A DATASET FOR FOREGROUND SPEECH ANALYSIS WITH SMARTWATCHES IN EVERYDAY HOME ENVIRONMENTS

Dawei Liang, Zifan Xu, Yinuo Chen, Rebecca Adaimi, David Harwath, Edison Thomaz

University of Texas at Austin, USA

ABSTRACT

Acoustic sensing has proved effective as a foundation for applications in health and human behavior analysis. In this work, we focus on detecting in-person social interactions in naturalistic settings from audio captured by a smartwatch. As a first step, it is critical to distinguish the speech of the individual wearing the watch (foreground speech) from all other sounds nearby, such as speech from other individuals and ambient sounds. Given the considerable burden of collecting and annotating real-world training data and the lack of existing online data resources, this paper introduces a dataset for foreground speech detection of users wearing a smartwatch. The data is collected from 39 participants interacting with family members in real homes. We then present a benchmark study for the dataset with different test setups. Furthermore, we explore a model-free heuristic method to identify foreground instances based on transfer learning embeddings.

Index Terms— foreground detection, transfer learning, wearable sensing, social interactions.

1. INTRODUCTION

Researchers have recently shown that off-the-shelf smartwatches can serve as an effective platform to recognize various forms of human behaviors and activities of daily living [1, 2]. In our research, we are particularly interested in exploring whether the acoustic sensing capability of smartwatches can be used to detect face-to-face social interactions. A critical step towards this is to distinguish speech instances of an individual (i.e., an individual wearing a smartwatch) from other sounds. Because of the dynamic noise conditions in real-world situations, this task is extremely challenging, and a typical voice activity detection system is not designed for modeling this speaker difference [3]. In wearable sensing, speech produced by the person wearing a smart device (e.g., a smartwatch) is usually referred to as the *foreground*, while all other forms of noise such as the speech of others, natural sounds, man-made sounds, music, and silence are often referred to as the *background* [4]. Hence, the problem of foreground detection in wearable sensing is binary classification of foreground speech versus background sounds.

Intuitively, one approach to detecting the foreground speech is to employ a speaker verification system so that one can map speech to one or more speakers [5]. Similar efforts include estimating a binary mask of target speech [6]. However, these techniques are based on specific user or channel information, i.e., a known *apriori*. This is often not possible in non-deterministic real-world settings, when conversations and interactions cannot be anticipated ahead of time [7]. Hence, the recent focus of foreground analysis for wearable sensing is to build supervised models that directly learn to identify foreground instances [3, 7, 8]. Due to the lack of annotated training sets in the relevant domain, however, such studies mostly aim at meeting scenarios or controlled environments. Public meeting datasets such as the ICSI [9] and the AMI [10] are not sufficient for studies in the wild where varying speech characteristics, unanticipated noise types, and human artifacts such as body movements are often expected.

The goal of this work is to mitigate the challenge of foreground detection for wearable sensing in unconstrained daily living situations. Specifically, we make the following contributions:

- A dataset[1] containing over 110K (31 hours) audio instances of foreground speech and background sound classes recorded by smartwatches, annotated at one second granularity. The data is collected from 39 participants, formulating 18 social groups, in real homes.
- A benchmark study for our dataset based on different setups. We specifically discuss the challenge of foreground detection in the wild and address the needs of model development on real-world datasets.
- Exploration of leveraging transfer learning embeddings for foreground detection, which inspires a method to identify the foreground instances without relying on supervised detectors.

2. DATASET

Our study is conducted towards real-world situations and with commercial smart wearables. This section describes the details of the dataset we collected in the study.

[1] https://doi.org/10.18738/T8/IKWZPW

Session	Description
Telephone Call	The wearer placed a telephone call.
Watch TV	The wearer watched content on a TV/laptop with the sound on for 10 min.
Chat Indoors	All participants played the NASA decision-making game [11] together.
Chat Outdoors	All participants walked outdoor for at least 5 min and chatted on any topics.
Meal	All household members cooked and ate together for dinner.

Table 1. The studied interaction sessions in homes.

2.1. Data Collection Protocol and Procedures

Our data collection was semi-controlled. This IRB-approved study took place in 18 distinct homes, each containing at least two participants. For each group, one of the participants was asked to wear a Fossil smartwatch for continuous audio recording, and this participant was referred to as the wearer of the study. The other participant(s) in the home engaged in a set of social activities with the wearer following a pre-defined study script of activity sessions (Table 1).

The study period was from 3:30 pm to 8 pm in a day, where there was a 30-minute gap between each session. To maintain the naturalistic aspect of the interaction process, the watch was left recording continuously throughout the entire study period, and we did not specify a set time for the conversations to end. However, the wearer was asked to note down the approximate end time of each session. The acoustic environments were left as usual, where sounds of home appliances and other non-participating household members were allowed to be included anytime throughout the study.

On the day of the study, the study materials, including the smartwatch, reading materials and study script were delivered by a researcher. Right before the study began, the researcher connected with the participants via Zoom to introduce the study and procedure. The researcher then logged out and collected the materials after the entire study was completed. An audio recording app was pre-installed on the watch to enable continuous audio recording. Our preliminary test showed that the app can continuously record audio for up to 8 hours, so the battery was not a problem. The data was saved on the watch in PCM format with one 16-bit channel at 16 kHz.

In total, we collected data from 18 homes (groups). Groups 8, 9 and 10 consist of three participants including the wearer. The remaining groups consist of two, resulting in a total of 39 participants. The total number of household members per group varies from two to five. The participants age from 15 to 59, with various occupations. Besides, 19 / 39 of the participants are male, and 15 / 18 of the wearers are right-handed. All study participants are fluent in English or native English speakers.

Sound Type	Percentage
Non-vocal noise	38%
Television	22%
Wearer speech	20%
Non-wearer speech	13%
Mixed (wearer&others) speech	3%
Telephone voice	2%
Ambiguous sounds	1%
Baby sounds	1%

Table 2. Instance size distribution of sound classes.

2.2. Annotation

We annotated the audio on a server after the collection by listening back to the audio clips. The gap between the sessions was not used. We annotated the audio with a temporal window of one second with no overlap. The annotation labels we used included *wearer speech*, *non-wearer speech*, *telephone voice*, *television*, *mixed speech*, *baby sounds*, *non-vocal noise* and *ambiguous sounds*. Specifically, class *wearer speech* indicated the case if the speech turn within an instance was exclusively held by the wearer, and *mixed speech* indicated an overlap between the wearer and other speech. *Non-wearer speech* was used when the speech turn was only held by physical participants not wearing the watch. Other observed vocal sounds included *telephone voice*, *television*, or *baby sounds*. Instances of silence or non-vocal background noise were labeled as *non-vocal noise*. *Ambiguous sounds* was used if the annotators were not confident about the sound type. Following the previous definition, we categorized classes *wearer speech* and *mixed speech* both as the foreground speech type, whereas instances of all other vocal and non-vocal classes were counted as the background type.

To annotate the audio, three human annotators were recruited, including a trained researcher. The quality of annotation was evaluated by comparing the pairwise inter-rater reliability between the researcher and the other annotators based on a randomly selected interaction session. Specifically, a mean of 0.907 Cohen's kappa was observed. The kappa value indicates a good agreement of annotation [12].

We obtained a total of 31 hours (111,423 1-second instances) of audio in our dataset. We categorized 23.5% of the total instances as the foreground speech instances. Ambiguous instances accounted for only 1% of the total. Table 2 shows the temporal size distribution of the collected sound classes based on the fine-grained annotation. The imbalance in temporal distribution of the sound classes indicates the nature of people's unconstrained social behaviors.

2.3. Released Data

To facilitate customized model development in relevant studies, we release the raw audio instances with our dataset. Due

to the IRB and study requirements, we applied certain post-processing steps to the data instead of publicizing the original smartwatch recordings. The first step was removing ambiguous instances during the annotation process. After this, the audio segments were randomized. Once the entire session-level recordings were annotated, specifically, we randomly shuffled the temporal order of the 1-second clips. This can be an effective approach to prevent the leakage of user sensitive speech information [13]. In the third step, we randomly mixed the participant groups' data to formulate three folds in the dataset. It is noted that there is no overlap of participant groups between folds so that they can be independently used for supervised model training and validation. The number of participant groups in each fold remains the same. In our dataset, the 1-second audio clips are segmented and released individually with a file name formatted as *label_idx*, where *label* is either 0 (background) or 1 (foreground), and *idx* is a pure count mark without any semantic meanings related to the sessions or participant IDs.

3. BENCHMARK STUDY

To understand the challenge of foreground speech detection based on smartwatch recordings from real users and to validate our collected dataset, we conducted a comparative study based on various settings:

- **Speaker voice match**: As discussed earlier, speaker verification addresses the task of partitioning an audio clip into segments according to who is talking. This problem is similar to foreground speech detection but requires extra information from the speakers and the environment. In this test, we used each participant's reading session as voice samples. We then applied PyAnnote [5] for user voice embedding generation and compared the similarity between embeddings of the tested sounds and embeddings of the sample voice. We applied cosine similarity to measure the feature distance and a sensitivity threshold of 0.5 as the classification boundary for all participants. The test was conducted for each participant group of our dataset, and we reported the overall performance.
- **AMI model**: In prior work, Nadarajan et al. [7] obtained promising results for foreground detection in meeting scenarios with a model trained and refined on a publicly available meeting corpus. Following the same approach, we developed the same supervised neural network classifier based on the raw fast Fourier transform (FFT) features extracted from the public AMI dataset. The dataset contains around 100 hours of close-proximity talk and far-field audio, recorded by multiple meeting participants using wearable headsets and desktop recorders. We then directly tested the model on individual participant groups of our dataset.
- **Personalized model with raw audio features**: In this setting, we developed the AMI model architecture based on our personalized smartwatch data. For a fair comparison, the model was developed following a leave-one-group-out (LOGO) scheme, where all but one group of participants were used to derive a checkpoint model, and the checkpoint model was tested for the remaining group. This process was repeated until results for all participant groups were obtained.
- **Personalized model with transfer learning embeddings**: In addition to the FFT features, we also developed and examined models based on transfer learning embeddings. Our motivation is based on the benefits of transfer learning in various acoustic detection tasks, especially with neural network embedding features [14, 15, 16]. In this test, we examined two types of speaker voice embeddings derived from the public VoxCeleb 1 & 2 [17, 18] datasets and the TIMIT [19] dataset. The test also followed the LOGO scheme.

3.1. Setup of Supervised Training

The FFT features are extracted based on a window size of 50 ms with the same hop and 64 output bins. This results in FFT features of shape (64×20) per second. We removed max pooling for the last convolutional layer of the AMI models to avoid feature size mismatch. For the embedding inputs, we switched the 2D convolutional layers to 1D. The kernel / stride is 3 / 1 for the 1D convolutional layers and 2 / 2 for the 1D max pooling. To train the models, we used the binary cross-entropy loss and the RMSprop optimizer with a learning rate of $1\times e^{-4}$, ρ=0.9, ϵ=0, and momentum=0. The batch size is 128. Besides, the patience for early-stopping is 15 epochs.

To extract the embedding features, we studied two strategies. The first strategy is based on a speaker model [20] trained on the public VoxCeleb 1 & 2 speaker datasets which consist of audio utterances of over 1K celebrities from public YouTube videos. The second strategy is based on a shallow neural network [21] that we trained on a random subset (65 males and 35 females) of the public TIMIT clean speech corpus. We also augmented the clean speech utterances by overlaying them with common household noise. The noise clips were collected by a researcher using a smartphone placed in the home. We obtained 1D embeddings of size 512 per audio instance with the above strategy 1 (*emb1*) and size 1,000 per audio instance with strategy 2 (*emb2*).

3.2. Results of the Benchmark Study

To evaluate the test performance, we leveraged the macro F1 score and the class-balanced accuracy (unweighted mean of class accuracy) metrics so that each of the foreground and the background class can contribute equally to the overall performance. Table 3 shows the results. First of all, we can

Evaluation Setup	Macro F1	Accuracy
Speaker Voice Match	62.7%	61.1%
AMI Model	61.4%	68.2%
Personalized + FFT	**81.5%**	**80.4%**
Personalized + emb1	77.2%	76.3%
Personalized + emb2	79.7%	78.6%

Table 3. The comparative study and benchmark results.

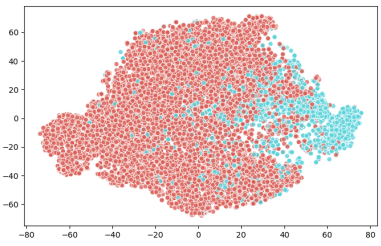

Fig. 1. Feature distributions of the transfer learning embeddings regarding the foreground (blue) and background (red) classes. Data is mixed from two participant groups.

see that the unsupervised verification setup is not reliable on our dataset, especially when no prior information can be accessed ahead of system deployment (e.g., the sensitivity of the feature distance measure cannot be fine-tuned beforehand). Secondly, the model trained with the AMI meeting dataset does not generalize well on our user data as well. This is expected because the training set was collected in meeting scenarios with little daily living noise (e.g., sounds of household appliances, music, public transportation) and human artifacts wearing a watch unobtrusively. These tests show the challenge of foreground speech detection in the real world with commercial wearable devices. As a comparison, models trained with the LOGO setup perform significantly better, which demonstrates the importance of using real-world data for foreground model development. Specifically, the FFT models show the best inference performance. This is reasonable because the raw FFT features fully represent the frequency and temporal patterns of the input audio. The speaker embeddings are slightly worse, since the embedding extractors are trained for a different task, i.e., speaker classification. Nevertheless, there are still benefits of using embeddings rather than raw audio features for foreground analysis, such as less computational burden for long audio segments [3] and better privacy protection, especially when the embedding extractors are non-reversible or unknown to attackers [22].

3.3. Further Analysis for Transfer Learning Embeddings

In an extra study, we further explored the reason why the transfer learning embeddings can enable foreground speech detection. Figure 1 presents the t-distributed Stochastic Neighbor Embedding (t-SNE) [23] plot of embedding features (emb2) for a mixture of two sample participant groups. We can see that embeddings of the foreground and background classes formulate two distinct types. This is interesting because the embedding extractor was trained for speaker classification only, yet the resulting embeddings between the foreground and background classes are highly distinguishable despite the mixture of speakers. A possible reason is that the far-field sound types tend to be a 'null class' to the extractor, which results in embeddings of a unique type.

Setup	Macro F1	Accuracy
emb1+K-Medoids	64.2%	69.3%
emb1+Spectral	**80.4%**	78.8%
emb2+K-Medoids	78.1%	78.5%
emb2+Spectral	79.0%	**78.9%**

Table 4. Foreground detection performance by using heuristic embedding clustering.

Given this finding, we explored a heuristic way of foreground speech detection without training a supervised model. Specifically, we applied K-Medoids and spectral clustering with two output clusters to embedding features extracted from each group of test data. Once the output clusters were formulated, we compared the average inner-cluster cosine similarity from centroid for each output cluster to identify its class label. For every group, we noticed that the output background cluster always came with a higher inner-cluster similarity than the corresponding foreground cluster, possibly because of the high similarity among the null-class embeddings, so we assigned the class labels accordingly. Table 4 shows the overall performance. We can see that the results are generally comparable to the best supervised results, and the results are consistent for both embedding types. This extra study provides a new direction for foreground speech detection when generalizing a supervised model is challenging, for example, because of the lack of training data in a new target environment.

4. CONCLUSION

This paper presents a dataset for speaker-agnostic foreground speech detection with smartwatches in unconstrained daily living situations. The dataset contains foreground and background audio of common in-person social event types, collected from 39 participants. We then conducted a comparative study as the benchmark and discussed the challenge of foreground speech detection in the wild. Furthermore, we explored the leverage of transfer learning embeddings, which inspires a heuristic method for foreground detection without relying on supervised detectors.

5. REFERENCES

[1] Vincent Becker, Linus Fessler, and Gábor Sörös, "Gestear: combining audio and motion sensing for gesture recognition on smartwatches," in *ISWC*, 2019.

[2] Edison Thomaz, Cheng Zhang, Irfan Essa, and Gregory D Abowd, "Inferring meal eating activities in real world settings from ambient sounds: A feasibility study," in *IUI*, 2015.

[3] Rajat Hebbar, Pavlos Papadopoulos, Ramon Reyes, Alexander F Danvers, Angelina J Polsinelli, Suzanne A Moseley, David A Sbarra, Matthias R Mehl, and Shrikanth Narayanan, "Deep multiple instance learning for foreground speech localization in ambient audio from wearable devices," *EURASIP*, vol. 2021, no. 1, pp. 1–8, 2021.

[4] Stuart N Wrigley, Guy J Brown, Vincent Wan, and Steve Renals, "Speech and crosstalk detection in multichannel audio," *IEEE Transactions on speech and audio processing*, vol. 13, no. 1, pp. 84–91, 2004.

[5] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill, "pyannote.audio: neural building blocks for speaker diarization," in *ICASSP*, 2020.

[6] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM TASLP*, vol. 26, no. 10, pp. 1702–1726, 2018.

[7] Amrutha Nadarajan, Krishna Somandepalli, and Shrikanth S Narayanan, "Speaker agnostic foreground speech detection from audio recordings in workplace settings from wearable recorders," in *ICASSP*, 2019.

[8] Bethany Little, Ossama Alshabrawy, Daniel Stow, I Nicol Ferrier, Roisin McNaney, Daniel G Jackson, Karim Ladha, Cassim Ladha, Thomas Ploetz, Jaume Bacardit, et al., "Deep learning-based automated speech detection as a marker of social functioning in late-life depression," *Psychological Medicine*, pp. 1–10, 2020.

[9] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al., "The icsi meeting corpus," in *ICASSP*, 2003.

[10] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al., "The ami meeting corpus: A pre-announcement," in *MLMI*, 2006.

[11] Hall and Watson, "Nasa exercise: Survival on the moon," `https://www.humber.ca/centreforteachingandlearning/assets/files/pdfs/MoonExercise.pdf`, 1970, Accessed: 2020-10-01.

[12] J Richard Landis and Gary G Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.

[13] Dawei Liang, Wenting Song, and Edison Thomaz, "Characterizing the effect of audio degradation on privacy perception and inference performance in audio-based human activity recognition," in *MobileHCI*, 2020.

[14] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM TASLP*, vol. 28, pp. 2880–2894, 2020.

[15] Dawei Liang and Edison Thomaz, "Audio-based activities of daily living (adl) recognition with large-scale acoustic embeddings from online videos," *IMWUT*, vol. 3, no. 1, pp. 1–18, 2019.

[16] Dawei Liang, Yangyang Shi, Yun Wang, Nayan Singhal, Alex Xiao, Jonathan Shaw, Edison Thomaz, Ozlem Kalinli, and Mike Seltzer, "Transferring voice knowledge for acoustic event detection: An empirical study," *arXiv preprint arXiv:2110.03174*, 2021.

[17] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[18] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[19] John S Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium, 1993*, 1993.

[20] Hervé Bredin, "TristouNet: Triplet Loss for Speaker Turn Embedding," in *ICASSP*, 2017.

[21] Yanick X Lukic, Carlo Vogt, Oliver Dürr, and Thilo Stadelmann, "Learning embeddings for speaker clustering based on voice equality," in *MLSP*, 2017.

[22] Joel Shor, Aren Jansen, Wei Han, Daniel Park, and Yu Zhang, "Universal paralinguistic speech representations using self-supervised conformers," in *ICASSP*, 2022.

[23] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *JMLR*, vol. 9, no. 11, 2008.