Poster: Speech in Smartwatch based Audio

Daniyal Liaqat
University of Toronto
Vector Institute

Robert Wu
University Health Network
University of Toronto

Andrea Gershon
University of Toronto
Sunnybrook Health Sciences Centre

Hisham Alshaer
Toronto Rehabilitation Institute

Frank Rudzicz
Toronto Rehabilitation Institute
University of Toronto
Vector Institute

Eyal de Lara
University of Toronto

CCS CONCEPTS

• **Human-centered computing → Mobile devices**; **Empirical studies in ubiquitous and mobile computing**;

1 INTRODUCTION

Speech from a microphone can be a rich source of information. The speech analysis and audio processing community has explored using speech to detect emotion, depression and even Alzheimer's disease. Audio data from these studies tends to be collected in more controlled environments with well placed, high-quality microphones. Applying these kinds of analyses to in-the- wild audio could have significant contributions, particularly in the context of health monitoring. However, it is expected that the quality of the data would drop in less controlled, mobile environments. As part of a health monitoring study, we collected in-the-wild audio from a smartwatch and in this paper we characterize speech in our collected audio. Our findings include that smartwatch based audio is good enough to discern speech. However, isolating speech is difficult because of the wide variety of noise in the signal and current tools perform poorly at dealing with this noise. We also find that a surprisingly high proportion of speech does not come from the user.

2 DATA COLLECTION

We built an Android Wear based application to record data from the built-in microphone. This data is transmitted to a remote server for analysis. We recruit participants and ask them to wear a smartwatch running our application for three months while data is recorded. We inform participants about the invasive nature of recording raw audio and the measures we take to protect their data. To our knowledge, this is the first study to record raw, unfiltered audio from an in- the-wild smartwatch.

3 ANALYSIS AND RESULTS

We have collected over 4,000 hours of audio across 15 participants. Our analysis of this audio so far has relied on manual annotation, with one goal being to develop automatic classifiers for detecting various sounds of interest. To assist annotation, we remove silence from the audio. We have found that roughly 60% of audio is silence, which means our annotators only need to listen to the 40% that is non-silence. To date, we have listened to 83 hours of silence-removed audio. Of the silence removed audio, 59% of it was identified as speech. Annotators also label the source of the speech, and interestingly, only 15% of the speech comes from the user. Another 15% comes from a 2^{nd} person and 70% comes from a TV/radio. This highlights that speech analysis must be applied carefully since the majority of speech does not belong to the user.

Manual annotation is difficult and time consuming, so we wanted to see how well existing Voice Activity Detection (VAD) tools perform on our data. We used VAD tools from WebRTC's[1], Loizou [1], Giannakopoulos[2], and LIUM SpkDiarization [2]. WebRTC's VAD has a parameter to control the aggressiveness ranging from 0 to 3, where 0 is the least aggressive about filtering out non-speech and 3 is the most aggressive. Interestingly, most of these tools were too lenient and classified around 90% of audio as speech (the exception being VAD(2) at 80% and VAD(3) being far too strict at 2%). One explanation for this is that these tools were developed and tested on more consistent audio sources. LIUM for example, was developed for TV and radio broadcasts, Loizou [1] used curated dataset of in-lab recordings and Giannakopoulos[2] assumes that the level of background noise is low.

4 CONCLUSION

We collected real-world audio traces from 15 people using a Smartwatch and analyzed the contents using manual annotation to better understand in-the-wild speech. We found that existing VAD tools do not perform well on this data and that only a small portion of speech comes from the user. These results highlight two problems that will need to be addressed in order for in-the-wild speech analysis to become viable. First, we need more robust VAD to better identify speech in noisy and inconsistent environments. Secondly, we need reliable methods for distinguishing the user's speech from other sources of speech. While both of these topics have been studied, applying them to wearable and mobile systems brings new opportunities and challenges.

REFERENCES

[1] P. C. Loizou. Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum. 13(5):857–869.

[2] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier. An open-source state-of-the-art toolbox for broadcast news diarization. Idiap.

[1] https://webrtc.org

[2] http://www.mathworks.com/matlabcentral/fileexchange/28826-silence-removal-in-speech-signals

MobiSys '18, June 10–15, 2018, Munich, Germany
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-5720-3/18/06.
https://doi.org/10.1145/3210240.3210817