Spectrally Selective Dithering for Distorted Speech Recognition

Michal Borsky, Petr Mizera and Petr Pollak

Faculty of Electrical Engineering,
Czech Technical University in Prague,
Czech Republic.

borskmic@fel.cvut.cz,mizerpet@fel.cvut.cz,pollak@fel.cvut.cz,

10.21437/Interspeech.2015-601

Abstract

The performance of speech recognition systems can be significantly degraded if the speech spectrum is distorted. This includes situations such as the usage of an improper recording device, enhancement technique or speech coder. This paper presents a front-end compensation method called spectrally selective dithering aimed at reconstructing the spectral characteristics of nonlinearly distorted speech. The technique is designed to detect the suppressed frequency bands in the speech signal and add a weighted amount of additive noise. The detection algorithm is based on the smoothness of the excitation signal spectrum obtained through analyzing LPC filtration. The gain of the added noise is estimated from the unaffected frequency bands. The practical usability of the algorithm has been studied in the task of MP3 speech recognition for very low bit-rates. The obtained results have demonstrated the advantage of using the proposed technique. We achieved up to 1.85% absolute WER reduction using the standard HMM-GMM architecture in LVCSR task.

Index Terms: distorted speech recognition, front-end compensation, spectrally selective dithering, HMM-GMM

1. Introduction

The common speech features currently used in Automatic Speech Recognition (ASR) systems are derived from the short-time estimation of the spectra. These features been shown to achieved high a accuracies in situations with acoustically clean conditions and little distortion to the speech wave. Precise recognition of the speech in the presence of ambient noise, distorted by the transmission channel or coded for efficient storage requires modifications to built a more robust system. It usually involves usage of a speech enhancement technique, normalization of extracted features, acoustic model (AM) adaptation or employment of a more robust architecture in general [1].

The MPEG-1/2 Layer-III format represents a lossy perceptual audio codec which introduces several forms of nonlinear distortions, which have been studied in detail in [2]. The main problem the authors identified was the bandwidth clipping and areas of comparatively low energy in the spectrogram (spectral holes), which have been known to degrade performance of the ASR systems dramatically.

Practical studies on recognition of MP3 recordings concluded that ASR systems can work with little difficulties if sufficiently high bit-rates are used[3], [4] or [5]. A common trend all authors reported was the a rapid drop in accuracy for bit-rates of 24kbit/s and lower. These findings correlate with the evaluation of the robustness of standard MFCC features done in [6]. While these results may indicate that the distortion for higher bit-rates is negligible, authors in [7] demonstrated that accurate automatic bit-rate detection can be done even for bit-rates $>$ 128kbit/s using a simple SVM classifier.

This works demonstrate that the recognition of speech coded by a lossy codec is still a challenge for current state-of-the-art ASR systems.

2. ASR for MP3 speech

The MP3 compression causes degradation to ASR systems due to changes to extracted features and mismatch at the level of acoustic model (AM). This section provides the analysis of the distortions, their impact on features and AM and historically proposed solutions. The description of designed compensation method follows afterwards.

2.1. MP3 spectral distortion

The bandwidth clipping is the distortion most easily identifiable from the spectrogram. Its influence on ASR is mainly due to the loss of information carried by higher frequencies. Therefore, it is expected to effect mainly speech units with rich high frequency structure, such as unvoiced consonants. On the other hand, the voiced speech units (vowels and voiced cons.) have a strong harming structure at low frequencies which makes them more robust against this type of distortion. If we consider the standard parameterization scheme for short-time spectral features, we can conclude that the distortion will effect always the same higher cepstral coefficients regardless of the neighboring context. Authors in [4] tried to account for this loss by artificially limiting the bandwidth of training data to match the MP3 coder but concluded that such approach yields only marginal improvement to the *WER*.

The Spectral Valleys Phenomenon (SVP) is present in the signals coded by perceptual audio coders when certain frequency components are removed from the spectra based on the principle of spectral masking. The psychoacoustic model selects the frequencies which are considered inaudible due to their low magnitude and position next to another, much stronger component and removes them. The effect can be distinguished in the spectrogram as the areas of energy quantized to zero, generally at low and middle freq. bands. The primary problem of SVP is its statistically random nature from both the time and spectral point of view, which means that only a part of training data for each speech unit is likely to be affected by it and even then not always the same coefficients. Therefore, the phonemes are randomly displaced in the acoustic domain which increases the chance of false recognition. Authors in [5] proposed a solution based on adding a uniformly distributed noise and showed significant improvements by about 45% absolutely for 16kbit/s

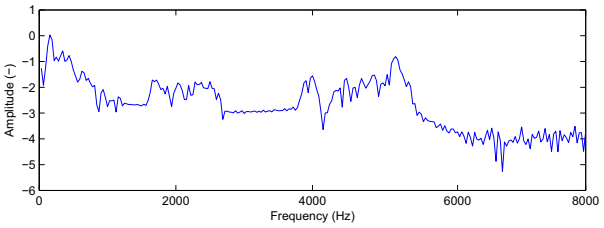

Figure 1: *An example of a frame from a signal with* $fs = 16kHz$ *distorted by MP3 coding.*

rate and MFCC features. A more recent study in [8] investigates the usage of the DNN-HMM approach for recognition of non-linearly distorted speech. The authors reported an 1.7% absolute *WER* improvement over the standard HMM-GMM architecture without any compensation for 16kbit/s coded speech. Once the features were compensated, the HMM-GMM system performed at exactly the same *WER*.

Figure 1 illustrates mentioned distortions in a frame from 12kbit/s coded signal. The frequency clipping occurred for $f > 5600$ *Hz* and the flat areas with the central freq. at $f = \{1400, 3300\}$ *Hz* resulted from SVP.

2.2. Spectrally Selective Dithering

Following the nature of the distortion described in the previous section, along with already achieved results in this task, lead us to the conclusion that the reconstruction of missing low and middle frequency components is the key to robust MP3 coded speech recognition. Numerous works on this topic have been published in the field of audio coding, i.e. [9] or [10], but to our knowledge, none have been tried in the field on ASR. Following the research presented in these works, we decided to design a similar algorithm, which could be easily incorporated into the current parametrization schemes for computing the short-time spectral based features (MFCC and PLP). In order to combine all of these aspects, we decided to modify the uniform dithering technique to dither only the selected frequency bands with automatically estimated amount of noise. To accomplish this approach we had to design two essential blocks : the zero-bands detector and gain estimation/compensation block.

Figure 2 presents the block scheme of the designed algorithm called Spectrally Selective Dithering (SSD). The principal idea was to use to the LPC model to decompose the signal into the spectral envelope and the residual signal (exc), detect the zero-energy bands in the residual signal and patch them to get the compensated signal.

The zero-energy band detector was composed of: LPC coefficient estimation, analysis filter, FFT, filter bank and criteria estimator. The block used for gain estimation and reconstruction was composed of: gain estimator, compensation block, multiplexer, IFFT and reconstruction filter.

2.2.1. Zero-band detection

The analysis of coded speech has shown that the discussed distortions effect both the spectral envelope and the residual signal as well. Although these changes were detectable in both spectral parts, we chose to base our detection algorithm on the residual signal. The residual signal for AR process is supposed to have the characteristics of white noise with zero mean value, but the actual values in the bands affected by the distortion had very little variance and their trend could be approximated a linear function with gradient $\rightarrow 0$. It allowed to employ a criteria function based on the smoothness of the spectral curve to classify distorted bands by a fixed threshold. The input frame was passed through analysis filter with the frequency response $1/H(z)$ to obtain the residual signal. Let's assume the frequency band b which contains spectral components f_1 through f_2 was extracted from the spectra of the excitation in a frame. The criteria $crit(b)$ used in the detector was computed as:

$$crit(b) = \sqrt{\sum_{f=f_1}^{f_2} (exc(f) - exc(f\text{-}1))^2}. \quad (1)$$

The masking function was then defined as:

$$mask(b) = \begin{cases} 0, & crit(b) \geq Thr, \\ 1, & crit(b) < Thr, \end{cases} \quad (2)$$

where *Thr.* was a fixed which was set as a constant for all bitrates. The threshold value was estimated empirically from the 12kbit/s data and was the only manually tunned variable in the algorithm. The number of spectral components in a band was crucial to robust estimation. In case the band was narrow the detector returned too many false alarms. On the other hand, if the band was too wide the location of zeroed bands became inaccurate. Our experiments showed that 4 spectral components gave a reasonably precise estimation for $16KHz$ sampled signals.

Figure 3 illustrates the spectrograms of the same signal coded at three bit-rates: 32kbit/s, 20kbit/s and 12kbit/s and their respective masking functions. It can be noted that the selected function was not only able to detect the suppressed bands accurately but did so only in the speech frames. These attributes enhanced the selectivity of the algorithm as only the speech frames were subject to subsequent compensation.

2.2.2. Gain estimation and compensation

Using the masking function we estimated the gain of noise as an average from the undistorted bands. The added noise had uniform distribution with zero mean value and unit variance. We can then express the compensated excitation band *EXC(b)* as follows:

$$exc(b) = \begin{cases} exc(b) + G * noise, & mask(b) = 1, \\ exc(b), & mask(b) = 0. \end{cases} \quad (3)$$

The excitation signal was then put together in the multiplexer block and the compensated frame was obtained through forward LPC filtering. The whole compensated signal was reconstructed using the OLA method and a new set of features was extracted from the compensated signals. The reconstruction filter was designed as a standard all-pole filter which frequency response was defined as:

$$H(z) = \frac{\sqrt{E_p}}{1 + \sum_{k=1}^{p} a_k z^{-k}}, \quad (4)$$

where E_p was the power of the prediction error.

3. Experimental evaluation

In this section we present the results achieved by the proposed front-end compensation technique and compare them to the baseline and uniformly dithered system. The experiments were

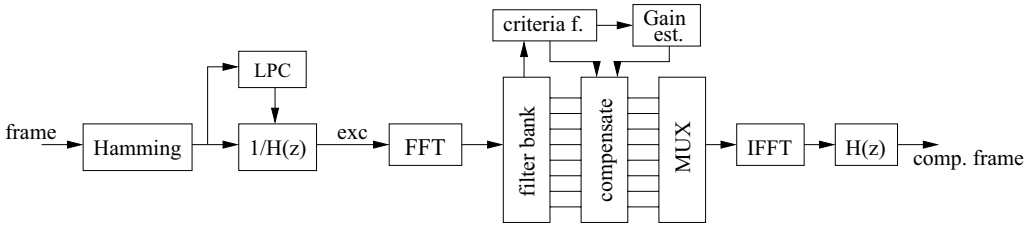

Figure 2: *Block diagram of the SSD compensation technique*

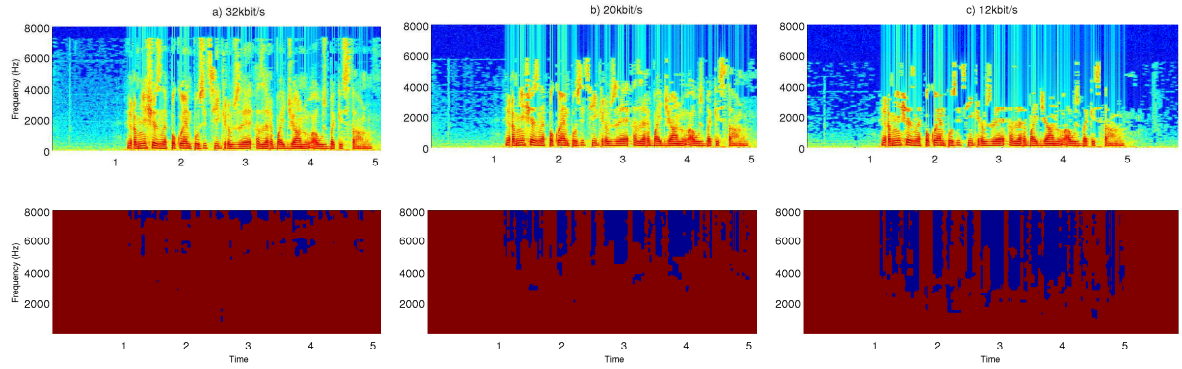

Figure 3: *Spectrogram of MP3 coded speech and their respective masks for SSD. Blue areas represent the positive detection.*

performed with the standard HMM-GMM architecture. The training and testing recognition tools came from the KALDI toolkit [11].

3.1. Experimental setup

The signals for the experiments come from Czech SPEECON and CZKCC databases. The selected subsets were recorded in acoustically similar environments with minimal background noise by a headset microphone and 16kHz sampling frequency. Speakers for train and test subsets were selected randomly but only the test subset was MP3 coded to induce distortions. The compression was simulated by Lame [12].

The 13-dimensional PLP feature vector was computed using CtuCopy tool [13] with 32*ms* frame and 16*ms* shift, normalized by Cepstral Mean Normalization (CMN) technique and then extended by Δ and Δ^2 features. Four preceding and following vectors were spliced onto the initial vector. The high-dimensional vector was reduced by LDA to 40 and then transformed by MLLT [14]. The coded signals were dithered with uniformly distributed values in $< -R, R >$ range. We manually increased R until we achieved the minimal error rate in recognition.

The AM was trained on 72 hours of speech using Viterbi algorithm and HMM-GMM architecture for cross-word triphones. The starting phone set consisted of 44 monophones and a single silence model, which also served as a garbage model for other non-speech events. The quality of AM was later improved by SAT, the combination of UBM + SGMM [15] and discriminative training using the *MPE* [16] criteria.

An internally created trigram LM with a 340k vocabulary was used. The LM [17] was created from publicly available resources of the Czech National Corpus [18] and had 1.5% OOV on our recognition task.

The recognition task consisted of 2 hours of speech containing only full sentences. The fMMLR adaptation [19] was performed in unsupervised, speaker specific fashion. Since the SI model was trained on uncompressed speech, the adaptation also served the purpose of fitting the AM to the compression. The performance was evaluated by *WER* criteria.

Table 1: *WER [%] for MP3 coding, LP-filtered speech on corresponding f_c and full-band speech.*

f_c	7.2 kHz		5.8 kHz		5.6 kHz	
Bit-Rate	32k	28k	24k	20k	16k	12k
WAV	13.8		14.2		14.3	
MP3	14.4	14.6	15.0	16.0	18.1	25.2
Full-band	13.8					

3.2. Results using HMM-GMM system

The *WER* of the baseline system on the uncompressed test subset reached 13.8%. In order to evaluate the contribution of distortions separately, we designed a FIR low-pass (LP) filter, filtered the uncompressed speech and compared these results to the MP3 coded speech. The correct cutoff frequency (f_c) for each bit-rate was determined from the spectra of the MP3 signals. Table 1 summarizes the results, showing that the filtration effects the *WER* only marginally as the absolute difference between full-band signals and signals with f_c = 5600 Hz was only 0.5%. Meanwhile, the error rates for the MP3 speech rose rapidly with the decreasing bit-rate up to 25.2% for 12kbit/s. This experiments documented the that major portion of performance drop was not caused by the freq. clipping alone, but rather by the SVP or the combination of the two.

The contribution of both discussed front-end compensation techniques is summarized in Table 2. The results obtained with uniform dithering were ambiguous as we observed an actual increase in *WER* for all rates aside from 20kbit/s and 12kbit/s.

Table 2: *WER [%] for baseline, SA system, uniform dithering with various R values and SSD*

MP3	Base (SI)	Base (SA)	Uniform Dith. (SA) 2.0	4.0	8.0	SSD (SA)
32k	18.70	14.45	14.5	14.64	14.77	**14.25**
28k	19.21	14.64	14.64	14.72	14.94	**14.52**
24k	19.87	15.02	15.07	15.12	15.31	**14.97**
20k	21.27	16.02	15.99	15.88	16.03	**15.72**
16k	25.19	18.15	18.27	18.42	18.43	**17.83**
12k	34.73	25.20	24.31	24.04	24.07	**23.35**

This behavior, along with the need to set the dithering value *R* properly, is the reason why the technique is used rarely, if at all.

On the other hand, the proposed SSD technique displayed an absolute *WER* reduction over baseline system ranging from 0.05% for 24kbit/s to 1.85% for 12kbit/s. The second important observation was that the SSD never increased the error rate. An absolute margin of 0.69% between uniform dithering and SSD was the highest for the lowest bit-rate. This results showed the proposed algorithm is able to accurately detect low-energy areas in the spectrogram and estimate the amount of noise needed to compensate the distortion. Although the presented improvements are relatively small, the nature of distortion lead us to the conclusion that more significant error reductions can be achieved by compensating SVP further.

The evaluation runs demonstrated the proposed algorithm was able to accurately detect low-energy areas and estimate the amount of noise needed to compensate the distortion. Although the presented improvements were relatively small, the nature of distortion lead us to conclusion more significant error reductions can be achieved by compensating SVP.

4. Conclusions

In this paper we have presented an algorithm named spectrally selective dithering which was designed to compensate the nonlinear distortions in MP3 coded speech. The algorithm was based on the principle of detecting the corrupted frequency bands and compensating them by adding a weighted amount of noise. The smoothness of the spectra was used as the criteria function for detection. In the task of MP3 coded speech recognition we demonstrated the contribution of the said algorithm when we observed 1.85% *WER* absolute reduction for the lowest 12kbit/s in HMM-GMM system.

While SSD is a modified version of an algorithm which dithers all frequency bands uniformly, it has displayed to have several advantages over the uniform dithering. Unlike the uniform dithering, SSD was always able to improve the recognition score and does not require to manually tune the weight of noise. While the detection of zero energy bands was based on fixed thresholding, its value was set to a constant for all bitrates. The experiments showed that SSD always outperformed the uniform dithering, although by a small margin. Finally, the algorithm was designed in a way which makes it relatively easy to implement into the established parametrization schemes for computing spectral-based features.

5. 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".

6. References

[1] M. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. of ICASSP*, 2013, pp. 7398–7402.

[2] C.-M. Liu, H.-W. Hsu, and W.-C. Lee, "Compression artifacts in perceptual audio coding," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 681–695, 2008.

[3] C. Barras, L. Lamel, and J. Gauvain, "Automatic transcription of compressed broadcast audio," in *Proc. of ICASSP*, 2001, pp. 265–268.

[4] L. Besacier, C. Bergamini, D. Vaufreydaz, and E. Castelli, "The effect of speech and audio compression on speech recognition performance," in *Proceedings of 2001 IEEE Fourth Workshop on Multimedia Signal Processing*, 2001, pp. 301–306.

[5] J. Nouza, P. Cerva, and J. Silovsky, "Adding controlled amount of noise to improve recognition of compressed and spectrally distorted speech," in *Proc. of ICASSP*, 2013, pp. 8046–8050.

[6] S. Sigurdsson, K. B. Petersen, and T. Lehn-Schiler, "Mel frequency cepstral coefficients: An evaluation of robustness of MP3 encoded music," in *Proceedings of the International Symposium on Music Information Retrieval*, 2006.

[7] B. D'Alessandro and Y. Q. Shi, "MP3 bit rate quality detection through frequency spectrum analysis," in *Proceedings of the 11th ACM Workshop on Multimedia and Security*, 2009, pp. 57–62.

[8] L. Seps, J. Malek, P. Cerva, and J. Nouza, "Investigation of deep neural networks for robust recognition of nonlinearly distorted speech," in *Proc. of INTERSPEECH*, 2014, pp. 363–367.

[9] H.-W. Hsu, C.-M. Liu, and W.-C. Lee, "Audio patch method in audio decoders - MP3 and AAC," in *Proc. of 116th Convention of Audio Engineering Society Convention*, 2004.

[10] M. Arora, J. Lee, and S. Park, "High quality blind bandwidth extension of audio for portable player applications," in *Proc. of 120th Convention of Audio Engineering Society Convention*, 2006.

[11] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Hilton Waikoloa Village, Big Island, Hawaii, US: IEEE Signal Processing Society, 2011.

[12] R. Hegemann, A. Leidinger, and R. Brito. (2011) Lame. [Online]. Available: http://lame.sourceforge.net

[13] P. Fousek, P. Mizera, and P. Pollak. (2014) Ctucopy feature extraction tool. [Online]. Available: http://noel.feld.cvut.cz/speechlab

[14] R. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *Proc. of ICASSP*, vol. 2, 1998, pp. 661–664.

[15] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiat, A. Rastrow, R. Rose, P. Schwarz, and S. Thomas, "Subspace gaussian mixture models for speech recognition," in *Proc. of ICASSP*, 2010, pp. 4330–4333.

[16] D. Povey and P. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Proc. of ICASSP*, 2002, pp. 105–108.

[17] V. Prochazka, P. Pollak, J. Zdansky, and J. Nouza, "Performance of Czech speech recognition with language models created from public resources," *Radioengineering*, vol. 20, pp. 1002–1008, 2011.

[18] "Ústav českého národního korpusu (Institute of Czech National Corpus) - SYN2006PUB," Prague, 2006, http://ucnk.ff.cuni.cz/english/syn2006pub.php.

[19] A. Ghoshal, D. Povey, M. Agarwal, P. Akyazi, L. Burget, K. Feng, O. Glembek, N. Goel, M. Karafiat, A. Rastrow, R. Rose, P. Schwarz, and S. Thomas, "A novel estimation of feature-space mllr for full-covariance models," in *Proc. of ICASSP*, 2010, pp. 4310–4313.