Toward Measuring Conversation Duration Using a Wristwatch-type Wearable Device

Yuki Komatsu* , Kazuki Shimojo† , Yuuki Nishiyama† and Kaoru Sezaki†

*Ritsumeikan Keisho Senior High School, Hokkaido, Japan , †The University of Tokyo, Tokyo, Japan

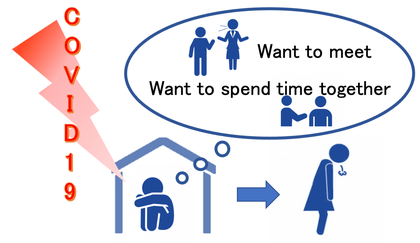
{yuukin,sezaki}@iis.u-tokyo.ac.jp

Introduction

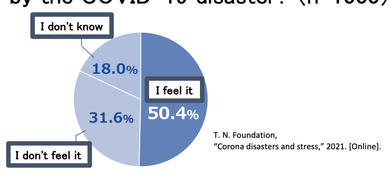
○The frequency of social contact positively correlated with self-rated physical health

The ability to record the frequency of social contact continuously and with little effort would enable a rapid assessment of an individual's situation and subsequent development of countermeasures.

Example: The spread of novel coronavirus infection (COVID-19) has resulted in lifestyle changes.



Q: Do you feel blocked by the COVID-19 disaster? (n=1000)



T. N. Foundation, "Corona disasters and stress," 2021. [Online].

Potential for secondary health effects!!

Assessment of situation → **Subsequent development of countermeasures**

Smartphone places in various locations

↓

Recognition accuracy might be decreased

Originality

Propose system : OHANASHI — System for measuring the duration of conversations in daily life by using an off-the-shelf wristwatch-type wearable device

- Detecting conversation events every second
- On-device Machine Learning
- Recording conversation duration
- Practicality in daily life

Measurement area

Evaluation

①Classification accuracy of the audio classification model

②Battery consumption of the system

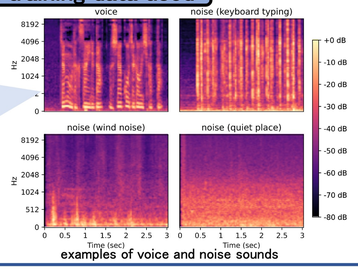
OHANASHI: Measuring Conversation Duration Using A Smartwatch

*1:Ohanashi means "Chatting" in Japanese

「Ohanashi」An overview of our proposed system

Spectrogram of training data used

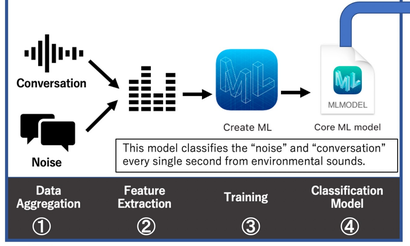
A human voice has characteristic waves around 100 to 1000 Hz



examples of voice and noise sounds

Conversations can be detected using ML

Learning Phase



②Useing Audio Feature Print

③Traininng 100 times

④Using Sound Classification model in CreateML (3.0) on macOS

①The training dataset(6K)

Voice(3K): male voices (1.5K)+ female voices(1.5K) from Common Voice Corpus

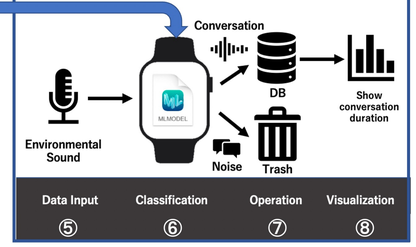
Noise(3K): Breakdown of sample:
- The Dataset for Environmental Sound Classification (2K)*1
- FSDnoisy18k (800)*2
- Radio and TV voices were collected independently with a smartwatch (200)

+ White noise

↓

12K samples

Prediction Phase

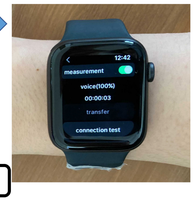


⑤Using AVAudioEngine

⑥SNAudioStreamAnalyzer working with the audio

⑦Determining whether it is classified (each second)

⑧Saving label + Timestamp → Local SQLite database



*1 K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," 2015. [Online]. Available: https://doi.org/10.7910/DVN/YDEPUT

*2 E.Fonseca *et al.*,"Learning sound event classifiers from web audio with noisy labels," in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 21–25.

Results

A. Accuracy of Conversation Detection

Situation

Length of 1 audio file : 1 minute
Number of audio data : 10 noise + 10 conversation
Number of subjects : 1 female + 3 males
Used Equipment : Microphone on Apple Watch SE
Condition : noisy and silent

The average value of the F1-score in silent and noisy conditions was 0.90 and 0.86!

Figure: Noise and Voice classification in two conditions

Conditio	Label	Precision	Recall	F1
Silent	coversation	1.00	0.73	0.84
	noise	0.94	1.00	0.97
	Average	0.97	0.85	0.90
Noisy	coversation	0.99	0.62	0.76
	noise	0.91	1.00	0.95
	Average	0.95	0.81	0.86

B. Battery Consumption

We monitored changes in the battery level by using Apple Watch series 7, 44mm, GPS model.

【Conditions】
①Natural battery consumption with the device on a desk. (Baseline)
②With microphone- activated conditions (With-mic)
③With the model applied condition (With-mic&ml)

Result: The battery life was
Baseline : 54.25hours, With-mic : 18 hours, With-mic&ml : 15.52 hours

Discussion and Conclusion

(F1-score>0.86)

Measuring conversation in controlled real-world settings accurately

Recall value: Low (0.62 in noisy conditions)

Optimized power consumption

Less than a day (15.52 hours)

↓

The power consumption could be optimized by executing conversation detection periodically at regular intervals and continuing the process when a conversation is detected.

Filler (Well..., Uh..., and Um...), Breath and a short Pause

→ Classify as noise → Add these audio clips to the learning dataset.

Our system can extend to detect multiple contexts from sounds

- Recognizing people
- Human-generated sounds (e.g., laugh, cough, and crying)
- Environmental sounds
- Emotion

In the future, we plan to integrate context estimation functions into Ohanashi.

○The accuracy of conversation duration detection in a natural environment needs to be further evaluated.
○The effect of conversation duration measurement and its visualization on the awareness of one's own frequency of social contact and behavioral changes are also issued for future study.