A Survey of the Effects of Data Augmentation for Automatic Speech Recognition Systems

Jose Manuel Ramirez(✉), Ana Montalvo(✉), and Jose Ramon Calvo(✉)

Advanced Technologies Application Center, 7th A Street, #21406, Havana, Cuba
{jsanchez,amontalvo,jcalvo}@cenatav.co.cu
https://portal.cenatav.co.cu/

Abstract. Data augmentation has been proposed as a method to increase the quantity of training data. It is a common strategy adopted to avoid over-fitting, reduce mismatch and improve robustness of the models. But, would the system performance improve if we add data of any nature? This paper presents a survey about data augmentation techniques and its effect on Automatic Speech Recognition systems, some experiments were carried out to support the hypothesis that adding noise is not allways help.

Keywords: Data augmentation · Speech recognition

1 Introduction

Data augmentation is a popular technique for increasing the size of labeled training sets by applying class-preserving transformations to create copies of labeled data points [3]. Data augmentation in Automatic Speech Recognition (ASR) is an effective method to reduce mismatch between training and testing samples, improve robustness of the models and to avoid over-fitting.

Which are the most used techniques to get your data augmented? In which stage of an ASR system should be added more data? Are there any available tools or databases to get training data transformed and augmented? How much does ASR performance benefits from data augmentation?

Mostly motivated by deep learning revolution and its data greedy approach [11], data augmentation has played an important role in ASR, where the main focus of research has been on designing better network architectures that tend to over-fit easily and require large amounts of labeled training data [2].

Data augmentation strategy depends strongly on the ASR architecture. The simplicity of "end-to-end" models and their recent success in neural machine-translation have prompted considerable research into replacing conventional ASR architectures with a single "end-to-end" model, which trains the acoustic and language models jointly rather than separately. Recently, state-of-the-art results have been achieved [2] using an attention-based encoder-decoder model

I. Nyström et al. (Eds.): CIARP 2019, LNCS 11896, pp. 669–678, 2019.
https://doi.org/10.1007/978-3-030-33904-3_63

trained on over 12K h of speech data. However, on large publicly available corpora, such as "Librispeech" or "Fisher English", which are one order of magnitude smaller, performance still lags behind that of conventional systems.

The goal of data augmentation in "end-to-end" ASR systems is to leverage much larger text corpora alongside limited amounts of speech datasets to improve performance. Various methods of leveraging these text corpora have improved "end-to-end" ASR performance [21], for instance: composes recurrent neural network output lattices with a lexicon and word-level language model, while [1] simply re-scores beams with an external language model [13,20] incorporate a character-level language model during beam search, possibly disallowing character sequences absent from a dictionary, while [8] includes a full word level language model in decoding by simultaneously keeping track of word histories and word prefixes.

On the other hand there is conventional ASR architecture: hybrid Hidden Markov Model - Deep Neural Network (HMM-DNN), on which this paper is focused. In HMM-DNN based ASR systems, data augmentation can be done to improve the Acoustic Modeling (AM) or the Language Modeling (LM), and there are several ways to do so. The authors propose a taxonomy of the most used data augmentation techniques for HMM-DNN based ASR systems.

The remainder of the paper is organized as follows: Sect. 2 presents the taxonomy proposed by the authors relative to data augmentation methods for acoustic modeling in HMM-DNN based ASR systems. As part of the Sect. 3 data sets, tools and experimental setup are shown. Section 4 gives the experimental results and Sect. 5 is devoted to the conclusions of the paper.

2 Data Augmentation for HMM-DNN Based ASR Systems

Given augmenting data technique, an important question is how to best exploit the augmented data. The answer to this question will ultimately depend on the particular architecture the speech recognizer adopts and the nature and amount of augmented data used.

Concerning data augmentation methods to improve acoustic models, the following taxonomy is proposed:

- semi-supervised training,
- transformation of acoustic data,
- speech synthesis.

2.1 Semi-supervised Training

The semi-supervised training approach assumes the use of the text produced by an automatic speech recognition system to train acoustic models. In other words the unlabeled data may be adopted by recognizing it with an existing or boot-strapped system, filtering out those utterances that fail to pass confidence threshold [6,34] and re-training the system on supervised and filtered unlabeled training data.

The main advantage of this approach is that it is generally possible to collect vast amounts of such data, e.g., radio and television news broadcasts, covering all sorts of speaker and noise conditions [19].

The main disadvantage of this type of data is the lack of correct transcriptions. This limits possible gains from the approaches particularly sensitive to the accuracy of transcriptions supplied, such as discriminative training [31] and speaker adaptation based on discriminative criteria [32].

To date the majority of work has considered individual data augmentation schemes, with few consistent performance contrasts or examination of whether the schemes are complementary. In [29] two data augmentation schemes, semi-supervised training and vocal tract length perturbation, are examined and combined.

2.2 Transformation of Acoustic Data

The methods based on transformation of acoustic features include the variation of the Vocal Tract Length (VTL) on the stage of extracting the standard features [14] and its extended version to large vocabulary continuous speech recognition presented in [3]. In [3] instead of randomly choosing a warping factor for each utterance of a speaker, they deterministically perturb the estimated VTL warping factor of a speaker.

They also proposed a novel data augmentation approach based on Stochastic Feature Mapping (SFM) for utterance transformation. SFM estimates a maximum likelihood linear transformation in some feature space of the source speaker against the speaker dependent model of the target speaker. Different from vocal track and length perturbation (VTLP) which perturbs a speaker, SFM explicitly maps the features of a speaker to some target speaker based on a statistically estimated linear transformation.

In [16] they did research into an elastic spectral distortion method to artificially augment training samples to help HMM-DNNs acquire enough robustness even when there are a limited number of training samples. Three distortion methods were proposed: vocal tract length distortion, speech rate distortion, and frequency-axis random distortion.

The family of techniques based on acoustic data transformations includes methods such as audio signal speed alteration [18], applying noises, introduction of artificial reverberation into the records [24].

In [18] experiments are conducted with audio speed perturbation, which emulates a combination of pitch perturbation and VTLP, but it shows to perform better than either of those two methods. It is particularly recommended to change the speed of the audio signal, producing versions of the original signal with different speed factors.

In [23] the authors propose SpecAugment, an augmentation method that operates on the log mel spectrogram of the input audio, rather than the raw audio itself. This method is applied on Listen, Attend and Spell networks for end-to-end speech recognition tasks, which even when it is beyond this paper scope it is a very interesting proposal that could be evaluated over conventional ASR systems. SpecAugment consists of three kinds of deformations of the log mel spectrogram. The first is time warping, a deformation of the time-series in the time direction. The other two augmentations, proposed in computer vision [5], are time and frequency masking, where it is masked a block of consecutive time steps or mel frequency channels.

2.3 Speech Synthesis

The synthesized data may refer to existing but perturbed in a certain way data as well as new artificially generated data. One major advantage of synthesized data is that, similar to semi-supervised case, it is possible to collect vast amounts of such data. Another important advantage of synthesized datasets lies in the ability to approximate the required recognition conditions and get the necessary amount of training data. In addition, this method allows to obtain a precise alignment of noised data using known text transcriptions and the corresponding clean recordings.

Furthermore, different to semi-supervised case, the correctness of associated transcriptions is usually guaranteed. A major disadvantage of this type of data could be its quality.

Corrupting clean training speech with noise was found to improve the robustness of the speech recognizer against noisy speech. In [7,9], noisy audio has been synthesized via superimposing clean audio with a noisy audio signal.

The use of an acoustic room simulator has been explored in [17]. This paper describes a system that simulate millions of different utterances in millions of virtual rooms, and use the generated data to train deep-neural network models. This simulation based approach was employed on Google Home product and brought significant performance improvement.

3 Experimental Setup

In order to evaluate the impact of the augmented data on the effectiveness of an ASR task, we decided to measure the Word Error Rate (WER) obtained by several ASR systems facing the same decoding scenario. The five systems used in the experimentation: tri1, tri2, tri3, sgmm and dnn; were trained over the same training set, with the same LM and evaluated on the same set of test data; but in each system a new AM and a transformation in the feature space was tested. All systems share the same 3-gram LM and the acoustic feature. The acoustic feature selected was Mel Frequency Cepstral Coefficients (MFCC) [4] with a Mel Filter Bank of 40 filters (8 filter per octave), discarding the value of the energy of the frame for a total of 13 coefficients. This will allow knowing if the new noise data used in training enhance the performance of the ASR systems used in the experimentation.

The first three systems tri1, tri2 and tri3 have an Hidden Markov Model - Gaussian Mixture Model (HMM-GMM) triphone based architecture [33], this kind of statistical systems have the characteristic that after a certain amount of data the accuracy remained constant, unlike HMM-DNN that improve its accuracy when the amount of data increased. The feature extraction phase in system tri1 starts with the 13-dimensional MFCC feature, then Cepstral Mean and Variance Normalization (CMVN) [33] is applied over MFCC features and concatenated with their first and second order regression coefficients [33], to obtain a 39-dimensional vector. The feature extraction phase in system tri2 starts with 13-dimensional MFCC features that are spliced across ±4 frames to obtain 117-dimensional vectors. Linear Discriminant Analysis (LDA) [15] is applied to reduce the dimensionality to 40, using context-dependent HMM states as classes for the acoustic model estimation. Maximum Likelihood Linear Transform (MLLT) [28] is applied to the resulting features, making them more accurately modeled by diagonal-covariance Gaussians. The feature extraction phase in system tri3 is the same of tri2, but with one extra step before applying MLLT: a feature-space maximum likelihood linear regression (fMLLR) [27] is applied to normalize inter-speaker variability of the features. Systems tri1 and tri2 have similar AM, unlike tri3 that uses a speaker adaptive training (SAT) [33].

The fourth system (sgmm) have a triphonic Hidden Markov Model - Subspace Gaussian Mixture Model (HMM-SGMM) architecture [25]. The feature extraction phases of sgmm and tri3 are equals.

The last system (dnn) have a HMM-DNN architecture and its feature extraction phase is the same of tri3 and sgmm, but with three extra steps: before fMLLR comes a new spliced across ±4 frames, then another LDA over the spliced frames is applied and last another splice across ±4 frames for a 160-dimensional final vector.

3.1 Kaldi Toolkit

Kaldi is a set of free and open source tools, developed by Daniel Povey et al. [26] for research in ASR area. Kaldi allows to build ASR systems through a series of well-documented shell command routines. ASR systems in Kaldi are based on weighted finite state transducers that optimize training and decoding processes [22].

3.2 Speech Data

The TC-STAR project, funded by the European Commission, represents a long-term effort focused on advanced research in language technologies such as ASR, automatic speaker recognition, automatic speech translation and speech synthesis [10].

The TC-STAR recordings used in the experimentation, which we will call from now on TC_STAR_USED, correspond to sessions of the European Parliament or sessions of the Spanish Court where the announcers speak only Spanish. The recorded sessions of TC_STAR_USED come from different scenarios (two

Table 1. TC_STAR_USED corpus description.

	Training data	Test data
Sessions	17 163	1 908
Words	241 413	33 492
Unique words	15 722	4 178
Speakers	53 f 100 m	9 f 14 m
Hours	26:38:59	3:36:16

auditorium-like scenarios), it has multiple speakers of both genders (60 female and 112 male) and different ages, there are sessions of spontaneous speech and since some speakers talk at different times of the session or sometimes on different days, this adds variability between recorded sessions of the same speaker, which makes ASR task extremely complex and real. Criterion 90–10 was followed for the creation of training and test sets; where the training data consists of a set of recorded sessions corresponding to 90% of the total time (including silences) of the TC_STAR_USED data and the remaining 10% to the test data. The test data set shares 2 female and 2 male speakers with the training data set. Table 1 provides a summary of the characteristics of TC_STAR_USED where the term word refers to the spoken words, the interjections and sounds without linguistic information labelled in the database as noise or throat clearing. The format of the audio files is the standard RIFF (.wav) encoded PCM, 16 bits signed, at 16 kHz without compression.

3.3 Noise Database and Fant Tool-Kit

We used a noise database and a tool to simulate noisy conditions with different Signal-to-Noise Ratio (SNR) levels in the TC_STAR_USED train set, to evaluate the impact of augmenting training data in a ASR system over the WER in our clean test scenario. The noise database selected in this research has been DEMAND for more detail see description in [30]. All DEMAND noises can be classified into two groups, according to the nature of the noise type: in-door and out-door.

The selected tool to simulate the audio files of the test data set with the noises of the DEMAND database was FaNT - Filtering and Noised Adding Tool [12]. This tool allows us to add noise to speech sessions recorded with a desired SNR.

Using FaNT and the two groups of noises provided by DEMAND, we were able to augment the training set many times. This is possible by choosing a specific SNR value or range and one of the two groups of noises in DEMAND. For this research we used two range of SNR, the first one, 5 dB–15 dB, and the second one, 15 dB–25 dB. All this provided us with a four times bigger training data set, but just in duration because the amount of unique words and unique phrases said did not increase. Table 2 provides a summary of the characteristics of the new training data set, called TC_STAR_AUGMENTED.

Table 2. TC_STAR_AUGMENTED corpus description.

	Training data	Test data
Sessions	85 815	1 908
Words	1 207 065	33 492
Unique words	15 722	4 178
Speakers	53 f 100 m	9 f 14 m
Hours	133:12:25	3:36:16

4 Results

In this section we present the result of decoding the test data set with and without data augmentation. Tables 3 shows the difference between the WER of the decoding processes for all models tested, training with TC_STAR_USED and TC_START_AUGMENTED. Because the WER is an error metric, lower values are more desired than higher ones, positive differences mean that the efficiency of the recognition using data augmentation is better than when it is not used.

Table 3. Differences of WER (TC_STAR_USED - TC_START_AUGMENTED)

Systems	WER	INS	DEL	SUB
tri1	0.30%	0.20%	−0.58%	1.36%
tri2	0.26%	−0.48%	0.75%	−0.27%
tri3	0.32%	1.38%	−2.01%	0.64%
sgmm	0.11%	1.47%	−0.34%	−1.13%
dnn	**1.10%**	−0.92%	0.75%	0.17%

In data presented in Table 3 the most important difference occurs in the dnn model, where the impact of data augmentation was better. This behavior is consistent with what was described in the previous sections. The interesting thing about this experiment is that the increase in the accuracy of DNN-based models is small (and only due to the large amount of data involved, statistically relevant). Our hypothesis about the cause of this behavior is that the augmentation of the data using different noises did not contribute to having new phonetic realizations in the data, but rather redundancy. This would explain why the data augmentation method using noise is usually used to improve robustness of the recognition, because the network can "learn" about the nature of noise processing the augmented data; but this doesn't entail improving the recognition of clean signals.

5 Conclusions

Data augmentation is intended as a procedure that starting from your available data, multiplies the amount of it by producing versions of the original. Data augmentation strategy depends on the goal of the classification. For ASR all different ways of doing data augmentation could be put in one of the three categories proposed.

From the experiments carried out, it can be concluded that noise-based data augmentation methods are not suitable for raising the recognition rate in ASR systems, because from the data augmented the network learns more about the nature of the noise than about the phonetic combinations present in the utterances; this knowledge allows a system to deal better with different acoustic conditions but does not help to obtain better recognition rates.

From the results obtained it can be inferred that data augmentation procedures based on the simulation of acoustic conditions different from those present in the training set, for example simulating different acoustic channels or increasing the reverberation in the new signals, will not have an impact on the recognition rate of ASR systems. Future studies, with other data augmentation methods, might allow us to generalize our hypothesis.

References

1. Chan, W., Jaitly, N., Le, Q.V., Vinyals, O.: Listen, attend and spell: a neural network for large vocabulary conversational speech recognition. In: ICASSP (2016). http://williamchan.ca/papers/wchan-icassp-2016.pdf
2. Chiu, C.C., et al.: State-of-the-art speech recognition with sequence-to-sequence models. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 2018. https://doi.org/10.1109/icassp.2018.8462105
3. Cui, X., Goel, V., Kingsbury, B.: Data augmentation for deep neural network acoustic modeling. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, vol. 23, pp. 5582–5586, May 2014. https://doi.org/10.1109/ICASSP.2014.6854671
4. Davis, S.B., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In: Readings in Speech Recognition, pp. 65–74. Morgan Kaufmann Publishers Inc., San Francisco (1990). http://dl.acm.org/citation.cfm?id=108235.108239
5. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout (2017)
6. Evermann, G., Woodland, P.: Large vocabulary decoding and confidence estimation using word posterior probabilities (2000)
7. Gales, M.J.F., Ragni, A., AlDamarki, H., Gautier, C.: Support vector machines for noise robust ASR. In: 2009 IEEE Workshop on Automatic Speech Recognition Understanding, pp. 205–210, November 2009. https://doi.org/10.1109/ASRU.2009.5372913

8. Graves, A., Jaitly, N.: Towards end-to-end speech recognition with recurrent neural networks. In: Proceedings of the 31st International Conference on International Conference on Machine Learning, ICML 2014, vol. 32, pp. II-1764–II-1772. JMLR.org (2014). http://dl.acm.org/citation.cfm?id=3044805.3045089
9. Hannun, A., et al.: Deep speech: scaling up end-to-end speech recognition (2014)
10. Van den Heuvel, H., Choukri, K., Gollan, C., Moreno, A., Mostefa, D.: TC-STAR: new language resources for ASR and SLT purposes. In: LREC, pp. 2570–2573 (2006)
11. Hinton, G., et al.: Deep neural networks for acoustic modeling in speech recognition. Sig. Process. Mag. **29**, 82–97 (2012)
12. Hirsch, H.G.: Fant-filtering and noise adding tool. Niederrhein Univ. Appl. Sci. (2005). http://dnt.kr.hsnr.de/download.html
13. Hori, T., Watanabe, S., Zhang, Y., Chan, W.: Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM. CoRR abs/1706.02737 (2017). http://arxiv.org/abs/1706.02737
14. Jaitly, N., Hinton, E.: Vocal tract length perturbation (VTLP) improves speech recognition. In: Proceedings of the 30th International Conference on Machine Learning (2013)
15. Kajarekar, S.S., Yegnanarayana, B., Hermansky, H.: A study of two dimensional linear discriminants for ASR. In: Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (Cat. No. 01CH37221), vol. 1, pp. 137–140, May 2001. https://doi.org/10.1109/ICASSP.2001.940786
16. Kanda, N., Takeda, R., Obuchi, Y.: Elastic spectral distortion for low resource speech recognition with deep neural networks. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, 8–12 December 2013, pp. 309–314 (2013). https://doi.org/10.1109/ASRU.2013.6707748
17. Kim, C., et al.: Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google home, pp. 379–383 (2017). http://www.isca-speech.org/archive/Interspeech_2017/pdfs/1510.PDF
18. Ko, T., Peddinti, V., Povey, D., Khudanpur, S.: Audio augmentation for speech recognition. In: INTERSPEECH (2015)
19. Lamel, L., Gauvain, J.L., Adda, G.: Lightly supervised and unsupervised acoustic model training. Comput. Speech Lang. **16**, 115–129 (2002). https://doi.org/10.1006/csla.2001.0186
20. Maas, A.L., Xie, Z., Jurafsky, D., Ng, A.Y.: Lexicon-free conversational speech recognition with neural networks. In: Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL) (2015)
21. Miao, Y., Gowayyed, M., Metze, F.: EESEN: end-to-end speech recognition using deep RNN models and WFST-based decoding. CoRR abs/1507.08240 (2015). http://arxiv.org/abs/1507.08240
22. Mohri, M., Pereira, F., Riley, M.: Speech recognition with weighted finite-state transducers. In: Benesty, J., Sondhi, M.M., Huang, Y.A. (eds.) Springer Handbook of Speech Processing. SH, pp. 559–584. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-49127-9_28
23. Park, D.S., et al.: SpecAugment: a simple data augmentation method for automatic speech recognition. CoRR abs/1904.08779 (2019). http://arxiv.org/abs/1904.08779
24. Peddinti, V., Chen, G., Povey, D., Khudanpur, S.: Reverberation robust acoustic modeling using i-vectors with time delay neural networks. In: INTERSPEECH (2015)

25. Povey, D.: A tutorial-style introduction to subspace Gaussian mixture models for speech recognition. Microsoft Research, Redmond, WA (2009)
26. Povey, D., et al.: The Kaldi speech recognition toolkit. Technical report, IEEE Signal Processing Society (2011)
27. Povey, D., Yao, K.: A basis representation of constrained MLLR transforms for robust adaptation. Comput. Speech Lang. **26**(1), 35–51 (2012)
28. Psutka, J.V.: Benefit of maximum likelihood linear transform (MLLT) used at different levels of covariance matrices clustering in ASR systems. In: Matoušek, V., Mautner, P. (eds.) TSD 2007. LNCS (LNAI), vol. 4629, pp. 431–438. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74628-7_56
29. Ragni, A., Knill, K.M., Rath, S.P., Gales, M.J.: Data augmentation for low resource languages (2014)
30. Thiemann, J., Ito, N., Vincent, E.: The diverse environments multi-channel acoustic noise database (DEMAND): a database of multichannel environmental noise recordings. In: Proceedings of Meetings on Acoustics, ICA 2013, vol. 19, p. 035081. ASA (2013)
31. Wang, L., Gales, M.J.F., Woodland, P.C.: Unsupervised training for mandarin broadcast news and conversation transcription. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, vol. 4, pp. IV-353–IV-356, April 2007. https://doi.org/10.1109/ICASSP.2007.366922
32. Wang, L., Woodland, P.C.: Discriminative adaptive training using the MPE criterion. In: 2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721), pp. 279–284, November 2003. https://doi.org/10.1109/ASRU.2003.1318454
33. Young, S.: HMMs and related speech recognition technologies. In: Benesty, J., Sondhi, M.M., Huang, Y.A. (eds.) Springer Handbook of Speech Processing. SH, pp. 539–558. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-49127-9_27
34. Zavaliagkos, G., Colthurst, T.: Utilizing untranscribed training data to improve performance. In: DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne, pp. 301–305 (1998)