), 2003, p. 255-260.

[19] UMESH, S., COHEN, L., NELSON, D. Fitting the Mel scale. In *Proc. ICASSP*, 1999, vol. 1, p. 217-220.

[20] HÖNIG, F., STEMMER, G., HACKER, C., BRUGNARA, F. Revising Perceptual Linear Prediction (PLP). In *Eurospeech 2005*, p. 2997-3000.

[21] ZOLNAY, A., SCHLÜTER, R., NEY, H. Acoustic feature combination for robust speech recognition. In *ICASSP'05*. Philadelphia (PA, USA), March 2005, vol. 1, p. 457-460.

[22] SCHLÜTER, R., BEZRUKOV, I., WAGNER, H., NEY, H. Gamma tone features and feature combination for large vocabulary speech recognition. In *ICASSP 2007*. Honolulu (HI, USA), April 2007, p. 649-652.

[23] LI, Q., SOONG, F. K., SIOHAN, O. An auditory system-based feature for robust speech recognition. In *Eurospeech 2001*, p. 619-622.

[24] PSUTKA, J., MÜLLER, L., PSUTKA, J. V. The influence of a filter shape in the telephone-based recognition module using PLP parameterization. In *TSD 2001*. Berlin, Springer-Verlag 2001, p. 222-228.

[25] PSUTKA, J., MÜLLER, L., PSUTKA, J. V. Comparison of MFCC and PLP parameterizations in the speaker independent continuous speech recognition task. In *Eurospeech 2001*, p. 1813-1816.

[26] FOUSEK, P., POLLÁK, P. Additive noise and channel distortion-robust parameterization tool – performance evaluation on Aurora 2&3. In *Eurospeech 2003*, p. 1785-1788.

[27] OKUNO, H. G., OGATA, T., KOMATANI, K. Computational auditory scene analysis and its application to robot audition: Five years experience. In *Proc. of the 2nd Int. Conf. on Informatics Research for Development of Knowledge Society Infrastructure*. ICKS. IEEE Computer Society, Washington, DC, 2007, p. 69-76.

[28] MARTIN, R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Tran. on Speech and Audio Processing*, July 2001, vol. 9, no. 5, p. 504 - 512.

[29] SOVKA, P., POLLÁK, P., KYBIC, J. Extended spectral subtraction. In *European Signal Processing Conference (EUSIPCO'96)*. Trieste (Italy), September 1996.

[30] NOVOTNY, J., MACHACEK, L. Noise reduction applied in real time speech recognition system. In *Polish-Czech-Hungarian Workshop on Circuit Theory, Signal Processing, and Telecommunication Networks*. Budapest (Hungary), September 2001.

[31] HTK speech recognition toolkit. [Online]. Ver. 3.3. July 2005. Available at: http://htk.eng.cam.ac.uk/

[32] CtuCopy. [Online]. Ver. 3.0.11. Available at: http: //noel.feld.cvut.cz/speechlab/en/download/CtuCopy_3.0.11.tar.bz2

[33] SPEECON database distributed through the European Language Resources Association [Online]. Available at: http://catalog.elra.info/search_result.php?keywords=speecon&language=en&osCsid=66

[34] HIRSCH, H. G., PEARCE, D. The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. In *ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium.* Paris (France), September 2000.

[35] RAJNOHA, J., POLLÁK, P. Voice activity detection based on perceptual cepstral analysis. In *Technical Computing Prague 2008* [CD-ROM]. Prague: HUMUSOFT, 2008, vol. 1, p. 1-9. (in Czech).

[36] ETSI Distributed speech recognition ES 202 050 standard [online]. Available at: http://www.etsi.org/WebSite/Technologies/ Distributed SpeechRecognition.aspx

About Authors...

Josef RAJNOHA was born in 1982 in Roudnice nad Labem, Czech Republic. He graduated from the Faculty of Electrical Engineering, Czech Technical University in Prague in 2006 (Ing.). During his Ph.D. studies at the same faculty, he has been interested in robust speech recognition with special focus on modeling of non-speech events. Now he is finishing his doctoral thesis and works as a business process consultant in SAP ČR.

Petr POLLÁK was born in 1966 in Ústí nad Orlicí, Czechoslovakia. After the graduation (Ing. 1989) he joined the Czech Technical University in Prague where he has also received his further degrees (CSc. 1994, Doc. 2003). He works as teacher and researcher in the Speech Processing Group at the Faculty of Electrical Engineering. His most important activities are in robust speech recognition, speech enhancement, speech database collection, and other related activities. He was responsible person for several EC project aiming at speech database collection realized in cooperation with leading European industrial partners (SpeechDat, SPEECON, LC-StarII, and others). He is responsible person for the grant GAČR 102/08/007 "Speech Recognition under Real-World Conditions" and he leads the "Signal Processing team" in Research activity MSM 6840770014 "Perspective Informative and Communications Technicalities Research".