Privacy Preserving Continuous Speech Recording using Throat Microphones

Tim Schneegans
schneegans@teco.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

Leon Simmon
simmon@teco.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

Michael Beigl
beigl@teco.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

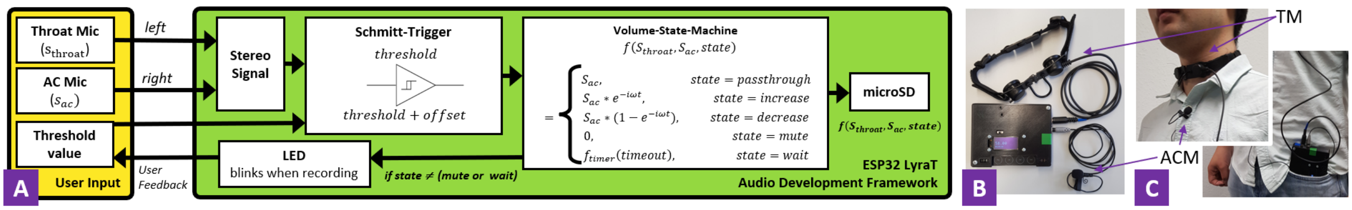

Figure 1: A) Functional Scheme, B) Prototype Device, C) Device worn on a person. ACM=Air-Conduction Mic., TM=Throat Mic.

ABSTRACT

A prerequisite for field-research on audio data are privacy preserving recordings that exclusively contain the target speaker who gave consent. For this purpose, we investigated the potential of a simple but robust wearable technology consisting of three parts: first, a standard air-conduction microphone provides the necessary audio quality for speech analysis; second, a throat microphone is used as a speech activity filter; third, a custom ESP32 based recording device enables on-device real-time processing. The system was evaluated in two challenging free discussion settings with two and four participants each (total N=16). Results from manual annotations show an Equal Error Rate of M=23.4-29.69 %. Based on simple instructions, our participants managed to maintain a False Acceptance Rate below 5 % while recording more than half of their utterances.

CCS CONCEPTS

• **Applied computing**; • **Law, social and behavioral sciences**;

KEYWORDS

speech processing; audio sensors; data privacy protection; Research on behaviour, psychology, and cognition; Biometrics and security

ACM Reference Format:
Tim Schneegans, Leon Simmon, and Michael Beigl. 2022. Privacy Preserving Continuous Speech Recording using Throat Microphones. In *The 2022 International Symposium on Wearable Computers (ISWC '22), September 11–15, 2022, Cambridge, United Kingdom.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3544794.3558480

1 INTRODUCTION

Speech is a major data source for understanding human behavior. Vocal features relate to psychiatric disorders, such as depression, schizophrenia, and bipolar disorder [16]. Everyday word use relates to personality traits [18] and cognitive ageing processes [9]. Wearable sensors help researchers to access such data [17, 25]. Aside of collecting data, speech is used for providing user-adaptive feedback, for example in voice assistants [11, 14].

A prerequisite for everyday audio recordings is to separate target speakers who actively agreed to data collection from non-target speakers who did not agree. Recording latter is legally prohibited in the EU [4, 5, 26]. Typically, an everyday life audio stream contains social situations with a mix of both of these groups. Mehl et al. [17, 19, 20] achieved approval from ethic committees for the US only. Their system collects random speech samples from 5 % of the participant's daytime, which suffices for examining certain research questions, such as social interactions [9, 18].

Throat microphones (TMs) offer a computationally efficient way for detecting the target speaker. TMs use a piezoelectric sensor to record speech via vibrations in the human body. This mechanism decrements any external sound from non-target speakers. This is not possible with *air-conduction microphones (ACMs)* even if they are placed near the user. ACM signals can be filtered using Speaker Recognition and Diarization systems [8, 24, 27, 29], which in practice often require cloud computing and raise additional privacy issues. Compared to earbuds with active noise cancelling [2], TMs allow ear-free interactions. However, TMs do not meet the data quality requirements for extensive speech analysis (but see [21, 22]).

Our contributions are of two kind: First, we propose a wearable system for privacy preserving and continuous audio recordings. For that purpose, we combine the filtering capability of TMs with the high audio quality of ACMs. Previous TM-ACM combinations had the purpose of reducing random background noise but not speech from non-target speakers [6, 28, 30]. Second, we evaluated our filtering principle in a group study in two challenging free discussion settings. While not recording any non-target speakers, we aim to collect more data from the target speaker than current sampling methods (e.g., 5 % [17]).

ISWC '22, September 11–15, 2022, Cambridge, United Kingdom

2 PROTOTYPE

We propose a combination of off-the-shelf ACM and TM that are processed in real-time using an ESP board (Figure 1). The user (target speaker) chooses a decibel-value for the TM signal that works as a threshold value passing or muting the synchronous ACM signal (Figure 1). For the calibration, the user is instructed to move their head, speak with a moderate volume, watch an LED that indicates the recording state, and set their personal threshold. Both microphones were plugged into an ESP32 LyraT v4.3 board, which integrates several audio processing components (e.g., Audio Codec Chip, control buttons). We replaced the built-in stereo microphones with 4-pin audio ports each in order to connect the ACM (left channel) and TM (right channel). We implemented additional software features to improve performance: a Schmitt-Trigger model for the threshold to prevent unstable trigger behaviour; an exponential smooth volume control for the mute/unmute function to remove crackling in the recordings; and a time based unique file name auto-save file writer for stable data storage.

3 EVALUATION

Procedure. We recorded audio data from four groups of four German-speaking participants each (N=16; n=4 female) from a sample of convenience. Most participants (n=14) were 18-25 years old (n=2 were 25-30 years) and had an alto voice (n=8; n=3 bariton/bass, n=4 tenor, n=1 sopran). In a balanced Within-Subject-Design they sat on a table and played a collaborative game (www.keeptalkinggame.com) in two conditions à 20 min: (i) in the entire group and (ii) in a dialogue. A ground truth audio was established by passing the unfiltered ACM signal through the ESP32 board to a smartphone. The study was preregistered on aspredicted.org (#89869). Prior to data collection, a data protection officer was consulted.

Preprocessing. Speaker segments and identification were manually annotated. The annotations were framed into non-overlapping windows of 25 ms each. Because human raters annotate with different levels of precision, labels within 0.25 s after a label change were excluded (15.9 %). Relabeling 13 files (20.3 %) showed an average Inter-Rater-Agreement (IRR) of Kappa=82.2 % (SD=15.5 %) [3], after removing one outlier file. Due to high annotation effort, only the first three minutes of each participant and condition were analyzed. Across participants, this results in 72 min of filtered and 72 min of ground truth annotated audio material. The users spoke for M=1.97 s (SD=0.78 s) in the filtered and M=2.23 s (SD=0.95 s) in the ground truth stream. User and other speakers spoke simultaneously for M=4.73 s (2.62 %) (SD=7.44 s; 4.12 %).

Metrics. Three standard performance metrics were calculated [15, 23]: The (i) False Acceptance Rate (FAR) is the error rate of recording a non-target speaker. We consider the FAR as the privacy level. The (ii) False Rejection Rate (FRR) is the error rate of rejecting a target speaker, i.e., the user. The (iii) Equal Error Rate is the minimum error rate when FAR and FRR are equal. These metrics were analyzed for the real-time filtered audio stream (RT) based on the participant's calibration and a post-hoc analysis (PH) for a range of threshold values based on the ground truth audio. Table 1 shows the FAR and FRR for the self-calibrated threshold values (M=51.8 db, SD=4.2 db) based on RT and PH analysis (also see Figure 2). On average, the participants managed to set a FAR of ca. 5 % or lower.

Results. Overall, the system showed better performance scores in the dialogue than the group condition. Figure 2 shows FAR and FRR for the Post-Hoc analysis and also the tradeoff between security for non-target speakers in speech recordings and coverage of the target speakers utterances.

Table 1: FAR and FRR [%] based on participant calibration.

	Group			Dialogue		
	M	SD	Range	M	SD	Range
FAR_{RT}	4.03	4.56	[0-17.4]	1.8	3.12	[0-12.7]
FAR_{PH}	7.0	7.19	[0.7-25.5]	6.85	9.35	[0-33.8]
FRR_{RT}	46.76	38.16	[0.8-100]	33.42	30.18	[2.8-97.0]
FRR_{PH}	79.54	18.51	[42.7-100]	74.38	18.95	[41.9-99.7]

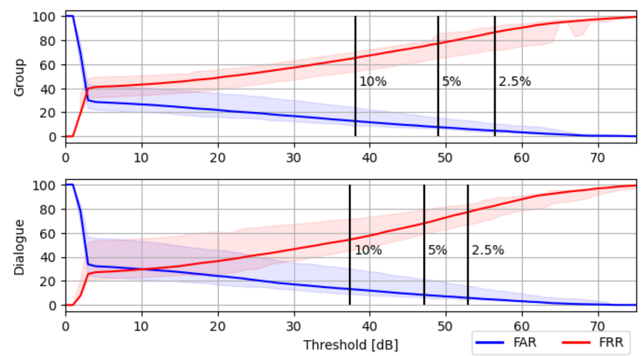

Figure 2: Mean and Standard Deviations for FAR and FRR across thresholds. Vertical lines highlight thresholds with FAR values of 2.5 %, 5 %, and 10 %.

EER was M=29.69 % (SD=17.23 %; group) and M=23.4 % (SD=13.2 %; dialogue). As a benchmark, we modified an i-vector and PLDA based Speech Recognizer [7, 12] to analyze 400 ms frames (approx. two English syllables [1]) in real-time, which performed on EER=22.19 % (dialogue) to 34.57 % (group). By combining both systems where i-Vectors are applied to pre-filtered audio from the TM, we could improve the EER to 20.70 % (dialogue) and 26.37 % (group).

The users showed a moderate comfort score on an adapted version of the comfort rating scale (M=1.9, SD=0.5, range=1.1-2.9) [13].

4 CONCLUSION

We introduced a wearable speech recording system that provides high quality data from the user while protecting the data privacy of non-target speakers. Based on simple instructions, our participants calibrated a small FAR<5 % while recording more than half of their speech. Although our system has a high FRR, it captures more data than current sampling methods [17]. Our study limitations leave room for further evaluation on (i) a larger sample with (ii) freely moving participants and (iii) other conversation scenarios with longer utterances. What remains a challenge is that non-target speakers can be recorded if they speak at the same time as the target speaker. We plan to address this by further augmenting the audio signals with embedded machine learning (e.g., [10]). Moreover, we plan to improve the user comfort of the TM. Finally, a perfect speaker filtering system does not guarantee perfect privacy if the target speaker's data contain personal details of their conversational partners. Future research should examine whether and to what extent non-target speakers can be identified in such a scenario.

ACKNOWLEDGMENTS

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK2739/1 – Project Nr. 447089431 – Research Training Group: KD²School – Designing Adaptive Systems for Economic Decisions

REFERENCES

[1] Melissa M Baese-Berk and Tuuli H Morrill. 2015. Speaking rate consistency in native and non-native speakers of English. *The Journal of the Acoustical Society of America* 138, 3 (2015), EL223–EL228.

[2] Ishan Chatterjee, Maruchi Kim, Vivek Jayaram, Shyamnath Gollakota, Ira Kemelmacher, Shwetak Patel, and Steven M Seitz. 2022. ClearBuds: wireless binaural earbuds for learning-based speech enhancement. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*. 384–396.

[3] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[4] U.S. Congress. 2002. 18 USC 2511: Interception and disclosure of wire, oral, or electronic communications prohibited. https://uscode.house.gov/view.xhtml?req=%28title:18%20section:2511%29. Accessed: 2022-07-09.

[5] Council of Europe. 2018. 128th Session of the Committee of Ministers (Elsinore, Denmark, 17-18 May 2018). Modernised Convention for the Protection of Individuals with Regard to the Processing of Personal Data – Consolidated text. https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=09000016807c65bf.

[6] Engin Erzin. 2009. Improving throat microphone speech recognition by joint analysis of throat and acoustic microphone recordings. *IEEE transactions on audio, speech, and language processing* 17, 7 (2009), 1316–1324.

[7] Daniel Garcia-Romero and Carol Y Espy-Wilson. 2011. Analysis of i-vector length normalization in speaker recognition systems. In *Twelfth annual conference of the international speech communication association*.

[8] Aleksei Gusev, Vladimir Volokhov, Alisa Vinogradova, Tseren Andzhukaev, Andrey Shulipa, Sergey Novoselov, Timur Pekhovsky, and Alexander Kozlov. 2020. STC-Innovation Speaker Recognition Systems for Far-Field Speaker Verification Challenge 2020.. In *INTERSPEECH*. 3466–3470.

[9] Maximilian Haas, Matthias R Mehl, Nicola Ballhausen, Sascha Zuber, Matthias Kliegel, and Alexandra Hering. 2022. The Sounds of Memory: Extending the Age–Prospective Memory Paradox to Everyday Behavior and Conversations. *The Journals of Gerontology: Series B* 77, 4 (01 2022), 695–703. https://doi.org/10.1093/geronb/gbac012 arXiv:https://academic.oup.com/psychsocgerontology/article-pdf/77/4/695/43224411/gbac012.pdf

[10] Simon Haykin and Zhe Chen. 2005. The cocktail party problem. *Neural computation* 17, 9 (2005), 1875–1902.

[11] Shin Katayama, Akhil Mathur, Marc Van den Broeck, Tadashi Okoshi, Jin Nakazawa, and Fahim Kawsar. 2019. Situation-Aware Emotion Regulation of Conversational Agents with Kinetic Earables. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 725–731.

[12] Patrick Kenny. 2010. Bayesian speaker verification with, heavy tailed priors. *Proc. Odyssey 2010* (2010).

[13] James F Knight and Chris Baber. 2005. A tool to assess the comfort of wearable computers. *Human factors* 47, 1 (2005), 77–91.

[14] Effie Lai-Chong Law, Asbjørn Følstad, Jonathan Grudin, and Björn Schuller. 2022. Conversational Agent as Trustworthy Autonomous System (Trust-CA)(Dagstuhl Seminar 21381). In *Dagstuhl Reports*, Vol. 11. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[15] David A van Leeuwen and Niko Brümmer. 2007. An introduction to application-independent evaluation of speaker recognition systems. In *Speaker classification I*. Springer, 330–353.

[16] Daniel M Low, Kate H Bentley, and Satrajit S Ghosh. 2020. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology* 5, 1 (2020), 96–116.

[17] Matthias R Mehl. 2017. The electronically activated recorder (EAR) a method for the naturalistic observation of daily social behavior. *Current directions in psychological science* 26, 2 (2017), 184–190.

[18] Matthias R Mehl, Samuel D Gosling, and James W Pennebaker. 2006. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology* 90, 5 (2006), 862.

[19] Matthias R Mehl, James W Pennebaker, D Michael Crow, James Dabbs, and John H Price. 2001. The Electronically Activated Recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior research methods, instruments, & computers* 33, 4 (2001), 517–523.

[20] Matthias R Mehl, Megan L Robbins, and Fenne große Deters. 2012. Naturalistic observation of health-relevant social processes: The Electronically Activated Recorder (EAR) methodology in psychosomatics. *Psychosomatic medicine* 74, 4 (2012), 410.

[21] Yoshitaka Nakajima, Hideki Kashioka, Nick Campbell, and Kiyohiro Shikano. 2006. Non-audible murmur (NAM) recognition. *IEICE TRANSACTIONS on Information and Systems* 89, 1 (2006), 1–8.

[22] Yoshitaka Nakajima, Hideki Kashioka, Kiyohiro Shikano, and Nick Campbell. 2003. Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, Vol. 5. IEEE, V–708.

[23] National Institute of Standards and Technology. 2021. NIST 2021 Speaker Recognition Evaluation Plan. https://sre.nist.gov/

[24] Wei Rao, Chenglin Xu, Eng Siong Chng, and Haizhou Li. 2019. Target speaker extraction for overlapped multi-talker speaker verification. *arXiv preprint arXiv:1902.02546* (2019).

[25] Timothy J Trull and Ulrich Ebner-Priemer. 2013. Ambulatory assessment. *Annual review of clinical psychology* 9 (2013), 151.

[26] European Union. 2016. Art. 6.1.a, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

[27] Chenglin Xu, Wei Rao, Eng Siong Chng, and Haizhou Li. 2019. Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6990–6994.

[28] Zhengyou Zhang, Zicheng Liu, Mike Sinclair, Alex Acero, Li Deng, Jasha Droppo, Xuedong Huang, and Yanli Zheng. 2004. Multi-sensory microphones for robust speech detection, enhancement and recognition. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3. IEEE, iii–781.

[29] Siqi Zheng, Weilong Huang, Xianliang Wang, Hongbin Suo, Jinwei Feng, and Zhijie Yan. 2021. A real-time speaker diarization system based on spatial spectrum. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7208–7212.

[30] Yanli Zheng, Zicheng Liu, Zhengyou Zhang, Mike Sinclair, Jasha Droppo, Li Deng, Alex Acero, and Xuedong Huang. 2003. Air-and bone-conductive integrated microphones for robust speech detection and enhancement. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*. IEEE, 249–254.