Borsky *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2015) 2015:20
DOI 10.1186/s13636-015-0064-7

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Advanced acoustic modelling techniques in MP3 speech recognition

Michal Borsky*, Petr Pollak and Petr Mizera

Abstract

The automatic recognition of MP3 compressed speech presents a challenge to the current systems due to the lossy nature of compression which causes irreversible degradation of the speech wave. This article evaluates the performance of a recognition system optimized for MP3 compressed speech with current state-of-the-art acoustic modelling techniques and one specific front-end compensation method. The article concentrates on acoustic model adaptation, discriminative training, and additional dithering as prominent means of compensating for the described distortion in the task of phoneme and large vocabulary continuous speech recognition (LVCSR). The experiments presented on the phoneme task show a dramatic increase of the recognition error for unvoiced speech units as a direct result of compression. The application of acoustic model adaptation has proved to yield the highest relative contribution while the gain of discriminative training diminished with decreasing bit-rate. The application of additional dithering yielded a consistent improvement only for the MFCC features, but the overall results were still worse than those for the PLP features.

Keywords: MP3 compression; Phoneme recognition; LVCSR; GMM-HMM; Acoustic modelling; Additional dithering; AM adaptation; Discriminative training

1 Introduction

The aim of automatic speech recognition (ASR) research is to develop an intermediary system for the purpose of human speech transcription where the construction and block architecture is often customized. The transcription of digitally stored data represents an example where the ASR systems has to be specifically tailored to perform optimally. Ordinary people, call centers, and media companies have large amounts of data stored for the purpose of further processing or simply for later accessibility. This data can consume large amounts of storage capacity. The obvious solution to this problem has been to compress these recordings using an audio coder with high compression rate. Although it may initially have been assumed that these compressed recordings would be accessed by a human listener, the introduction of automatic speech processing systems has changed this paradigm.

MPEG-2 AudioLayer III, also known as MP3, belongs to the group of perceptual audio codecs, which are based on the physiology of human hearing. Its main advantage is a relatively high compression rate while at the same time retaining good intelligibility for human listeners. This characteristic is the reason for its wide-spread use in the personal, commercial, and public sphere. On the other hand, the compression introduces severe distortions which limit its use for audio professionals or for automatic speech recognition. The application of auditory masking functions and the quantization in the compression scheme results in speech distortion. The exact nature of the distortion has been studied in [1] and [2]. The two main problems identified by the authors were the bandwidth limitation and spectral valleys.

This article investigates the performance of current state-of-the-art acoustic modelling (AM) and feature extraction techniques in the task of phoneme and large vocabulary continuous speech recognition of MP3 compressed speech. It is organized as follows: the next section gives a short overview of related works on this topic, followed by a theoretical analysis of distortion for spectral-based speech features, a description of used techniques, and a section detailing the experimental setup and achieved results. The article concludes with a discussion.

*Correspondence: borskmic@fel.cvut.cz
Department of Circuit Theory, Czech Technical University, Technická 2, 166 27, Prague 6, Czech Republic

2 MP3 speech recognition

Studies on practical usability of MP3 recordings in automatic speech recognition concluded that the system can perform without degradation for sufficiently high bit-rates. The authors consistently reported a significant drop in accuracy for bit-rates lower than 24 kbps, i.e., [3–7]. Several solutions have been proposed to improve the recognition for lower bit-rates, starting with limiting the training signal bandwidth, using perceptual linear prediction (PLP) features or adding a controlled amount of noise.

2.1 Related works

The MP3 format actively limits the spectral bandwidth of the compressed data, which can create the problem of training and testing data mismatch. To solve this problem, and to avoid compressing and decompressing the whole training subset, the authors in [3] proposed a parameterization scheme with a low-pass filter in the block of the front-end processing. A specific cutoff frequency was assigned to each bit-rate, and the AMs were trained on the filtered speech and then tested on the compressed speech. The method yielded only a marginal decrease in word error rate (*WER*) by 1–2 % in a simple digit recognition task, depending on the bit-rate. The second problem of bandwidth limitation is the loss of information carried by higher frequencies. This is expected to mainly affect the speech units without strong low-frequency harmonic structure, such as unvoiced consonants, and subsequently result in increased likelihood of false recognition of these units.

In [7], the authors studied the effect of spectral valleys on speech features and concluded that the main portion of degradation could be attributed to the spectral valleys. The valleys act as a step energy change between neighboring frames, which increases the values of Δ and Δ^2 parameters and randomly displaces their position in the feature space. The proposed solution was based on adding a controlled amount of noise to "fill in"? the valleys and to reduce the features' variance. The report demonstrated that the application of this technique could bring significant WER reduction by up to 45 % for low bit-rates and Mel-frequency cepstral coefficient (MFCC) features. In general, better results were obtained for lower bit-rates while the results for higher bit-rates were only slightly degraded due to the introduction of additional noise.

A comparative study of spectral-based features for MP3 speech recognition [6] demonstrated the advantage of using PLPs over MFCCs where the PLP features outperformed the MFCCs by 11 % absolute at a bit-rate of 24 kbps. The reported WER differences for higher bit-rates were much lower, at 4–6 %. The authors concluded that this behavior could be attributed to the application of the equal loudness curve and the psychoacoustic scaling of the analysis filter bank.

2.2 Robust front-end processing

Cepstral mean normalization (CMN) is a well-established technique for robust speech recognition. The principle is based on the assumption that if the convolutional noise in a short time-frame is stationary, then it can be subtracted from the extracted features in the logarithmic spectral domain. Although it is a fairly simple method, it has been proven to provide robustness against environmental and channel distortions and speaker variability.

Linear discriminant analysis (LDA) aims to improve separability among classes by transforming the feature vector of dimension n to the vector of dimension m. It is typically used in ASR systems to reduce the dimensionality of spliced feature vectors and to decorrelate the features. However, it has also been shown to improve the performance of standard features in the presence of noise [8].

Modifications of the signal can occasionally result in a sequence of zeros in the time domain which, if not treated properly, can cause the extraction algorithm to fail. If the standard procedure to avoid the *Inf.* values in logarithmic spectra is to add small amounts of uniformly distributed noise, then the addition of relatively strong noise has been shown to improve the recognition of spectrally distorted speech [7]. This technique is referenced as additional dithering later in the text.

2.3 MP3 acoustic modelling

Since MP3 recognition failure is primarily influenced by the changes at the feature extraction and acoustic modelling levels, additional refinements are required to compensate for the decline. This section provides an overview of methods used in a situation with adverse conditions or, in general, in order to increase the quality of AM.

The AM adaptation has been documented to perform well in situations with a training and testing data mismatch [9, 10] when the commonly used approach is to adapt the existing AM to the new conditions. The technique employed in this article was the feature maximum likelihood linear regression (fMLLR). Its main advantage is its robustness against the lack of adaptation data and the ability to estimate new model parameters even from the erroneous transcription.

The conventional system based on Gaussian mixture models can contain several hundred thousand of mixtures. A subspace Gaussian mixture model (SGMM) has been proposed as an alternative approach in which the model parameters are typically initialized from a clustered model, i.e., universal background model, and then retrained and shared among multiple models. The result is a reduction in model parameters, which allows estimation

of SGMM parameters using a smaller amount of data and is expected to better model the acoustic variabilities.

Discriminative training has become a common modelling method when the main principle is based on formulating an objective function tied to the classification and minimizing the recognition error directly instead of maximizing the observation likelihood. Some published works have reported on its usage for noisy speech recognition, e.g., [11] or [12], and concluded that its usage can increase the robustness of the system. The main drawback of discriminative training may be the lack of generalization to the test set, which is likely to occur if the uncompressed, discriminatively trained AM is deployed to recognize the compressed speech. Despite this obvious disadvantage, it was expected that the modelling improvement would outweigh the generalization problem for our task.

3 Experimental evaluation

This section describes a series of experiments investigating the influence of various acoustic modelling and modified feature extraction techniques in the task of MP3 recognition. The experiments were performed using the Kaldi [13] toolkit, and the described recognition system was based on the Gaussian mixture model-hidden Markov model (GMM-HMM) approach.

3.1 Common experimental setup

The signals for experiments came from the Czech SPEECON and TEMIC database, were recorded in 16-bit precision and 16-kHz sampling frequency by a headset microphone, and were manually transcribed. The data were split randomly into train and test subsets. The MP3 compression for the test subset was simulated by Lame [14] software, and SOX was used for the decompression. The compression rates were selected with the intention of evaluating the performance of the system for bit-rates of 128, 32, 28, 24, 20, 16, and 12 kbps.

The 39-dimensional PLP and MFCC features were computed using the CtuCopy extraction tool [15] with 32 ms window and 16 ms shift. The CMN technique was applied in a speaker-specific fashion and on static features only. In the first stage of the experiments, the signals were dithered with uniformly distributed random values from the $< -1, 1 >$ range. The effect of additional dithering for the test subset was studied in the later stages, when the dithering value $< -R, R >$ was gradually increased.

The AMs were trained on uncompressed speech using the Viterbi training algorithm from 72 h of speech and 555 speakers. The baseline uncompressed system consisted of continuous density hidden Markov models for context-dependent crossword triphones. The basic phonetic alphabet contained 44 Czech monophones and silence. The quality of the baseline AM was later improved by LDA, speaker adaptive training (SAT), SGMM framework, and discriminative training (DT). The initial feature vector for LDA was extended by three neighboring vectors, and its dimensionality was then reduced to 40. The fMLLR adaptation was used in the SAT scheme to produce the speaker-independent AMs. The final AMs were adapted in an unsupervised, speaker-specific fashion, and the maximum mutual information (MMI) criterion was used for DT. The weighted finite state machine decoder was used for recognition.

3.2 Results of optimized acoustic modelling

The preparation of the test subset involved compression and decompression from which the features were extracted afterwards. The phoneme test subset contained 45 speakers and 8.5 h of speech of varying content, and the recognition was performed using a bigram phoneme model. The results were evaluated by phone error rate (*PER*) and the phone error rate reduction (*PERR*) criteria:

$$\begin{aligned} \mathrm{PER} &= \frac{S+D+I}{N} \times 100\,[\,\%], \\ \mathrm{PERR} &= \frac{\mathrm{PER}_{base} - \mathrm{PER}_{imp}}{\mathrm{PER}_{base}} \times 100\,[\,\%], \end{aligned} \tag{1}$$

where S, D, I, and N represent the number of substituted, deleted, inserted, and total number of phones, respectively. PER_{imp} represents the improved error rate for the particular modelling technique and PER_{base} the base error rate against which the relative improvement was computed. In addition, the recognized transcription was remapped into three phonetic classes: voiced consonants, unvoiced consonants and vowels, and the phone error rate contribution for a particular phonetic class ($\mathrm{PERC}_{\mathrm{cl}}$) was computed using the following formula:

$$\mathrm{PERC}_{\mathrm{cl}} = \frac{\mathrm{PER}_{\mathrm{cl}}}{\mathrm{PER}_{\mathrm{all}}},\ [\,\%] \tag{2}$$

The purpose was to evaluate the effect of compression on each phonetic class separately.

The initial study of the behavior of PLP and MFCC features for MP3 speech recognition was performed with all previously discussed AM refinement techniques but without any non-standard modifications to the feature extraction process. This analysis served as the benchmark for the subsequent modification in the form of additional dithering and its potential contribution.

Table 1 presents results for the PLP-based system. Significant differences in absolute PER were observed between compressed and uncompressed data for initial baseline AMs, but implementation of each subsequent modelling technique decreased the ΔPER. The studied bit-rates were selected to have a linear trend, but the achieved results marked the 20 kbps bit-rate as a

Table 1 PER [%] for PLP-based system for progressively refined AM

Data	Baseline	LDA	SAT	SGMM	MMI
Raw	17.9	16.1	12.3	10.2	7.2
128k	18.8	17.1	13.5	10.8	7.9
32k	19.3	17.3	13.2	11.0	8.2
28k	19.6	17.7	13.5	11.3	8.6
24k	20.7	18.7	14.1	11.9	9.3
20k	23.9	21.1	15.5	12.9	10.5
16k	36.2	30.4	18.6	15.5	13.3
12k	62.5	52.9	26.8	22.2	20.2

breakpoint after which the error started to rise exponentially. This conclusion held true for all levels of AM development.

Another point of interest was the reduction of error as a function of the employed modelling technique. The fMLLR adaptation achieved the highest PERR for compressed speech in general, and its gain rose with decreasing bit-rate. These results indicate that the AM adaptation represents a crucial part of any system intended for MP3 recognition. On the other hand, the gain of discriminative training was the highest for raw data and decreased with decreasing bit-rate. This finding is consistent with the theoretical premise that discriminative training fits the AM on the training set but not necessarily on the testing set. In the case of our experiment, the MMI criteria optimized the AM for uncompressed signals and as the bit-rate decreased, so did the match between the testing and training signals. It should be noted, however, that the overall PERR computed for the baseline and final MMI models were in the ⟨59 %, 67 %⟩ range.

The same set of experiments for an MFCC-based system is summarized in Table 2. The system behaved similarly and displayed the same trends as far as the contribution of specific acoustic modelling techniques went. The PER displayed a tendency to rise rapidly after passing the 20 kbps breakpoint and AM adaptation proved to be crucial as it contributed the most to the overall PERR. The major difference was the overall increase of error rate for uncompressed signals. This finding leads to the conclusion that the MFCC features are not suitable for low bit-rate MP3 speech recognition.

Table 2 PER [%] for MFCC-based system for progressively refined AM

Data	Baseline	LDA	SAT	SGMM	MMI
Raw	17.8	15.9	12.3	10.3	7.3
128k	18.9	17.2	13.7	10.8	8.1
32k	20.9	17.9	13.6	11.3	8.7
28k	24.6	19.5	14.3	11.9	9.4
24k	33.9	25.9	16.4	13.5	11.4
20k	47.1	31.9	19.0	15.6	13.6
16k	62.2	49.4	25.8	20.4	18.7
12k	73.0	68.7	46.0	32.6	30.2

A more detailed study of the nature of error confirmed the theoretical assumptions about the compression distortions and their effect on particular phones outlined in Section 3. Figure 1 documents a decrease in PERC for voiced phones at the expense of unvoiced phones. PERC for unvoiced phones at the expense of voiced phones. While the reference PERC distribution was dominated by the vowels, the PERC for unvoiced constants steadily increased up to 34.2 %. Later experiments with partial contribution of each distortion showed that this negative effect occurred due to the combined presence of both the bandwidth limitation and spectral valleys.

3.3 Results of additional dithering

All the above experiments used features without any further optimization. Since the previous runs showed the detrimental effect on higher frequency bands, it was important to investigate the proposed patch methods. The dithering value R was gradually increased by a factor of 2. Tables 3 and 4 present the best PER together with the optimal R.

Additional dithering for PLP features, Table 3, yielded consistent improvement for the lowest 12-kbps rate and some improvement for the 16-kbps rate. Its application was particularly useful for baseline and LDA models, but the reduction for more advanced acoustic modelling techniques was only marginal and higher bit-rates were mainly unaffected by the method. In cases where the dithering value was too high, the additional noise degraded the features further, which resulted in worse PER than for the undithered system. The process of estimating the R value included several iterations of feature extraction and decoding and thus consumed a lot of time and resources. When all of these factors were taken into consideration, we came to the conclusion that the usage of additional dithering cannot be advised for PLP features.

The next main point of interest was to investigate whether the dithered MFCCs can match the PLPs. These experiments showed more convincing results as positive PERR was obtained for all bit-rates and levels of AM refinement, as summarized in Table 4. The generally observed trend was that the lower bit-rates gained more from the additional dithering than the higher bit-rates. It should be noted, however, that MFCCs still did not manage to outperform the PLPs. The error rates were somewhere between the original MFCCs and PLPs.

Since it was confirmed that additional dithering can improve the recognition with MFCCs, the last analysis was focused on phonetic composition of the error. The

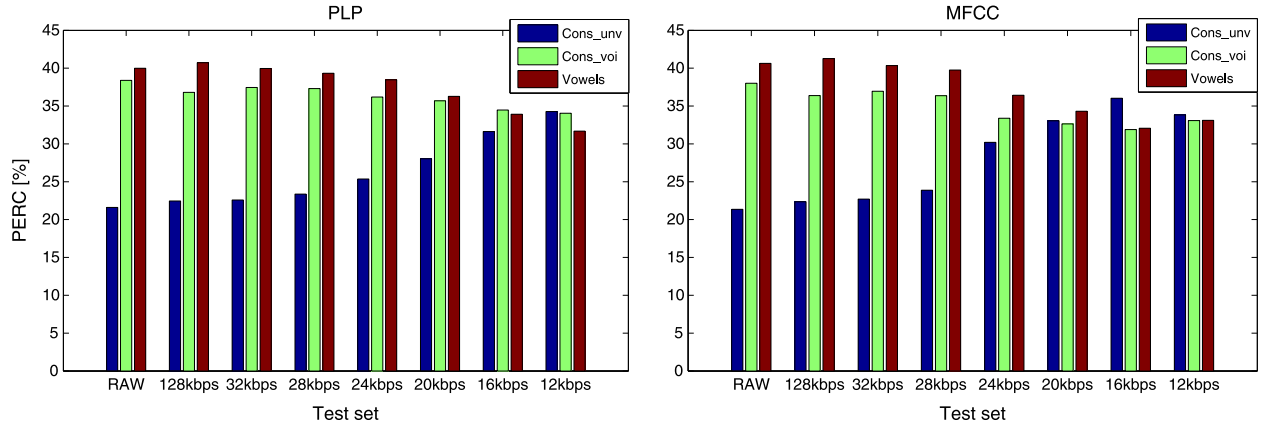

Fig. 1 PERC [%] for PLP and MFCC features. The contribution of three phonetic groups: unvoiced consonants, voiced consonants, and vowels to the total phone error. The contribution of unvoiced consonants increased with the decreasing bit-rate

initial hypothesis was that the addition of relatively weak noise would reconstruct the spectra of unvoiced phones, but the results showed that the reduction was spread evenly among phonetic classes or slightly towards the voiced phones. Figure 2 compares PERC for dithered and undithered MFCCs at 16/12 kbps as these were the only bit-rates which displayed a statistically relevant improvement. The values for the 16-kbps test sets with and without dithering were basically the same, which indicated a uniform contribution. The error for the dithered 12 kbps was dominated by the unvoiced phones, but the PERCs for the original MFCCs were even, which means that voiced phones gained more than the unvoiced phones. Based on these results, it can be speculated that the key principle of the method lies more in enhancing the feature's suitability for statistical modelling and less in actual reconstruction of their spectral characteristics.

3.4 Results of large vocabulary continuous speech recognition

Since MP3 speech recognition is generally intended for applications such as off-line transcription of recorded speech or indexing of audio archives, the following experiments were aimed at analyzing the described acoustic modelling techniques at the standard large vocabulary continuous speech recognition (LVCSR) task. The used AMs were trained in the described manner, the decoding graph was constructed from the bigram language model [16] of 340k vocabulary size created from the Czech National Corpus [17], and the test subset with an overall length of 1 h contained only signals with a full sentence structure. The results were evaluated by the standard word error rate (WER) metric.

Table 5 compares the LVCSR results for the baseline, adapted, and MMI trained AMs. Since the dithered PLPs displayed practically the same error rates as the undithered ones, they were not included in the next set of experiments. The observed trends of error rate corresponded to the conclusions drawn from the phoneme experiments. The error started to rise exponentially after passing the 20-kbps threshold, and the AM adaptation was the main source of error reduction for lower bit-rates. The advanced acoustic modelling techniques displayed a trend of increasing relative gains as the bit-rates decreased, a conclusion which was in contrast the previous results. This development occurred most likely due

Table 3 Results for dithered PLPs, the dithering value, and the corresponding error rate *R*/PER [%]

Bit-rate	Baseline	LDA	SAT	SGMM	MMI
128	2/18.7	2/17.0	4/13.2	4/10.7	4/ 7.9
32k	4/19.1	2/17.3	2/13.2	2/11.0	2/8.2
28k	4/19.4	4/17.7	4/13.5	4/11.2	4/8.6
24k	2/20.8	2/18.8	4/14.2	2/11.8	2/9.3
20k	4/23.5	2/21.0	4/15.6	2/12.9	2/10.5
16k	16/32.0	8/27.9	8/18.5	4/15.4	2/13.4
12k	32/44.7	32/39.4	16/24.9	8/21.2	4/19.8

Table 4 Results for dithered MFCCs, the dithering value, and the corresponding error rate *R*/PER [%]

Data	Baseline	LDA	SAT	SGMM	MMI
128k	2/18.9	2/17.1	4/13.1	4/10.7	4/8.0
32k	8/20.1	4/17.8	4/13.4	4/11.1	2/8.6
28k	8/21.6	4/18.6	8/13.9	8/11.5	8/9.1
24k	16/30.1	2/26.2	8/15.8	8/13.1	8/10.9
20k	16/34.7	8/30.0	8/17.7	8/14.8	8/12.8
16k	32/41.2	32/35.9	16/21.3	8/17.9	8/16.4
12k	64/49.9	64/45.5	16/29.0	16/24.2	16/23.0

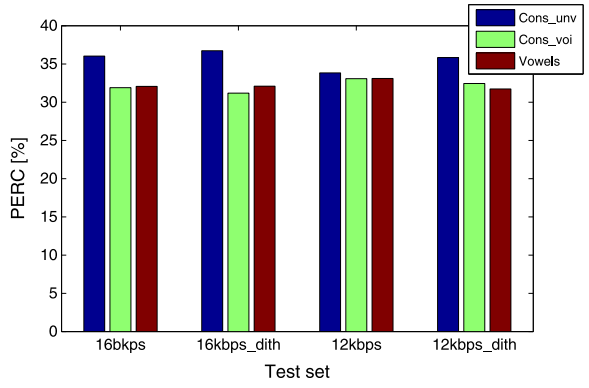

Fig. 2 The comparison of PERC [%] for classic and additionally dithered MFCC features. The analysis showed little difference in PERC distribution between classic and modified features. The contribution of additional dithering was constant for every class

to the usage of word level LM in combination with progressively better AMs. The PLP features achieved better results than MFCC and marginally better than the dithered MFCC (dMFCC) features.

Previous experiments determined that lossy compression affects unvoiced phonemes more strongly than the voiced ones. In order to determine which compression degradation, bandwidth limitation, or spectral valleys is more detrimental to the overall performance, we decided to estimate their partial contributions separately. The uncompressed signals were filtered by a FIR low-pass filter at corresponding cutoff frequencies, which were estimated for each bit-rate from their spectrograms. The spectral valleys' contribution was estimated by replacing the suppressed spectral bins at higher frequencies of a coded signal with the spectral bins from an uncoded signal. The replaced bins were selected using the same cut-off frequencies. It should be noted however, that other compression artifacts (pre-echo, birdie, etc.) might have also degraded the lower bands and thus affected the obtained results. This approach allowed us to quantify the contributions as if they affected only the selected parts of the spectra. Table 6 summarizes the results of recognition in the LVCSR task and gives an overview of the used cutoff frequencies of the filters. The results demonstrate that the spectral valleys degraded the speech more significantly on average, but that their contribution to the overall degradation was only marginal, with the exception of the 12-kbps bit-rate and MFCC features. The performance drop generally observed in the MP3 speech was the result of the non-linear combination of both distortions.

Table 5 WER [%] for LVCSR task with 340k bigram LM

Features	Raw	128k	32k	28k	24k	20k	16k	12k
PLP_base	23.74	24.5	24.5	25.57	25.98	28.19	38.79	68.57
PLP_adapt	18.19	19.45	19.4	19.59	19.84	20.67	23.56	33.19
PLP_MMI	14.25	14.43	14.55	14.54	15.21	16.15	18.57	25.23
MFCC_base	23.72	25.07	25.13	26.67	31.75	38.46	62.45	91.43
MFCC_adapt	18.44	19.06	19.11	19.92	20.7	22.51	28.11	44.82
MFCC_MMI	14.22	14.72	14.92	15.12	15.82	17.57	21.48	31.54
dMFCC_base	–	24.85	25.05	26.01	31.77	36.69	48.83	70.03
dMFCC_adapt	–	18.75	19.01	19.61	20.32	21.71	25.17	34.75
dMFCC_MMI	–	14.25	14.78	15.06	15.5	16.84	19.47	26.41

4 Conclusions

This paper studied the current state-of-the-art acoustic modelling and a specific feature compensation technique in the task of MP3 speech recognition. More precisely, linear discriminant analysis in conjunction with acoustic model adaptation, subspace Gaussian mixture model framework, discriminative training, and additional dithering was described. The baseline system was trained on uncompressed data and tested on both the uncompressed and the compressed signals in the task of phoneme recognition and LVCSR.

The evaluation runs documented that the usage of PLP features and application of AM adaptation and discriminative training can reduce the error rate of the system. The MMI-trained AMs performed at 14.24 % on the reference test set, but the WER dropped to 18.57 % for 16 kbps and 25.23 % for 12 kbps rates. For comparison, the MFCC system performed at 14.22, 21.48, and 31.54 % WERs for the same subsets. Adapting the AMs to the specific speaker and bit-rate yielded the highest mean improvement out of all the analyzed modelling techniques and proved to be essential for recognition of compressed speech. Our preliminary experiments on usage of DNN-HMM displayed results which were slightly worse (by approximately 1 %) than GMM-HMM and, for this reason, were not presented in article.

The phoneme-level recognition confirmed the theoretical hypothesis that the MP3 compression affects the unvoiced phonemes more significantly than voiced ones phonemes. The contribution of the unvoiced phonemes to the total phone error rose from 21.6 % for the reference test set to 34.2 % for the 12-kbps set. A more

Table 6 The partial contribution of LP filtering and spectral valleys (SV) on LVCSR task, WER [%]

	Bit-rate	128k	32k	28k	24k	20k	16k	12k
	f_{cut}		7200 Hz			5800 Hz		5600 Hz
PLP	SV	14.16	14.28	14.45	14.22	14.43	14.96	16.68
	Low-pass		14.22			14.34		14.54
MFCC	SV	14.35	14.91	14.89	14.94	15.56	16.18	19.7
	Low-pass		14.15			14.45		14.86

detailed study of bandwidth limitation and spectral valleys showed that the observed increase of the error rate occurred as a result of the non-linear combination of both distortions.

While the observed results justified the usage of additional dithering for the MFCC features, the error rates for dithered MFCCs were still slightly higher than those for PLPs. However, the main problem of this approach was the need to manually tune the dithering value to achieve the best results. The results of detailed phoneme accuracy showed that the technique was not able to compensate for the loss of information due to the low-pass filtering and spectral valleys phenomena.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
Research described in the paper was supported by internal CTU Grant SGS14/191/OHK3/3T/13 "Advanced Algorithms of Digital Signal Processing and their Applications".

Received: 6 March 2015 Accepted: 7 July 2015
Published online: 28 July 2015

References

1. C-M Liu, H-W Hsu, W-C Lee, Compression artifacts in perceptual audio coding. IEEE Trans. Audio Speech Lang. Process. **16**(4), 681–695 (2008). doi:10.1109/TASL.2008.918979
2. RJJH Van Son, A study of pitch, formant, and spectral estimation errors introduced by three lossy speech compression algorithms. Acta Acustica United Acustica. **91**, 771–778 (2005)
3. C Barras, L Lamel, J Gauvain, in *Acoustics, Speech, and Signal Processing. Proceedings of 2001 IEEE International Conference on*. Automatic transcription of compressed broadcast audio, vol. 1 (Salt Lake City, USA, 2001), pp. 265–268
4. L Besacier, C Bergamini, D Vaufreydaz, E Castelli, in *Multimedia Signal Processing. Proceedings of 2001 IEEE Fourth Workshop on*. The effect of speech and audio compression on speech recognition performance (Cannes, France, 2001), pp. 301–306
5. PS Ng, I Sanches, in *Proceedings of 2004 Conference Speech and Computer, SPEECOM*. The influence of audio compression on speech recognition systems (St. Petersburg, Russia, September 2004)
6. P Pollak, M Borsky, Communications in Computer and Information Science, in *E-Business and Telecommunications*. Small and large vocabulary speech recognition of MP3 data under real-word conditions: experimental study, vol. 314 (Springer Berlin, 2012), pp. 409–419
7. J Nouza, P Cerva, J Silovsky, in *Acoustics, Speech and Signal Processing. Proceedings of 2013 IEEE International Conference on*. Adding controlled amount of noise to improve recognition of compressed and spectrally distorted speech (Vancouver, Canada, May 2013), pp. 8046–8050
8. H Abbasian, B Nasersharif, A Akbari, M Rahmani, MS Moin, in *Communications, Control and Signal Processing. 3rd International Symposium on*. Optimized linear discriminant analysis for extracting robust speech features, (March 2008), pp. 819–824
9. S Tamura, S Hayamizu, in *Signal Information Processing Association Annual Summit and Conference. 2012 Asia-Pacific*. Multi-stream acoustic model adaptation for noisy speech recognition (Hollywood, USA, December 2012), pp. 1–4
10. U Remes, KJ Palomäki, M Kurimo, in *EUSIPCO. Proceedings of 16th European Signal Processing Conference*. Missing feature reconstruction And Acoustic Model Adaptation Combined For large vocabulary continuous speech recognition (Lausanne, Switzerland, 2008)
11. D Yu, L Deng, Y Gong, A Acero, in *Proceedings of the Interspeech*. Discriminative training of variable-parameter HMMs for noise robust speech recognition (International Speech Communication Association Brisbane, Australia, September 2008)
12. J Du, P Liu, F Soong, J-L Zhou, R-H Wang, Lecture Notes in Computer Science, in *Chinese Spoken Language Processing*. Noisy speech recognition performance of discriminative HMMs, vol. 4274 (Springer Berlin, 2006), pp. 358–369
13. D Povey, A Ghoshal, G Boulianne, L Burget, O Glembek, N Goel, M Hannemann, P Motlicek, Y Qian, P Schwarz, J Silovsky, G Stemmer, K Vesely, in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. The Kaldi speech recognition toolkit (IEEE Signal Processing Society Hilton Waikoloa Village, Big Island, Hawaii, US, 2011)
14. R Hegemann, A Leidinger, R Brito, LAME (2011). http://lame.sourceforge.net
15. P Fousek, P Mizera, P Pollak, CtuCopy feature extraction tool (2014). http://noel.feld.cvut.cz/speechlab
16. V Prochazka, P Pollak, J Zdansky, J Nouza, Performance of Czech speech recognition with language models created from public resources. Radioengineering. **20**, 1002–1008 (2011)
17. Ústav Českého národního korpusu FF UK Praha, Český národní korpus - SYN2006PUB (2006). http://www.korpus.cz