ADDING CONTROLLED AMOUNT OF NOISE TO IMPROVE RECOGNITION OF COMPRESSED AND SPECTRALLY DISTORTED SPEECH

Jan Nouza, Petr Cerva, Jan Silovsky

SpeechLab, Technical University of Liberec, Studentska 2, 461 17 Liberec, Czech Republic

ABSTRACT

This paper deals with the recognition of speech whose spectrum is notably distorted by lossy compression (namely MP3) or by some implementations of 'speech enhancement' techniques. We show that these non-linear treatments can introduce gaps in spectrum that significantly change the distribution of MFCCs and degrade performance of ASR. We propose a method that measures the level of spectrum distortion and use it for adding a controlled amount of noise to the signal. It effectively masks the gaps and helps namely in situations where the source and parameters of the distortion are not known and hence we cannot use a properly matched acoustic model. In spite of its simplicity, the method can improve significantly speech recognition of highly compressed or spectrally distorted signals. We demonstrate it in several large experiments conducted on publicly available speech databases, in two languages and for two types of spectral distortion.

Index Terms— speech recognition, compressed speech, MP3

1. INTRODUCTION

The amount of digital audio documents containing speech has increased significantly during the last two decades. All major broadcasters archive their recently produced programs and digitize the historical ones from the analog epoch. Other large sources of digital speech are data kept by telecom operators and call centers, or recordings of parliament sessions, meetings, presentations, lectures, etc. As the size of the data is huge, they are often stored in a compressed form. For many years, the MP3 format has been the most popular and most widely used one. Even if it had not been intended for speech compression, its overall quality is sufficient for human perception. For ASR systems, however, it can cause serious performance degradation. We have faced this problem in projects whose goal is to automatically transcribe and index large audio archives. The largest one belongs to the Czech Radio company [1] and contains more than 100.000 spoken documents covering 90 years of broadcasting. Most recordings were digitized during the last decade and the MP3 format (with various settings) was used as the major standard. Not only the historical records but also the contemporary ones have been stored as compressed audio. The problem is that their compression rates and other parameters are not always same and may vary according to the studio and radio station where they were produced. There are also documents that were created by external collaborators (e.g. correspondents or free-lance news people) who used their own recording and sound editing tools. Moreover, some documents passed through multiple compression steps (e.g. GSM and MP3). All these facts complicate the transcription process and introduce additional ASR errors.

2. RELATED WORK

There exist several research papers that study the impact of MP3 compression on automatic speech recognition, e.g. [2-6]. Their authors agree that MP3 coders operating with 32 kb/s and higher bitrates have almost negligible effect on word error rate (WER) values. For lower bitrates, however, the impact is considerable and WERs can increase by more than 10 %, which may be critical for most applications. To eliminate it, the authors offer three solutions: 1) to use (if possible) compression algorithms that better fit speech characteristics (e.g. OGG, [6]), 2) to train acoustic models (AM) on band-limited speech [3], or 3) to train bit-rate-specific AMs on data passed through MP3 coders [2],[3].

The first approach is not applicable when audio documents have been already compressed and original recordings are not available anymore. The second approach relies on the assumption that the main effect of the MP3 compression is band limitation. In this paper, we show that this is just a minor source of performance degradation. The third proposed approach solves the problem of the mismatch between the acoustic model and compressed speech in a more general way. However, it assumes that we know which compression type (and setting) was used for each document, and that we have a set of AMs trained for each particular setting. As explained in Introduction, in practice, this is not always true.

In our paper, we investigate the impact of signal compression on ASR performance in a more detailed way. We show that the main source of mismatch between an AM trained on common speech data and a highly compressed audio signal are spectral 'holes' introduced by the masking effect of the MP3 algorithm. Initial experiments proved that this type of spectrum distortion could be mitigated by noise added to the signal. In section 4, we try to answer the question how much noise should be added to get the optimal results. The proposed method is evaluated in LVCSR tests conducted on real broadcast data, in two languages (Czech and Polish), and on recordings distorted by a 'noise suppression' technique implemented on some mobile devices. In most cases, our method yields a considerable WER reduction up to tens of percent.

3. IMPACT OF MP3 COMPRESSION

3.1. Impact on Spectrum and Cepstrum

Let us compare four spectra depicted in Fig. 1. The first has been computed for uncompressed speech signal, the others belong to the same source signal that was compressed by mp3PRO coder (within Adobe Audition software [7]) using 3 predefined bitrates (32, 24, and 16 kb/s). It is clearly visible that the main difference between the spectra is the effective width of the band, which gets narrower for lower bitrates. Yet, there is also another, less evident distortion

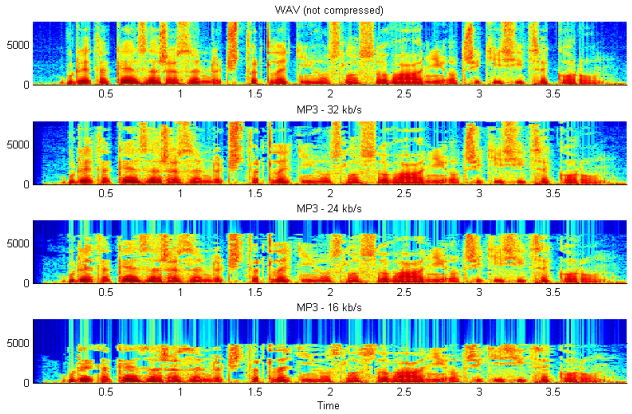

Figure 1: *Spectra of original speech recording and its MP3 compressed versions with different bitrates*

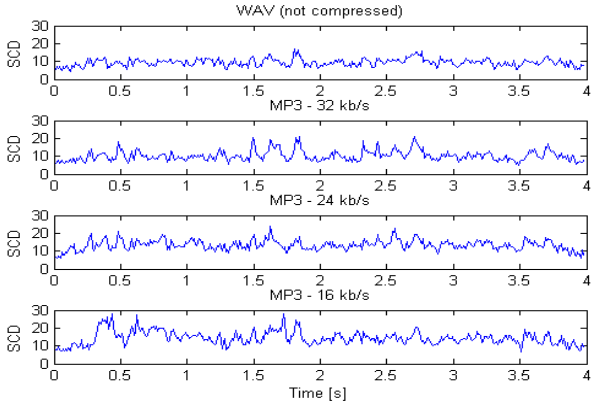

Figure 3: *Spectrum channel difference (SCD) plots computed for signals in Fig.1*

in the spectra: It is the bins and channels with almost zero energy, which make 'holes' occurring particularly in the spectra of highly compressed speech. These holes are artifacts of the masking effect in the MP3 algorithm.

When mel-frequency coefficients (MFCC) are computed via the discrete cosine transform (DCT), these holes will change the positions of MFCCs in the cepstrum domain. Moreover, as the holes become deeper (for lower bitrates), the difference between MFCCs in adjacent frames gets larger, which will result in larger values of Δ and ΔΔ coefficients. This is shown in Fig. 2 where histograms of DMEL5 and DDMEL5 coefficients are compared for an original and 16 kb/s compressed utterance. The latter have significantly larger variance. This has a large impact on GMM likelihoods values and consequently on speech decoding.

3.2 Impact on Recognition

Now, let us try to quantify the impact on ASR, using a real LVCSR system and a speech dataset compressed in a controlled manner.

3.2.1 Employed ASR system

The system used in all the experiments is our LVCSR system designed for highly inflected languages with very large lexicons. It has been employed for broadcast news (BN) transcription in Czech since 2005 [8], and later also for several other Slavic languages. Recently, it has been intensively used in the project aimed at the transcription of the Czech radio archive [1].

The system processes 16 kHz audio data and converts them into a stream of 39-dimensional MFCC vectors computed every 10 ms in 25-ms long frames. These are normalized by the CMS (Cepstral

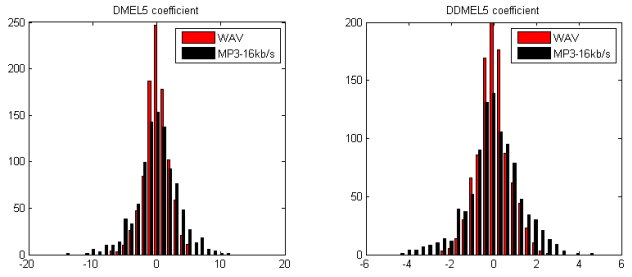

Figure 2: *Distributions of two dynamic MFCCs for an uncompressed and the same MP3 compressed utterance*

Mean Subtraction) technique within a 2-second long sliding window. Next, the square-matrix HLDA transformation is applied and all the 39 features are passed to the decoder. The acoustic model uses triphones with 5562 physical states and 180864 gaussians in total. It has been trained on 320 hours of speech (120 hours of manually annotated microphone and BN data + 200 hours of lightly supervised BN data). The currently used lexicon contains 483K items (with 527K pronunciations). The LM is based on bigrams trained on 40 GB corpus of multi-genre Czech texts. As the lexicon includes also 4200 multi-word expressions (frequently collocated word strings), almost one third of the bigrams actually covers sequences that are three-, four-, five- or even six-word long. The unseen bigrams are backed-off by the Kneser-Ney method

3.2.2 Data used in experiments

The experiments were conducted on publicly available Czech speech databases. We have used 919 utterances from the Czech part of GlobalPhone database [9] for development works and the Czech BN recordings from the COST278 database [10] as an independent evaluation set. Both are well suited for the task, as: a) they were recorded using the classic PCM format (no digital codecs), b) they consist of short as well as long utterances (from 3 to 30 words), and c) they cover many speakers and various speaking styles. The size of the data and the baseline results achieved with the above described system are shown in Table 1.

Table 1. *Experimental data and baseline results*

Data	# words	WER [%]	OOV [%]
GlobalPhone - dev.	13453	10.46	0.96
COST278 - eval.	8451	14.08	0.54

3.2.3 Initial tests with simulated compression

All the recordings from the development set were compressed using two types of MP3 coders: Lame coder bundled in FFmpeg software [11] and mp3Pro coder available in Adobe Audition software [7]. Both offer several predefined bitrates, from which we used the following ones: 48, 32, 24 and 16 kb/s for Lame and 32, 24, 20, 16 kb/s for mp3Pro. After the compression, each file was converted back to WAV format as it is the native input format for the employed ASR system. After that we run a series of recognition experiments whose results are presented in Table 2. As expected, the higher bitrates had a small impact on the performance, while

Table 2. *WER achieved on GlobalPhone development data compressed by two coders with selected bitrates*

Format, coder, bitrate	WER [%]
WAV (uncompressed)	10.46
MP3, Lame - 48 kb/s	10.58
MP3, Lame - 32 kb/s	11.51
MP3, Lame - 24 kb/s	21.14
MP3, Lame - 16 kb/s	72.33
MP3, mp3Pro - 32 kb/s	12.79
MP3, mp3Pro - 24 kb/s	16.33
MP3, mp3Pro - 20 kb/s	21.48
MP3, mp3Pro - 16 kb/s	30.36

the lower ones (16 kb/s in particular) led to very high error rates. We may also notice a large difference between WERs achieved for 16 kb/s Lame and mp3Pro compression.

3.2.4 Experiments with added noise

To verify our hypothesis that the main source of poor ASR performance (in case of lower bitrates) are the spectrum holes mentioned in section 3.1, we run another series of experiments, in which we added a controlled amount of noise to a signal distorted by the compression. This was done by adding a randomly generated integer number to each signal sample:

$$s'[n] = s[n] + rnd(R) \quad (1)$$

where R specifies (uniform random) generator's interval <-*R*,*R*>.

We have applied a large range of R values from 1 (the default value commonly used in most ASR front-ends) to 256. In Table 3 we present a selection of the results. The bold values denote the best WERs for each compression. We can see that adding a larger amount of noise to highly compressed speech helps to reduce WER values. Especially for 16 kb/s data, it is significant (from 72.3 to 26.5 % in case of Lame16 compression). On the other side, a small amount of noise does not harm ASR of lightly distorted signals. The observed impact of the added noise is in accord with the theory presented e.g. in [12], where it is shown that additive noise has a larger impact on MFCC values if its power dominates over that of speech. In our special case, this dominance is applied locally by 'filling' those mel-frequency spectrum channels that are related to the holes introduced by the compression algorithm.

4 SPECTRUM DISTORTION MEASURE AND ITS USE FOR CONTROLLED NOISE ADDITION

To benefit from the above described observation, we must be able to control the amount of noise that is to be added. It should be related to the abrupt local changes in the spectrum. We have proposed and tested several measures. The one presented below fitted best to the standard MFCC implementation.

4.1 Spectrum Channel Difference (SCD) Measure

Let us take mel-frequency spectral channels as the basic units of spectrum and denote their log energy as E_c. We define Spectrum Channel Difference (*SCD*) measure in frame *f* as:

$$SCD(f) = \sum_{c=1}^{C-1} | E_{c+1} - E_c | \quad (2)$$

Table 3. *WER achieved on GlobalPhone development data after compression and adding noise*

Coder, bitrate	WER [%] for increasing R values						
	1	4	8	16	32	64	128
Lame48	**10.6**	10.8	10.7	11.4	12.7	15.3	20.6
Lame32	11.5	**11.4**	11.7	11.9	13.5	16.2	21.5
Lame24	21.1	21.2	**21.0**	21.5	22.5	23.9	29.4
Lame16	72.3	60.9	51.4	40.1	32.2	**26.5**	28.8
Pro32	12.8	12.4	**12.2**	12.4	13.6	16.4	21.1
Pro24	16.3	**16.2**	16.8	18.0	20.3	25.9	24.9
Pro20	21.5	21.6	20.9	**20.1**	20.3	21.6	22.3
Pro16	30.4	28.6	27.2	26.4	**26.2**	27.8	31.0

where C is the total number of channels (e.g. 24 as in many implementations). For each utterance, we compute the average value of *SCD* (over all its F frames) and denote it as ASCD:

$$ASCD = [\sum_{f=1}^{F} SCD(f) \cdot VAD(f)] / F_{VAD} \quad (3)$$

where *VAD(f)* is a binary decision from a voice activity detector (1 for speech frame, 0 otherwise) and F_{VAD} is the total number of speech frames found by the VAD. The *SCD* values computed for compressed and uncompressed speech differ, as illustrated in Fig. 3, and so does the *ASCD* parameter, which can be used as an indicator of the spectral distortion caused by the compression.

4.2 Controlled Noise Addition (CNA)

Using the complete set of results obtained in the experiments described in section 3.2.4, we constructed a scatter plot for pairs of *ASCD* values and optimal R values corresponding to them. As the relation between the 2 parameters exhibited two saturation levels, we decided to fit the data to a logistic function:

$$R = Int[\frac{K}{1 + \exp(-G \cdot (ASCD - L))}] \quad (4)$$

Its parameters were determined on the dev. set using the method described in [13]. (In our case, K=220, G=0.6 and L=16).

4.3 Experiments on Development and Evaluation Data

We run a new set of experiments, in which for each compressed utterance, *ASCD* and R values were computed using eq. (3) and (4), and noise was added according to eq. (1). The results achieved for the development and evaluation sets are presented in Table 4. The former can be directly compared to those in Table 3. We can see that the automatically estimated amount of noise helped to reach either optimal or close-to-optimal WERs. (The R values varied in a wide range, from 1 to 180). Again, the lightly compressed data remained almost untouched, while the strongly distorted ones were modified in the way that significantly improved WERs. The results achieved on the independent evaluation set followed the same trends. The largest WER reduction (almost 40 %) was observed for Lame16 compression.

5. TESTS ON REAL DATA

So far, all the experiments were conducted on data that were distorted intentionally (for study purposes). Now we want to show

Table 4. *WER for GlobalPhone and COST278 data after compression and noise added according to eq.(1) and (4)*

Coder, bitrate	GlobalPhone		COST278	
	Base	CNA	Base	CNA
Lame48	10.58	10.70	14.15	14.22
Lame32	11.51	11.42	14.77	14.79
Lame24	21.14	20.65	18.76	18.35
Lame16	72.33	25.58	67.81	27.98
Pro32	12.79	12.20	14.56	14.68
Pro24	16.33	16.16	16.84	16.74
Pro20	21.48	19.91	20.89	19.74
Pro16	30.36	25.50	40.02	29.41

how the proposed CNA method performs on real data. Results from all the below described tests are summarized in Table 5.

5.1 Czech Radio Recordings

In the Czech Radio archive [1], several thousands of audio documents come from regional studios. For many years, each studio editor has had his/her own standard for data storing. Usually they used MP3 with 44,1 kHz sampling frequency, but bitrates varying in a larger range. When listening to these records one can easily distinguish between low-rate and high-rate compressed files by noticing the compression artifacts. If the data are down-sampled to 16 kHz (required by the ASR system), their spectrum exhibit similar 'holes' like those described in previous sections (see Fig.4).

We have created a test set of mixed regional broadcast news that included recordings with and without those audible artifacts. Using the standard Czech LVCSR system, we obtained 8.14 % WER. When the CNA method was applied, the WER decreased to 7.4 %.

5.2 Polish Radio Recordings

Recently, we have been developing also a broadcast news transcription system for Polish. It employs the same modules as the Czech one, except of the lexicon (the Polish has 297K words), a language model and an acoustic model. The system is regularly tested on recently broadcast data. For this purpose we use short discussion programs produced by Jedynka radio station, which are available on web [14] as audio files accompanied with almost verbatim transcriptions. Unfortunately, the audio files are compressed, with clearly audible compression artifacts. The difference between the WER values achieved for the original (compressed) recordings and those processed by the automatically applied CNA method is more than 3.3 %, as shown in Table 5.

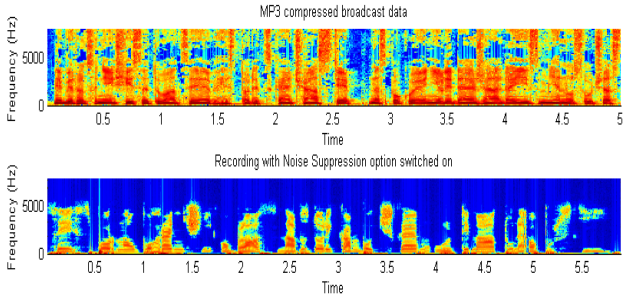

Figure 4: *Spectra of real signals tested in section 5*

Table 5. *Results achieved on real data acquired from Czech and Polish audio archives and from a sound device switched to a 'noise suppression' mode.*

Test speech data (language)	Test set size (#words)	WER [%]	
		Original	CNA
News from regional broadcast stations (CZ)	6809	8.14	7.40
Conversation speech - Jedynka radio (PL)	8451	21.30	17.93
Speech recorded with noise suppression (CZ)	12780	55.36	22.77

5.3 Data Recorded with Noise Suppression Option

The last mentioned application of the proposed method deals with speech data recorded on devices (notebooks, smartphones) that have a 'noise suppression' option. It is known that signal manipulation techniques like this can harm the performance of ASR systems. A signal passed though a 'denoising' procedure can look like that in Fig. 4. Here, the 'holes' are even larger than 'islands' belonging to speech.

What to do if one has already recorded a large amount of audio in this way? The WER for this type of distortion is very high again. Anyway, it can be significantly reduced if the proposed Controlled Noise Addition method is applied. We demonstrate it on a large database that was (non-intentionally) recorded on a modern notebook with the 'noise suppression' option switched on. While the original data achieved 55.4 % WER, after applying the CNA method, the WER was reduced to 22.8 %.

6. CONCLUSIONS

In this paper, we investigate the impact of non-linear spectral distortions caused by the MP3 audio compression and by some implementations of noise suppression techniques on speech recognition performance. We show that one source of the mismatch between a distorted speech signal and an acoustic model trained on common (not necessarily clean) speech are spectral holes, i.e. channels with extremely low power. Their impact can be mitigated by adding an automatically controlled amount of noise that helps to fill the low-power channels while having only a negligible effect on those parts of spectrum where speech dominates. Our experiments proved that the proposed Controlled Noise Addition (CNA) method can reduce the WER even by tens of percent. If we compare our results to those published in [3], where the authors used matched AMs trained either on band limited speech or on one passed through MP3 coders with various bit rates, we see that our method yields improvements in a similar range. Our approach does not require prior information about the type of compression applied to the processed signal and does not need a set of AMs trained for each coder setting. Moreover, it is applicable also to other types of non-linear distortions, like e.g. the 'noise suppression' technique used on some mobile devices.

ACKNOWLEDGEMENTS

This research work was supported by Czech Ministry of Culture (project no. DF11P01OVV013 in program NAKI).

REFERENCES

[1] Nouza, J., Blavka, K., Bohac, M., Cerva, P., Zdansky, J., Silovsky, J., Prazak, J.: "Voice Technology to Enable Sophisticated Access to Historical Audio Archive of the Czech Radio", In Multimedia for Cultural Heritage. Springer Berlin Heidelberg, 2012, CCIS vol. 247, pp.27-38.

[2] Besacier, L., Bergamini, C., Vaufteydaz, D., Castelli, E.: "The effect of speech and audio compression on speech recognition performance" Proc. of IEEE Multimedia signal Processing Workshop, Cannes, France, October 2001

[3] Barras, C., Lamel, L., Gauvain, J.L.: "Automatic Transcription of Compressed Broadcast Audio". Proc. of ICASSP'2001, Salt Lake City, May 2001

[4] Ng, P.,S., Sanches, I: "The Influence of Audio Compression on Speech Recognition Systems", SPECOM 2004, St. Petersburg.

[5] Pollak, P., Behunek, M.: "Accuracy of MP3 speech recognition under real-word conditions: Experimental study", Proc. SIGMAP 2011, Seville, July 2011, pp. 5-10

[6] Van Son, R.J.J.H.: "A study of pitch, formant, and spectral estimation errors introduced by three lossy speech compression algorithms", Acta Acustica united with Acustica, 91 (4), 2005, pp. 771-778.

[7] http://www.adobe.com/cz/products/audition.html

[8] Nouza, J., Zdansky, J., Cerva, P., Kolorenc, J.: "Continual On-line Monitoring of Czech Spoken Broadcast Programs". Proc. of Interspeech 2006, Pittsburgh, Sept., 2006, pp. 1650-1653

[9] ELRA catalogue (http://catalog.elra.info), GlobalPhone Czech, catalogue reference: ELRA-S0196

[10] Vandecatseye, A., et al., "The COST278 pan-european broadcast news database," in Proc. of LREC 2004, Lisbon, May 2004.

[11] http://ffmpeg.org/

[12] Pettersen, S.G.S.: "Robust Speech Recognition in the Presence of Additive Noise". PhD thesis. NUST, Norway 2008

[13] Arnold D.: "Fitting a Logistic Curve to Data". Article at redwoods.cc.ca.us/instruct/darnold/diffeq/logistic/logistic.pdf

[14] http://www.polskieradio.pl/7,Jedynka