Noise, Device and Room Robustness Methods for Pronunciation Error Detection

1st Ville-Veikko Eklund
Tampere University
Tampere, Finland
ville.eklund@hotmail.com

2nd Aleksandr Diment
Tampere University
Tampere, Finland
aleksandr.diment@tuni.fi

3rd Tuomas Virtanen
Tampere University
Tampere, Finland
tuomas.virtanen@tuni.fi

***Abstract*—In this work, we address the problem of audio classification operating on signals recorded with various mobile devices in challenging environments. We propose a method for device, room and noise robust pronunciation error detection. It involves a data augmentation pipeline of convolution operations with room impulse responses and mobile device microphone impulse responses, and addition of background noise. A dataset of impulse responses of a diverse set of mobile devices, rooms and noises is collected. The method is evaluated in a pronunciation error detection task. The data consists of Finnish people uttering various English words accompanied by expert annotations of pronunciation errors. Classification accuracy is shown to improve by up to 12.9 percentage points as the amount of generated training data is increased. Given the large diverse set of collected impulse responses, we demonstrate that robustness is achieved consistently for new rooms and devices, excluded from the training set.**

***Index Terms*—data augmentation, robust classification, additive noise, impulse response, pronunciation error detection.**

I. Introduction

The quick development of machine learning methods has lead to an increasing need of large amounts of data. While collecting large datasets is a tedious and time-consuming task, the quality of data also has a great impact on the performance of a model. Obtaining real-life data is essential, when machine learning is being integrated with a growing rate into smartphones and other mobile devices.

The ability of a machine learning model to cope with noise and distortions, i.e. robustness, can be improved with a number of methods, one of which is data augmentation. In data augmentation, additional training data is created by altering the existing data for example by adding noise or by applying filters to it. A model trained with the augmented dataset is expected to be less susceptible to distortions and therefore more robust because the model learns to ignore unimportant variations.

The surrounding environment distorts an acoustic signal in a number of ways. These distortions can be divided into additive and convolutional noises [1]. The model is formulated as $y(m) = x(m) * h(m) + n(m)$, where $y(m)$ is the distorted signal, $x(m)$ is the clean signal, $h(m)$ is the convolutional noise or linear channel, $n(m)$ is the additive noise, m is the discrete time index and $*$ denotes convolution. Typically, the linear channel can be modelled with a room impulse response (RIR) and the additive noise with acoustic scene recordings.

Besides environmental distortions, a recording device can also distort a signal during its capture. All microphones have their own non-ideal frequency responses and the capture process may cause also introduce other kinds of distortions such as clipping, aliasing, and data loss [2, Chapter 3].

Several audio data augmentation techniques have been presented in the literature. The use of additive acoustic scene recordings had a positive impact on the accuracy of an environmental sound classifier in [3]. However, additive acoustic scenes did not improve significantly the performance of a musical instrument recognizer in [4]. Background noise consisting of different types of music, technical noises and non-technical noises from the MUSAN Noise dataset [5] was used in [6] to augment speech data. When tested against clean test data, additive noise lowered the character error rate only marginally, but the baseline was outperformed when evaluating with noisy data. In [7], even a small amount of additive Gaussian noise only increased the classification error in a singing voice detection task. Gaussian noise has not been lately used as much in augmentation of audio data as acoustic scenes, but it has been shown [8] to improve the generalization performance of other regression and classification problems.

RIRs were beneficial for a speech recognition task in reverberant environments [9]. When tested against reverberant test data, the word error rate (WER) decreased from 59.7 % to 41.9 % by convolving the training data with RIRs, while with non-reverberant test data the WER increased from 19.1 % to 26.2 % instead.

In [10] real room impulse responses yielded better results than simulated room impulse responses on a speech recognition task with several evaluation sets consisting of reverberated speech. When adding point-source noise to the augmentation routine, the performance gap between simulated and real impulse responses vanished. Combining clean and augmented data in the training set was noted to be more useful than using only augmented data.

Using simulated room impulse responses created from very basic room information improved the performance in speaker identification and mood detection tasks [11]. The evaluation data was collected in real reverberant environments and the system performed within 5%-10% of a non-reverberant baseline.

An impulse response of a smartphone microphone together with a RIR were used for convolutions in a musical instrument

recognition task [12]. For majority of the instruments in the task, the two-step convolution improved the performance of the recognizer over a nonaugmented baseline. However, robustness against new devices or rooms was not tested.

Other augmentation methods were successfully used in audio analysis tasks. Pitch shifting was shown helpful in a singing voice detection [7] and sound event classification [3]. Vocal tract length perturbation (VTLP) improved phoneme error rate in a speech recognition task by at least 0.5 %-points [13]. Dynamic range compression (DRC) was found [3] to be the most helpful technique in classification of gunshots, and was most harmful for classifying noise-like air conditioner sounds. In [7], the effect of dropout, loudness, random frequency filtering, and mixing were studied, showing that only random frequency filtering improved the performance of the detection system. Blocks mixing was also used in [14] to augment data for sound event detection, showing considerable performance improvement in car and stadium contexts. Speed perturbation was used in [15] with VTLP and time stretching in training a speech recognition system. Speed perturbation was found to lower the WER more than the other tested techniques.Stochastic feature mapping (SFM) was implemented in [16] to improve speech recognition of small languages with limited data. A GSM codec was used in [17] to emulate phone line channel effects on clean speech data with added background noise for the task of whispering detection. In multiple-width frequency-delta (MWFD) [18], delta features are extracted from spectrograms with varying widths to create additional data samples. MWFD with a convolutional neural network beat the compared baselines in nearly all acoustic scenes.

To our best knowledge, studies on robustness against distortions produced both by the capture devices and the environment have so far been performed without consideration of the diversity of the devices. In this paper, we propose a data augmentation pipeline to improve room, device and noise robustness of a pronunciation error detector. The method consists of convolution operations with room and device impulse responses as well as addition of background noise. We propose that given sufficient number of diverse devices and rooms in the dataset of impulse responses, robustness can be achieved towards new, unseen devices and rooms. The procedure of the impulse response collection and the augmentation pipeline are outlined in Section II, followed by evaluation in Section III, with conclusions drawn in Section IV.

II. Methodology

Making a machine learning model robust to environmental distortions and recording device effects requires a large dataset recorded with several devices in varying locations. To reduce the effort of collecting such a dataset, the distortions can be applied on clean data by convolution and summation, if the distortions are assumed to be linear and time-invariant.

The proposed method consists of simulating the process of recording a sound with a mobile device in varying locations. The process (Fig. 1) is split into three steps: convolution with a room IR, background noise addition, and convolution with a device IR.

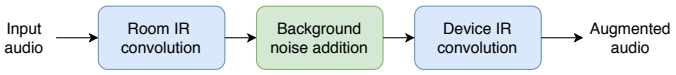

Fig. 1. Flow diagram of the proposed augmentation process.

Every time convolution or noise addition is performed, the IR sample or the background noise recording is selected randomly. To accommodate for the difference in duration of the payload signal and the background noise, the segment with the desired length is also randomly selected from the background noise recording. The magnitude of additive noise is controlled by and scaled according to a signal-to-noise ratio (SNR) drawn randomly from a set of predefined SNRs. The SNR of clean signals to the background noise segment are computed in terms of root mean square of the waveforms. Due to the importance of training on non-augmented data too [10], we use a probabilistic threshold to decide whether each of the operations is to be performed for each speech sample. For each operation, a random number between 0 and 1 is drawn from a uniform distribution, and the operation is applied if the number is within the interval $[0, 0.3]$.

To allow for generalisation to new rooms and devices, an IR dataset of sufficient diversity needs to be collected. The following sections outline the procedure.

A. Room measurements

The RIRs were measured with the Farina method [19] using a sweep length of 10 seconds and a frequency range from 80 Hz to 20 kHz. The following equipment was used: an Earthworks Audio M30 measurement microphone, a Genelec G Two loudspeaker, and a Focusrite Scarlett 18i20 1st gen audio interface.

RIRs were measured in five locations at Tampere University: small office, medium-size meeting room, medium-size lounge, large lecture hall, and a very large underground bomb shelter. The locations were selected to cover sufficient variation in their acoustical characteristics. The reverberation times were estimated to be from hundreds of milliseconds in the smaller rooms and medium rooms to several seconds in the bomb shelter. In each room, IRs were measured from five positions: one in each corner and one in the middle of the room. In the corner measurements, the microphone was in the corner facing the center of the room. The speaker was set to a distance of 100 cm towards the center facing the microphone. In the middle of the room the same distance was used. In each position, three measurements were made by changing the direction of the speaker between -15°, 0°, and 15° in the horisontal plane. A total of 78 RIRs were collected.

B. Mobile device measurements

The mobile device IRs were measured using the similar method and sweep settings as with the rooms. The capture was performed by passing the microphone signal of the mobile

devices through their output jacks, ensuring that no additional processing or compression was performed on the device. The signal repeated through the output jack was recorded externally to measure the impulse responses. The measurements were carried out in an anechoic chamber to minimize the contribution of the room. The equipment was the following: Genelec 1029A loudspeaker and a Focusrite Scarlett 2i2 2nd gen audio interface.

Eleven devices were used to measure IRs: Huawei Mate 10 lite, iPhone SE, iPhone 6S+, iPhone 8, LG G4, Motorola Moto C, Motorola Moto G (3rd gen), Samsung Galaxy J5, iPad Pro 12.9", iPhone 8 headset, and Huawei Mate 10 lite headset. IRs were mainly measured from eight positions with some exceptions with the headsets and the iPad. First, the loudspeaker was positioned in the corner of the anechoic chamber. A person standing behind the loudspeaker as if the speaker were their head held the phone in front of the speaker at a 30 cm distance first on chest level and then on mouth level. On both levels, the phone was held horizontally, with a 45° incline and vertically to measure a total of six impulse responses. Additionally, two measurements were performed with a person sitting at a table: first holding the phone in hand, and then having the phone lying on the table. The use of the person and the table in the measurements was motivated by the realistic reflections caused by them. Five people assisted in holding the phones to account for the diversity of reflections. Since most of the smartphones have multiple microphones, separate measurements were made with the same settings for each microphone. In total, 148 device IRs were collected.

C. Background noises

An ambient noise dataset was crowdsourced with mobile devices to be used in evaluation of the system. It consists of 715 five-second-long clips that were recorded by 80 unique subjects in uncontrolled locations using 41 unique smartphone models (12 iOS and 29 Android). The recorded clips contain mostly ambient noise and noises coming from handling the device, with occasional babble and music-type of noises present as well.

III. Evaluation

We evaluate the proposed method on a task of pronunciation error detection. Given an utterance of an English word made by a native Finnish speaker, the classifier is to detect whether a certain typical pronunciation error is present. In the following sections, we outline briefly the speech dataset and the classifier architecture. More details are found in the previous work [20].

A. Data

The dataset consists of recordings of 120 mostly Finnish subjects pronouncing 80 different English words two or three times. The words were selected by English teachers to contain most of the errors Finnish speakers make when speaking British English. The data was collected in a noise-insulated room with a reverberation time of 0.26 s and dimensions 4.53 m $\times$ 3.96 m $\times$ 2.59 m. A Røde NT55 condenser microphone and a Focusrite Scarlett 2i2 audio interface were used to record the audio with a 40 cm distance from the speaker and a 44.1 kHz sampling rate.

TABLE I
Room experiment results.

description	room	accuracy
Aug. train, aug. test	Bomb shelter (VL)	0.845
	Lecture hall (L)	0.857
	Living room (M)	0.847
	Meeting room (M)	0.848
	Office (S)	0.855
Aug. train, clean test	All	0.890
Clean train, clean test	All	0.886
Clean train, aug. test	All	0.762
Maj. class predictor	All	0.737

Each of the samples was assigned a label indicating whether the target phoneme in the uttered word is pronounced correctly or with a specific error. Secondary errors were not taken into account due to their scarcity. Samples with disagreeing annotations between the two annotators were discarded to have a fixed ground truth for all samples. Five words were selected for evaluation based on the availability of sufficient data of both erroneous and correct samples: hit, job, join, pull, and worse. Four different target phonemes appear in the selected words.

B. Classifier

The classifier architecture consisted of an recurrent neural network (RNN) with three bidirectional long short-term memory (BiLSTM) layers with 100 nodes in each and an output layer with a single node using sigmoid activation. BiLSTM was used because pronunciation errors may affect both the preceding and the following parts of the words, and BiLSTM allows information to flow between past and future frames. Mel-frequency cepstral coefficients (MFCC) with 128 mel bands and 20 coefficients were used with standardization and zero padding to maximum length of samples. A five-run Monte Carlo cross-validation setup was implemented with 0.6-0.2-0.2 split intro training, validation and test subsets.

C. Experiments

Four experiments were designed to evaluate the proposed method with a focus on (1) rooms, (2) backgrounds, (3) devices, and (4) all three steps combined into a pipeline. In addition to using original test data, the test data was augmented to simulate noisy test conditions. To facilitate a fair evaluation, when creating training, validation and test subsets, we ensured that the information leakage was avoided both for the speech and augmentation data. Such partitioning was performed in the following manner. The underlying speech data was split by speakers. Background noise samples were split based on the unique user id's of the people recording them. RIRs were split based on the recording rooms (experiment 2) and the measurement points inside the rooms (experiment 4). Device IRs were split based on the device models. Headsets were treated as their own categories.

TABLE II
DEVICE EXPERIMENT RESULTS.

description	device	acc.
Aug. train, aug. test	Apple headset	0.848
	Huawei Mate 10 Lite	0.859
	Huawei headset	0.855
	LG G4	0.841
	Motorola Moto C	0.851
	Motorola Moto G	0.854
	Samsung Galaxy J5	0.852
	iPad Pro 12.9”	0.855
	iPhone 6S Plus	0.856
	iPhone 8	0.846
Aug. train, clean test	All	0.897
Clean train, clean test	All	0.886
Clean train, aug. test	All	0.745
Maj. class predictor	All	0.737

TABLE III
ADDITIVE NOISE EXPERIMENT RESULTS.

description	acc.
Aug. train, clean test	**0.892**
Clean train, clean test	0.886
Aug. train, aug. test (mism. SNRs, test less noisy)	0.863
Aug. train, aug. test (matched SNRs)	0.848
Clean train, aug. test	0.758
Majority class predictor	0.737

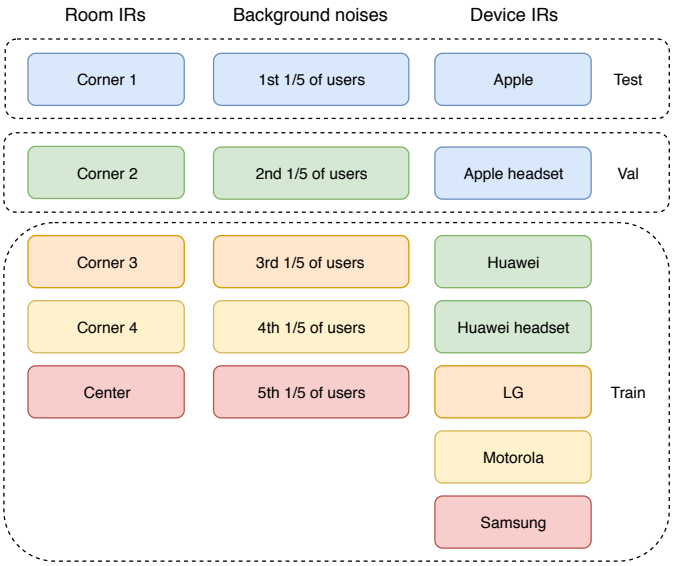

Fig. 2. Partitioning of the augmentation data in one of the runs of the combined experiment.

For each speech sample, the IR convolutions were performed with a 30% probability drawn from a uniform distribution. The background noise addition operation was performed for each speech sample with a target SNR value selected randomly and uniformly from the pre-defined list of SNRs. In the combined augmentation experiment (4), a measurement point split, id split, and a device manufacturer split was used for rooms, noises, and devices, respectively.

In the first experiment, the effect of augmentation by convolution with RIRs on the classifier performance was studied. Robustness against unseen rooms was tested in mismatched conditions: for each test room, the training set was augmented with the remaining RIRs. The effect of augmenting only the training data was also evaluated and compared with a case where only the test data was augmented. As a baseline, the classifier was trained and tested without augmentation. In addition, a majority class predictor (zero rule classifier) performance was measured, since accuracy alone is not a sufficient metric for an unbalanced problem.

The results are presented in Table I. The letters in the room column denote the size of the room, from small (S) to medium (M), large (L), and very large (VL). Matched training without robustness requirement (the so-called clean train, clean test case) is among the easiest, as expected. Interestingly, applying RIRs for increasing the amount of training data without the robustness requirement (the so-called augmented train, clean test) yields even better performance than the matched clean conditions. Finally, with robustness requirements, despite mismatched conditions, reasonable and consistent performance is achieved across the test rooms, presumably due to the sufficient diversity of the RIRs in the training set that allow generalisation to new rooms.

The second experiment studies the effect of device IRs in a similar manner. The results are shown in Table II. Similar behaviour was observed: given the large and diverse set of device IRs used in training, the model becomes robust to unseen devices. No particularly challenging test device was found in the experiments, which suggests good generalisation properties of the method. As with rooms, performing the proposed augmentation was useful also without the robustness requirement, for the sake of increasing the diversity of training data (the augmented train, clean test case).

The third experiment consists of different additive noise scenarios. The seen/unseen noise test cases from previous experiments was replaced with matching/mismatching SNRs. The SNR values were randomly selected from 0, 6, 12, 24 and 96 dBs in all but the mismatched noise test case, where test SNRs were selected from 12 and 24 dBs. The varying SNR adds another dimension to the augmentation process. The results are presented in Table III. Augmenting the training set with additive noise improves the performance on the original test set. When augmenting both training and test sets, the performance expectedly depends on the test set SNR values.

The fourth experiment combines all implemented augmentation techniques into a three-step routine. All the word data subsets in each run are augmented with non-overlapping and rotating impulse response and background noise partitions as illustrated in Fig. 2 for the case of one of the runs. The room, background noise, and device partitions in test and validation sets are rotated for each run, while rest of the partitions are used for training. In the results shown in Table IV, the presented

TABLE IV
AUGMENTATION COUNT EXPERIMENT RESULTS.

augmentation count	accuracy	
	mean	std
1	0.759	0.014
5	0.818	0.023
10	0.833	0.009
30	0.878	0.021
50	0.878	0.020
100	0.888	0.009

standard deviations take into account only the variation between the five runs, which have been averaged across words. The augmentation count stands for to the number of replications of the original dataset performed with the proposed augmentation method. A count of one corresponds to the amount of data being the same as original. The test data was fixed to an augmentation count of 100. The classifier was trained with augmentation counts 1, 5, 10, 30, 50, and 100. A total gain of 12.9 percentage points was achieved with the augmentation count 100, with performance starting to saturate at the augmentation count of 30.

IV. CONCLUSIONS

A method for augmenting audio signals to have characteristics of having been recorded with mobile devices in various environments was presented. The method can improve the robustness of classification models to convolutional and additive noises with a focus on mobile device effects. It consists of convolving audio with a room impulse response, adding background noise, and convolving the result with a mobile device microphone impulse response. A system implementing the method was evaluated with a pronunciation error detector. The augmentation steps were studied individually and in a combined three-step augmentation routine.

Augmenting the data improved the performance of the classifier on both augmented and original data in all tests scenarios. Consistent robustness towards unseen rooms and devices was observed. The combined augmentation consistently improved performance until data was augmented to 30 times the original amount of data. Further research possibilities include using data generators to augment data on the fly enabling unlimited augmentations during training. The proposed method could also be evaluated with different learning tasks and datasets.

V. ACKNOWLEDGEMENTS

We would like to thank CSC—IT Center for Science, Finland, for providing the computing resources.

REFERENCES

[1] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*. Cambridge University Press, 1993.

[2] T. Virtanen, R. Singh, and B. Raj, *Techniques for Noise Robustness in Automatic Speech Recognition*. John Wiley & Sons, 2012.

[3] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[4] B. McFee, E. Humphrey, and J. Bello, "A software framework for musical data augmentation," in *16th International Society for Music Information Retrieval Conference (ISMIR)*, 2015.

[5] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.

[6] D. Liang, Z. Huang, and Z. C. Lipton, "Learning noise-invariant representations for robust speech recognition," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 56–63.

[7] J. Schlüter and T. Grill, "Exploring Data Augmentation for Improved Singing Voice Detection with Neural Networks," in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)*, 2015.

[8] G. An, "The effects of adding noise during backpropagation training on a generalization performance," *Neural Computation*, vol. 8, no. 3, pp. 643–674, 1996.

[9] M. Ritter, M. Mueller, S. Stueker, F. Metze, and A. Waibel, "Training deep neural networks for reverberation robust speech recognition," in *Speech Communication; 12. ITG Symposium*, 2016, pp. 1–5.

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

[11] R. F. Dickerson, E. Hoque, P. Asare, S. Nirjon, and J. A. Stankovic, "Resonate: reverberation environment simulation for improved classification of speech models," in *IPSN-14 Proceedings of the 13th International Symposium on Information Processing in Sensor Networks*, 2014, pp. 107–117.

[12] S. Bhardwaj, "Audio data augmentation with respect to musical instrument recognition," Master's thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2017.

[13] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013.

[14] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6440–6444.

[15] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[16] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 9, pp. 1469–1477, 2015.

[17] A. Diment, T. Virtanen, M. Parviainen, R. Zelov, and A. Glasman, "Noise-Robust detection of whispering in telephone calls using deep neural networks," in *2016 24th European Signal Processing Conference (EUSIPCO)*, 2016.

[18] Y. Han and K. Lee, "Acoustic scene classification using convolutional neural network and multiple-width frequency-delta data augmentation," *CoRR*, vol. abs/1607.02383, 2016. [Online]. Available: http://arxiv.org/abs/1607.02383

[19] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio Engineering Society Convention 108*, 2000.

[20] A. Diment, E. Fagerlund, A. Benfield, and T. Virtanen, "Detection of typical pronunciation errors in non-native english speech using convolutional recurrent neural networks," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2019.