Small and Large Vocabulary Speech Recognition of MP3 Data under Real-Word Conditions: Experimental Study

Petr Pollak and Michal Borsky*

Faculty of Electrical Engineering, Czech Technical University in Prague,
Technicka 2, 166 27 Prague, Czech Republic
{pollak,borskmic}@fel.cvut.cz

Abstract. This paper presents the study of speech recognition accuracy both for small and large vocabulary task with respect to different levels of MP3 compression of processed data. The motivation behind the work was to evaluate the usage of ASR system for off-line automatic transcription of recordings collected from standard present MP3 devices under different levels of background noise and channel distortion. Although MP3 may not be an optimal compression algorithm, the performed experiments have prooved that it does not distort speech signal significantly for higher compression rates. Realized experiments showed also that the accuracy of speech recognition (both small- and large-vocabulary) decreased very slowly for the bit-rate of 24 kbps and higher. However, slightly different setup of speech feature computation is necessary for MP3 speech data, mainly PLP features give significantly better results in comparison to MFCC.

Keywords: Speech recognition, Small vocabulary, Large vocabulary, LVCSR, MPEG compression, MP3, Noise robustness.

1 Introduction

Automated speech recognition (ASR) represents a field which is nowadays more present in everyday human life in growing number of applications as in voice operated control of consumer devices, automated information services, or general conversion of uttered speech into text record. The systems for automatic transcription of speech to text are currently well developed for all important world languages. It is possible to meet today dictation software for standard PC enabling users to input texts into documents without using the keyboard, e.g. Dragon dictate, or for mobile devices, e.g. [1]. Further, the transcription of broadcast news is currently a very important task solved by many research teams [2], [3]. Probably the most popular is the transcription of news, but there are also other applications such as automated subtitling of TV programmes, e.g. parliament meetings [4] or sportscasts [5]. Special attention is also paid to the transcription and indexing of large audio archives enabling better search within them in the future [6], [7].

When audio records are transcribed on-line, e.g. the above-mentioned subtitling of TV programmes, ASR systems work with full quality input signal. On the other hand,

* This research was supported by grant GAČR 102/08/0707 "Speech Recognition under Real-World Conditions".

M.S. Obaidat, J.L. Sevillano, and J. Filipe (Eds.): ICETE 2011, CCIS 314, pp. 409–419, 2012.

when they work off-line, recordings can be saved in formats of different quality and typically, MP3 format (more precisely MPEG Layer III) represents one of the most frequently used formats for the saving of sound files in compressed form [8], [9]. It is well known that this format uses psychoacoustic models reducing the precision of components less audible to human hearing so it makes it possible to decrease the size of the sound file to 10% while CD quality is preserved. Although this format has been developed especially for saving the music, it is standardly used also in simple dictation devices or mobile phones enabling recording and saving audio speech files. Some works in MP3 speech data recognition have been already done. The recognition of spontaneous speech from large oral history archives published in [7] used signals saved in MP3 format but rather high bit-rate (128 kbps) was used in this case. In [10] the study of automatic transcription of compressed broadcast audio with different bit-rates was done. The comparison of various compression schemes was realized in this study, however, the quality of the signal was rather better.

This paper is an extension of experimental study published at the conference SIGMAP 2012 [11]. Commonly with previously presented results for small-vocabulary task, accuracy analysis of basic Large Vocabulary Continuous Speech Recognition (LVCSR) with respect to different quality of compressed data is additionally presented within this paper. As current ASR systems need to work accurately under real conditions, often under presence of additive background noise or channel distortion, the analysis is performed on signals from different channels with different signal quality depending mainly on the position of the microphone used. Standard features used most commonly in ASR systems have typically modified versions increasing their robustness in real environment [12], [13], [14]. But these methods are designed usually for uncompressed data, so our study focuses mainly on the analysis of the information loss in compressed speech signals when varying levels of additive noise and channel distortion appear in speech signal. The results of this study are supposed to be helpful for further application of automated speech recognition from MP3 speech files.

2 MP3 Speech Recognition

Within this study, the behavior of both small and large vocabulary recognition tasks are analyzed. When we want to analyze mainly the sensitivity of ASR system to loss of information after MP3 compression and without a further dependency on complex blocks such as statistic language model, it is convenient to use much simpler small vocabulary digit recognizer for this purpose in the first step. Finally, complex LVCSR system is used for selected cases as a confirmation of previously obtained results within simpler recognition task.

2.1 Speech Compression Scheme

MP3 compression was developed for the compression of music audio files [9], [8] and it is known that it gives slightly worse results for speech signal. The masking and attenuation of some frequency components can yield to a suppression of a phone at the beginning or at the end of the word, sometimes inter-word pause shortening can appear.

Less naturalness of decoded utterance is then the main effect of this fact and consequently, the accuracy of speech recognition can decrease too. Algorithms, which have been designed and optimized for speech signal, are represented mainly by G.729 [15], AMR [16], or Speex [17]. These encoders are based typically on CELP algorithm, but they are used rather in telecommunications and they do not appear so frequently in standard audio devices.

Consequently, although speech signals can be compressed in a better way, our attention was paid just to MP3 compression in this study because the long-term goal of our work was mainly in off-line mode of ASR operation on compressed speech data from wide-spread audio consumer devices. The study was realized with signals from the database SPEECON recorded in real environment with full-precision PCM coding. The MP3 compression was then simulated by successive encoding and decoding of signals from SPEECON database using publicly available software LAME [18] which made it possible to simulate also different levels of MP3 compression bit-rate.

2.2 Speech Recognition Setup

Current ASR systems are usually based on Hidden Markov Models (HMM). HMM based recognizer consists typically of 3 principal function modules: feature extraction (parameterization), acoustic modelling, and decoding (see block scheme in Fig 1). Generally known principles of HMM based recognizer are not explained in this paper, as they are known or can be found in many sources, e.g. in [19], [20], and others. Only a brief description of our ASR setups, which is relevant for the next parts of this paper, is presented.

Feature Vector

Concerning the parameterization, two sets of features are most standardly used in ASR systems: *Mel-Frequency Cepstral Coefficients* (MFCC) and *Perceptually based Linear Predictive cepstral coefficients* (PLP). All our experiments were carried out just with these two feature sets. Both of them use their own non-linear filter-bank which in both cases models the perceptual properties of human auditory system, for more detail see [19], [20], [21], or [22]. Finally, feature setup can be summarized in the following points:

- 12 cepstral coefficients plus logarithm of frame energy form vector of static features,
- 1st and 2nd derivatives of static parameters are added,
- settings of short-time analysis originates from typically used values, i.e. 16-32 ms for frame length and 8-16 ms for frame period.

Acoustic Modelling

Acoustic modelling in our study was slightly different for mentioned two tasks, according to requirements of small and large vocabulary system respectively.

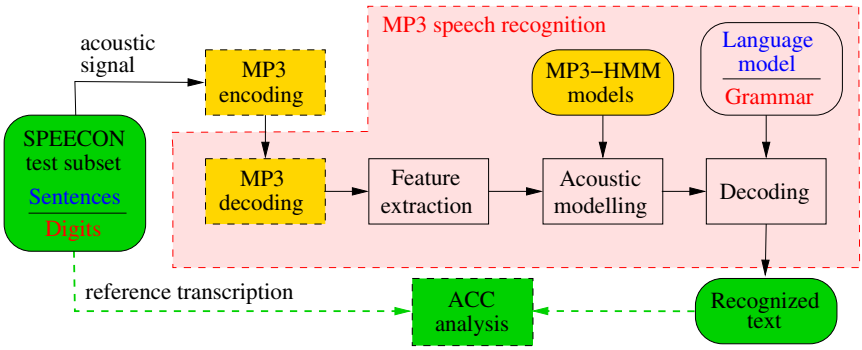

Fig. 1. Block scheme of experimental setup for HMM-based ASR testing

– *small vocabulary digit recognizer*
This system was based on simpler monophone HMM models, i.e. phones modelled without any context to neighboring phones. Finally, HMM models of 44 Czech monophones with 3 emitting states were used as the simplest sub-word acoustic models in ASR. As used phone models were context independent, their higher variability was modelled by 32 mixtures of Gaussian emitting function of HMM models and 3 streams were also used for modelling of static, dynamic and acceleration features respectively.
– *large vocabulary continuous speech recognizer*
This system was based on cross-word triphone HMM models, i.e. phones modelled with left and right context to neighboring phones. Finally, all HMM models had again 3 emitting states, triphone variability was modelled by 16 mixtures of Gaussian emitting function and all (static, dynamic and acceleration) features were processed in 1 stream.

Language Modelling
Language modelling is principally different for small and large vocabulary system, i.e.

– *grammar in small vocabulary digit recognizer*
Connected digit ASR uses simple grammar in the phase of decoding. On the other hand, though just basic digits from 0 to 9 can appear in the utterance, the number of digits in the utterance can vary and they can be pronounced with or without pauses and with possible repetitions. Finally, it means that our digit recognizer should be sufficiently general for our experiments and we can assume that it simulates well operating mode of target practical application, see Fig 1.
– *statistical language model in LVCSR*
Our LVCSR system works with the simplest statistical language models (bigram) and moderate size of vocabulary (240k words). Such simple setup enables off-line processing with tools from the HTK Toolkit [20]. Language model (LM) were created from SYN2006PUB 5-gram corpus from Czech National Corpus (CNC) [23],

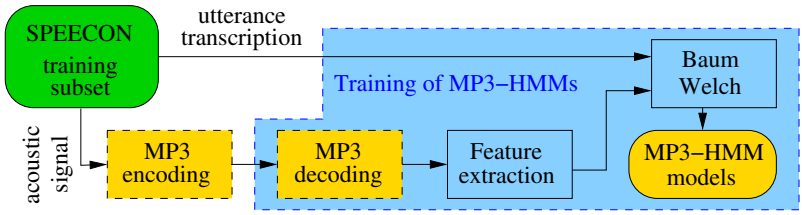

Fig. 2. Block scheme of HMM-based acoustic model training

[24]. More details about the creation of this LM and its performance in LVCSR under standard conditions (using full-precision PCM data) can be found in [25].

2.3 Training of Speech Recognizer

Training of HMM models, which are composed of Gaussian emitting functions representing probability distribution of a feature in given state and from probabilities of transitions between particular states, is performed on the basis of iterative embedded training procedure from large training data (Baum-Welch algorithm). The size of the training database containing speech signals with precisely annotated content must guarantee sufficient appearance of each acoustic element. The training procedure, which has been based on Czech SPEECON data in our case, is illustratively represented by block scheme in Fig. 2, more details can be found in [20] or [19].

Our acoustic models (monophone- and triphone-based described above in the section 2.2) were trained iteratively with flat start for each parameterization and also for many operating conditions. As training data had to match these conditions, we have finally obtained comprehensive set of HMM models for particular channels, for different bit-rates in MP3 encoding, and for selected feature vector used.

2.4 Implementation Tools

The training of acoustic HMM models, small vocabulary ASR and also LVCSR, were realized by tools from publicly available HTK Toolkit [20] which is often used worldwide for the realization of HMM based recognition. For readers without detail knowledge of the HTK Toolkit, typical and core tools of the HTK Toolkit are *HCopy* as parameterization tool, *HERest* as the tool for the training by Baum-Welch algorithm, or *HVite* as Viterbi based word recognizer for digit recognizer and finally *HDecode* tool for cross-word triphone-based LVCSR decoding.

The computation of PLP cepstral coefficients was performed by *CtuCopy* tool [12] and [26], providing some extensions to *HCopy* from the standard set of HTK Toolkit.

3 Experiments

Experiments described in this part comprise the core contribution of this study, which is mainly in the analysis of ASR performance for MP3 compressed speech data under different real-word conditions.

3.1 Speech Data Description

All experiments were carried out with signals from Adult Czech SPEECON database [27]. It is the database (DB) of sentences, digits, command, names, etc. recorded by 580 speakers under different conditions, i.e. in offices or home environment, at public places, or in the car. For this study, only well-balanced subset of adult speaker data from office environment were used, i.e. 90% of data for training and 10% for testing (digits or sentences). It contains signals with similar and not so strong background noise.

Speech data in SPEECON DB are raw 16 bit linear PCM sound files sampled by 16 kHz sampling frequency. These signals were then encoded into MP3 sound files with different bit-rates. The MP3 compression was simulated by successive encoding and decoding by the above-mentioned *lame* software encoder/decoder [18].

Although only data from one environment were used in our experiments, the influence of additive noise and channel distortion could be analyzed, because signals in SPEECON DB were recorded in 4 channels which differed in microphone type and its position, see [28]. Following Tab. 1 describes the properties of particular channels. Although different types and quality of microphones were used, it was mainly the distance from the speaker's mouth that played the key role in the quality of recorded speech signal. Finally, signals from close talk head-set channel CS0 are then almost without any additive noise and reasonable channel distortion appears only in signals from channels CS2 and CS3. These data can simulate well real MP3 recordings made by standard devices in various environments.

Table 1. Description of channels recorded in SPEECON database

Channel ID	*Microphone type*	*Distance*	*Additive noise*	*Channel distortion*
CS0	head-set	2 cm	-	-
CS1	hands-free	10 cm	+	-
CS2	middle-talk	0.5-1 m	++	+
CS3	far-talk	3 m	+++	++

3.2 Recognition Accuracy

The accuracy (ACC) of a recognizer was measured standardly on the basis of errors on word level. It was defined according to [20] as

$$ACC = \frac{N - D - S - I}{N} \cdot 100 \quad [\%]\,, \tag{1}$$

where N is the total number of words in the testing set while D, S, and I are numbers of word deletions, substitutions, and insertions respectively. The following sections describe in details obtained results.

3.3 Analysis of Optimum Segmentation for MP3 Speech Data

Within the first experiment, the influence of short-time analysis setup on target accuracy of MP3 recognition was analyzed. In accordance with phonetically based assumptions

as well as default settings used in [20], the optimum length of the frame for short-time acoustic analysis is 25 ms with the period of 10 ms for uncompressed speech data, while for MP3 compressed data the segmentation with frame length 32 ms and frame period 16 ms gives the best results for both studied feature sets.

The reasons for this effect lie in the first modules of both feature extraction algorithms which realize short-time Fourier analysis followed by non-linear filter banks computing perceptually based power spectra. Due to the decrease of short-time frame length, frequency resolution of Discrete Fourier Transform decreases too and consequently the masking and deletions of some frequency components within the MP3 compression scheme increase the estimation error of power spectrum at the output of the filter bank.

The results of this experiment for both MFCC and PLP features are in Tab 2. MP3 compression was realized with bit-rate of 160 kbps and results are presented for the CS0 channel. It can be supposed that this error at the output of filter bank increases for shorter frame length also when uncompressed speech signal is more corrupted by additive noise, which is the case of channels CS1, CS2, and CS3.

Abbreviations used in the following tables describe the feature extraction used, e.g. MFCC_3216 means MFCC features computed from 32 ms long frame with the period of 16 ms.

Table 2. ASR accuracy (ACC) dependence on varying segmentation for WAV or MP3 signals: (a) MFCC features, (b) PLP features

(a)

Features	*WAV*	*MP3*
MFCC_1608	95.55	54.39
MFCC_2510	96.89	76.31
MFCC_3216	95.22	93.21

(b)

Features	*WAV*	*MP3*
PLP_1608	95.11	72.64
PLP_2510	96.33	81.76
PLP_3216	95.11	93.10

Table 3. Digit ASR accuracy for varying MP3 bit-rate: (a) MFCC features, (b) PLP features

(a)

MP3 bit-rate	CS0	CS1	CS2	CS3
WAV	95.22	92.44	89.54	61.18
160 kbps	93.21	83.31	42.83	30.03
64 kbps	93.21	84.43	43.60	31.59
48 kbps	93.33	85.54	43.83	32.26
40 kbps	93.33	86.43	43.94	31.59
32 kbps	89.77	88.54	44.75	33.48
24 kbps	89.32	37.71	38.71	27.92
8 kbps	21.02	16.35	12.57	6.90

(b)

MP3 bit-rate	CS0	CS1	CS2	CS3
WAV	95.11	89.43	88.88	64.63
160 kbps	93.10	82.09	78.75	27.36
40 kbps	92.66	87.32	83.09	28.48
24 kbps	92.32	88.21	80.09	28.70
8 kbps	62.40	36.15	24.25	11.01

3.4 Results of MP3 Digit ASR

Within the second experiment, the influence of digit recognition accuracy in particular channels on different MP3 bit-rates was analyzed. All these experiments were realized

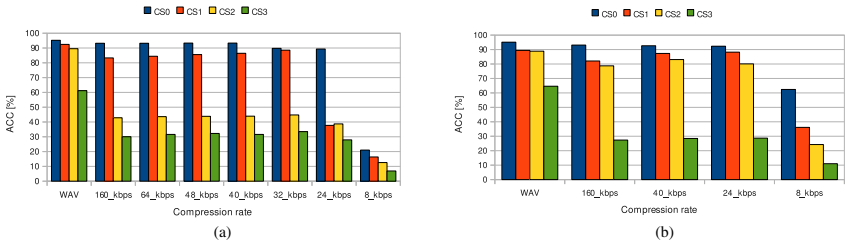

Fig. 3. ASR accuracy (ACC) dependence on varying MP3 bit-rate for all channels: (a) MFCC features, (b) PLP features

with optimum segmentation parameters, i.e. 32 ms frame length and 16 ms frame period. Achieved results are presented numerically in the following tables and for quick illustrative overview also in figures showing the same data in graphical form.

Tab. 3(a) and Fig. 3(a) present results obtained with mel-frequency cepstral coefficients. Looking at the results achieved for CS0 channel, we can see that for rather high quality signal the MP3 compression has just minimum effect for bit-rate of 24 kbps and higher. For other channels the trend is always similar but the absolute values of the achieved ACC are lower according to our assumptions. We can also see that for channels CS2 and CS3, containing higher level of background noise and stronger channel distortion, the ACC falls rapidly already for rather high bit-rates of MP3 compression. Such results disable in principle the recognition of MP3 data collected under similar conditions. On the other hand it must be mentioned that all experiments in this study were carried out with basic feature setup, i.e. no algorithm for additive noise suppression or channel normalization was used.

Tab. 3(b) and Fig. 3(b) show similar results for the recognition with perceptually based linear predictive cepstral coefficients. The same trends have been observed again, so in the end the experiments were realized just with 4 different bit-rates. In comparison with MFCC, better performance can be observed for PLP for channels CS1 and CS2. Especially the results for channel CS2 represent acceptable values of ACC for MP3 compressed data for the bit-rate of 24 kbps and higher (80.09% as for MFCC it was 38.71%) which is similar to the high quality CS0 channel. In principle it allows the practical usage of ASR of MP3 compressed data collected by middle-distance microphone, e.g. it can be the case of MP3 recorder placed on the table.

Finally, we computed sizes of compressed data so we could compare the level of compression (in percent) with the achieved accuracy of speech recognition. These results are shown in Fig. 4 where we can observe minimum decrease of ASR accuracy for 20% compression of sound file and higher. Strong downfall appears as far as beyond 10% compression. These results were obtained for MFCC features and the high quality CS0 channel.

3.5 Results of MP3 LVCSR

Finally, the experiments with LVCSR were performed. As it was mentioned above, these experiments were realized using the tool HDecode from the HTK Toolkit. As this

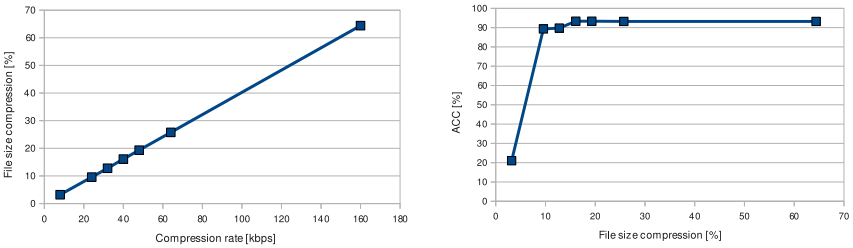

Fig. 4. Dependence of file size reduction and ASR accuracy on varying MP3 bit-rate (MFCC features and channel CS0)

tool is not working in the real-time, we used mentioned very basic setup of LVCSR (bigram LM and dictionary of 240 kwords). Nevertheless, the real-time factor for this setup ranged from 5 to 10 for signals with higher quality. When we started working with MP3 data with high compression level (8kbps) the accuracy began falling down rapidly and the real-time factor increased strongly due to bad acoustic models.

Table 4. LVCSR accuracy for varying MP3 bit-rate for MFCC (a) and PLP (b) features

(a)

MP3 bit-rate	CS0	CS1
WAV, see [25]	*67.82*	-
160 kbps	61.87	46.63
40 kbps	59.09	46.51
24 kbps	49.57	4.67
8 kbps	1.43	1.27

(b)

MP3 bit-rate	CS0	CS1
160 kbps	65.26	55.13
40 kbps	65.10	54.60
24 kbps	60.81	48.67
8 kbps	10.91	2.69

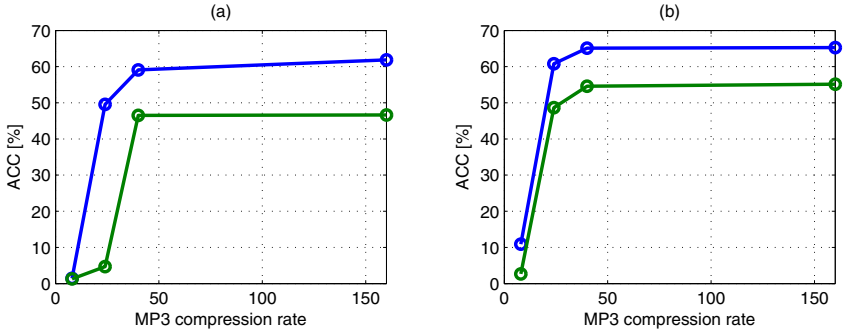

Fig. 5. LVCSR accuracy dependence on varying MP3 bit-rate for channels CS0 (blue line) and CS1 (green line): (a) MFCC features, (b) PLP features

For LVCSR tests we worked only with data from channels CS0 and CS1, i.e. channels with the most typical quality of signals used in LVCSR. Also only compression rates of 160, 40, 24, and 8 kbps were used (160 kbps can be assumed very close to full-precision data). The results were very similar to digit recognition task and they are

summarized in the table 4 and in the figure 5. Finally, the accuracy is significantly lower in comparison with digit recognition but it is given by limits of used dictionary size and bigram LM. The reference accuracy presented in [25] was computed with more precise acoustic models and for MFCC features the value was 67.82%. Very similar results, around 65%, can be achieved also with MP3 recognition using PLP features from 40 kbps compression rate.

4 Conclusions

The analysis of speech recognition using MP3 compressed audio data was done. The achieved results confirmed acceptable accuracy of speech recognition of MP3 compressed speech data when reasonable compression rate is used. The most important contributions can be summarized in the following points.

- ACC decreases rapidly for shorter frame length of short-time features when MP3 speech is recognized. It is affected by perceptual masking in MP3 compression scheme and decreasing of short-time Fourier analysis frequency resolution used in computation of MFCC and PLP features. It means that the usage of longer short-time frame for Fourier analysis is necessary.
- Generally, the loss of accuracy is very small from bit-rate of 24 kbps. It was proved both with small vocabulary and large vocabulary recognition task. The size of compressed data for 24 kbps is just 10% of full precision linear PCM and ACC decreased for digit recognition by 6% for MFCC and 3% for PLP features. For LVCSR the difference in accuracy between results for signals with very high compression rate 160 kbps and low 24 kbps was approximately 11% for MFCC and only 5% for PLP.
- The results are worse for noisy channels where 50% decrease of ACC can be observed for MFCC features, comparing the standard PCM and 24 kbps MP3 speech signal from desktop microphone for digit recognition. This decrease is just about 8% for PLP features.
- Realized experiments proved that MP3 compressed speech files used in standardly available consumer devices such as MP3 players, recorders, or mobile phones, can be used for off-line automatic conversion of speech into text without critical loss of accuracy. PLP features seem to be preferable for speech recognition in this case.

References

1. Nouza, J., Červa, P., Ždánský, J.: Very large vocabulary voice dictation for mobile devices. In: Proc. of Interspeech 2009, Brighton, UK, pp. 995–998 (2009)
2. Chen, S.S., Eide, E., Gales, M.J.F., Gopinath, R.A., Kanvesky, D., Olsen, P.: Automatic transcription of broadcast news. Speech Communication 37(1-2), 69–87 (2002)
3. Gauvain, J.-L., Lamel, L., Adda, G.: The LIMSI broadcast news transcription system. Speech Communication 37(1-2), 89–108 (2002)
4. Vaněk, J., Psutka, J.: Gender-dependent acoustic models fusion developed for automatic subtitling of parliament meetings broadcasted by the Czech TV. In: Proc. of Text, Speech and Dialog, Brno, pp. 431–438. Czech Republic (2010)

5. Psutka, J., Psutka, J., Ircing, P., Hoidekr, J.: Recognition of spontaneously pronounced TV ice-hockey commentary. In: Proc. of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, pp. 83–86 (2003)
6. Makhoul, J., Kubala, F., Leek, T., Liu, D., Nguyen, L., Schwartz, R., Srivastava, A.: Speech and language technologies for audio indexing and retrieval. Proc. of the IEEE 88(8), 1338–1353 (2000)
7. Byrne, W., Doermann, D., Franz, M., Gustman, S., Hajič, J., Oard, D., Pichney, M., Psutka, J., Ramabhadran, B., Soergel, D., Ward, T., Zhu, W.J.: Automatic recognition of spontaneous speech for access to multilingual oral history archives. IEEE Trans. on Speech and Audio Processing 12(4), 420–435 (2004)
8. Bouvigne, G.: MP3 standard. Homepage (2007), http://www.mp3-tech.org
9. Brandenburg, K., Popp, H.: An introduction to MPEG layer 3. EBU Technical Review (June 2000)
10. Barras, C., Lamel, L., Gauvain, J.L.: Automatic transcription of compressed broadcast audio. In: Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, USA, pp. 265–268 (2001)
11. Pollak, P., Behunek, M.: Accuracy of MP3 speech recognition under real-world conditions. Experimental study. In: Proc. of SIGMAP 2011 - International Conference on Signal Processing and Multimedia Applications, Seville, Spain, vol. 1, pp. 5–10 (July 2011)
12. Fousek, P., Pollák, P.: Additive noise and channel distortion-robust parameterization tool. performance evaluation on Aurora 2 & 3. In: Proc. of Eurospeech 2003, Geneve, Switzerland (2003)
13. Bořil, H., Fousek, P., Pollák, P.: Data-driven design of front-end filter bank for Lombard speech recognition. In: Proc. of ICSLP 2006, Pittsburgh, USA (2006)
14. Rajnoha, J., Pollák, P.: ASR systems in noisy environment: Analysis and solutions for increasing noise robustness. Radioengineering 20(1), 74–84 (2011)
15. ITU-T: International Telecommunication Union Recommendation G.729, coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction(CS-ACELP) (2007), http://www.itu.int/ITU-T
16. ETSI: Digital cellular telecommunications system (Phase 2+) (GSM). Test sequences for the Adaptive Multi-Rate (AMR) speech codec (2007), http://www.etsi.org
17. Valin, J.M.: The speex codec manual. version 1.2 beta 3 (2007), http://www.speex.org
18. Cheng, M., et. al.: LAME MP3 encoder 3.99 alpha 10 (2008), http://www.free-codecs.com
19. Huang, X., Acero, A., Hon, H.-W.: Spoken Language Processing. Prentice-Hall (2001)
20. Young, S., et al.: The HTK Book, Version 3.4.1, Cambridge (2009)
21. Psutka, J., Müller, L., Psutka, J.V.: Comparison of MFCC and PLP parameterization in the speaker independent continuous speech recognition task. In: Proc. of Eurospeech 2001, Aalborg, Denmark (2001)
22. Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech. Journal of the Acoustical Society of America 87(4), 1738–1752 (1990)
23. Institute of the Czech National Corpus: Homepage (2010) http://www.korpus.cz.
24. Institute of the Czech National Corpus: SYN2006PUB - corpus of newspapers and magazines from 1989 - 2004 (2006), http://ucnk.ff.cuni.cz/english/syn2006pub.php
25. Prochazka, V., Pollak, P., Zdansky, J., Nouza, J.: Performance of Czech speech recognition with language models created from public resources. Radioengineering 20(4), 1002–1008 (2011)
26. Fousek, P.: CtuCopy-Universal feature extractor and speech enhancer (2006), http://noel.feld.cvut.cz/speechlab
27. ELRA: Czech SPEECON database. Catalog No. S0298 (2009), http://www.elra.info
28. Pollák, P., Černocký, J.: Czech SPEECON adult database. Technical report (April 2004)