A Vision for Adaptive and Generalizable Audio-Sensing Systems

Akhil Mathur[‡†], Fahim Kawsar[‡], Nadia Berthouze[†], Nicholas D. Lane[◇]
[‡]Nokia Bell Labs, [†]University College London, [◇]University of Oxford

ABSTRACT

We present our vision and key research directions for next generation audio and speech-sensing systems, to make them robust against variabilities in sensing hardware and operating conditions.

KEYWORDS

audio sensing, domain adaptation, user personalization

1 VISION

In the last few years, audio has emerged as a promising modality for personal [2], social [5] and community-level sensing [9]. This has been made possible by the advancements in sensing hardware as well as the computational models which process raw audio to infer user contexts [3]. For example, it is now straightforward to create a custom hardware device [1] similar to Amazon Echo by connecting an off-the-shelf low-cost microphone with an embedded platform such as Raspberry Pi, and using local or cloud-based models to build audio-sensing applications.

The next grand challenge would be to take audio-sensing systems outside the lab and make them work robustly on millions of devices in-the-wild. This in turn would enable application developers to combine audio with other sensing modalities (e.g., motion, vision) to extract much richer contextual information about user behavior than what is possible today. We envision two major research challenges in this aspect as shown in Figure 1.

Audio Domain Adaptation. Current audio-based computational models are prone to failure when they are deployed with unseen microphone hardware[4], in new acoustic environments [7], or when they encounter a new class of speakers [8] – thereby severing limiting their widespread adoption. A fundamental reason for poor robustness of audio models is the mismatch between their training and deployment conditions. The aforementioned real-world variabilities cause the distribution of the test data to differ from the training data (also referred to as *domain shift*), which leads to poor generalization of the ML models. Domain Adaptation has been an emerging topic of research in the vision community with the goal of minimizing domain shift, and we believe that developing domain adaptation techniques for audio sensing models could improve their robustness in real-world situations.

Lifelong Learning for Audio Models. Personalization of audio models to a target user's preferences and operating conditions (e.g., ambient noise profile of their surroundings) would be critical for their success. For instance, general-purpose keyword detection models are trained for a large number of classes, however a given user may be interested in detecting a small subset of the keywords. Moreover, user preferences and operating conditions may also vary over time. A such, it is imperative for audio models to learn from their mistakes or refine their knowledge over time. In this regard, research on topics such as continuous learning and model personalization [6] would be particularly useful.

2 PROGRESS

We are exploring both data-level and feature-level alignment techniques for audio domain adaptation. Currently, we have proposed a generative modeling solution (based on CycleGANs) which can reduce microphone-induced domain shift in speech systems using unlabeled data. There are however a number of open-research questions from a systems perspective: (a) will domain adaptation and continuous learning updates to the model take place centrally on the cloud, or can they be done locally on-device in a privacy preserving manner? (b) what are the best ways to acquire and possible label the sensing data from the end-users? (c) can the solutions to these problems generalize to multiple audio sensing tasks? We hope to discuss these challenges at the workshop.

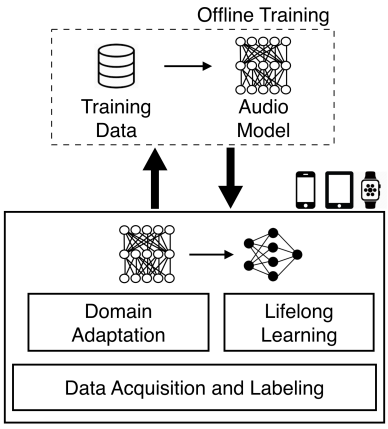

Figure 1: Vision for on-device domain adaptation and personalization of audio models.

REFERENCES

[1] 2015. DIY hardware to emulate Amazon Echo. http://www.instructables.com/id/Build-DIY-Amazon-Alexa-With-a-MATRIX-Creator-on-Ha/.
[2] Anagnostopoulos et al. 2015. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review* 43, 2 (2015), 155–177.
[3] Lane et al. 2015. DeepEar: Robust Smartphone Audio Sensing in Unconstrained Acoustic Environments Using Deep Learning. In *Proceedings of Ubicomp '15.*
[4] Mathur A. et al. 2018. Using Deep Data Augmentation Training to Address Software and Hardware Heterogeneities in Wearable and Smartphone Sensing Devices. In *IPSN.* IEEE.
[5] Lee et al. 2013. Sociophone: Everyday face-to-face interaction monitoring platform using multi-phone sensor fusion. In *Proceeding of Mobisys '13.* ACM
[6] McGraw et al. 2016. Personalized speech recognition on mobile devices. *arXiv preprint arXiv:1603.03185* (2016).
[7] Qian et al. 2016. Very deep convolutional neural networks for noise robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*
[8] Senior and Lopez-Moreno 2014. Improving DNN speaker independence with i-vector inputs. In *ICASSP '14.* IEEE, 225–229.
[9] Xu et al. 2013. Crowd++: unsupervised speaker count with smartphones. In *Proceedings of Ubicomp '13.* ACM, 43–52.

SocialSense'19, April 15, 2019, Montreal, QC, Canada

ACM ISBN 978-1-4503-6706-6/19/04…$15.00
https://doi.org/10.1145/3313294.3313936