Data Augmentation for Training of Noise Robust Acoustic Models

Tatiana Prisyach[1,2], Valentin Mendelev[2(✉)], and Dmitry Ubskiy[3]

[1] STC-Innovations Ltd., St. Petersburg, Russia
[2] Speech Technology Center, St. Petersburg, Russia
{prisyach,mendelev}@speechpro.com
[3] ITMO-University, St. Petersburg, Russia
ubskiy@speechpro.com

Abstract. In this paper we analyse ways to improve the acoustic models based on deep neural networks with the help of data augmentation. These models are used for speech recognition in a priori unknown possibly noisy acoustic environment (with the presence of office or home noise, street noise, babble, etc.) and may deal with both the headset and distant microphone recordings. We compare acoustic models trained on speech corpora with artificially added noises of different origins and reverberation. At various test sets, word recognition accuracy improvement over the baseline model trained on clean headset recordings reaches 45%. In real-life environments like a meeting room or a noisy open space, the gain varies from 10 to 40%.

Keywords: Data augmentation · Robust speech recognition · Deep neural network

1 Introduction

Recently, multilayer neural networks (deep neural networks, DNNs) have found a widespread use for acoustic modeling in speech recognition [1]. In many cases the DNNs demonstrate better generalization capabilities as compared with the conventional Gaussian mixture models (GMMs). But in the case where the conditions for training and testing (usage) of the DNN mismatch the recognition quality may degrade significantly. In order to compensate this mismatch, various techniques are used to increase the quality of the speech and decrease the influence of noises.

This research is concerned with methods to improve the DNN based acoustic models using bottleneck features [2] and speech data augmentation [3].

The initial training dataset includes clean headset recordings, whereas the trained acoustic model is intended to be used for recognition in noisy open space or in a meeting room.

The general problem which arises in the case where the training and testing corpora mismatch is to construct a recognition system which is robust to acoustic environment variability.

D.I. Ignatov et al. (Eds.): AIST 2016, CCIS 661, pp. 17–25, 2017.
DOI: 10.1007/978-3-319-52920-2_2

To solve that problem, techniques are utilized which compensate the mismatch between the testing and training corpora with the help of:

1. special features (application of noise robust features such as PNCC [4] and RASTA [5], feature normalisation [6], feature compensation—correction of features in the frequency domain—spectral subtraction [7], Wiener filtering [8]) or acoustic model parameters transformation (standard statistical techniques such as the maximum a posteriori (MAP) estimators [9], SAT+CMLLR [10]);
2. a priori knowledge about the environment (utilization of stereo data [11] to train the mapping from the noisy to clean speech; here the advantage depends on how close the training corpora is to the testing environment; multi-condition training, construction of noise dictionaries (cluster adaptive training, CAT [12]); combination of pre-trained acoustic models with the use of non-negative matrix factorisation (NMF [13]));
3. application of explicit and implicit noise models (vector Taylor series [14]);
4. addition of various kinds of noise with different SNRs, which may occur in the testing corpus (data augmentation) [15–17].

Many of the above approaches use a priori information to estimate the parameters for specific conditions and fail when no environment-specific data are present. The data augmentation based approach provides a considerable advantage because it works well even when no target data is available.

There are several ways to augment the training data:
semi-supervised training [15], multi-lingual training [18], transformation of acoustic data [19], speech synthesis [20,21].

The semi-supervised training approach assumes the use of the text produced by an automatic speech recognition system to train acoustic models. The advantage of this approach is that we are able to use, say, radio or TV broadcasts featuring various kinds of speakers and noises; the obvious drawback is the presence of recognition errors in the texts.

The 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.

The methods based on transformation of acoustic features include the variation of the vocal tract length on the stage of extracting the standard features [17] and stochastic feature mapping (SFM) [20].

The family of techniques based on recording transformations includes such methods as the audio signal speed alteration [19], applying noises, introduction of artificial reverberation into the records [22].

To transform the data we apply the artificial reverberation with the use of binaural room impulse response (BRIR) [21] and several kinds of noise (street noise, office or home noise, babble) with various signal-to-noise ratio (SNR). The initial training dataset includes headset recordings. The problem consists of training the acoustic model which can be applied both to headset and to distant microphone recordings under various noises and reverberation conditions. We

demonstrate that the bottleneck feature extractor trained on the augmented train datasets is more robust to the noise and increases the recognition accuracy.

In the second section, we describe acoustic features and the DNN structure used in training. The third section includes the description of the train and test datasets, as well as the datasets resulting from data augmentation. The fourth section presents the results and discussion of the study, and the conclusion follows in the fifth section.

2 Bottleneck Features and DNN Structure

The bottleneck features extracted from a multilayer neural network have found a wide use in automatic speech recognition systems. Such features have been successfully used in [23,24] to solve the recognition problem under the testing and training corpora mismatch conditions. All acoustic models in our presentation are trained on this kind of features. The bottleneck features are generated from the DNN which has a hidden layer of smaller dimension as compared with the other layers.

In this paper we consider two bottleneck feature extractors:

1. the extractor trained on the initial training dataset including clean headset voice records only;
2. the extractor trained on the same corpora after applying data augmentation.

In Fig. 1, the general structure of deep neural networks used for training is shown. The first DNN is trained on plain MFCC features [25] (the left and right context length is equal to 15) to produce the bottleneck features. The network contains four fully connected hidden layers of dimension 2048 and a bottleneck layer of dimension 80.

The second DNN is trained on bottleneck features with context of length 5 and left/right spacing 3. The network contains four fully connected layers of dimension 2048 and a final classification layer with 2857 outputs.

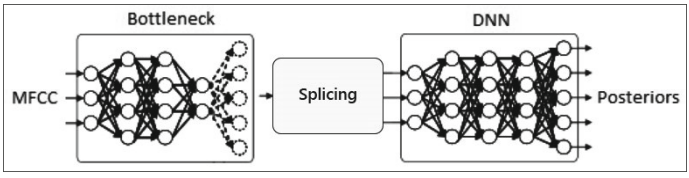

Fig. 1. The general DNN structure used to train the acoustic model

3 Speech Datasets for Training and Testing

In order to decrease the mismatch between the training and testing conditions, we make use of various transformations of the initial sound files preserving the state alignment unaltered. The difficulty consists in constructing a corpus which matches the reverberation and noise conditions which are unknown at the training phase. Since this objective is unattainable, we augment the training dataset with some variations to make our acoustic model more robust.

The training and test datasets are compiled from the recordings made by the Speech Technology Center. The sets contain phonetically rich sentences recorded with the use of a headset and distant microphones.

We consider the following ways to augment the training dataset:

1. application of noises corresponding to certain acoustic conditions (babble, office, home, car, street) with SNR from a fixed interval;
2. artificial reverberation of speech recordings.

For convenience we label the training datasets by abbreviations that reflect the properties of data containing in them. The training set `C` (clean data) contains only clean headset recordings of more than a thousand of different speakers. The set `NB` (noise, babble) includes a subset of recordings from `C` mixed with office, street, car noises and background speech (babble). The background recordings were scaled before mixing them with the clean data to produce the desired signal to noise ratio.

For artificial reverberation, we use BRIR, which contains the information about the size of the room where the recording is carried out, the distance to the sound source and its direction. BRIR includes three basic components:

$$h(t) = h_{\mathrm{dp}}(t) + h_{\mathrm{ee}}(t) + h_{\mathrm{rev}}(t),$$

where

$h_{\mathrm{dp}}(t)$ reproduces the sound passing directly from the source to the microphone; it depends on the azimuth and height of the source and the microphone; its energy decreases as r^2, where r is the distance between the source and the microphone;

$h_{\mathrm{ee}}(t)$ is the early echo related to reflection; it contains the information concerning the geometry of the room, its volume, number and positions of the walls;

$h_{\mathrm{rev}}(t)$ is the echo induced by reverberation, it contains a large number of reflections and dispersions of higher order.

We use two kinds of BRIR:

1. the distance from the source to the microphone is equal to 3 m, the azimuth is 0, the room parameters are $24 \times 15 \times 4.5$, which makes the reverberation time equal to 0.5 s;

Table 1. The description of the training and test datasets

Dataset	Duration, hours	SNR, dB	RT60, sec.
Train datasets			
`C (clean data)`	353	[15; 30]	[0.1; 0.3]
`R (reverb)`	222	[15; 30]	[0.5; 1.5]
`NB (noise, babble)`	250	[−5; 10]	[0.2; 1.5]
Test datasets			
`T1`	1.4	[15; 30]	[0.1; 0.3]
`T2`	1.4	[7; 10]	[0.1; 0.3]
`T3`	1.4	[−5; 7]	[0.1; 0.3]
`T4`	1.5	[15; 30]	[0.1; 0.5]
`T5`	0.4	[20; 45]	[0.1; 0.3]

2. the distance from the source to the microphone is equal to 5.5 m, the azimuth is 90, the room parameters are $24 \times 15 \times 4.5$, which makes the reverberation time equal to 0.8 s.

Detailed description of the training and test datasets is presented in Table 1.

The test datasets are divided into five groups based on the SNR and noise types. Each test dataset contains recordings of several dozens of speakers which were not included in the training sets. The first three groups contain the recordings made with the use of the close (`T1`), medium (`T2`) and long (`T3`) range microphone respectively. The environments are the office, domestic, and street. `T4` contains background speech. `T5` contains headset recording with a high SNR. The SNR is calculated as in [27] with decisions made by our voice activity detection (VAD) algorithm. RT_{60} denotes the reverberation time which is the time required for reflections of a direct sound to decay 60 dB.

The concluding table in this paper contains the results of comparison of acoustic models trained with the use of data augmentation on real-life datasets, which consist of recordings of dialogues in a meeting room and in a noisy open space at a peak rush of people. The recordings are characterized by a low SNR (10 dB on average), presence of background speech and noise of various kinds (the sales register printer, electronic queue alerts, phone rings, etc.).

The information concerning the datasets compiled from real-life data is presented in Table 2.

Table 2. The description of the train and test datasets derived from real recordings

Dataset	Type of microphone	SNR, dB	Rev-time, sec.
`R1`	Headset	[20; 35]	[0; 0.3]
`R2`	Distant (1 m)	[10; 15]	[0; 0.5]
`R3`	Distant (1 m)	[−10; 15]	[0; 0.6]

R1 and R2 are done at the same time and at the same place but with the use of different devices.

4 Experimental Results

In order to test the acoustic models which utilize the data augmentation techniques, we train several DNNs on bottleneck features. All networks contain 4 fully connected hidden layers of dimension 2048 and are trained with the use of discriminative pre-training [29]. In Table 3, we show how the word accuracy (recognition accuracy, WAcc) depends on the properties of the train datasets compiled with the use of clean, noisy and reverberated recordings. Only the most interesting results were included in Table 3.

Word accuracy is defined as follows:

$$WAcc = 1 - WER = \frac{N - S - D - I}{N},$$

where WER – word error rate, N is the number of words in the reference, S is the number of substitutions, D is the number of deletions, I is the number of insertions.

Table 3. The recognition accuracy dependence from the datasets properties

N	Name	Training data				Features			Test accuracy				
		C	R	NB	Hrs	MFCC	bn_C	bn_N	T1	T2	T3	T4	T5
1	Baseline	+			280	+			75.2	27.9	1.4	67.7	83.3
2	C_bn_C	+			353		+		78.9	51.2	8.4	78.6	87.2
3	CR_bn_C	+	+		575		+		83.7	63.5	25.5	78.7	86.3
4	CNBR_bn_C	+	+	+	825		+		**84.6**	73.6	41.1	76.9	86.9
5	CNBR_mfcc	+	+	+	825	+			82.8	74.1	46.3	78.9	**87.8**
6	CNBR_bn_N	+	+	+	825			+	84.3	**74.3**	**47.2**	**79.6**	87.2

As a baseline we used the model trained with plain MFCC features on a subset of the C dataset (280 of 353 h).

From the Table 3 it is obvious that adding augmented data improves recognition accuracy a lot and that bottleneck features are more robust to the speaker and environment variability.

In Table 4, comparison results on real-life test sets are given.

One can see that at different test cases the increase of the recognition accuracy as compared with the baseline model is substantial and varies from 12 to 40%.

The test set Real_3 is a more challenging one, so the recognition accuracy gain obtained with the proposed methods is less than on Real_2. Recordings

Table 4. The recognition accuracy on real-life test cases

N	Name	R1	R2	R3
1	Baseline	51.5	2.4	8.6
2	C_bn_C	60.5	14.4	25.7
3	CR_bn_C	62.2	29.7	37.9
4	CNBR_bn_C	62.3	34.4	**38.6**
5	CNBR_mfcc	59.5	38	37.9
6	CNBR_bn_N	**63.9**	**42.7**	37.9

in the `Real_3` contain specific kinds of noise which we didn't use during the augmentation process and background speech. The latter is loud enough to be passed by the voice activity detection algorithm so the acoustic models recognize it as they become more robust to noisy environment and since the reference texts contain only words belonging to a target speaker a larger number of insertions occurs. Some reduction in WER may be achieved with a VAD algorithm tuned to work in adverse noisy environments.

The presented recognition accuracy values are low but they allow to successfully perform keyword search and solve certain speech analytics tasks.

We publish a Kaldi recipe[1] for building a speech recognition system for the Russian language. It is based on publicly available speech corpus (Voxforge) and may well serve as a starting point to study data augmentation and other techniques aimed at producing effective ASR solutions.

5 Conclusions

In this research, it has been shown experimentally that the application of data augmentation methods increases substantially the robustness of the DNN-based acoustic models. The bottleneck features themselves are more robust to perturbations of acoustic conditions, but when the extractor is trained on the augmented datasets the recognition accuracy increases even more. The increase of the recognition accuracy has been found to be as high as 45% at some test cases. Experiments with real-life recordings in a quiet meeting room and in a noisy open space with low SNR demonstrate that even in the case where we have only clean recordings from a low-range microphone for training purposes, certain data transformations allow us to significantly increase the recognition accuracy.

Acknowledgements. This work was financially supported by the Ministry of Education and Science of the Russian Federation, Contract 14.579.21.0057 (ID RFMEFI57914X0057).

[1] https://github.com/freerussianasr/recipes.

References

1. Hinton, G., Deng, L., Yu, D., Dahl, G.E.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Sig. Process. Mag. **29**, 82–97 (2012)
2. Yaman, S., Pelecanos, J.W., Sarikaya, R.: Bottleneck features for speaker recognition. Odyssey **12**, 105–108 (2012)
3. Ragni, A., Knill, K.M., Rath, S.P., Gales, M.J.F.: Data augmentation for low resource languages. In: Proceedings of Interspeech 2014, pp. 810–814 (2014)
4. Kim, C., Stern, R.M.: Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring. In: Proceedings of ICASSP 2010, pp. 4574–4577 (2010)
5. Hermansky, H., Morgan, N., Bayya, A., Kohn, P.: Compensation for the effect of communication channel in auditory-like analysis of speech (RASTA-PLP). In: Proceedings of European Conference on Speech Technology 1991, pp. 1367–1370 (1991)
6. Viikki, O., Bye, D., Laurila, K.: A recursive feature vector normalization approach for robust speech recognition in noise. In: Proceedings of ICASSP 1998, pp. 733–736 (1998)
7. Boll, F.: Suppression of acoustic noise in speech using spectral subtraction. IEEE T-ASSP **27**(2), 113–120 (1979)
8. Mauuary, L.: Blind equalization in the cepstral domain for robust telephone based speech recognition. In: Proceedings of EUSPICO 1998, vol. 1, pp. 359–363 (1998)
9. Gauvain, J.-L., Lee, C.-H.: Maximum a posteriori estimation of multivariate Gaussian mixture observations of Markov chains. IEEE T-SAP **2**(2), 291–298 (1994)
10. Gales, M.J.F.: Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. **12**, 75–98 (1998)
11. Deng, L., Acero, A., Jiang, L., Droppo, J., Huang, X.D.: High-performance robust speech recognition using stereo training data. In: Proceedings of ICASSP 2001, pp. 301–304 (2001)
12. Gales, M.J.F.: Cluster adaptive training of hidden Markov models. IEEE T-SAP **8**(4), 417–428 (2000)
13. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Proceedings of NIPS 2000, pp. 556–562 (2000)
14. Deng, J., Li, L., Yu, D., Gong, Y., Acero, A.: High-performance HMM adaptation with joint compensation of additive and convolutive distortions via vector Taylor series. In: Proceedings of ASRU 2007, pp. 65–70 (2007)
15. Lamel, L., Gauvain, J.-L.: Lightly supervised and unsupervised acoustic model training. Comput. Speech Lang. **16**, 115–129 (2002)
16. Gales, M.J.F., Ragni, A., AlDamarki, H., Gautier, C.: Support vector machines for noise robust ASR. In: Proceedings of ASRU 2009, pp. 205–210 (2009)
17. Jaitly, N., Hinton, G.E.: Vocal tract length perturbation (VTLP) improves speech recognition. In: Proceedings of ICML 2013 (2013)
18. Burget, L., Schwarz, P., Agarwal, M., Akyazi, P.: Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models. In: Proceedings of ICASSP 2010, pp. 4334–4337 (2010)
19. Ko, T., Peddinti, V., Povey, D., Khudanpur, S.: Audio augmentation for speech recognition. In: Proceedings of Interspeech 2015 (2015)

20. Cui, X., Goel, V., Kingsbury, B.: Data augmentation for deep neural network acoustic modeling. In: Proceedings of ICASSP 2014 (2014)
21. Jeub, M., Schaefer, M., Vary, P.: A binaural room impulse response database for the evaluation of dereverberation algorithms. In: Proceedings of 16th International Conference on Digital Signal Processing (DSP), Santorini, Greece (2009)
22. Peddinti, V., Chen, G., Povey, D., Khudanpur, S.L.: Reverberation robust acoustic modeling using i-vectors with time delay neural networks. In: Proceedings of Interspeech 2015, pp. 2440–2444 (2015)
23. Yu, D., Seltzer, M.L.: Improved bottleneck features using pretrained deep neural networks. In: Proceedings of Interspeech 2011, pp. 237–240 (2011)
24. Karafiát, M., Grézl, F., Burget, L., Szőke, I., Černoský, J.: Three ways to adapt a CTS recognizer to unseen reverberated speech in BUT system for the ASpIRE challenge. In: Proceedings of Interspeech 2015, pp. 2454–2458 (2015)
25. Picone, J.W.: Signal modeling techniques in speech recognition. Proc. IEEE **81**(9), 1215–1247 (1993)
26. Dean, D.B., Kanagasundaram, A., Ghaemmaghami, H., Rahman, M., Sridharan, S.: The QUT-NOISE-SRE protocol for the evaluation of noisy speaker recognition. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association, Interspeech 2015, pp. 3456–3460 (2015)
27. Pollák, P.: Efficient and reliable measurement and simulation of noisy speech background. In: 2002 11th European Signal Processing Conference, pp. 1–4 (2002)
28. Löllmann, H.W., Yilmaz, E., Jeub, M., Vary, P.: An improved algorithm for blind reverberation time estimation. In: Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC) (2010)
29. McDermott, E., Hazen, T., Roux, J.L., Nakamura, A., Katagiri, S.: Discriminative training for large vocabulary speech recognition using minimum classification error. IEEE Trans. Speech Audio Process **15**(1), 203–223 (2007)