Reviews of 5523 - *"ProtoSound: A Personalized, Scalable Sound Recognition System for d/Deaf and Hard-of-Hearing Users"*

Reviewer 4 (1AC)

Expertise

Expert

Originality

High originality

Significance

High significance

Rigor

High rigor

1AC: Recommendation

We recommend Accept with Minor Revisions.

1AC: The Meta-Review

This paper focuses on improving sound recognition for people who are D/HoH. The authors introduce a system called ProtoSound which is a system which allows D/HoH users personalised sound recognisers by recording some examples. The authors motivate the work with a large participant survey (N=472) to understand personalization preferences of DHH users. The authors evaluate ProtoSound by two quantitative evaluations on real-world datasets and a field study using a built real-time mobile application. Notably, the study does not include an evaluation with DHH users which the authors are planning on doing in the future.

1. EDITORIAL SCORE: We recommend Accept with Minor Revisions.

2. REVIEW SYNTHESIS/ANALYSIS:
POSITIVES:

All reviewers found the contributions to be important and novel.
Well written (R3, R1, R2)
Strong technical work using state-of-the-art ML techniques on an Accessibility problem (R3). R1 appreciated the availability of the model and code.

Well motivated (R3, R1, R2)
Solution nicely grounded with user centred research (R3, R1)
Experiments show promise (R3)
User survey with hundreds of participants is useful (R1, R2)
Advantages over previous work (R2)

NEGATIVES:

R3 wanted a more clearer comparison to ListenLearner as the most closely related competitor (R2 mentions this too)
& how would certain features from ListenLearner affect performance in ProtoSound
More details on what happens when sound changes over time, or more error labels are present (R3, R2)
Interface might not be designed for DHH people (R3)
R1and R2 sought a number of clarifications of certain features within the system, and a mismatch of their description within the paper
R1 asks how context can be taken into consideration (e.g., people only care about baby cries when it's their own baby or the baby is in distress without a caregiver etc)
R2 asks for more understanding of different perspectives for people who are D/HoH
R2 found some missing work that should be cited

==== POST PC MEETING COMMENTS ====

The paper was discussed at the PC meeting. Both 1AC and 2AC found there was great value in this work and the contribution was strong. As such, we decided to conditionally accept this work with minor revisions.

1AC: The Summary of Revisions Required

>> Required:
Remove the claim that hard-of-hearing individuals connect to deafness audiologically (see R2's review)
Clarify the comparisons to ListenLearner and discuss how would certain features from ListenLearner affect the performance of ProtoSound
Add more discussion on what happens when a sound changes over time, and how this uncertainty is presented to users
Be more consistent in word choice and clarify a number of terms within the paper when discussing the system (see R1 and R2's reviews)

Add to the discussion details on how context can be taken into consideration mentioned by R1 (e.g., people only care about baby cries when it's their own baby or the baby is in distress without a caregiver etc)
Add missing work suggested by R2

Please consider all other changes as suggested by the reviewers

Reviewer 3 (2AC)

Expertise

Knowledgeable

Originality

High originality

Significance

High significance

Rigor

High rigor

Recommendation

I recommend Accept with Minor Revisions.

Review

This paper presents ProtoSound, a sound recognition system that allows DHH users to create personalized sound recognizers by providing only a few example recordings. The authors motivated their work using a combination of prior literature, as well as a survey with ~500 participants. Then the authors presented their system architecture, and open source Python and Android implementations of their technique. System performance was evaluated in a series of three experiments, which focused on existing datasets with hearing people, datasets collected by DHH people, and a field study for pre-deployment validation of the feasibility.

This is a strong technical piece, applying state-of-the-art ML techniques on an Accessibility problem. The problem is well motivated. The solution is nicely grounded with evidence from user research. Three experiments show great promise of the approach, which set the precursor of a larger scale deployment that extends the utility of prior work. The discussion section was to the point. The paper is overall nicely written, and the video figure was effective in demonstrating how the system

nicely written, and the video figure was effective in demonstrating how the system works in action. This paper makes a strong contribution to the Accessibility community, and I recommend acceptance.

Below are more detailed comments and suggestions:

With ListenLearner as the closest related work, I was hoping to see a more detailed comparison between the two approaches to help characterize the benefit of ProtoSound, perhaps as part of the technical evaluation. How many samples are needed in the two approaches to reach the same level of performance? Can the method of prompting the users for feedback be integrated into ProtoSound, which might increase the utility of ProtoSound for additional unexpected events? As mentioned in the Discussions section, alternative kinds of representation would be needed to visualize the sound and link that to the visual and contextual cues. On the other hand, if integrating some of the techniques of ProtoSound (e.g., the active few shot labeling technique, among others) in ListenLearner, how would the performance change?

More broadly, while the technical evaluations of ProtoSound against existing datasets and models show improved performance, I'm hoping to understand which specific system components contributed to the improvements. Additional ablation studies would help.

The field evaluation provided quantitative evidence that the technique can work in the wild, with hearing people collecting and validating the labels. However, the mobile UI is not really applicable for DHH people to directly use, and I think that was not immediately clear in the paper, not until the discussion section 8.1.

Another question I have is about the ability of ProtoSound to adapt to errors and changes over time. The authors discussed providing better interfaces for labeling which would reduce errors, but how might the system handle cases where the same event varies over time? E.g., consider how a baby cry might be quite different, or when playing piano the notes would be different from one second to the next. How does the system detect these changes?

As discussed by the authors in 8.4, to expand on the number of classes ProtoSound can support in a variety of contexts, GPS, perhaps also additional sensors can be used to narrow down the most relevant events to focus on. I'm also curious whether background noise or the sound profile can be used as a signal to first identify the context as well. Perhaps sound events can also have the property of public vs. private, which the application can enable sharing across users and automatically downloads model files as users move around (similar to loading map content).

The List of Revisions Required

The List of Revisions Required

Please refer to my review

Reviewer 1 (reviewer)

Expertise

Knowledgeable

Originality

High originality

Significance

High significance

Rigor

Very high rigor

Recommendation

I recommend Accept with Minor Revisions.

Review

The paper presents a strong contribution in the area of sound recognition systems for DHH individuals. The paper makes a clear contribution to HCI (as well as to ML), that will be valuable to other researchers (due to the availability of their model and code), with the ultimate goal of improving the accessibility of environmental sounds to people with limited ability to hear them for themselves.

The paper makes a clear contribution to the field by introducing a robust personalization mechanism (allowing users to configure the tool/app to recognize sounds of their own choosing). The authors take a user-centred approach to incorporating personalization by consulting with hundreds of potential DHH users via a survey, and then evaluating various aspects (including but not limited to accuracy) of both capturing and identifying personalization sounds.

I have a few recommendations and comments (detailed under 'List of Revisions'), but otherwise the paper was enjoyable to read and very well presented. The included figures (except 2, see 'List of Revisions') and tables support the story well, and relevant previous literature is incorporated.

I was going to raise the 'one second sample' limitation. This seemed like an arbitrary choice when I was reading the paper - there are sounds much shorter (e.g., dings,

bings, whacks, knocks) and much longer (e.g., longer ringtones/doorbells/alarms, washer/dryer ending cycles, thunder) than this. However, the authors do a great job of discussing this limitation in the Discussion (8.2), so bravo.

The only other comment I have is on how you would consider really difficult cases like a baby's cry. Baby cries are very important sounds in ANY context (in which the user AND the baby might be in different rooms/locations), they are difficult to capture (don't tend to just leave babies crying while we fiddle with our apps), they have lots of variation (babies have different cries for hunger, tired, scared, hurt), they are HIGHLY personal (I really only care about MY baby's cries), and they have a cycle that lasts much longer than one second (depending on the type of cry). How would you propose making such sounds work with your system?

The List of Revisions Required

Section 3.1:
* I found this section difficult to follow, especially Figure 2. What I struggled with the most is understanding how the system is trained (which is one step) and how it is used to identify a given sound (which is a separate step). These are presented together in the figure, which is confusing.
* In addition, there are some terms such as 'User Recordings', 'Library of Infrequent Sounds', and 'Query Sound' which are not well reflected in the paper (e.g., 'Library of Infrequent Sounds' is called 'Library of Difficult-to-Produce Sounds' in Section 3.1.2).
* Section 3.1.3 talks about sounds being 'Ignored' which does not map clearly to the output of 'Unknown'. Confusion.
* Likewise, the first paragraph of Section 3.2 talks extensively about the generation of a set of sounds (which I think are the 'Support Set' in Figure 2), but I'm not 100% sure how or where these fit into the ML system.
* When exactly is the classifier regenerated? Does the system come with a classfier for the 'Infrequent Sounds', which is then retrained whenever a user records a sample sound (to be classified later)?
* Finally, are query sounds (i.e., sounds that the system tries to identify) rolled back into the machine learning tool? I don't think so (as Figure 2 doesn't show this [but I still don't understand/trust Fig 2] and the discussion under 'Human-in-the-Loop' suggests it doesn't) but making this clearer would be useful.

All of the above needs to be taken with a grain of salt. I'm not an ML person. That said, CHI isn't an ML conference so I think that the above needs to be cleared up to be more accessible to a general HCI audience.

Minor edits:
1. First paragraph under Discussion (...which do not support:) has a list that goes 1,2,2 instead of 1,2,3.
2. Survey Q2 has 'Next of the above' instead of 'None of the above'.

Reviewer 2 (reviewer)

Expertise

Knowledgeable

Originality

High originality

Significance

High significance

Rigor

High rigor

Recommendation

I can go with either Accept with Minor Revisions or Revise and Resubmit.

Review

This paper makes an important contribution in the domain of sound recognition systems for Deaf and Hard-of-hearing users both within HCI and applied AI research. The paper provides details about the implementation and evaluation of a system that trains and scales on the fly and in various contextual settings. ListenLearner is the closest prior work. ProtoSound has an advantage over ListenLearner in terms of quicker adaptation to new environments and more personalization. It consists of a survey with 472 DHH users to uncover personalization preferences of DHH users followed by two quantitative evaluations on real-world datasets and a field study using a built real-time mobile application. Notably, the study does not include an evaluation with DHH users which the authors are planning on doing in the future. The authors did a good job structuring and writing the paper, which made for a smooth reading experience. The only concern was that there were no line numbers in the margin which makes it harder to point to specific paragraphs and sentences. Nonetheless, the presentation of the system implementation, study designs, and results largely provide clear and ample evidence for the need for personalization in sound recognition systems and the effectiveness of the proposed system.

Prior work is largely well organized and helps situate the work in the body of prior work in this domain well. Section 2.1 of the literature review describes the similarities and differences in the sound awareness needs of DHH users. The differences may be due to social contexts, physical locations, or the identity of a DHH individual. Section 2.2 highlights that while there are a lot of existing sound recognition systems

Section 2.2 highlights that while there are a lot of existing sound recognition systems (including personalized recognition systems), they are limited by their scalability and personalization. Section 2.3 describes prior machine learning approaches for similar tasks. The authors use meta-learning which has advantages over traditional supervised or semi-supervised learning approaches, especially in this context.

There are some avenues for further improvement. There are some prior works that the authors should consider citing. For example, the one below:

Lu, H., Pan, W., Lane, N. D., Choudhury, T., & Campbell, A. T. (2009, June). Soundsense: scalable sound sensing for people-centric applications on mobile phones. In Proceedings of the 7th international conference on Mobile systems, applications, and services (pp. 165-178).

Similarly, in section 2.1, the authors said that the hard-of-hearing individuals connect to deafness audiologically. I would recommend authors use less rigid, (e.g. some or a lot of) wording since hard-of-hearing individuals can choose an audiological or cultural perspective.

Section 3 described the system design and implementation in sufficient detail. However, I was confused about some sentences, e.g. the first sentence on page 6.

The supplementary video file was also very useful to understand the system, especially user-centric features like a searchable sound library, cross-setting generalization, and computational efficiency of training.

The key finding from the survey was the number of sounds that participants wanted to be recognized in each context. Although, the authors later state that prior work already provides a similar number in section 5.1 (3-5 medium-to-high priority sounds per context). Findings of evaluation using real-world sounds collected from hearing participants and real-world sounds collected by DHH participants were well documented and visualized. In comparison to the manual labels section, it might be useful to get more statistics about different types of sounds. The authors claim that some sounds were hard to classify because they were too similar to the rest or due to background noise. It would be interesting to explore average accuracies for each sound.

Protosound learned sound events in a diverse set of environments with an average precision of 87.4% in 56 locations. The field evaluation of the mobile application and the discussion sections were well written. The discussion not only ties to prior work on GUIs and algorithmic improvements but also discusses socio-cultural implications briefly. The limitation section was also thorough and well organized. The references are largely cleaning and consistent but have some missing fields.

There are a few minor grammatical and other writing issues in the paper:

1. In the caption of figure 1, it should be supports instead of support.
2. The authors use hard of hearing (without hyphens, e.g. in the introduction) in some cases and hard-of-hearing in others.
3. In the discussion section, there is a list with two items 2s.

Overall, the paper makes an important contribution to the field. I am confident about the validity of the results and the writing is largely clear.

The List of Revisions Required

1. Remove the claim that hard-of-hearing individuals connect to deafness audiologically.
2. Cite more relevant prior work at least the paper mentioned in the review.
3. Correct minor grammatical and other writing issues.
4. Provide average accuracies for recognition of each sound, especially the ones that were hard to classify.
5. Although section 3 was understandable to me, there is a need to make it more understandable (especially 3.1) and easy to follow for the broader HCI community who might not be familiar with some of the machine learning concepts mentioned. One strategy could be to include a very descriptive caption for figure 2.

Return to submission and reviews