On Robustness of Cloud Speech APIs: An Early Characterization

Akhil Mathur
Nokia Bell Labs and
University College London
Cambridge, UK

Robert Smith
University College London
London, UK

Anton Isopoussu
Nokia Bell Labs
Cambridge, UK

Nicholas D. Lane
University of Oxford
Oxford, UK

Fahim Kawsar
Nokia Bell Labs
Cambridge, UK

Nadia Berthouze
University College London
London, UK

UbiComp/ISWC'18 Adjunct,, October 8–12, 2018, Singapore, Singapore
ACM 978-1-4503-5966-5/18/10.
https://doi.org/10.1145/3267305.3267505

Abstract

The robustness and consistency of sensory inference models under changing environmental conditions and hardware is a crucial requirement for the generalizability of recent innovative work, particularly in the field of deep learning, from the lab to the real world. We measure the extent to which current speech recognition cloud models are robust to background noise, and show that hardware variability is still a problem for real-world applicability of state-of-the-art speech recognition models.

Author Keywords

Audio Sensing; Machine Learning; Robustness; ASR

Introduction

Recent advances in machine learning algorithms have accelerated the development of consumer devices and sensing applications which aim to infer user context, activities and behavior from a variety of sensor data collected from a user. Particularly, audio sensing has emerged as a promising driver for inferring user context and behavior such as subjective states (e.g., emotion [22]), eating episodes (e.g., chewing [9]), and speech characteristics (e.g., speaker verification [25], keyword spotting [14]). Fortuitously, as audio-based inference models become more accurate, it is also becoming easier to deploy them in different real-world situations owing to the availability of low-cost embedded microphones and processors. For instance, it is straightforward to create a custom audio sensing device similar to Amazon Echo using an off-the-shelf low-cost microphone, an embedded platform such as Raspberry Pi, and cloud-based audio sensing models [4].

As sensory inference systems move out of the laboratory setting into the wild, questions remain about their robustness in unconstrained real-world scenarios [8, 12]. Particularly for audio inference models, it is critical to be robust to the following two types of variabilities:

Acoustic Environment Noise: Ideally, a user would expect an audio-sensing application to make accurate inferences irrespective of where and when it is used. However, the environment (e.g., cafe, gym, train station) and environmental conditions (e.g., raining, ambient music) in which an audio signal is captured add background noises to the signal that may confuse the underlying inference models and impact their accuracy. As such, one of the key desired properties for audio-based inference systems is their robustness in diverse acoustic environments.

Microphone Heterogeneity: Audio inference models, once developed, are expected to work on a diverse set of mobile and wearable devices, often from different manufacturers. This is challenging, because different manufacturers may use different hardware components (i.e., microphones) and may also have variations in the software that process the raw audio signal before exposing them to user applications. Therefore, inference models need to be robust against these forms of microphone heterogeneity.

In this paper, we investigate the robustness of state-of-the-art automatic speech recognition (ASR) models against these two forms of real-world noise. Our result show that ASR models can counter moderate amount of background noise, but show higher errors as the noise power increases. Unexpectedly, we also find significant variance in model performance when they are exposed to different microphones. Both these preliminary findings suggest the need for further research on improving the robustness of machine learning models in real-world scenarios.

Related Work

A number of works have shown how sensor data variability can impact the accuracy of mobile sensing models. Stisen et al. [23] studied sampling rate heterogeneity in inertial sensors of smart devices and how it impacts the accuracy of shallow HAR classifiers. Chon et al. [15] found that sound classification models show poor accuracies when deployed in unconstrained environments. Similar findings were shown by Lee et al. [17] about the adverse impact of acoustic environments on speaker turn-taking detection. Vision models are also impacted by environmental variabilities such as lighting conditions [27], various forms of object occlusion [13], and operation variabilities such as blurry, out-of-focus images due to unstable cameras.

Experiments

We now discuss our methodology for evaluating the robustness of audio models in real-world scenarios.

Audio Task and Dataset: We focus on Automatic Speech Recognition (ASR) as a representative audio processing task. ASR is a fundamental component of audio- or speech-processing systems and recent advances in the field of deep learning have significantly improved the performance of ASR models [16]. Our experiments are conducted on the Librispeech-clean [20] dataset, which is a widely-used ASR benchmark dataset for comparing the accuracy of different ASR models. We use 1000 randomly selected test audios from the Librispeech-clean dataset, with an average duration of 7.95

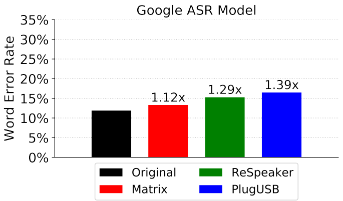

Figure 1: Impact of microphone variability on Google ASR model. Values on the bars illustrate the increase in WER over the Original audio WER (black bar).

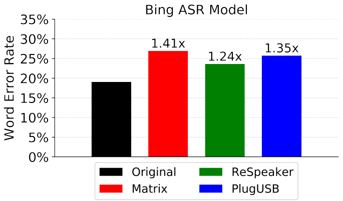

Figure 2: Impact of microphone variability on Bing ASR model. Values on the bars illustrate the increase in WER over the Original audio WER (black bar).

seconds and sampling rate of 16,000 Hz. In the rest of the paper, we refer to this dataset as *Librispeech-clean-1000*.

Experiment Conditions: As discussed earlier, our investigation of audio model robustness focuses on two key sources of noise observed in audio signals in real-world scenarios:

Microphone Heterogeneity: To evaluate how audio models cope against microphone variability, we needed to record a large-scale test dataset from different microphones under the same environment conditions. For this, we replayed the *Librispeech-clean-1000* dataset on a JBL LSR 305 monitor speaker [1] and recorded the entire dataset simultaneously on three different microphones namely Matrix Voice [5], ReSpeaker [7] and PlugUSB in a quiet environment. While the first two microphones are multi-channel microphone arrays commonly used in consumer devices such as Amazon Echo, the last microphone is a low-cost USB microphone compatible with embedded platforms such as Raspberry Pi. The microphones were kept at a distance of 10cm from the speaker in order to minimize the effect of room acoustics on the recorded audio. In effect, we created four variants of the *Librispeech-clean-1000* dataset, including the original dataset and the three recordings that we did with off-the-shelf embedded microphones.

Acoustic Environment noise: To simulate the effect of different acoustic environments, we mix the speech audios from Librispeech dataset with examples of real-world background noise taken from the ESC-50 dataset [21]. To this end, we randomly sampled 200 audios from the *Librispeech-1000* dataset and augmented them with background audios of *Rain* and *Wind* from the ESC-50 dataset.

ASR Models: We conducted our experiments on ASR models from Google (using the Google Cloud Speech API [2]) and Microsoft (using the Bing Speech API [1]). The models use a CNN-bidirectional LSTM model structure [26] and have shown near-human accuracy on ASR tasks [6, 3]. Audios from the *Librispeech-clean-1000* dataset under both experiment conditions were passed to the models through REST APIs, and Word Error Rate (WER) was computed on the ASR transcripts.

Results

Figures 1 and 2 show the effect of microphone variability on the accuracy of the ASR models. Firstly, we observe that for all three microphones, the word error rate (WER) increases over the baseline (i.e., the original Librispeech audios) by as high as 1.41 times. More importantly, the model performance varies across different microphones (e.g., from 1.24x to 1.41x WER increase in the case of Bing ASR model), which suggests that the ASR models are not completely robust to microphone variability.

Further, in Figure 3, we plot the spectrograms of an audio segment from the Librispeech-1000 dataset in its original form (3a) as well as when it is captured by different microphones (3b-d). Subtle variabilities in how different microphones capture the same audio signal can be observed from the figures, and we hypothesize that the ASR models are not trained to account for these variabilities, which in turn leads to varying levels of increase in the WER. In future work, we plan to deeply investigate the causes of these subtle variations in microphone response and explore how to make inference models robust to these variations.

Next, Figures 4 and 5 illustrate the findings on acoustic environment robustness. We varied the power of the background noise that is added to the speech signal (effectively

[1] We chose this speaker due to its flat frequency response in the human speech frequency range.

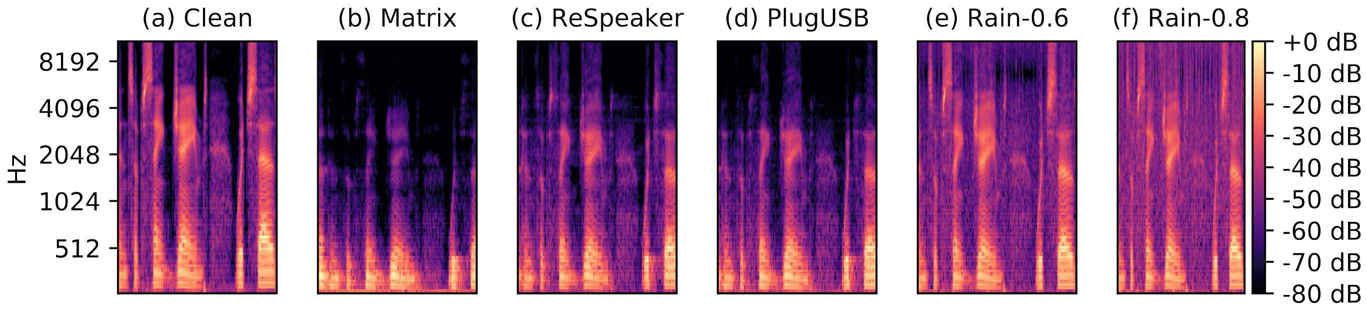

Figure 3: Mel-Scale Spectrograms of an audio segment under different experiment conditions.

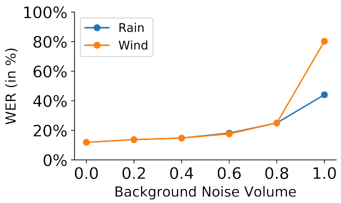

Figure 4: Effect of two types of background noise on Google ASR model.

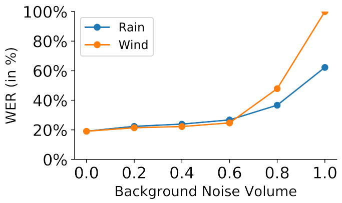

Figure 5: Effect of two types of background noise on Bing ASR model.

the signal-to-noise ratio) and measured the WER in each configuration. For example, background noise of 0.0 corresponds to the clean signal and background noise volume of 1.0 means that the signal and noise have the same power in the audio.

We observe that the ASR models can cope up with moderate amount of background noise – e.g., when the speech signal is mixed with 'Wind' and 'Rain' audios at 0.4 relative noise power, the increase in WER is less than 1.25x for both noise power, the increase in WER is less than 1.25x for both Google and Bing ASR models. However, when the relative noise power is increased to 0.8, the WER increases by more than 2x above the baseline for both the models.

Finally, we make the following observation on the comparative robustness of the ASR models to microphone variability and environment noise. In Figure 3, although the Rain-0.6 spectrogram (3e) looks visibly more noisy than the spectrograms collected from different microphones (3b-d), the performance of ASR models on Rain-0.6 dataset is similar to that on various microphones. This indicates that the ASR models are able to cope with background noise in the speech much better than the subtle variabilities caused by different microphones. Further research is needed to uncover the underlying causes behind this behavior.

Discussion and Conclusion

Our experiments show that deep ASR models are not robust to real-world noise caused by microphone variability and different acoustic environments. In this section, we broadly discuss the research directions that could be explored to solve this problem.

In the context of machine learning, the problems of microphone heterogeneity and environmental noise can be interpreted as instances of *dataset shift*[24] – in both cases, the training data does not accurately reflect the test data, violating a basic assumption made for machine learning models. Two broad solution approaches are used to address this problem, namely *domain adaptation* [11] and *domain generalisation* [10]. *Domain adaptation* attempts to address the problem by adapting an existing model by making use of either unlabeled data, or alternatively, small amounts of labeled data from the test domain. The latter scenario can be seen as an example of *transfer learning*. Methods that attempt to make the classifier behave consistently under dataset shift with no information about the test set fall under *do-*

main generalization. The easiest way to achieve this consistency is by finding features which are invariant under the dataset shift [19]. This could be done by designing specialized denoising algorithms which minimize the effect of noise sources on the learned features. Alternatively, the training of the speech recognition algorithm may itself be changed by augmenting the training data with a representative range of types of noise [18].

As a future work, we plan to explore solutions in both these areas to address the issue of audio model robustness. More generally, we plan to explore how domain adaptation and generalization could be applied to other sources of data variabilities observed in mobile sensing, such as variations in accelerometer data caused by how users carry their devices.

REFERENCES

1. Bing Speech API. `https://azure.microsoft.com/en-us/services/cognitive-services/speech/`.
2. Google Speech API. `https://cloud.google.com/speech-to-text/`.
3. Google speech recognition. `https://tinyurl.com/y7dm37vw/`.
4. Hardware to emulate amazon echo. `https://tinyurl.com/y84d6r2n/`.
5. Matrix Voice. `https://www.matrix.one/products/voice/`.
6. Microsoft speech recognition. `https://tinyurl.com/ybnm9zdj/`.
7. ReSpeaker. `https://respeaker.io/`.
8. O. Amft. On the need for quality standards in activity recognition using ubiquitous sensors. In *How To Do Good Research In Activity Recognition. Pervasive Workshop*, 2010.
9. O. Amft, M. Stäger, P. Lukowicz, and G. Tröster. Analysis of chewing sounds for dietary monitoring. In *Ubicomp*, pages 56–72. Springer, 2005.
10. G. Blanchard, G. Lee, and C. Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in neural information processing systems*, pages 2178–2186, 2011.
11. J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128, 2006.
12. H. Blunck et al. On heterogeneity in mobile sensing applications aiming at representative data collection. In *Proceedings of the 2013 ACM Ubicomp*, pages 1087–1098. ACM, 2013.
13. B. Chandler and E. Mingolla. Mitigation of effects of occlusion on object recognition with deep neural networks through low-level image completion. *Computational intelligence and neuroscience*, 2016, 2016.
14. G. Chen, C. Parada, and G. Heigold. Small-footprint keyword spotting using deep neural networks. In *ICASSP*, pages 4087–4091. IEEE, 2014.
15. Y. Chon et al. Understanding the coverage and scalability of place-centric crowdsensing. In *Ubicomp*, pages 3–12. ACM, 2013.
16. A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.
17. Y. Lee, C. Min, C. Hwang, J. Lee, I. Hwang, Y. Ju, C. Yoo, M. Moon, U. Lee, and J. Song. Sociophone: Everyday face-to-face interaction monitoring platform using multi-phone sensor fusion. In *Proceeding of Mobisys '13*, pages 375–388. ACM, 2013.
18. A. Mathur et al. Using deep data augmentation training to address software and hardware heterogeneities in wearable and smartphone sensing devices. In *IPSN*. IEEE, 2018.
19. K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *ICML*, pages 10–18, 2013.
20. V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *ICASSP*, pages 5206–5210. IEEE, 2015.
21. K. J. Piczak. Esc: Dataset for environmental sound classification. In *ACM Multimedia*, pages 1015–1018. ACM, 2015.
22. K. Rachuri et al. Emotionsense: a mobile phones based adaptive platform for experimental social psychology research. In *Proceedings of Ubicomp '10*, pages 281–290. ACM, 2010.
23. A. Stisen et al. Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition. In *Proceedings of Sensys*, pages 127–140. ACM, 2015.
24. M. Sugiyama, N. D. Lawrence, A. Schwaighofer, et al. *Dataset shift in machine learning*. The MIT Press, 2017.
25. E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez. Deep neural networks for small footprint text-dependent speaker verification. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 4052–4056. IEEE, 2014.
26. W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke. The Microsoft 2017 Conversational Speech Recognition System. *ArXiv e-prints*, Aug. 2017.
27. S. Yang, A. Wiliem, and B. C. Lovell. To face or not to face: Towards reducing false positive of face detection. In *Image and Vision Computing New Zealand (IVCNZ), 2016 International Conference on*, pages 1–6. IEEE, 2016.