Improving of LVCSR for Causal Czech Using Publicly Available Language Resources

Petr Mizera(✉) and Petr Pollak

Faculty of Electrical Engineering, Czech Technical University in Prague,
K13131, Technicka 2, 166 27 Praha 6, Czech Republic
{mizera,pollak}@fel.cvut.cz
http://www.fel.cvut.cz
http://noel.feld.cvut.cz/speechlab

Abstract. The paper presents the design of Czech casual speech recognition which is a part of the wider research focused on understanding very informal speaking styles. The study was carried out using the NCCCz corpus and the contributions of optimized acoustic and language models as well as pronunciation lexicon optimization were analyzed. Special attention was paid to the impact of publicly available corpora suitable for language model (LM) creation. Our final DNN-HMM system achieved in the task of casual speech recognition WER of 30–60% depending on LM used. The results of recognition for other speaking styles are presented as well for the comparison purposes. The system was built using KALDI toolkit and created recipes are available for the research community.

Keywords: Speech recognition · LVCSR · Spontaneous speech · Casual speech · Czech · NCCCz · KALDI

1 Introduction

Nowadays, Large Vocabulary Continuous Speech Recognition (LVCSR) is very well developed for all major world languages as well as for the majority of other languages, spoken mostly by smaller amount of native speakers. The main research in the field of automatic speech recognition (ASR) is focused on the development of systems for low resources languages and improvements to the existing systems deployed under adverse conditions [1,2]. The recognition of spontaneous speech is a typical example of an ASR system intended for real-life environments. It represents a very challenging task, mainly because the accuracy of spontaneous speech recognition is still rather low in comparison with generally high accuracy of standard LVCSR systems [3–7]. Spontaneous or colloquial speech recognition deals with problems similar across all languages, the most typical ones being: strong variability in the pronunciation (mainly strong pronunciation reduction), changes in word morphology, free word order in the sentence, sentence breaks, etc. [8,9].

Many authors have presented solutions for the above mentioned tasks and achieved results different for various languages, speaking styles, or recording

A. Karpov et al. (Eds.): SPECOM 2017, LNAI 10458, pp. 427–437, 2017.
DOI: 10.1007/978-3-319-66429-3_42

conditions, e.g. the authors in [10] worked with transcriptions of oral interviews of survivors and witnesses of the Holocaust and they reported 39.60% Word Error Rate (WER) for English and 39.40% for Czech. However, when the level of speech spontaneity is higher, typically for very informal speaking style, the accuracy of speech recognition falls. Authors in [3] worked with the recordings of telephone conversations and reported 48% WER for the Czech language. Similarly in [6], authors presented results around 31–56% WER for the case of very informal speech recognition task.

The purpose of this paper is to present the results of very informal speech recognition performed on Nijmegen Corpus of Casual Czech (NCCCz) using the current state-of-the-art setup of LVCSR and publicly available language resources. It is a part of the research focused on understanding very informal speaking styles. The paper is organized as follows. In Sect. 2, we summarize the current state-of-the art of Czech LVCSR and we describe our approach applied to Czech casual speech recognition. Section 3 describes the setup of our experiments realized on casual speech data from NCCCz. In Sect. 4, the results of particular experiments are discussed in the context of other results obtained also for other speaking styles. The paper is concluded with the summary of achieved results and the information about the availability of used tools and recipes is presented.

2 Casual Speech Recognition for Czech

Due to the intensive studies of several research groups in the Czech Republic during last decades, available LVCSR systems for Czech language reach results similar to other languages spoken by a significantly higher number of native speakers. Concerning spontaneous speech recognition, several systems were presented e.g. in [3,6,9] or [7] and achieved accuracy is significantly lower then for LVCSR working under standard conditions.

Casual speech is defined as a way of talking used within a conversation among close people. Our investigation of casual speech recognition is done for the Czech language and it is based on exploiting the data from the Nijmegen Corpus of Casual Czech (NCCCz) [11] which consists of 30 h of spontaneous conversations of 60 speakers (30 males and 30 females recorded always in groups of 3 speakers of the same gender). The amount of available data is huge, since every group of three speakers was conversing for approximately 90 min. Also the recording procedure of NCCCz (the same one as used for the collection of similar Dutch, French, or Spanish corpora [12]) and the first analysis presented in [11] guarantee that NCCCZ contains highly casual speech. A lot of pronunciation reduction, extremely fast speed of talking, free grammar, word cutting, sentence restarts, etc. can be observed within speech data in NCCCz. Consequently, the recognition results using LVCSR with standard setup failed significantly [13].

2.1 LVCSR Architecture

The solution for improving the accuracy of casual speech recognition for Czech language is based on LVCSR based on Hidden Markov Models (HMM) and

Deep Neural Networks (DNN). Especially, the DNN-HMM based approach has been recently shown to increase the performance of LVCSR systems significantly [14,15]. Encouraged by these results, we compared both the conventional GMM-HMM system and the modern DNN-HMM hybrid approach. For both architectures, we used a rather standard training procedure without any special modifications related to the spontaneous speech because the conversations available in NCCCz were recorded in the quiet environment.

2.2 Front-End Processing

As basic features, the Mel-frequency cepstral coefficients (MFCC) with the standard setup are used in our system. Standardly, pre-emphasis with the coefficient of 0.97 is applied, short-time frame has the length of 25 ms and is moved with the step of 10 ms. Mel-filter bank contains 30 bands in the frequency range 100–7940 Hz and 12 cepstral coefficients with additional $c[0]$ are computed. Cepstral mean normalization (CMN) over the speaker is applied and these features with delta and delta-delta parameters are used for the creation of initial acoustic models.

In the next steps only static and normalized MFCC features are extended with the both-side context of 5 frames to a higher dimension vector which is then reduced and decorrelated by LDA+MLLT transforms. This target feature vector of the size 40 is further speaker-adapted using feature-space Maximum Likelihood Linear Regression (fMLLR) and these features are used in designed LVCSR with GMM-HMM architecture as standard setup used nowadays in modern advanced LVCSR systems.

System based on DNN-HMM hybrid architecture works with above described features used for GMM-HMM system but for the purpose of an application at the input of DNN mean and variance normalization (MVN) is applied and further both-side context of 5 frames is used again. We obtain 440 dimension vector which is directly applied to the input of DNN. Illustrative block scheme of feature extraction procedure is in Fig. 1.

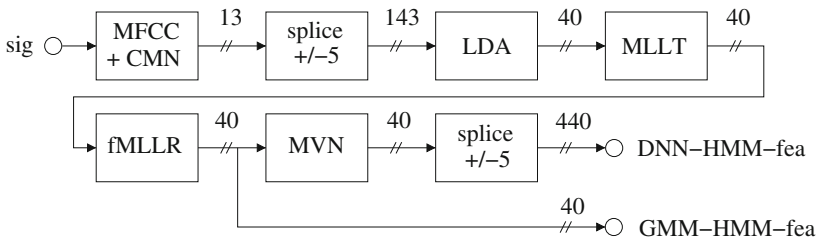

Fig. 1. Feature extraction used in GMM-HMM and DNN-HMM LVCSR

2.3 Acoustic Modeling

Acoustic models are also built using standardly used approach in modern ASR systems. The set of 45 Czech phones expanded to the context-dependent crossword triphones is used. Concerning *GMM-HMM approach*, the basic conventional GMM-HMM system is created first using the above mentioned LDA+MLLT features. It is followed by feature-space maximum likelihood linear regression (fMLLR) per each speaker (speaker adaptive training - SAT). Next iteration is based on UBM (Universal Background Model) in the combination with SGMM (Subspace GMM) and the system is finally retrained discriminatively using bMMI.

Concerning *DNN-HMM hybrid approach*, DNN topology consisted of input layer with 440 units (the context of 5 frames with 40 dimensional fMLLR features with MVN) followed by 6 hidden layers with 2048 neurons per layer and the sigmoid activation function. The process of building of DNN-HMM system started with the initialization of hidden layers of used network by Restricted Boltzmann Machines (RBMs) and then the output layer was added. The process continued by the frame cross-entropy training and ending with sMBR sequence-discriminative training.

Both AMs were trained using the utterances from SPEECON database [16] and CZKCC (private database of car speech).

2.4 Language Models for Casual Czech

Concerning language modeling, we work with standard n-gram-based statistical language models (LMs). According to the preliminary assumptions and also on the basis of experimental evaluations, Witten-Bell discounting was used for the smoothing of created LMs. This procedure is rather standard, the significant problem which had to be solved was in the choice of suitable resource for coverage casual nature of speech which should have been recognized.

We have analyzed the suitability of five general LMs collected from three different publicly available resources, i.e. from Czech National Corpus (CNC) [17], Google n-grams distributed by Linguistic Data Consortium (WEB1T) [18], and from the corpora ORAL 2006, ORAL 2008, and ORAL 2013. While the corpora CNC same as WEB1T contain general text, corpora of ORAL family contain spontaneous conversations and they were produced also by the Institute of Czech National Corpus [19].

General LMs from CNC and WEB1T corpora were built for 340 k word forms and the steps of their creation were described in [18]. These models should cover general nature of Czech language. Similarly, we built ORAL n-gram LMs which should cover spontaneous nature of recognized conversations. The number of word forms obtained from spontaneous conversations was 162 k for ORAL corpus and 29 k for NCCCz which maent a contribution of 73 k specific words from ORAL and 9 k words from NCCCz approximately.

Finally, to cover the maximum vocabulary for our task we have also created LMs from NCCCz, first, from defined training part of NCCCz containing the transcription of 60% utterances per each recorded session which were not used

the for the evaluations later. It represents slightly more realistic scenario as the content of recognized utterances has not been seen before. Second, we created also for comparison purposes optimum LM for causal speech from all available NCCCz transcriptions.

2.5 Modeling of Pronunciation Variation

Finally, the modeling of pronunciation variation in casual speech (mainly its reduction) was taken into account. We apply particular rules, some of them known from other works, e.g. [9] or [20], others obtained on the basis of results of realized psycholinguistic study of pronunciation reduction in NCCCz [21]. In the end, we have used approximately 6700 additional pronunciation variants. The illustrative examples of several rules are

"v[sSzZ]→[sSzZ]" - e.g. "*vždyt', vstát*" ("*but, to stand up*"),
"[td]J → [cJ\J]" - e.g. "*letní*" ("adj. *summer*"),
"cons_1-t-cons_2 → cons_1-cons_2" - e.g. "*jestli*" ("*if*"),
"js → s" - e.g. "*jsem*" ("*I am*"),
"j[eai] → [eai]" - e.g. "*jestli, jinam*" ("*if, elsewhere*"),
"zj → z" - e.g. "*zjistíš*" ("*You will find*"),
"t-S → t_S" - e.g. "*většina*" ("*majority*"),
"nsk → nt_sk" - e.g. "*č]ínský*" ("*Chinesse*"),
"vZd → vd" - e.g. "*vždycky*" ("*always*").

3 Experimental Part

Within the experimental part of this study, the behavior of designed systems on the principal task of spontaneous and casual speech recognition was analyzed. For the comparison purposes the results obtained for standard read speech recognition are also presented in the paper.

3.1 Used Speech Corpora

Experiments were performed on utterances from the following Czech databases: SPEECON (Czech database from SPEECON family which contains mainly standard read speech), CtuTest (private database of read journal sentences of various topics), CzLecDSP (recording of technical lectures from the field of DSP, these data have spontaneous nature but they are more formal [22]), and finally with NCCCz with strongly informal (casual) utterances. For the training of AMs also CZKCCC database was used (car-speech data). The following particular setups were used:

- *TA1 - read speech recognition*
  (a) read sentences, phonetically rich (SPEECON database),
  (b) journal sentences (CtuTest database),

– *TA2 - spontaneous speech recognition* recordings of lectures (CzLecDSP database),
– *TA3 - casual speech recognition*
recordings of highly informal conversations (NCCCz database).

Signals from all used databases were available at 16 kHz sampling frequency in 16-bit PCM format and final amounts of data in particular train and test subsets are summarized in Table 1.

Table 1. Evaluation data subsets for training and testing

Training subsets				Testing subsets			
Database	Speakers	Utterances	Hours	Database	Speakers	Utterances	Hours
SPEECON	225	60877	53.6	SPEECON	24	699	1.1
CZKCC	302	12771	20.6	CtuTest	40	577	1.1
NCCCz	40	10975	21.0	CzLecDSP	8	1417	1.7
Total	567	84623	95.2	NCCCz	20	890	1.1

3.2 Used Tools

Designed LVCSR systems were built using KALDI toolkit [23], while SRILM toolkit was used for the creation of particular LMs. The process of feature extraction was performed by our internal tool *CtuCopy* [24]. In comparison to *compute-mfcc-feats* available in KALDI, *CtuCopy* enables to extract features as MFCC, PLP, DCT-TRAP, and also to apply frequency-domain noise reduction, various cepstral normalizations. Recently, the compatibility with KALDI tools has been also added [25]. All recipes created for described experiments with Czech casual speech recognition using KALDI toolkit are publicly available under APACHE 2.0 license in "Download" section at "http://noel.feld.cvut.cz/speechlab".

3.3 Results and Discussion

The achieved results for previously established recognition tasks are presented from the following points of view: the *optimization of acoustic modeling*, the impact of *language modeling* and *pronunciation variation*.
I. The impact of AM
The first results describe the quality of used AM, i.e. from basic GMM-HMM approach to the best AM based on DNN-HMM architecture. General bigram LM from CNC with 340 k words was used for all these experiments with results in Table 2. Particular acronyms represent the following systems:

- "tri2" - triphone GMM-HMM with LDA+MLLT features,
- "tri3" - triphone GMM-HMM with LDA+MLLT followed by SAT,
- "SGMM" - subspace GMM,

- "bMMI" - discriminatively trained models,
- "DNN" - DNN-HMM system.

Achieved results show that our target DNN-HMM LVCSR system works with the accuracy comparable to the current state-of-the art, i.e. 15.2% of WER for standard read speech. For spontaneous speech we received WER of 37.4% for the transcription of lectures (i.e. with slightly more formal speaking style) and 72.0% for very informal speech from NCCCz.

Table 2. WERs of LVCSR in the phase of AM optimization

Tasks	tri2	tri3	SGMM	bMMI	DNN
TA1a	29.8	23.4	22.2	21.8	21.1
TA1b	24.0	17.0	15.9	15.3	15.2
TA2	49.9	41.3	39.9	38.0	37.4
TA3	82.5	76.1	74.9	74.2	72.0

II. The impact of LM

The results shown in Table 3 present the influence on used various LMs in analyzed tasks. The first part summarizes the results of recognition for all speaking styles using general CNC and WEB1T LMs where the strong decrease for the case of casual speech is clearly shown. The second part of Table 3 presents the results for TA3 task (casual speech) and LMs from ORAL a NCCCz (transcribed spontaneous speech corpora). The reduction of out-of-vocabulary (OOV) confirmed better modeling of casual speech and led to results around 60–70% WER. An exceptional case is LM *NCCCzAll* created from all available data including the test set so that it had OOV of 0%. We present these results for this non-realistic situation as a limit case which can be achieved using ideal setup.

The next experiments were focused on minimization OOV and WER in the TA3 task by merging of various bigram LMs. The results for merged LMs with uniform interpolation weight are summarized in Table 4. Using various merged LMs reduced the level of OOV significantly but the WER decreased just a little as the setup of the interpolation weights (λ) was not optimal. Therefore, we optimized the value of λ for particular LMs and the best result was obtained with the following setup: 0.2 weight for ORAL LM, 0.15 for CNC 0.15 for WEB1T and 0.5 for NCCCz achieving WER about 59.7%. The final investigation was based on merging various LMs with the NCCCz-based LM. The contributions of various interpolation weights λ to the final WER are summarized in Table 5. The best results were achieved for the setup with $\lambda = 0.75$.

In the end, the combination of all LMs brought an improvement in target OOV but the decrease of WER was smaller. The results proved that general LMs (CNC and WEB1T) did not contain proper information describing the causal speech in NCCCz, however, LMs ORAL corpus covered casual speech

Table 3. WERs of LVCSR with various 2-gram a 3-gram LMs on particular tasks

Tasks	LM	OOV	PPL	2-gram	3-gram
TA1a	CNC	1.6	3572	21.1	21.8
TA1b	CNC	1.8	2034	15.2	14.7
TA2	CNC	4.8	2937	37.4	37.2
TA3	CNC	4.6	2065	72.0	72.2
	WEB1T	4.5	4427	68.9	-
TA3	ORAL06	6.5	389	67.1	66.4
	ORAL08	6.7	445	66.8	66.3
	ORAL13	4.7	475	66.1	65.4
	ORALall	4.0	426	63.6	62.5
	NCCCz60	7.2	248	61.4	61.2
	NCCCzAll	*0*	*69*	*41.3*	*28.4*

Table 4. DNN-HMM casual speech recognition (TA3) with merged bigram LMs

Bigram LMs	OOV	WER
CNC+WEB1T	4.3	69.8
CNC+WEB1T+ORALall	2.8	64.7
CNC+WEB1T+ORALall+NCCCz60	1.5	61.2

Table 5. DNN-HMM with various weights of NCCCz in merged LMs on TA3 task

LMs	OOV	NCCCz weight λ				
		0.0	0.25	0.50	0.75	1
CNK340+NCCCz60	2.2	72.0	62.8	60.8	59.4	61.4
ORALall+NCCCz60	2.5	63.6	60.9	59.8	58.9	61.4
WEB1T+NCCCz60	2.1	68.9	62.3	60.6	60.0	61.4

very similarly as LM created directly from NCCCz, of course, except the case when used LM NCCCzAll was created also using the test data.

III. The impact of pronunciation reduction

The final results describe the achieved WER for three approaches of pronunciation modeling in casual speech. Firstly, we used automatically generated pronunciation for all words in analyzed LMs (which is used always for any new word not present in a available dictionary). Secondly, we used approved canonic pronunciation of all words from NCCCz which was created by manual checks by two independent experts. Thirdly, the dictionary with the additional pronunciation variants with phone reductions on the basis of rules described in Sect. 2 was used and obtained results are in Table 6. According to the assumptions, the

Table 6. Impact of pronunciation variation in DNN-HMM system

LM	Lexicon	WER
0.25 ORALall + 0.75 NCCCz60	Automatic	59.8
	Canonic checked	58.9
	Reduction variants	58.4

recognition accuracy has improved for the most proper case taking into account the reduced pronunciation, however, its decrease was only about 1.4%

4 Conclusions

In this paper, we described an optimization of DNN-HMM based LVCSR for casual speech recognition in Czech and its performance on speech from the Nijmegen Corpus of Casual Speech (NCCCz). Achieved results proved possible usage of these systems for casual speech recognition, however, the results are significantly worse than for the recognition of more formal speech. It was also demonstrated that publicly available corpora ORAL with transcriptions of spontaneous conversations commonly with available corpora of formal Czech can be used for the creation of basic LMs for the task of casual speech recognition.

Concerning obtained results, the best setup of DNN-HMM system with merged language model and pronunciation variation modeling achieved 58.4% WER, which is comparable to results of other authors. The built system was also tested on other spontaneous data (lecture recordings which were slightly more formal) where it achieved better WER of 37.2%, similarly in the task of formal read speech recognition where WER of 14.7% was achieved. The observed margin between the casual and formal speech recognition illustrated the challenge for the research in the field of more informal speech recognition.

Finally, created KALDI recipes for the recognition of Czech casual speech from NCCCz are publicly available in the Download section at the WEB-page "http://noel.feld.cvut.cz/speechlab". These scripts can be easily modified especially for the data from the family of Nijmegen casual speech corpora and SPEECON databases for other languages.

Acknowledgments. The research described in this paper was supported by internal CTU grant SGS17/183/OHK3/3T/13 "Special Applications of Signal Processing".

References

1. Cui, J., Ramabhadran, B., Cui, X., Rosenberg, A., Kingsbury, B., Sethy, A.: Recent improvements in neural network acoustic modeling for LVCSR in low resource languages. In: Proceedings of Interspeech 2014: 15th Annual Conference of the International Speech Communication Association, Singapore (2014)

2. Seltzer, L.M., Dong, Y., Yongqiang, W.: An investigation of deep neural networks for noise robust speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, Canada (2013)
3. Korvas, M., Plátek, O., Dušek, O., Žilka, L., Jurčíček, F.: Free English and Czech telephone speech corpus shared under the CC-BY-SA 3.0 license. In: Proceedings of LREC 2014: 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland, pp. 365–370 (2014)
4. Barras, C., Lamel, L., Gauvain, J.L.: Automatic transcription of compressed broadcast audio. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, USA, pp. 265–268 (2001)
5. Nouza, J., Ždánský, J., Červa, P.: System for automatic collection, annotation and indexing of Czech broadcast speech with full-text search. In: Proceedings of 15th IEEE MELECON Conference, La Valleta, Malta, pp. 202–205 (2010)
6. Nouza, J., Blavka, K., Bohac, M., Cerva, P., Málek, J.: System for producing subtitles to internet audio-visual documents. In: 38th International Conference on Telecommunications and Signal Processing, TSP 2015, Prague, Czech Republic, pp. 1–5, 9–11 July 2015
7. Psutka, J., Psutka, J., Ircing, P., Hoidekr, J.: Recognition of spontaneously pronounced TV ice-hockey commentary. In: Proceedings of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, pp. 83–86 (2003)
8. Lehr, M., Gorman, K., Shafran, I.: Discriminative pronunciation modeling for dialectal speech recognition. In: Proceedings of Interspeech 2014, Singapore, pp. 1458–1462 (2014)
9. Nouza, J., Silovský, J.: Adpating lexical and language models for transcription of highly spontaneous spoken Czech. In: Proceedings of Text, Speech, and Dialogue, LNAI, vol. 6231, Brno, Czech Republic, pp. 377–384 (2010)
10. Byrne, W., et al.: Automatic recognition of spontaneous speech for access to multilingual oral history archives. IEEE Trans. Speech Audio Process. **12**(4), 420–435 (2004)
11. Ernestus, M., Kočková-Amortová, L., Pollák, P.: The Nijmegen corpus of casual Czech. In: Proceedings of LREC 2014: 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland, pp. 365–370 (2014)
12. Torreira, F., Adda-Decker, M., Ernestus, M.: The Nijmegen corpus of casual French. Speech Commun. **52**, 201–221 (2010)
13. Prochazka, V., Pollak, P.: Conversational speech from Nijmegen corpus of casual Czech by general ASR language models. In: Production and Comprehension of Conversational Speech, pp. 34–35 (2011)
14. Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Sig. Process. Mag. **29**(6), 82–97 (2012)
15. Vesely, K., Karafiat, M., Grezl, F.: Convolutive bottleneck network features for lVCSR. In: 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), December 2011
16. Pollak, P., Cernocky, J.: Czech SPEECON adult database. Technical report (2004)
17. Institute of the Czech National Corpus: SYN2006PUB corpus (2006). http://ucnk.ff.cuni.cz/english/syn2006pub.php
18. Prochazka, V., Pollak, P., Zdansky, J., Nouza, J.: Performance of Czech speech recognition with language models created from public resources. Radioengineering **20**, 1002–1008 (2011)

19. Institute of the Czech National Corpus: Corpus oral 2006 and oral 2008 and oral 2013, Institute of the Czech National Corpus FF UK. http://www.korpus.cz
20. Schuppler, B., Adda-Decker, M., Morales-Cordovilla, J.A.: Pronunciation variation in read and conversational Austrian German. In: Proceedings of Interspeech 2014, Singapore (2014)
21. Kolman, A., Pollak, P.: Speech reduction in Czech. In: Proceedings of LabPhone 14, The 14th Conference on Laboratory Phonology, Tokyo, Japan (2014)
22. Rajnoha, J., Pollák, P.: Czech spontaneous speech collection and annotation: the database of technical lectures. In: Esposito, A., Vích, R. (eds.) Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions. LNCS, vol. 5641, pp. 377–385. Springer, Heidelberg (2009). doi:10.1007/978-3-642-03320-9_35
23. Povey, D., et al.: The Kaldi speech recognition toolkit. In: Proceedings of ASRU 2011, IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (2011)
24. Fousek, P., Pollak, P.: Efficient and reliable measurement and simulation of noisy speech background. In: Proceedings of EUROSPEECH 2003, 8-th European Conference on Speech Communication and Technology, Geneve, Switzerland (2003)
25. Borsky, M., Mizera, P., Pollak, P.: Noise and channel normalized cepstral features for far-speech recognition. In: Proceedings of SPECOM 2013, The 15th International Conference on Speech and Computer, Pilsen, Czech Republic (2013)