Article

A Study of Data Augmentation for ASR Robustness in Low Bit Rate Contact Center Recordings Including Packet Losses [†]

María Pilar Fernández-Gallego * and Doroteo T. Toledano

AUDIAS—Audio, Data Intelligence and Speech, Department of Electronics Technology and Communications, Universidad Autónoma de Madrid, 28049 Madrid, Spain; doroteo.torre@uam.es

* Correspondence: mariapilar.fernandezg@estudiante.uam.es

[†] This paper is an extended version of our paper published in the conference IberSPEECH 2020, Valladolid, Spain, 18–20 November 2020.

Abstract: Client conversations in contact centers are nowadays routinely recorded for a number of reasons—in many cases, just because it is required by current legislation. However, even if not required, conversations between customers and agents can be a valuable source of information about clients or future clients, call center agents, markets trends, etc. Analyzing these recordings provides an excellent opportunity to be aware about the business and its possibilities. The current state of the art in Automatic Speech Recognition (ASR) allows this information to be effectively extracted and used. However, conversations are usually stored in highly compressed ways to save space and typically contain packet losses that produce short interruptions in the speech signal due to the common use of Voice-over-IP (VoIP) in these systems. These effects, and especially the last one, have a negative impact on ASR performance. This article presents an extensive study on the importance of these effects on modern ASR systems and the effectiveness of using several techniques of data augmentation to increase their robustness. In addition, ITU-T G.711, a well-known Packet Loss Concealment (PLC) method is applied in combination with data augmentation techniques to analyze ASR performance improvement on signals affected by packet losses.

Keywords: data augmentation; packet losses; speech recognition; Fisher Spanish; G.711; packet loss concealment

Citation: Fernández-Gallego, M.P.; Toledano, D.T. A Study of Data Augmentation for ASR Robustness in Low Bit Rate Contact Center Recordings Including Packet Losses. *Appl. Sci.* **2022**, *12*, 1580. <https://doi.org/10.3390/app12031580>

Academic Editors: Francesc Alías, Valentin Cardeñoso-Payo, David Escudero-Mancebo, César González-Ferreras and António Joaquim da Silva Teixeira

Received: 22 December 2021

Accepted: 29 January 2022

Published: 1 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Currently, most of the calls in call centers are recorded, in many cases just because it is mandatory with the current legislation. Besides legal requirements, these recordings constitute a rich source of information about users, the call center operators, the efficiency of the campaigns, and market trends, which can be translated into valuable information on the business. In call centers, hundreds or thousands of hours are recorded daily, so they are typically stored in a highly compressed way to save space. An undesired consequence is that these recordings are normally stored with very limited quality.

Voice-over-IP (VoIP), the transmission of speech over IP packets, is nowadays mainstream in call centers and their recording systems. VoIP can make use of different speech codecs, and depending on the selected speech codec, the length of the packet used for the transmission of the speech signals can change. Normally, the packet length is between 20 and 40 ms.

Given that IP-based networks do not provide secure transmission in real time, packets can be lost or delayed. If the delay is small, the receiver can wait a small amount of time by introducing some latency to give some time for the packet to arrive. However, the latency has to be limited to allow a natural human-to-human communication. Consequently, a packet that is delayed more than a threshold is considered lost, and speech needs to be reconstructed without that package, which obviously introduces important degradation

in the reconstructed speech. This degradation is added to the degradation introduced by speech coding and other common problems such as echoes, noise, etc. in VoIP [1].

Packet Loss Concealment (PLC) techniques have been introduced to mitigate losses or delay issues and are crucial in VoIP systems. These techniques can vary from the most basic forms such as filling the lost packets by repeating the last frame or adding noise to fill the lost frame, to some more complex forms such as using interpolation methods to reconstruct the signal [2,3]. Given the importance of PLC in VoIP, some speech codec standards have included PLC techniques, such as ITU-T G.711 Appendix I [4], ITU-T G.729 [5] or ITU-T G.722 in Appendix III and Appendix IV [6]. All these methods have two limitations in common: the decoder needs to know when a packet is lost to be capable to mitigate it, and the lost packet can not be in the first frames because mitigation is based on the last packets received.

Different methods have been proposed in the field of Packet Loss Concealment. The first techniques were waveform level approaches, such as those included in some of the VoIP codecs and works such as [7–9]. Later, statistical models such as Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) have been proposed as possible solutions to the problem of PLC. These statistical models provide a more complex and richer description of the speech signal (provided that enough speech material is available for training), which results in better PLC results [10–12]. In the last few years, with the advent of the era of deep learning, the use of Deep Neural Networks (DNN) has become common in many aspects of speech processing. Deep learning techniques have proved to be able to extract higher level features from less processed data and to be capable of modeling, predicting, and generalizing better than classical (statistical) techniques. PLC has not been an exception, and several approaches have been proposed in the last few years for deep learning PLC based on regression either in waveform or time-frequency domains [13–17], based on autoencoders [18,19] and based on Generative Adversarial Networks (GANs) [20–23].

The field of Automatic Speech Recognition (ASR) has also benefited from deep learning approaches in the last few years, and currently all state-of-the-art systems include deep neural networks either combined with HMM in a hybrid Hidden Markov Model-Deep Neural Network (HMM-DNN) system [24] or as an end-to-end neural model [25]. When ASR systems operate in a real environments, they have to face many problems such as overlapping, background noise, echoes, speaker variability, etc. If the training data are abundant and contain the expected variability [26], the resulting ASR system can be very robust to those sources of variability. However, it can be very difficult to collect a large amount of training speech covering all the possible situations. Therefore, it is very common to resort to the use of data augmentation methods to generate artificial data representative of the possible situations that a system may face in real operation. This approach, in combination with deep learning approaches, has proved to provide large improvements in multiple research works to compensate speech variability due to noise, reverberation, speaking rate, etc. [27,28].

However, there are very few works in which data augmentation is applied to deal with the problem of packet losses and reduced bit rate codification. There are some related works such as [29], which tries to use data augmentation to increase robustness to deformations applying partial loss of information in time and frequency on the log mel spectrogram directly. It does not aim at reducing the effect of packet losses, but the removal of segments in time is similar to packet losses. Other works such as [30] apply data augmentation using audio codecs with different bit rates, sampling rates, and bit depths. The works that we have found more similar to our approach of applying data augmentation techniques to deal with the problem of packet losses are [31], which presents a study of different training approaches, including data augmentation approaches, to deal with packet losses in a emotion recognition task, and [17], which proposes and evaluates a deep learning based PCL system using ASR measured in terms of WER, and compares results without the PCL system with and without data augmentation in ASR training. We will discuss

in more detail these works in comparison with our results at the end of the Results and Discussion section.

The main motivation for this study on the possibility of increasing ASR robustness against packet losses and low bit-rate speech codecs using data augmentation techniques is based on two reasons. The first one is a real need to improve ASR accuracy in speech data recorded in call centers that presents these two types of issues, packets losses, and low bit-rate speech codecs, aside from all the other typical problems in these real situations. 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 being a very common issue in the industry.

The rest of the paper is organized as follows: Section 2 explains the methods used to train and evaluate the systems built, as well as the data used in each case and type of acoustic model applied. Section 3 presents the results obtained and finally Section 4 presents the conclusions obtained in this study.

2. Methods

Several issues affect the precision of a speech to text system when it is used to transcribe recordings from real call centers, which may result in large degradation of the results compared to those obtained in a controlled scenario. There are possibly many factors causing this degradation. In this article, we have identified two of the main reasons for the degradation in our particular call center recordings, the low bit-rate codecs used, and the packet losses during transmission in Voice-over-IP (VoIP) networks. This article tries to analyze these issues and to propose possible techniques to reduce such degradation.

The method followed to achieve these goals has been the following. First, each possible scenario has been simulated to analyze the effects produced, thus assessing how the original system behaves in each situation and how much degradation is found in each case. Next, several types of data augmentation models have been applied to build a more robust speech to text system, with the goal of analyzing the effect of data augmentation for each possible degradation and trying to find the best one to apply in real data. Finally, we have applied this more robust speech to text system to real data to check if the system selected using simulated data also works better in real data.

The detailed steps taken to train the robust speech to text models and to evaluate them are the following:

1. Encode and decode each audio using codecs commonly used in the industry to record calls in call centers (GSM-FR, MP3-8Kbit/s, MP3-16Kbit/s, ...).
2. After encoding and decoding, remove several packets using different packet loss simulation strategies. Three different strategies have been applied. The first one includes only single packets losses; the second contains only burst packet losses (three contiguous packets losses); and the third one includes individual and burst packet losses comprising two to three contiguous packet losses.
3. Each speech to text system trained has been evaluated in test data sets which contain each simulated degradation separately to analyze the results obtained in each case and each circumstance, and also to make a decision about which data augmentation techniques could be used to improve the performance obtained.
4. ITU-T G.711 packet loss concealment standard has been applied in test data sets that contain packet losses and each speech to text system has been evaluated to analyze if it can be a good approach to have a better performance of the system.
5. The last step has been to evaluate from real call center recording data the best system obtained based on our study on simulated data.

2.1. Automatic Speech Recognition Models

The KALDI open-source toolkit [24] has been used for training the Automatic Speech Recognition (ASR) models. This toolkit includes many different recipes for various public data sets. In this work, ASR models are trained based on the Fisher/Callhome Spanish

recipe. This recipe follows a hybrid Deep Neural Network–Hidden Markov Model (DNN–HMM) approach.

The Deep Neural Network (DNN) used consists of 13 Time Delay Neural Networks (TDNN) [32] layers with 1024 units each one and 128 bottleneck dimensions. After the TDNN layers, it has a pre-final linear layer with 192 dimensions and a last softmax layer with as many outputs as the number of tied HMM states, resulting in the previous training of tied-state HMM triphone models. TDNN layers include connections between its units at different times, being capable of taking nonlinear decisions considering the value of the input at a quite long time span around the current time. In this type of network, each layer works with a time context wider than the previous layer, and the whole network uses a long temporal context.

Mel frequency cepstral coefficients (MFCCs) with high frequency resolution (hires) with 40 dimensions adding a ± 1 temporal context and concatenated to an i-vector with 100 dimensions have been used as input features.

Additionally, the recipe uses a type of data augmentation, speed-perturbation, with two factors 0.9 and 1.1 multiplying the quantity of data by three times, original data and two speed-perturbed copies [27].

2.2. Data Description

Two types of data have been used. On the one hand, the Fisher Spanish database made up with 163 h of telephone speech from 136 native Caribbean and non-Caribbean Spanish speakers. It contains around 4 h reserved for the test set and another 4 h for development, keeping the rest of the corpus for training.

To augment the data for training with simulated distortions (packet losses and/or speech coding), a Fisher Spanish training data set has been replicated by artificially adding the following distortions. First, one of the three possible speech codecs considered (MP3 with 8 or 16 Kbit/s and GSM-FR) has been applied randomly to each recording. Then, a packet loss simulation system is applied that randomly applies a 0%, 5%, 10%, 15% or 20% packet loss percentage randomly selecting among three different methods (no bursts, only bursts or mixed).

For evaluation, we have used both simulated and real data. For the simulated data, the test and development data sets reserved in Fisher Spanish have been replicated several times by adding the same types of distortions explained above. However, in this case, the same type of codec or the same type and amount of packet losses are applied to all the files in a single copy, thus obtaining three copies including the different types of codecs and eight copies including different types and amounts of packet losses. This allows us to analyze the results and robustness against each particular type of distortion.

Finally, real data gathered from actual recordings of three real call centers with different types of data and different topics have been used to validate the use of data augmentation to compensate the effects of these distortions in real data. The data have been chosen in this way to have more variability and provide stronger conclusions about the techniques used.

2.3. Degradation Models

2.3.1. Packets Losses

We have considered a fixed frame size of 20 ms in all cases and simulate packet losses by making the signal zero for the whole duration of the lost packet. The amount of packets lost is tuned by a parameter that establishes the percentage of packets lost. In all experiments, the values used for this parameter have been 0%, 5%, 10%, 15%, and 20% for train and test data. The details of the three methods applied are described below:

- Individual packet losses: Packets along the audio file are randomly chosen and removed, assuring they are not consecutive packets.
- Burst packet losses: Batches of three consecutive frames are randomly selected and removed (a total of 60 ms) until the packet loss percentage is reached.

- Single and burst packet losses: The two previous modes are merged to have a more realistic simulation, since a real scenario can include both types of losses. Batches of one, two, or three packets are randomly selected and removed (segments of 20, 40 or 60 ms) until the packet loss percentage is reached.

2.3.2. Speech Coding

To simulate the effects of different low bit rate speech codecs, three speech coding schemes found in our real data and also commonly found in the call recording industry have been chosen: GSM-FR [33] and MP3 with bit rates of 8 and 16 Kbps [34].

- MP3 is a lossy audio codec designed originally to compress the audio channel of videos. It is therefore not specifically designed for speech, but it is a very popular standard in Internet audio applications and streaming because it achieves large compression rates, which can even be adjusted to the particular needs of the application. This makes it very convenient for audio transmission and storage. In this paper, MP3 with 16 Kbit/s and 8 Kbit/s bit rates have been applied using the standard FFmpeg tool (<https://ffmpeg.org/>).
- Full rate GSM 06.10 is a digital speech codec used in digital mobile telephony. The bit rate is 13 Kbit/s with an 8 KHz of sampling rate. The codec works with packets of 160 samples (20 ms). We have applied the full rate GSM coded available in the SoX tool (<http://sox.sourceforge.net/>).

2.4. Data Augmentation Strategies

Different data augmentation strategies have been used to train different speech to text system models. In particular, the following five different data augmentation strategies (corresponding to different speech to text models) have been used.

- *da_model1*: The augmented training data set is split into three equal-size parts, each one including a different codec: MP3 8 Kbit/s, MP3 16 Kbit/s, and GSM-FR.
- *da_model2*, *da_model3* and *da_model4*: First, speech is either kept unaltered or modified by applying one of the three possible codecs used in *da_model1*. Each of these four possibilities is selected randomly for each recording, with a probability of 1/4. The resulting speech is later distorted again by applying packet losses including only single packet losses (*da_model2*), only burst packet losses (*da_model3*) or both individual and burst packet losses (*da_model4*). The packet loss percentage is randomly selected independently for each recording among 0%, 5%, 10%, 15%, and 20% on a file-by-file basis. The possibility of having 0% packet losses and not applying any codec is discarded, since it will mean no data augmentation.
- *da_model5*: This models follows the same strategy as *da_model4*, but, after applying the distortions, it also applies the ITU-T G.711 Packet Loss Concealment (PLC) model to try to reconstruct the lost packets. We use this model to generate data augmentation that includes the artifacts introduced by this PLC model, in order to test if applying a PLC model before the use of a robust speech to text system provides increased robustness against packet losses.

2.5. Evaluation Techniques

Two different measures are used to evaluate results:

- Word Error Rate (WER). This is the most common metric used to evaluate the accuracy of speech recognition systems. In the article, we will use it to evaluate the accuracy of each of the systems trained with the different data augmentation strategies. The evaluation will be performed on the development and evaluation sets, distorted with different effects, and also on the real data. All WERs in this article have been computed using the `compute_wer` tool included as part of the KALDI toolkit [24]. To analyze the statistical significance of WER results, the bootstrapping method described in [35] has been used. In particular, the implementation of these methods in the `compute_wer_bootci` tool included

in KALDI have been used to compute 95% confidence intervals of WER estimates, and to estimate the probability of improvement (*POI*) of one method over another as evaluated for the same test, allowing us to determine if the difference between two results is statistically significant at the 99% level.

- Mean Opinion Score (MOS). This metric is common in the evaluation of speech quality in fields such as speech enhancement and speech coding. In the article, we will use it to evaluate the quality of the data, both original and distorted and both simulated and real. Given that estimating the subjective mean opinion score is costly, we will use the objective estimation of MOS defined by an ITU-T P.563 single ended method for objective speech quality assessment in narrowband telephony applications [36]. In particular, we have used the reference software published along with the ITU-T P.563 standard to estimate the MOS value for each audio file. The software takes as input the audio file and produces the estimated MOS for the file. The MOS scale ranges from 1 to 5, 1 being the worst speech quality and 5 the best.

3. Results and Discussion

The experiments performed try to analyze the effectiveness of each data augmentation type to deal with the problems of packet losses and low bit-rate speech coding. The main goal is to improve the Word Error Rate (WER) of speech recognition tested on simulated data and real data coming from call centers. The different data augmentation strategies have been evaluated on twelve data sets, each one with a different condition. Table 1 shows the Mean Opinion Score (MOS) (estimated by ITU-T P.563 [36]) for each of these twelve conditions and the original data without any degradation. The results obtained show that single packet losses get worse MOS than data with burst packet losses. This can be surprising, but it may probably be related to the fact that, when the packet loss percentage is the same, for individual packet losses, the number of interruptions is three times higher than for burst packet losses. There is also a somewhat surprising result comparing MP3 coding at 8 and 16 Kbit/s. Besides this, the variations shown by MOS are as expected. MOS decreases as the distortion effect increases (e.g., higher packet loss percentage).

Table 1. Mean Opinion Score (MOS) for the original data and distorted data in Test and Dev sets of Fisher Spanish database. PL means packet losses.

Distortion	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

3.1. Speech Recognition Results with Low Bit Rate Coding

In Figure 1, a large difference in terms of WER can be appreciated between the baseline model and the rest of the models with data augmentation when the three different codecs are applied to test data. The statistical significance of these differences have been analyzed by computing the probability of improvement (*POI*) as defined in [35], showing that all differences between the baseline and the data augmentation models have a *POI* > 0.99, which means that the probability of obtaining a result in which the baseline is better is lower

than 1%. The performance achieved on distorted data with data augmentation techniques is relatively close to the results obtained with the baseline system on the non-degraded scenario. There is still a small difference of about 1% in all cases, except when coding with MP3 at 8 Kbit/s where there is a difference of about 2% in absolute WER. These small differences between the baseline on the non-degraded scenario and data augmentation results on degraded scenarios are always statistically significant ($POI > 0.99$). Accordingly, it seems that degradation produced in speech recognition performance by low bit-rate codecs, even several of them, can be almost solved with any of these data augmentation strategies. It must be noted that the five different data augmentation models include data coded with the three speech codecs.

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).

3.2. Speech Recognition Results with Packet Losses

Figures 2 and 3 show the results obtained with the baseline system and the systems trained with several data augmentation techniques in development and test data sets, respectively, for several packet loss percentages of individual packet losses and burst packet losses.

The *da_model2*, which only contains data augmentation with single packet losses, obtains the best results in scenarios with only individual losses. There is a similar situation with *da_model3*, a data augmentation strategy that only considers burst packet losses, obtaining the best results in the scenario with burst packet losses. With the aim to apply these techniques in real world scenarios where both types of losses can occur, the *da_model4* data augmentation strategy, which considers both single and burst losses, is proposed as a more realistic model. The results obtained with this strategy are close to the best results reached by the rest of the models in both situations, individual and burst packet losses. Therefore, this approach provides a large improvement in terms of robustness against both individual and burst packet losses. The statistical significance of the differences shown in Figures 2 and 3 have been analyzed by estimating the probability of improvement (POI) of

one method over the others. For the case of individual losses, both *da_model2* and *da_model4* are better than the rest of the models (with $POI > 0.99$ in all cases with packet losses). Comparing both approaches, the differences between *da_model2* and *da_model4* are only statistically significant ($POI > 0.99$) for packet loss rates of 20% in test data and 15% and 20% in development data. For the case of burst packet losses, both *da_model3* and *da_model4* are better than the rest of the models (with $POI > 0.99$ in all cases with packet losses, except for a 5% packet loss rate in development data, for which $POI = 0.8507$ between *da_model4* and *da_model5*). Comparing both approaches, the differences between *da_model3* and *da_model4* are not statistically significant ($POI < 0.99$) in all cases (the maximum is $POI = 0.9659$ in test data with a 20% packet loss rate).

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

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

Experiments so far show that, when there are individual losses and the probability is low (5% or 10 %), Automatic Speech Recognition (ASR) results including a *da_model4* data augmentation strategy are not far from the baseline results without packet losses (but we must highlight that the difference between the baseline result and any result including some percentage of packet losses with any data augmentation strategy is still statistically significant with $POI > 0.99$). However, for the rest of the scenarios (more individual packet losses or burst packet losses), there is still a significant gap in performance compared to results without packet losses, even after having obtained large improvements from the *da_model4* data augmentation strategy. Obviously, these larger differences are also statistically significant. This gap in performance reaches a maximum of about 7% in terms of absolute WER in the worst scenario considered. To try to mitigate these effects even more, we have applied the ITU-T G.711 Packet Loss Concealment (PLC) standard to the development and test data sets to analyze how the different data augmentation strategies behave. Figures 4 and 5 show the results obtained. In these figures, it is quite clear that

only the *da_model5* strategy, which has been trained with speech previously processed by the ITU-T G.711 PLC system, provides clear improvements. In fact, the difference between *da_model5* and any of the other models is statistically significant ($POI > 0.99$) in all cases with packet losses. The rest of the data augmentation strategies have been trained with speech not processed by the PLC system, and therefore fail to provide consistent improvements on speech processed with this PLC system.

Figure 4. Development results comparing the baseline with data augmentation systems for single packet losses (a) and burst packet losses (b) for different packet loss probabilities and applying ITU-T G.711 PLC system to recover the losses in development data.

Figure 5. Test results comparing the baseline with data augmentation systems for single packet losses (a) and burst packet losses (b) for different packet loss probabilities and applying a ITU-T G.711 PLC system to recover the losses in test data.

Figures 4 and 5 were not fair comparisons since all models except *da_model5* suffer from a large mismatch between training and evaluation data. To show a fairer comparison among the different data augmentation strategies, Figures 6 and 7 show the results obtained with all models on the original development and test data sets and the results obtained with *da_model5* both when evaluated on the original data sets (denoted as *da_model5* in the figures) and when evaluated on data processed by the ITU-T G.711 PLC system (denoted as *da_model5** in the figures). These experiments show that, for individual packet losses, performances achieved by *da_model2* (the best performing system) and *da_model5** are quite close, although performance is slightly worse for *da_model5**. The difference in performance between these two models is statistically significant ($POI > 0.99$) only for packet loss rates of 15% and 20% in test and development data. However, when there are burst packet losses, the accuracy is clearly better for *da_model5**. In this case, all differences between *da_model5** and the second best performing system, *da_model3* are statistically significant ($POI > 0.99$) in all cases with a packet loss rate over 5%, while, for a 5% packet loss rate, the POI is 0.9696

in test data and 0.9887 in development data, showing a high probability of improvement. Therefore, our results show that combining a standard packet loss concealment system such as the one included in ITU-T G.711 standard and a data augmentation strategy provides additional benefits over just using data augmentation.

Figure 6. Development results comparing the baseline with data augmentation systems for single packet losses (a) and burst packet losses (b) for different packet loss probabilities and applying ITU-T G.711 to recover the losses in the *da_model5** model. *da_model3* is hidden behind *da_model4* in the bottom figure.

Figure 7. Test results comparing the baseline with data augmentation systems for single packet losses (a) and burst packet losses (b) for different packet loss probabilities and applying ITU-T G.711 to recover the losses in the *da_model5** model. *da_model3* is hidden behind *da_model4* in the bottom figure.

Thus far, all the experiments were performed on data with simulated degradations. At this point, our goal is to corroborate the effectiveness of these data augmentation strategies in real data. An experiment using real call-center data was designed to compare results using the baseline model (which only included data augmentation based on speed perturbation) and *da_model4*, the best data augmentation strategy found in data with simulated degradations without the packet loss concealment system. It is worth mentioning that the ITU-T G.711 PLC system cannot be applied to this real data because we do not have information about which packets were lost. To do these experiments, the language model and the lexicon have been adapted to the call center data and its particularities, being exactly the same for the baseline and the *da_model4* data augmentation strategy.

Table 2 shows the results on real call center data. Each data set comes from a different call center (of different companies). In all of them, the improvement is quite clear using the *da_model4*, achieving an improvement of 10% in terms of absolute WER in test_call_center_3, compared with the baseline model. Table 2 also shows the 95% confidence intervals of the results obtained according to the method defined in [35]. The confidence intervals of the baseline and the *da_model4* do not overlap for any of the data sets, which shows that the improvement is statistically significant in all cases. These results strongly suggest that,

besides producing better results on simulated data, these strategies also provide important improvements in real data.

Table 2. Results in real data obtained from three different call centers obtained with the baseline system and with the best data augmentation strategy (*da_model4*) found using simulated data. Results show the WER and the 95% confidence interval.

Data Set	Duration (h)	WER% [95% Conf. Int.]	
		Baseline	da_model4
test_call_center_1	3.5	39.51[37.89, 41.15]	32.07[30.63, 33.53]
test_call_center_2	3	40.68[37.16, 44.21]	34.20[31.30, 37.13]
test_call_center_3	2	47.83[45.77, 49.90]	37.94[36.26, 39.62]

Direct comparison of our results with other results published in the literature is not possible due to differences in the databases used and in our particular approach. In particular, we assume that information about packet losses is not available as is customary in typical PCL settings. A notable exception to this is audio inpainting, which can be applied without knowing the position of the missing audio (blind audio inpainting) [16,20], but these works do not use data augmentation and speech recognition for evaluation. Despite the difficulties in making comparisons, we have found that our results are in line with a few recently published works.

The closest work we have found is [17], where the authors evaluate a PCL system using ASR measured in terms of WER. Although the comparison is not possible since the database is different (Librispeech read speech in English in [17] vs. Fisher conversational telephone speech in Spanish in this work), they found that there is a large improvement in terms of WER when applying data augmentation techniques (no additional information is provided about the data augmentation techniques used), which allows going from a baseline 19.99% WER to a 7.43% WER on data with packet losses. Using the advanced PCL technique proposed in [17], they reach a smaller improvement, lowering the WER to 6.19%. The difference in terms of WER is due to the difficulty of the data (read speech in [17] vs. conversational telephone speech in our case). In any case, and even with these differences, results are similar in the sense that they show the importance and capability of data augmentation techniques to deal with packet losses, and also that, by using a specific PCL approach, results can be further improved (although to a lesser extent).

Other related work [31] presents a study of different training approaches for dealing with packet losses in a different task, emotion recognition in speech, showing that matched training (training with the same packet losses degradation as in test) is the best option if that is possible, and data augmentation (including packet losses at different rates in training) is an effective way to improve results when packet losses are present, but the packet loss rate is not known in advance. Although the task is different, again the conclusion is similar to our study.

An important difference between our study and the aforementioned studies is that, in our case, we also evaluate our results in real data, while studies such as [17,31] only present results on artificially introduced packet losses.

4. Conclusions

In this work, we have tried to mitigate reduced Automatic Speech Recognition (ASR) accuracy when working with low bit-rate codecs and audios including packet losses, by applying data augmentation techniques. During this study, it has been found that data augmentation can improve robustness in simulated data in these types of scenarios, which can be very common in real call center recordings. On the one hand, the experiments have demonstrated the effectiveness of data augmentation when it is necessary to work with several low bit-rate codecs or audios containing relatively infrequent individual packet losses, achieving a huge improvement compared to the baseline model. On the other hand, on audios containing frequent packet losses or burst packet losses, the data augmentation

techniques improve the results in terms of Word Error Rate (WER), but there is still an important gap in performance. For that reason, other alternatives have been analyzed, showing that, when the audios have burst packet losses or many individual packet losses, a combination of a Packet Loss Concealment (PLC) system (such as ITU-T G.711) and a data augmentation strategy can provide further improvements. Finally, we have checked that the best data augmentation strategy found on data including simulated distortions also provides important improvements in real data taken from three different call centers.

As future work, considering the very promising results obtained, we plan to continue studying techniques to recover signals containing packet losses without the limitation of exactly knowing where the packet losses are, with the goal to achieve better accuracy in ASR applied to real recordings, and mainly focusing on cases with very frequent losses or burst losses that still have a large impact on ASR performance.

Author Contributions: The individual contributions are provided as follows. Contributions: Conceptualization, M.P.F.-G. and D.T.T.; methodology, M.P.F.-G. and D.T.T.; software, M.P.F.-G.; validation, M.P.F.-G. and D.T.T.; formal analysis, M.P.F.-G. and D.T.T.; investigation, M.P.F.-G.; resources, D.T.T. and M.P.F.-G.; data curation, M.P.F.-G.; writing—original draft preparation, M.P.F.-G.; writing—review and editing, D.T.T.; visualization, M.P.F.-G.; supervision, D.T.T.; project administration, D.T.T.; funding acquisition, D.T.T. All authors have read and agreed to the published version of the manuscript.

Funding: Partly funded by project RTI2018-098091-BI00, Ministry of Science, Innovation and Universities (Spain) and the European Regional Development Fund (ERDF).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data used in this study are available at <https://catalog.ldc.upenn.edu/LDC2010S01> with the exception of real call center data, which cannot be made publicly available due to confidentiality constraints.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASR	Automatic Speech Recognition
VoIP	Voice over IP
PLC	Packet Loss Concealment
GMM	Gaussian Mixture Models
HMM	Hidden Markov Models
DNN	Deep Neural Network
TDNN	Time Delay Neural Network
MFCCs	Mel Frequency Cepstral Coefficients
WER	Word Error Rate
MOS	Mean Opinion Score

References

1. Takahashi, A.; Yoshino, H.; Kitawaki, N. Perceptual QoS assessment technologies for VoIP. *IEEE Trans. Audio Speech Lang. Process.* **2004**, *42*, 28–34. [\[CrossRef\]](#)
2. Perkins, C.; Hodson, O.; Hardman, V. A survey of packet loss recovery techniques for streaming audio. *IEEE Netw.* **1998**, *12*, 40–48. [\[CrossRef\]](#)
3. Wah, B.W.; Su, X.; Lin, D. A survey of error-concealment schemes for real-time audio and video transmissions over the Internet. In Proceedings of the International Symposium on Multimedia Software Engineering, Taipei, Taiwan, 11–13 December 2000. [\[CrossRef\]](#)
4. Recommendation G.711 Appendix I (09/99); G.711: A high quality low-complexity algorithm for packet loss concealment with G.711; ITU Telecommunication Standardization Sector: Paris, France, 1999.
5. Recommendation G.729 (03/96); Coding of speech at 8 kbits using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP); ITU Telecommunication Standardization Sector: Paris, France, 2007.

6. Kovesi, B.; Ragot, S. A low complexity packet loss concealment algorithm for ITU-T G. 722. In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 31 March–4 April 2008; pp. 4769–4772.
7. Lindblom, J.; Hedelin, P. Packet loss concealment based on sinusoidal modeling. In Proceedings of the 2002 IEEE Speech Coding Workshop: A Paradigm Shift Toward New Coding Functions for the Broadband Age, SCW 2002, Ibaraki, Japan, 9 October 2002. [\[CrossRef\]](#)
8. Aoki, N. Modification of two-side pitch waveform replication technique for voip packet loss concealment. *IEICE Trans. Commun.* **2004**, *87*, 1041–1044.
9. Nielsen, J.K.; Christensen, M.G.; Cemgil, A.T.; Godsill, S.J.; Jensen, S.H. Bayesian interpolation in a dynamic sinusoidal model with application to packet-loss concealment. In Proceedings of the European Signal Processing Conference, Aalborg, Denmark, 23–27 August 2010.
10. Lindblom, J.; Samuelsson, J.; Hedelin, P. Model based spectrum prediction. In Proceedings of the 2000 IEEE Workshop on Speech Coding—Proceedings: Meeting the Challenges of the New Millennium, Delavan, WI, USA, 17–20 September 2000. [\[CrossRef\]](#)
11. Rødbro, C.A.; Murthi, M.N.; Andersen, S.V.; Jensen, S.H. Hidden Markov model-based packet loss concealment for voice over IP. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 1609–1623. [\[CrossRef\]](#)
12. Borgström, B.J.; Borgström, P.H.; Alwan, A. Efficient HMM-based estimation of missing features, with applications to packet loss concealment. In Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010, Makuhari, Chiba, Japan, 26–30 September 2010.
13. Lee, B.K.; Chang, J.H. Packet Loss Concealment Based on Deep Neural Networks for Digital Speech Transmission. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *24*, 378–387. [\[CrossRef\]](#)
14. Lotfidereshgi, R.; Gournay, P. Speech Prediction Using an Adaptive Recurrent Neural Network with Application to Packet Loss Concealment. In Proceedings of the ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, 15–20 April 2018. [\[CrossRef\]](#)
15. Mohamed, M.M.; Schuller, B.W. Concealnet: An end-to-end neural network for packet loss concealment in deep speech emotion recognition. *arXiv* **2020**, arXiv:2005.07777.
16. Kegler, M.; Beckmann, P.; Cernak, M. Deep speech inpainting of time-frequency masks. *arXiv* **2019**, arXiv:1910.09058.
17. Lin, J.; Wang, Y.; Kalgaonkar, K.; Keren, G.; Zhang, D.; Fuegen, C. A Time-Domain Convolutional Recurrent Network for Packet Loss Concealment. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 7148–7152. [\[CrossRef\]](#)
18. Marafioti, A.; Perraudin, N.; Holighaus, N.; Majdak, P. A context encoder for audio inpainting. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 2362–2372. [\[CrossRef\]](#)
19. Chang, Y.L.; Lee, K.Y.; Wu, P.Y.; Lee, H.Y.; Hsu, W. Deep long audio inpainting. *arXiv* **2019**, arXiv:1911.06476.
20. Shi, Y.; Zheng, N.; Kang, Y.; Rong, W. Speech loss compensation by generative adversarial networks. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 347–351.
21. Zhou, H.; Liu, Z.; Xu, X.; Luo, P.; Wang, X. Vision-infused deep audio inpainting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 283–292.
22. Marafioti, A.; Majdak, P.; Holighaus, N.; Perraudin, N. GACELA: A Generative Adversarial Context Encoder for Long Audio Inpainting of Music. *IEEE J. Sel. Top. Signal Process.* **2020**, *15*, 120–131. [\[CrossRef\]](#)
23. Pascual, S.; Serra, J.; Pons, J. Adversarial auto-encoding for packet loss concealment. In Proceedings of the 2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 17–20 October 2021; pp. 71–75.
24. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M. The KALDI speech recognition toolkit. In Proceedings of the IEEE 2011 workshop on automatic speech recognition and understanding, Big Island, HI, USA, 11–15 December 2011.
25. Watanabe, S.; Hori, T.; Karita, S.; Hayashi, T.; Nishitoba, J.; Unno, Y.; Soplin, N.E.Y.; Heymann, J.; Wiesner, M.; Chen, N.; et al. ESPnet: End-to-End Speech Processing Toolkit. *arXiv* **2018**, arXiv:1804.00015.
26. Yu, D.; Seltzer, M.L.; Li, J.; Huang, J.T.; Seide, F. Feature learning in deep neural networks—studies on speech recognition. In Proceedings of the International Conference on Learning Representation, Scottsdale, AZ, USA, 2–4 May 2013.
27. Ko, T.; Peddinti, V.; Povey, D.; Khudanpur, S. Audio augmentation for speech recognition. In Proceedings of the INTERSPEECH 2015—16th Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015; pp. 3586–3589.
28. Ko, T.; Peddinti, V.; Povey, D.; Seltzer, M.L.; Khudanpur, S. A study on data augmentation of reverberant speech for robust speech recognition. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5220–5224.
29. Park, D.S.; Chan, W.; Zhang, Y.; Chung-Cheng Chiu, B.Z.; Cubuk, E.D.; Le, Q.V. SpecAugment: A simple data augmentation method for automatic speech recognition. In Proceedings of the INTERSPEECH 2019—The 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15–19 September 2019; pp. 2613–2617. [\[CrossRef\]](#)
30. Hailu, N.; Siegert, I.; Nürnberger, A. Improving Automatic Speech Recognition Utilizing Audio-codecs for Data Augmentation. In Proceedings of the 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), Tampere, Finland, 21–24 September 2020; pp. 1–5. [\[CrossRef\]](#)

31. Mohamed, M.M.; Schuller, B.W. “I have vxxx bxx connexxxn!”: Facing Packet Loss in Deep Speech Emotion Recognition. *arXiv* **2020**, arXiv:2005.07757.
32. Peddinti, V.; Povey, D.; Khudanpur, S. A time delay neural network architecture for efficient modeling of long temporal contexts. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015.
33. ETSI EN 300 961; Digital cellular telecommunications system (Phase 2+) (GSM); Full rate speech; Transcoding (GSM 06.10 version 8.1.1 Release 1999); European Telecommunications Standards Institute: Sophia Antipolis, France, 2000.
34. Noll, P. MPEG digital audio coding. *IEEE Signal Process. Mag.* **1997**, *14*, 59–81. [[CrossRef](#)]
35. Bisani, M.; Ney, H. Bootstrap estimates for confidence intervals in ASR performance evaluation. In Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, QC, Canada, 17–21 May 2004; Volume 1, p. I-409.
36. Malfait, L.; Berger, J.; Kastner, M.P. 563—The ITU-T standard for single-ended speech quality assessment. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 1924–1934. [[CrossRef](#)]