Chapter 4
Towards the Design and Evaluation of Robust Audio-Sensing Systems

Akhil Mathur, Anton Isopoussu, Fahim Kawsar, Robert Smith, Nadia Berthouze and Nicholas D. Lane

Abstract As sensor-based inference models move out of laboratories into the real-world, it is of crucial importance that these models retain their performance under changing hardware and environment conditions that are expected to occur in-the-wild. This chapter motivates this challenging research problem in the context of audio sensing models, by presenting three empirical studies which evaluate the impact of hardware and environment variabilities on cloud-scale as well as embedded-scale audio models. Our results show that even the state-of-the-art deep learning models show significant performance degradation in the presence of ambient acoustic noise, and more surprisingly under scenarios of microphone variability, with accuracy losses as high as 15% in some scenarios. Further, we provide intuition on how this problem of model robustness relates to the broader topic of dataset-shift in the machine learning literature, and highlight future research directions for the mobile sensing community which include the investigation of domain adaptation and domain generalization solutions in the context of sensing systems.

4.1 Introduction

In recent years, we have witnessed a rapid increase in consumer devices and mobile sensing applications which aim to infer user context, activities and behavior from a variety of sensor data collected from the users. A number of commercial smart devices have already been launched in the market for monitoring a user's sleep

A. Mathur (✉)
Nokia Bell Labs and University College London, London, England
e-mail: akhil.mathur@nokia-bell-labs.com

A. Isopoussu · F. Kawsar
Nokia Bell Labs, London, England

R. Smith · N. Berthouze
University College London, London, England

N. D. Lane
University of Oxford, Oxford, England

N. Kawaguchi et al. (eds.), *Human Activity Sensing*,
Springer Series in Adaptive Environments,
https://doi.org/10.1007/978-3-030-13001-5_4

(Nokia sleep tracker 2018), physical activity (FitBit 2017), dietary actions (New technology tracks food intake by monitoring wrist movements 2017), stress (Empatica wristband 2018), daily activities (FitBit 2017), ambient environments (Narrative Clip 2017; Google Glass 2016), and emotional well-being (Empatica wristband 2018). In addition to these dedicated consumer devices, a range of mobile sensing applications have been proposed to detect context and activities such as sleep (Hao et al. 2013), exercise (Lu 2019) and transportation mode (Blunck 2013). In particular, owing to the recent breakthroughs in machine learning techniques for audio-processing, a number of promising audio sensing systems and applications have been proposed, including those which infer a user's emotion (Rachuri 2010), eating episodes (e.g., chewing) (Amft et al. 2005), and speech characteristics (e.g., speaker verification) (Variani et al. 2014), keyword spotting (Chen et al. 2014). The advancements in audio-based inference models are also ushering in the design of open audio-based hardware platforms which allow developers to create powerful audio services for the end-users. For instance, by integrating off-the-shelf microphone arrays with embedded platforms such as Raspberry Pi and cloud-based audio sensing models, developers can rapidly create their own version (Hardware to emulate amazon echo 2019) of a speech processing device such as an Amazon Echo.

As sensory inference systems move out of the laboratory setting into the wild, it is imperative that they work robustly on thousands and millions of end-user devices in unconstrained real-world scenarios. Indeed, prior research has highlighted it as a major research challenge to make sensory systems robust against hardware, software, environment and user variabilities. Blunck et al. (2013) demonstrated how GPS sensor variability can impact the data quality and the performance of inference models on smartphones. Stisen et al. (2015) studied sampling rate heterogeneity in inertial sensors of smart devices and found that software factors such as instantaneous CPU loads can cause a large variability in the accelerometer outputs of smartphones and smartwatches. Chon et al. (2013) found that sound classification models show poor accuracies when deployed in unconstrained environments. Similar findings were shown by Lee et al. (2013) about the adverse impact of acoustic environments on speaker turn-taking detection. Vision models are also impacted by environmental variabilities such as lighting conditions (Yang et al. 2016), various forms of object occlusion (Chandler and Mingolla 2016), and operation variabilities such as blurry, out-of-focus images due to unstable cameras.

In this chapter, we focus our attention on the robustness of audio-sensing models in unconstrained real-world scenarios. While a number of factors can influence the performance of an acoustic inference model in practice, this chapter explores two forms of prominent noise that these models are expected to encounter in the real-world:

Acoustic Environment Noise: An audio-sensing application should ideally make accurate inferences irrespective of where and when it is used. However in practice, the environment (e.g., cafe, 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 desired properties for audio-based inference systems is their robustness to diverse acoustic environments.

Microphone Heterogeneity: Audio inference models, once trained, are deployed on numerous mobile and wearable devices, many of which are not known while the models are trained and could come from different hardware manufacturers. This is a challenging scenario because different manufacturers use different hardware components (i.e., microphones) and may also have different software pipelines which process the raw audio signal before exposing them to user applications. Therefore, inference models need to be robust against these forms of microphone heterogeneity expected in the wild.

In this chapter, we build upon our prior work (Mathur et al. 2018) and study the performance of two widely-used general-purpose audio models under scenarios of real-world noise. First, we study how microphone heterogeneity impact cloud-based automatic speech recognition (ASR) models and thereafter, we extend this analysis to a small footprint keyword detection model. For our experiments, we use off-the-shelf microphones ranging from mid-range microphone arrays to low-cost USB microphones. In addition to the microphone heterogeneity problem, we also study the impact of background noise on cloud-based ASR models.

Quite unexpectedly, we find significant difference in the performance of our target audio models when they are exposed to different microphones, with observed accuracy drops as high as 15% when models trained on one microphone are deployed on another. Our results also reveal that cloud-based ASR models are more tolerant to ambient acoustic noise and show reasonable performance under moderate amounts of ambient noise, however the error rates increase significantly as the noise power is increased. We conclude the chapter by introducing the general topic of domain adaptation in machine learning and ways of leveraging domain adaptation techniques for improving the robustness of sensing models. More specifically, using microphone heterogeneity as a use-case, we discuss techniques which could be used at training-time and inference-time—depending on the system requirements—to improve the robustness of audio models when the training and test microphone differ. Taken together, our analysis and findings suggest the need for more rigorous evaluation of sensor-based inference systems, going beyond the conventional evaluation techniques such as train-test split and cross-validation.

4.2 Methodology

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

Audio Tasks and Datasets: Two representative audio tasks and datasets are used in our analysis:

- Automatic Speech Recognition (ASR): 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 (Hannun et al. 2014). Our experiments are conducted on the Librispeech-clean (Panayotov et al. 2015) 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 s and sampling rate of 16,000 Hz. In the rest of the chapter, we refer to this dataset as *Librispeech-clean-1000*.

- Keyword Detection: In this task, the goal is to identify the presence of a certain keyword class (e.g., Hey Alexa) in a given speech segment. We use the *Speech Commands* dataset containing 65,000 one-second long utterances of 30 short keywords (Speech Commands Dataset 2018) for our experiments. Instead of using all 30 classes, we used a subset of 12 classes (yes, no, up, down, left, right, on, off, stop, go, zero, one) for our analysis.

Audio Models: We now describe the two audio models on which the above mentioned datasets were evaluated:

- ASR Models: We conduct our experiments on ASR models from Google (using the Google Speech API 2019) and Microsoft (using the Bing Speech API 2019). The models use a CNN-bidirectional LSTM model structure (Xiong et al. 2017) and have shown near-human accuracy on ASR tasks (Microsoft speech recognition 2019; Google speech recognition 2019). 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.
- Keyword Detection Model: We use a small-footprint keyword detection architecture proposed in Zhang et al. (2017) to train the model. The input to this model is a two-dimensional tensor extracted from the one-second long keyword recording, consisting of time frames on one axis and 24 MFCC features on the other axis. The model outputs a probability of a given audio recording belonging to a certain keyword class (e.g., Yes, No) or to an Unknown class.

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* and *Speech Commands* datasets on a JBL LSR 305 monitor speaker[1] and recorded them simultaneously on three different microphones namely Matrix Voice (2019), ReSpeaker (2019) 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 10 cm from the speaker in

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

order to minimize the effect of room acoustics on the recorded audio. In effect, we created four variants each of the *Librispeech-clean-1000* and *Speech Commands* datasets, including the original dataset and the three re-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 (Piczak 2015). 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.

4.3 Results

Figures 4.1 and 4.2 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. Similar trends are observed with the keyword detection model. Figure 4.3 shows that when the training and test devices are the same, the keyword detection model provides the highest accuracy. However, when there is a mismatch between the training and test devices, it causes a significant degradation in accuracy as high as 15%.

Further, in Fig. 4.4, we plot the spectrograms of an audio segment from the Librispeech-1000 dataset in its original form Fig. 4.4a as well as when it is captured by different microphones Fig. 4.4b–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.

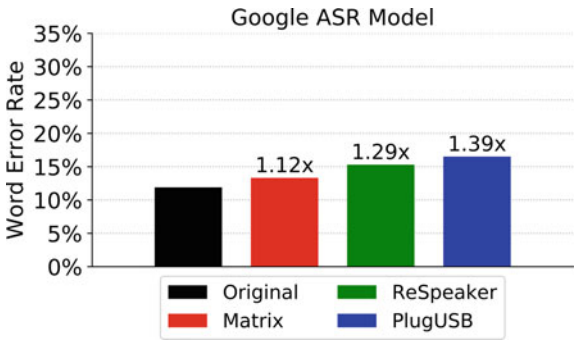


Fig. 4.1 Impact of the microphone variability on Google ASR model. Values on the bars illustrate the increase in WER over the original audio WER (black bar)

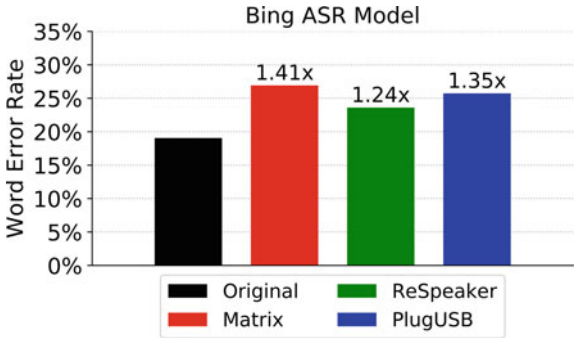


Fig. 4.2 Impact of microphone variability on Bing ASR model. Values on the bars illustrate the increase in WER over the original audio WER (black bar)

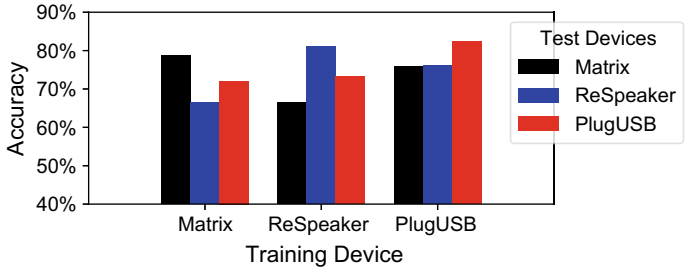


Fig. 4.3 Impact of microphone variability on the keyword detection model

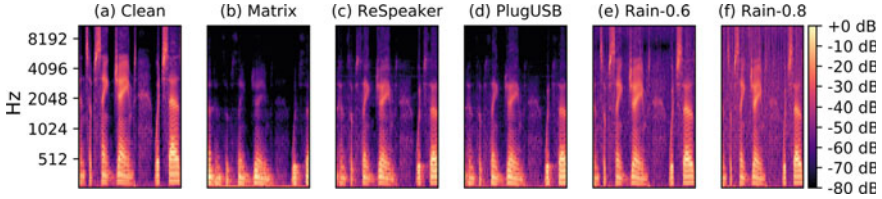


Fig. 4.4 Mel-Scale spectrograms of an audio segment under different experiment conditions

Next, Figs. 4.5 and 4.6 illustrate the findings on acoustic environment robustness. We varied the power of the background noise that is added to the speech signal (effectively the signal-to-noise ratio) and measured the WER of the ASR models 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 Google

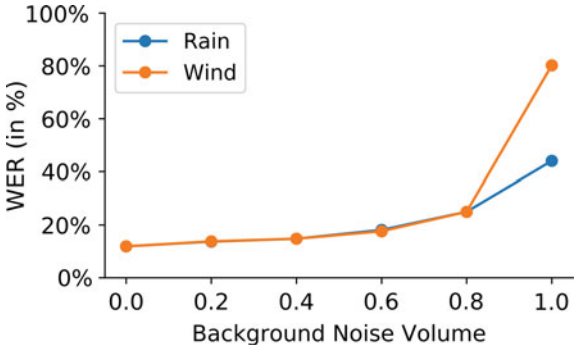


Fig. 4.5 Effect of two types of background noise on Google ASR model

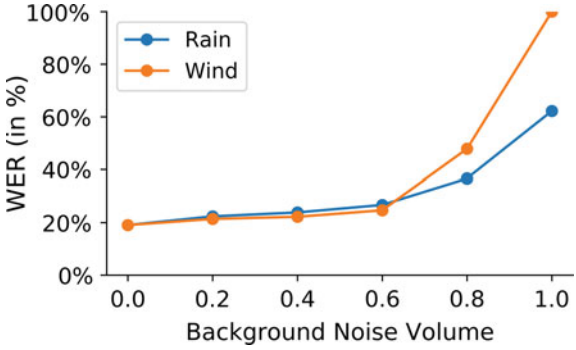


Fig. 4.6 Effect of two types of background noise on Bing ASR model

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 Fig. 4.4, although the Rain-0.6 spectrogram (Fig. 4.4e) looks visibly more noisy than the spectrograms collected from different microphones (Fig. 4.4b–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.

4.4 Discussion and Future Directions

Our experiments show that deep learning based audio 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* (Sugiyama et al. 2017)—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* (Blitzer et al. 2006) and *domain generalisation* (Blanchard et al. 2011). *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 *domain generalization*. The easiest way to achieve this consistency is by finding features which are invariant under the dataset shift (Muandet et al. 2013). 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 (Mathur 2018).

We propose that the application of domain adaptation and domain generalization techniques on audio-sensing models could be a promising research direction for the mobile sensing community. In an ongoing work, we are exploring the feasibility of formulating the issue of microphone variability as a data translation problem, i.e., given an audio from a microphone A, can we translate it to a different microphone's (e.g., B) domain? If a translation function can indeed be learned between a pair of microphones, it can subsequently be used to convert any audio training data across microphone domains, and audio models could be trained on such diverse training datasets. One key challenge however is that generating large aligned audio datasets to train the translation models discussed above can be hard. Each time a new microphone is considered, in order to produce an aligned version of the dataset, the entire dataset needs to be recorded using the new microphone. This leads to major issues in scaling the approach to multiple devices.

As such, we are exploring solutions which can learn the mapping between two microphones without requiring time-aligned data from them. To this end, we propose to use the CycleGAN architecture introduced in Zhu et al. (2017), which involves simultaneously training two translation models, one mapping the training domain to the test domain, and another one in reverse. The model also uses a cycle loss as a way to improve the performance of both translation models. Once trained, our translation model named Mic2Mic can be used in two different ways.

- Training time: As shown in Fig. 4.7, Mic2Mic can be used to translate the entire training dataset from microphone A to microphone B, as a way to augment the training data to add awareness about the properties of the test microphones into the audio model training process. The original training dataset is then combined with the translated dataset to generate an augmented dataset, upon which the task-specific audio model is trained.

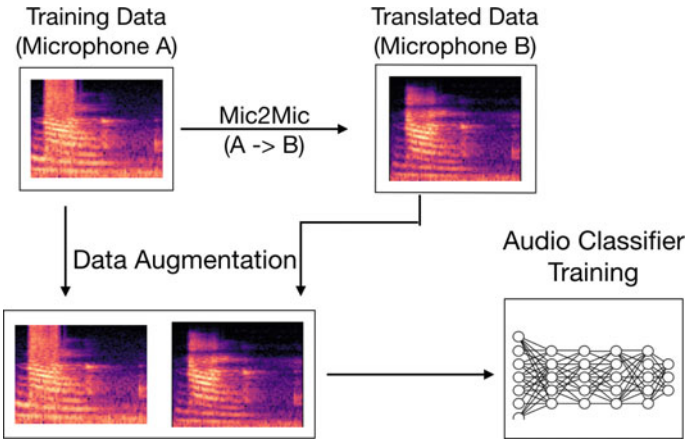
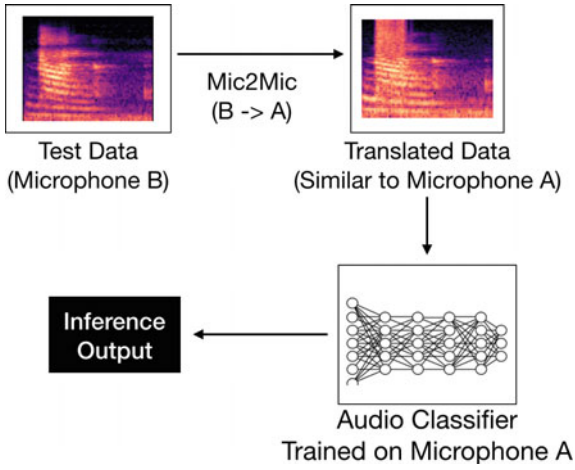


Fig. 4.7 Translation model deployed as a data augmentation to correct for microphone variability

Fig. 4.8 Translation model deployed to translate test time data to the training data distribution to correct for microphone variability



- Inference time: Alternatively, Mic2Mic can be deployed in the inference pipeline of audio models as a real-time translation component. As shown in Fig. 4.8, audio data from the test microphone at inference-time is first passed to Mic2Mic, which transforms it to bring it closer to the training data distribution. Thereafter, the translated data is inputted to the pre-trained audio classifier to obtain the inferences.

While our initial results are promising, there remain a number of open research challenges to make Mic2Mic and other domain adaptation challenges work in uncontrolled settings. In real-world settings, it is likely to encounter a combination of different variabilities in the sensor data. For instance, microphone variability can combine with acoustic environmental noise and user-specific behavior, and will need a much more complex solution than the single-variability adaptation approaches such

as Mic2Mic. Further, many domain adaptation approaches do pairwise adaptation, e.g., source microphone to target microphone adaptation. Clearly, this pairwise adaptation is not scalable for the thousands and millions of devices in the market. As such, there is a clear need to explore the feasibility of domain generalization techniques which can work on a larger scale.

4.5 Conclusions

We evaluated the robustness of embedded-scale and cloud-scale audio models to microphone and acoustic environment variability. To facilitate our first experiment on microphone variability, we collected speech samples from three different embedded microphones simultaneously for two common speech-related tasks: Automatic Speech Recognition (ASR) and Keyword Detection. Our results demonstrate significant performance degradation in both cloud-scale and small footprint embedded-scale models, with absolute accuracy drops of up to 7% and 15% in the ASR and Keyword Detection models respectively. For the acoustic background noise scenario, we also observe a moderate degradation in accuracy of the audio models, which becomes more severe as the intensity of the background noise increases.

Overall, this chapter highlighted the need to design better evaluation techniques for mobile sensing models, which take into account the real-world noise that the models are expected to encounter in practice. Further, we discussed that the challenges of model robustness are related to the wider problem of dataset-shift in the machine learning literature, and provided intuition on how domain adaptation and domain generalization approaches can be leveraged—both at training-time and inference-time—to adapt sensory inference models to new operating scenarios.

References

Nokia sleep tracker (2018). https://health.nokia.com/uk/en/sleep/
FitBit (2017). https://www.fitbit.com
New technology tracks food intake by monitoring wrist movements. http://gadgetsandwearables.com/2017/03/29/food-tracking/ (2017). Accessed 20 June 2019 10:48:03
Empatica wristband (2018). https://www.empatica.com/en-gb/research/e4/. Accessed 20 June 2019 10:48:03
Narrative clip (2017). http://getnarrative.com/narrative-clip-1. Accessed 1 Sept 2017
Google glass (2016). https://developers.google.com/glass/distribute/glass-at-work. Accessed 20 June 2019 10:48:03
Hao T, Xing G, Zhou G (ACM, 2013), SenSys'13. https://doi.org/10.1145/2517351.2517359
Lu HEA (2019) Proceedings of Sensys '10. ACM, pp 71–84
Blunck H et al (2013) In: Proceedings of the 2013 ACM Ubicomp. ACM, pp 1087–1098
Rachuri K et al (2010) In: Proceedings of Ubicomp'10. ACM, pp 281–290
Amft O, Stäger M, Lukowicz P, Tröster G (2005) In: Ubicomp Springer, pp 56–72

Variani E, Lei X, McDermott E, Moreno IL, Gonzalez-Dominguez J (2014) In: 2014 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 4052–4056
Chen G, Parada C, Heigold G (2014) In: ICASSP. IEEE, pp 4087–4091
Hardware to emulate amazon echo. https://tinyurl.com/y84d6r2n/
Stisen A et al (2015) In: Proceedings of Sensys. ACM, pp 127–140
Chon Y et al (2013) In: Ubicomp. ACM, pp 3–12
Lee Y, Min C, Hwang J, Lee I, Hwang Y, Ju C, Yoo M, Moon U, Lee J, Song J (2013) In: Proceeding of Mobisys'13. ACM, pp 375–388
Yang S, Wiliem A, Lovell BC (2016) In: 2016 international conference on image and vision computing New Zealand (IVCNZ). IEEE, pp 1–6
Chandler B, Mingolla E (2016) Computational intelligence and neuroscience
Mathur A, Isopoussu A, Kawsar E, Smith R, Lane ND, Berthouze N (2018) In: Proceedings of the 2018 ACM international joint conference and 2018 international symposium on pervasive and ubiquitous computing and wearable computers. ACM, New York, NY, USA, 2018) UbiComp'18, pp 1409–1413. https://doi.org/10.1145/3267305.3267505
Hannun A, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, Prenger R, Satheesh S, Sengupta S, Coates A, et al (2014) arXiv:1412.5567
Panayotov V, Chen G, Povey D, Khudanpur S (2015) In: ICASSP. IEEE, pp 5206–5210
Speech Commands Dataset. https://research.googleblog.com/2017/08/launching-speech-commands-dataset.html. Accessed 20 June 2019 10:48:03
Google Speech API. https://cloud.google.com/speech-to-text/
Bing Speech API. https://azure.microsoft.com/en-us/services/cognitive-services/speech/
Xiong W, Wu L, Alleva F, Droppo J, Huang X, Stolcke A (2017) ArXiv e-prints
Microsoft speech recognition. https://tinyurl.com/ybnm9zdj/
Google speech recognition. https://tinyurl.com/y7dm37vw/
Zhang Y, Suda N, Lai L, Chandra V (2017) arXiv:1711.07128
Matrix Voice. https://www.matrix.one/products/voice/
ReSpeaker. https://respeaker.io/
Piczak KJ (2015) In: ACM multimedia. ACM, pp 1015–1018
Sugiyama M, Lawrence ND, Schwaighofer A et al (2017) Dataset shift in machine learning. The MIT Press
Blitzer J, McDonald R, Pereira (2006) In: Proceedings of the 2006 conference on empirical methods in natural language processing, pp 120–128
Blanchard G, Lee G, Scott C (2011) Advances in neural information processing systems 2178–2186
Muandet K, Balduzzi D, Schölkopf B (2013) ICML 10–18
Mathur A et al (2018) In: IPSN. IEEE
Zhu JY, Park T, Isola P, Efros AA (2017) CVPR 2223–2232