IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

A study of data augmentation for increased ASR robustness against packet losses

María Pilar Fernández-Gallego[1], *Doroteo T. Toledano*[1]

[1]AUDIAS - Audio, Data Intelligence and Speech
Universidad Autónoma de Madrid

mariapilar.fernandezg@estudiante.uam.es, doroteo.torre@uam.es

Abstract

Nowadays a large amount of companies record conversations, calls, sales or even meetings, in many cases to comply with the current legislation. Apart from the legal need, these recordings constitute an invaluable source of information about clients, call center operators, marketing campaigns, markets trends, etc. The current state of the art in Automatic Speech Recognition (ASR) allows to exploit this information in a very efficient way. However, the recordings at these repositories tend to present very low quality because the audio is typically recorded in a highly compressed way to save storing space. Besides, since it is very common to use Voice over IP (VoIP) in these systems, it is usual to have short interruptions in the speech signal due to packet losses. Both effects, and particularly the last one, have an impact in ASR performance.

This paper presents an extensive study of the influence of these effects and the effectiveness of different data augmentation strategies to increase the robustness of ASR systems in these circumstances, and in particular when packet losses degrade the speech signal.

Index Terms: Data augmentation, Packet losses, Speech recognition, Fisher Spanish

1. Introduction

Nowadays many companies record conversations, calls, sales or even meetings for several reasons and, in many cases, just to comply with the current legislation. Apart from the legal need, these recordings constitute an invaluable source of information about clients, call center operators, marketing campaigns, markets trends, etc. Typically the speech recorded in call centers is recorded in a very reduced bit rate, and hence limited quality. The main reason for it is to save storing space because in some cases hundreds or thousands of hours daily need to be recorded. Therefore reduced bit rate speech coding is a common scenario in Automatic Speech Recognition (ASR) applied to this type of data.

Voice over IP (VoIP) communications have become very common and are currently mainstream in call centers and voice recording systems. In VoIP, speech signals are transmitted as packets of a fixed length that depends on the selected speech codec. Normally the packet length is between 20 and 40 ms. During the transmission of these packets several issues can occur, the most common being packet delays and packet losses, which may degrade the speech quality beyond the degradation introduced by speech coding and other common problems such as echoes [1]. To mitigate losses or delay issues, Packet Loss Concealment (PLC) techniques are applied, which can fill the lost packets by adding redundant information such as repeating frames, adding noise, etc; or alternatively using interpolation methods to reconstruct the the signal [2] [3]. PLC techniques are included in some standards, such as ITU-G711 Appendix I, which is a high quality low-complexity algorithm for packet loss concealment [4].

In the last years, deep learning techniques and deep neural networks (DNNs) have proved to be able to extract higher level features from less processed data and also to model, predict and generalize better than classical techniques. In ASR in particular, DNNs are included in all state-of-the-art systems, either combined with the Hidden Markov Model (HMM) machinery in hybrid HMM-DNN systems or completely replacing HMMs with end-to-end neural approaches. One of the reasons for the popularity of DNNs in ASR is the fact that, when training data is abundant and representative of the application, they are very robust against variability, such as bandwidth, environment or speaker [5]. However, to attain these advantages it is necessary to have a large amount of training data. For that reason, data augmentation methods have become commonly used to artificially generate more data, trying to represent the possible scenarios that a system may face in real operation. In a good number of works, data augmentation has been used to increase model robustness against variations such as noise, speed, reverberation, etc., with significant improvements with respect to other techniques [6][7]. However there are very few works in which it is applied to deal with the problem of packet losses and reduced bit rate codification. Some works like [8] try to use data augmentation to increase robustness to deformations in the time direction and partial loss of frequency information by working on log mel spectrogram directly. Other works such as [9] apply data augmentation using audio codecs with changed bit rate, sampling rate, and bit depth.

Therefore, the motivation for this study on data augmentation to increase ASR robustness against packet losses and reduced bit rate coding is based on two main reasons. The first one is the need to develop a robust ASR system capable of dealing with packet losses and limited bit rate speech codification, which are very common in call center scenarios, without a huge degradation in performance. The second one is that after studying the available scientific literature, we noticed that the research community has not extensively addressed this problem, despite it is a very common issue in the industry.

The rest of the paper is organized as follows: Section 2 explains the types of data augmentation used. Section 3 describes the ASR system used and data sets used to train it. Section 4 describes the data used to carry on the experiments and the results obtained. Finally, Section 5 presents conclusions and future work.

2. Augmentation types

To make ASR models more robust against packet losses and low quality coding we need to artificially introduce these dis-

10.21437/IberSPEECH.2021-39

tortions in the training database for the ASR acoustic models. This section describes how these distortions are introduced in the training data.

2.1. Packets losses

To simulate packet losses we have considered a fixed frame size of 20 ms. We consider three modes for the packet losses: individual packet losses, burst packet losses and mixed (individual and burst) packet losses. A tunable parameter controls the percentage of packets lost, which in our experiments take the values 5%,10%,15% or 20%.

2.1.1. Individual packet losses

To simulate this type of packet loss, randomly chosen packets along the audio file are removed (by making the signal 0), assuring that the removed packets are not together.

2.1.2. Burst packet losses

To recreate burst packet losses we remove batches of three consecutive frames randomly located along the audio until we reach the loss percentage chosen.

2.1.3. Single and burst packet losses

A real scenario can include both types of packet losses. For this reason, a more realistic simulation has been considered merging the two previous modes (single and burst losses), by removing randomly located batches of one, two or three packets along the audio until the loss percentage selected is reached.

2.2. Speech coding

To simulate the effect of speech coding we have used the two more common codecs in our real data: MP3 (with FFmpeg [10]) and full rate GSM (with SoX [11]), with several variations.

2.2.1. MP3

MP3 is a perceptual audio (not speech specific) codec designed for audio transmission and storage that became very popular for Internet audio applications and streaming in particular. Its encoding efficiency is defined by the bit rate, which can be adjusted to the particular needs of the application among several possible choices [12]. In this paper, two bit rates have been applied: 16 Kbit/s and 8 Kbit/s. To obtain 8 Kbit/s files, each channel of the original audio has been to converted to 8Kbit/s MP3 format and then to WAV-PCM. For the 16 kbit/s files, each channel has been converted to 16 kbit/s MP3 and later converted to WAV-PCM. In both cases FFmpeg [10] has been used.

2.2.2. Full rate GSM 06.10

GSM FR is a speech codec designed for digital mobile telephone use. The speech signal is divided into blocks of 20 ms and has an average bitrate of 13 kbps using a 8 kHz sampling rate [13]. Each channel is converted to GSM FR format and then coverted to WAV-PCM format using SoX [11].

3. Automatic Speech Recognition Models

KALDI [14] has been used for training the ASR models. In particular, the acoustic model is a Hybrid Deep Neural Network Hidden Markov Model (DNN-HMM) which uses Time Delay Neural Networks (TDNN) [15]. This type of network includes connections between its units not only at the current time, but also at different times, and is capable of taking nonlinear decisions taking into account the value of the input at a relatively long time span, normally around the current time. Each layer works with a time context wider than the previous layer thus using an increasingly amount of temporal context. For language modelling we have used a relatively simple n-gram statistical language model. Our recipe is based on the Fisher/Callhome Spanish recipe included in Kaldi without applying the re-scoring last stage.

The architecture of the TDNN consists of 13 TDNN layers with 1024 units each and 128 bottleneck dimensions, a pre-final linear layer with 192 dimensions and a last softmax layer. Input features are Mel frequency cepstral coefficients (MFCCs) with high frequency resolution (hires) with 40 dimensions, adding a ± 1 temporal context and an i-vector with 100 dimensions to model speaker characteristics.

In addition, the recipe applies speed-perturbation for data augmentation with two factors 0.9 and 1.1 obtaining a data set three times bigger than the original (including the original and two speed-perturbed copies). [6]

3.1. Baseline

The data used to develop our baseline system has been the Fisher Spanish data set. This data set is made up with 163 hours of telephone speech from 136 native Caribbean and non-Caribbean Spanish speakers. Around 4 hours have been reserved for the test set and another 4 hours for development. The rest of the corpus has been used for training. The baseline includes speed-perturbation data augmentation. The language model has been trained only on the transcriptions of this corpus.

3.2. Data augmentation experiments

We have retrained the system using different data augmentation strategies to deal with the problems of low bit rate coding and packet losses. In all of them, the amount of data used to train the models are exactly twice the one used for the baseline model. We refer to the different data augmentation strategies explored as da_model1-da_model4.

3.2.1. Low bit rate codec

This strategy (da_model1) trains the system using the baseline data plus another transformed Fisher Spanish data set obtained by applying each of the three codecs described in Section 2.2 to one third of the original corpus. The goal of this data augmentation strategy is to measure the gain obtained by including low bit rate coding effects.

3.2.2. Packets losses and low bit rate coding

To deal with both problems, these strategies double the training data by generating and additional copy in which each file has randomly suffered one or two transformations, low bit rate coding and/or packet losses of different types, depending on the model of packet losses applied:

- da_model2: Individual packet losses (Section 2.1.1).
- da_model3: Burst packet losses (Section 2.1.2).
- da_model4: Individual and burst packet losses (Section 2.1.3).

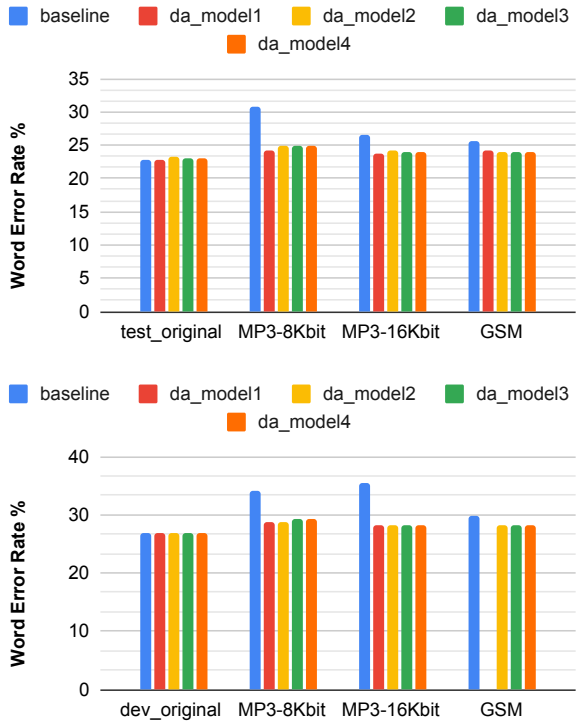
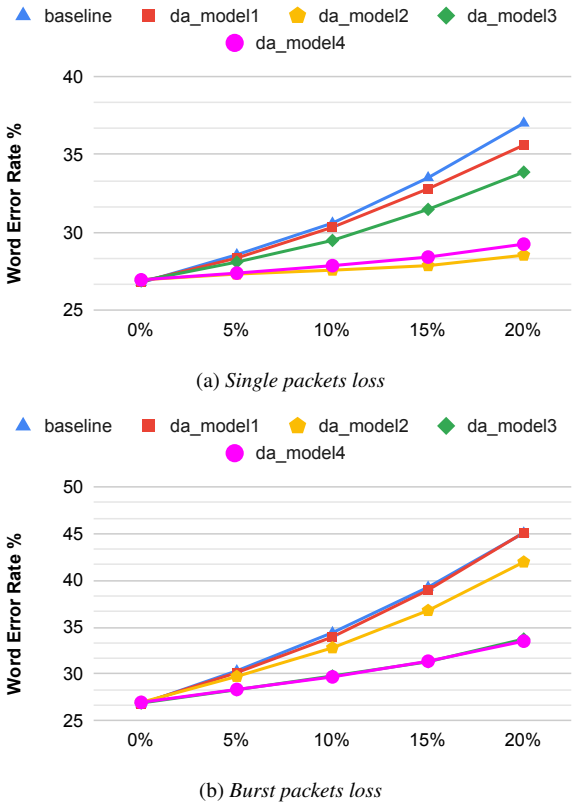


Figure 1: *Baseline vs. different data augmentation systems in test (top) and development (bottom) sets in original format and with three different low bit rate codecs (MP3-8Kbit/s, MP3-16KBit/s and GSM-FR).*



(a) *Single packets loss*

(b) *Burst packets loss*

Figure 2: *Development results comparing the baseline with data augmentation systems for single packet losses (up) and burst packet losses (bottom) for different packet loss probabilities. da_model3 is hidden behind da_model4 in bottom figure.*

4. Experiments and Results

The experiments performed try to measure the effectiveness of the different data augmentation approaches to deal with the problems of packet losses and low bit rate coding. We aim to improve the Word Error Rate (WER) of ASR both on simulated data and also on real data coming from real call centers.

4.1. Evaluation database

The main evaluation database is derived from the test and development datasets reserved from Fisher Spanish. These datasets have been transformed with the same transformations used in the different data augmentation strategies described above.

The three different codecs explained in Section 2.2, have been used to obtain three copies coded with MP3 with 8 and 16 Kbit/s and with GSM.

The different packet loss simulation systems have been used to obtain eight different copies of the data sets in each case applying 5%, 10%, 15% and 20% loss percentage including individual and burst packet losses, so that we can evaluate the worse scenario when the audio file only contains burst packet losses and a better scenario where there are only single packet losses.

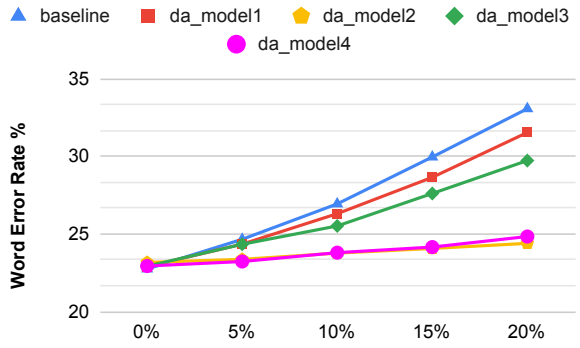
As a result, 12 conditions with different degradations are obtained. Table 1 shows the Mean Opinion Score (MOS) (as estimated by ITU-T P.563 single-ended method for objective speech quality assessment in narrow-band telephony applications [16]) in each one of this conditions, and also for the non-degraded baseline. The MOS scale is from 1 to 5, being the worst score 1 and the best 5. It is somewhat surprising that single packet loss data get worse MOS than burst packet loss data, probably because the percentage is the same but there are three times more interruptions for the individual packet losses.

4.2. Low bit rate coding results

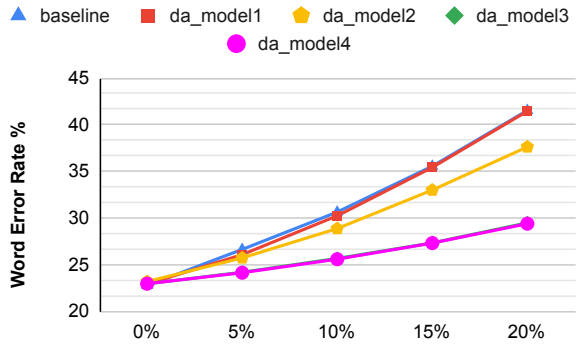
Figure 1 shows a big difference in terms of WER for low bit rate coding data sets, particularly between the baseline and all the data augmentation models, the latter having small differences among them. A much better WER is obtained when data augmentation techniques are applied, achieving almost the same results obtained with the baseline system in the non-degraded scenario. There is a small difference of $\sim 1\%$ in all cases, except when coding with MP3 at 8 Kbit/s where the difference is $\sim 2\%$. Therefore, it seems that these types of low bit rate coding scenarios can be almost solved with any of these data augmentation techniques.

4.3. Packet loss results

Figures 2 and 3 show the results obtained with the baseline and the different systems trained with different data augmentation strategies in development and test data sets, respectively, both for individual packet losses and burst packet losses at different packet loss probabilities. The data augmentation strategy including single losses only (da_model2) has better results in scenarios with only individual losses. Similarly, the data augmentation strategy containing burst losses (da_model3) has bet-



(a) *Single packets loss*



(b) *Burst packets loss*

Figure 3: *Test results comparing the baseline with data augmentation systems for single packet losses (up) and burst packet losses (bottom) for different packet loss probabilities. da_model3 is hidden behind da_model4 in bottom figure.*

ter results in the scenario with burst losses, showing in this case a great improvement compared with the rest of the models. Finally, the data augmentation strategy with single and burst losses (da_model4) have similar results in both cases than the best of the former models, and therefore shows an improved robustness against all types of packet losses.

When packet loss probability is relatively low (5%, 10%) and only individual losses are considered, ASR can be very accurate using data augmentation techniques, having similar results as without packet losses. However, when the number of packet losses increases or burst packet losses occur, data augmentation techniques greatly improve performance, but still a notable decrease in performance is observed. In the worst-case scenarios, performance could be about $\sim 7\%$ worse in terms of absolute WER than with non-degraded data, even for the best performing data augmentation strategies.

So far all the experiments were performed on simulated data including simulated degradations. With the aim to corroborate the effectiveness of these data augmentation strategies in real data, we have applied the baseline (no data augmentation besides the speed perturbation) and the best performing data augmentation strategy of the previous experiments, the da_model4, in real call-center data. Language model and lexicon have been adapted to the particularities of the specific call center data. In all cases the language model and lexicon used in the baseline and the da_model4 are exactly the same. Table 2 shows the results obtained on real call center data. Three data sets coming

Table 1: *Mean Opinion Score (MOS). PL means packet loss*

Transformation	**Test**	**Dev**
Original	2, 27	2, 42
MP3 8 Kbit/s	2, 03	2, 19
MP3 16 Kbit/s	1, 95	1, 87
GSM	1, 85	1, 95
5% Indiv. PL	1, 27	1, 43
10% Indiv. PL	1, 14	1, 24
15% Indiv. PL	1, 06	1, 11
20% Indiv. PL	1, 03	1, 04
5% Burst PL	1, 40	1, 61
10% Burst PL	1, 19	1, 33
15% Burst PL	1, 10	1, 19
20% Burst PL	1, 06	1, 10

from three different call centers have been used to compare ASR performance using the baseline and the data augmentation strategy. In all of them, results are clearly better with the da_model4 model, even 10% in the test_call_center_3 data set compared to the baseline model. This indicates that these strategies are not only adequate to model simulated low bit rate coding and packet losses problems, but also to improve recognition performance in real data.

Table 2: *Results in real data.*

Dataset	**duration (h)**	**baseline**	**da_model4**
		WER %	
test_call_center_1	3.5	39, 51	32, 07
test_call_center_2	3	40, 68	34, 20
test_call_center_3	2	47, 83	37, 94

5. Conclusions

In this work, we have applied data augmentation techniques to mitigate reduced ASR accuracy in audio including low bit rate coding and packet losses, having found that data augmentation by itself can greatly improve robustness in these scenarios, which are very common in the industry.

Experiments have shown that data augmentation can be very effective when audios include low bit rate codecs or contain relatively infrequent individual packet losses, obtaining an important improvement compared to the baseline model and reaching results close to those obtained without these degradations. However, for more frequent packet losses or burst losses, degradations in terms of WER remain important despite data augmentation mitigation, making it advisable to research other alternatives to try to compensate even more these degradations.

As future work, we plan to study techniques to recover loss packets to try to attain better accuracy in ASR when there are very frequent packet losses, including burst losses.

6. Acknowledgements

Partly funded by project RTI2018-098091-BI00, Ministry of Science, Innovation and Universities (Spain) and FEDER.

7. References

[1] A. Takahashi, H. Yoshino, and N. Kitawaki, "Perceptual QoS assessment technologies for VoIP," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 42, no. 7, pp. 28–34, 2004.

[2] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet loss recovery techniques for streaming audio," *IEEE network*, vol. 12, no. 5, pp. 40–48, 1998.

[3] B. W. Wah, X. Su, and D. Lin, "A survey of error-concealment schemes for real-time audio and video transmissions over the Internet," in *Proceedings - International Symposium on Multimedia Software Engineering*, 2000.

[4] "Recommendation, I. T. U. T. G. 711, Appendix I, a high quality low-complexity algorithm for packet loss concealment with G. 711," *Int. Telecom. Union (ITU)*, 1999.

[5] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks-studies on speech recognition," in *Proceedings of International Conference on Learning Representation*, 2013.

[6] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition." in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association, September 6-10, Dresden, Germany, Proceedings*, 2015, pp. 3586–3589.

[7] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[8] D. S. Park, W. Chan, Y. Zhang, B. Z. Chung-Cheng Chiu, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *INTERSPEECH 2019 – The 20th Annual Conference of the International Speech Communication Association, Graz, Austria, Sep. 15-19, Proceedings*, 2019, pp. 2613–2617.

[9] N. Hailu, I. Siegert, and A. Nürnberger, "Improving automatic speech recognition utilizing audio-codecs for data augmentation," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 2020, pp. 1–5.

[10] S. Tomar, "Converting video formats with FFmpeg," *Linux Journal*, vol. 2006, no. 146, p. 10, 2006.

[11] C. Bagwell. (1996) SoX - Sound eXchange. [Online]. Available: http://sox.sourceforge.net/ (Accessed: 4 February 2021)

[12] P. Noll, "MPEG digital audio coding," *IEEE Signal Processing Magazine*, vol. 14, no. 5, pp. 59–81, 1997.

[13] "ETSI EN 300 961 Digital cellular telecommunications system (Phase 2+)(GSM); Full rate speech; Transcoding, GSM 06.10 version 8.1. 1 Release 1999." in *Proceedings of International Conference on Learning Representation*, vol. 8, no. 1, 2000.

[14] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, and M. Hannemann, "The KALDI speech recognition toolkit," *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[15] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association, September 6-10, Dresden, Germany, Proceedings*, 2015, pp. 3214–3218.

[16] L. Malfait, J. Berger, and M. Kastner, "P. 563—The ITU-T standard for single-ended speech quality assessment." *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1924–1934, 2006.