TU Delft

Delft University of Technology

For Your Voice Only

Exploiting Side Channels in Voice Messaging for Environment Detection

Cardaioli, Matteo; Conti, Mauro; Ravindranath, Arpita

DOI
10.1007/978-3-031-17143-7_29

Publication date
2022

Document Version
Final published version

Published in
Computer Security – ESORICS 2022 - 27th European Symposium on Research in Computer Security, Proceedings

Citation (APA)
Cardaioli, M., Conti, M., & Ravindranath, A. (2022). For Your Voice Only: Exploiting Side Channels in Voice Messaging for Environment Detection. In V. Atluri, R. Di Pietro, C. D. Jensen, & W. Meng (Eds.), *Computer Security – ESORICS 2022 - 27th European Symposium on Research in Computer Security, Proceedings* (pp. 595-613). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 13556 LNCS). Springer. https://doi.org/10.1007/978-3-031-17143-7_29

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' - Taverne project

https://www.openaccess.nl/en/you-share-we-take-care

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.



For Your Voice Only: Exploiting Side Channels in Voice Messaging for Environment Detection

Matteo Cardaioli[1,3(✉)], Mauro Conti[1,2], and Arpita Ravindranath[2]

[1] University of Padua, Padua, Italy
matteo.cardaioli@phd.unipd.it
[2] Delft University of Technology, Delft, The Netherlands
[3] GFT Italy, Milan, Italy

Abstract. Voice messages are an increasingly popular method of communication, accounting for more than 200 million messages a day. Sending audio messages requires a user to invest lesser effort than texting while enhancing the message's meaning by adding an emotional context (e.g., irony). Unfortunately, we suspect that voice messages might provide much more information than intended to prying ears of a listener. In fact, speech audio waves are both directly recorded by the microphone and propagated into the environment, and possibly reflected back to the microphone. Reflected waves along with ambient noise are also recorded by the microphone and sent as part of the voice message.

In this paper, we propose a novel attack for inferring detailed information about user location (e.g., a specific room) leveraging a simple WhatsApp voice message. We demonstrated our attack considering 7,200 voice messages from 15 different users and four environments (i.e., three bedrooms and a terrace). We considered three realistic attack scenarios depending on previous knowledge of the attacker about the victim and the environment. Our thorough experimental results demonstrate the feasibility and efficacy of our proposed attack. We can infer the location of the user among a pool of four known environments with 85% accuracy. Moreover, our approach reaches an average accuracy of 93% in discerning between two rooms of similar size and furniture (i.e., two bedrooms) and an accuracy of up to 99% in classifying indoor and outdoor environments.

1 Introduction

Modern chats have replaced feature-poor SMS by adding text images, video, audio, and emoticons. This has allowed instant messaging apps to attract more and more users over the years. In 2020, more than 2.7 billion users used at least one instant messaging app[1]. Nowadays, the most used instant messaging app with over 2 billion users worldwide is WhatsApp[2]. One of the most used functions

[1] https://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps/.
[2] https://www.whatsapp.com/.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
V. Atluri et al. (Eds.): ESORICS 2022, LNCS 13556, pp. 595–613, 2022.
https://doi.org/10.1007/978-3-031-17143-7_29

by WhatsApp users are voice messages, considering that over 200 million are sent every day[3]. Sending a voice message requires even lesser effort for a user compared to texting. Moreover, voice messages allow enriching the message's meaning by adding an emotional context (e.g., irony). Given the appreciation of users, this feature has become common in other messaging apps as well [21], but does a voice message send more than we intend to?

As can be seen in Fig. 1 when a person speaks, the voice signals travel in different paths, some of which undergo reflection. The reflected paths depend on the shape, dimension, furniture that are present in the room. Reflected audio waves end up back at the speaker, causing the persistence of noise called reverberation. In addition, other ambient noises are also present, such as noises from secondary audio sources. The combination of reverberation, noises and the audio message gets picked up by the smartphone during voice messaging. In this work, we aim to use these physical measures that are readily accessible and inadvertently shared during WhatsApp audio messaging to gain intelligence about the victim's whereabouts. To the best of our knowledge, this is the first study that, leveraging short audio messages, identifies the location from which the message was sent. The main contributions we propose in this paper are:

- We propose a novel attack for inferring a specific user location (e.g., a specific room) leveraging simple WhatsApp voice messages.
- We collected a dataset of 15 people and four different environments (i.e., three indoor and one outside) for a total of 7200 recordings (i.e., 480 per participant). We will make the dataset public, available to the research community upon acceptance. We believe it will be useful in studying the problem further and developing countermeasures.
- We performed an extensive analysis of our attack simulating three different real attack scenarios based on the knowledge available to the attacker. We demonstrated that our attack can distinguish the location of the message among a pool of known environments (i.e., three bedrooms and a terrace) with an accuracy of up to 85%. Moreover, we show that our approach reaches an average accuracy of 93% in discerning the voice message location of two rooms of similar size and furniture (i.e., two bedrooms). We further inferred the specific position of a user within a room (e.g., a corner); for this task, we achieved an accuracy of up to 64%.

The structure of the rest of our paper is as follows - In Sect. 2, we discuss previous works related to environment inference using audio signals and location detection. In Sect. 3, we introduce our system and adversary model. Section 4 presents our $\mathcal{F}orY ourV oiceOnly$ attack. The experimental setup and results are discussed in Sects. 5 and 6 respectively. We discuss the limitations, potential future research directions, and concluding remarks in Sect. 7.

[3] https://www.thesun.co.uk/tech/6815812/texts-voice-messages-whatsapp-imessage-switching/.

2 Related Work

Sound classification represents a field of increasing interest in several areas and applications such as, surveillance [26], medicine [33], emotion recognition [34], music genre classification [27], and forensics [31]. The three main disciplines involved in sound classification are: Music Information Retrieval (MIR), [32,36], Automatic Speech Recognition (ASR) [28,37], and Environmental Audio Scene Recognition (EASR) [29,35]. Music and speech can be well described by features such as MFCC (Mel-frequency cepstral coefficients), bandwidth, zero-crossing rate (ZCR), and spectral flux [8,10]. While for the recognition of environments, the problem is more challenging since the sound, in this case, does not present any tonal or harmonic structure [15].

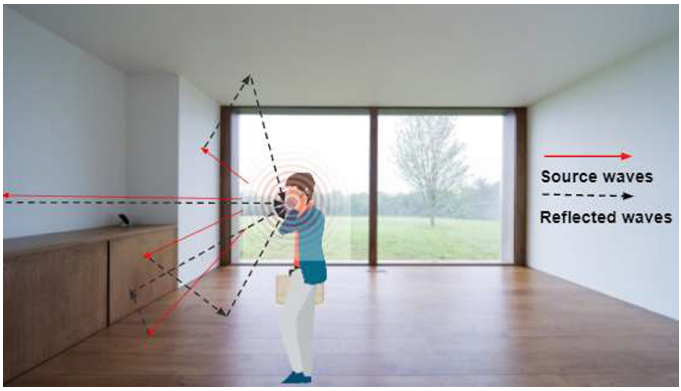


Fig. 1. Voice propagation when sending a voice message

A first comprehensive study on EASR was carried out by Cowling et al. [6]. In this work, the authors explore different feature extraction and classification techniques on EASR, achieving a 70% accuracy leveraging dynamic time warping classification techniques. One of the primary tasks in the EASR domain is the distinction between indoor and outdoor environments. Khonglah et al. [25] proposed the use of foreground speech segmentation to obtain foreground and background segments of an audio recording. Then from the obtained segments, the MFCCs were extracted and used to train an SVM classifier to perform indoor-outdoor classification. In this study, the authors highlighted that the primary cause of misclassification was the presence of speech in the background. Not only speech but also other background noises can induce classification errors. In real-world scenarios, it is quite common to have complex environment sound (i.e., environments with multiple sound sources). To mitigate the impact of complex sounds on environmental prediction performance, Delgado et al. [16] introduced a feature reduction strategy using a Chi-Squared Filter [2]. Unfortunately, a similar approach cannot be applied to the classification of similar locations. Both

speech reverberation and background noise are important sources of information that can describe the environment in which the voice message is recorded.

Recently, many works on EASR have leveraged deep learning algorithms to perform feature extraction and classification [20,23,24] Based on the work conducted by Chandrakala et al. [29] deep learning approaches show better performance compared to traditional machine learning techniques. However, these approaches cannot be applied in our case since they require large amounts of data to train the models.

Additional factors that affect EASR are the recording device's quality and the format in which the sound signal is saved (i.e., lossy audio formats). In this regard, several works have focused on recognizing environments from sounds recorded with resource-constrained devices (e.g., smartphones). Gomes et al. [22] present an application for the smartphone device to classify audio recorded on the device using a combination of SAX-based multiresolution motif discovery in combination with MFCC. The work by Peltonen et al. [5] aims to perform context-based audio scene recognition. However, the data used in this work were obtained using a stereo setup and stored in a digital audio tape recorder. To the best of our knowledge, there are no works in the literature that attempt to identify a specific location (e.g., a specific room) from a voice message recorded by a smartphone.

3 System and Adversary Model

In this section, we describe the system and the adversarial model of our attack. We further discuss the different types of realistic attack scenarios that we identified based on varying levels of information available to the attacker.

System Model. We assume that the victim has a smartphone device with WhatsApp installed and an internet connection. We further assume that the software on the victim device and the device itself is *not compromised* in any manner. While recording the audio messages, we assume that the phone is held at a distance of approximately 15 cm [4,14] from the face of the speaker at an upright position (see Fig. 2). This is one of the most common positions where a phone is held either during video calls or while sending audio messages. Moreover, we conducted an additional preliminary study by placing the phone close to the ear. Results showed that the location inference accuracy was nearly the same across both considered positions.

Adversarial Model. We assume that the attacker has access to the WhatsApp audio message of the victim. The attacker is a user who seeks to learn the location information of the victim. Depending on the attack scenario, the attacker may also have the target's recordings from the same or different positions at

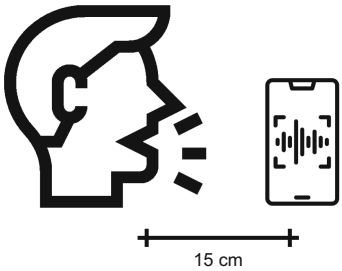


Fig. 2. Recording position

specific locations. Also, the victim is assumed to be in one of these selected locations when recording the audio message. For our experiment, we consider three different scenarios for the attacker:

- *Complete Profiling*: This scenario occurs when the attacker asks the victim to send voice messages from specific locations. For example, an investigator (i.e., the attacker) might ask a suspect (i.e., the victim) to stand in a specific part of a room to verify that the suspect was there or elsewhere at the time a voice message was sent. In this scenario, the attacker has recordings of the victim in all the selected locations. Moreover, the attacker also knows the victim's specific position in the selected locations (e.g., a room corner). In this scenario, the attacker has the highest knowledge to execute his attack.
- *Location Profiling*: In this scenario, the attacker cannot access any of the victim's voice messages other than the one he wants to infer the location. The attacker knows that the victim has sent the voice message from a selected location (e.g., the attacker knows that the victim is in a specific building). Therefore, the attacker can have WhatsApp audio recordings of different speakers but the victim. The speakers are assumed to have recorded their messages at the same locations where the victim is sending the voice message. Hence, the victim is "unknown" while the location position is "known" to the attacker.
- *User Profiling*: This scenario occurs when the attacker owns the victim's voice messages and knows the recording location but does not know the specific position in the location (e.g., a corner of a room) from which they were recorded. The attacker wants to infer the location of a new voice message sent by the victim. Different from the *Complete Profiling* scenario, the attacker cannot ask the victim to send more voice messages from specific positions of the selected locations (e.g., the victim is no longer reachable). The victim is "known" while the position is "unknown" to the attacker in this situation.

Based on the described scenarios, we can identify two main application fields: i) forensics and ii) malicious inference of user information. The *forensic field* is probably the one that would find the most significant benefits both for the wide range of applications (e.g., investigations, evidence in court) and for the high chance of being in the scenario with the highest knowledge (i.e., Complete Profiling). Commonly in forensics, there are no limitations in obtaining additional

voice messages from specific locations. Further, inferring the specific position in a location (e.g., a corner) from which a voice message was sent is of particular interest in forensics. This information can be crucial in understanding whether the suspect or witness could have taken action (e.g., interacted with something nearby) or could see something (e.g., through a window). Malicious inference of user information is another field in which inference of a victim's location from their voice messages finds application. In this case, an attacker can exploit this knowledge to understand whether the victim is in a location (e.g., home or office) and take specific actions (e.g., perform a theft) based on this. A practical application would be an employer who wants to monitor whether an employee is smart working from home or another location. This behavior would be highly invasive of workers' privacy and illegal (since it would occur without the employee's consent) while difficult to detect. Moreover, the malicious inference of user location could allow additional information such as habits, interests, activities, and relationships to be obtained, posing severe privacy concerns.

4 *ForYourVoiceOnly* Attack

Our attack consists of four phases: Data Acquisition, Data Processing, Model Training, and Location Inference. In Fig. 3 we provide an overview of how the attacker conducts the attack. Each of the four phases is discussed in detail in the following sections.

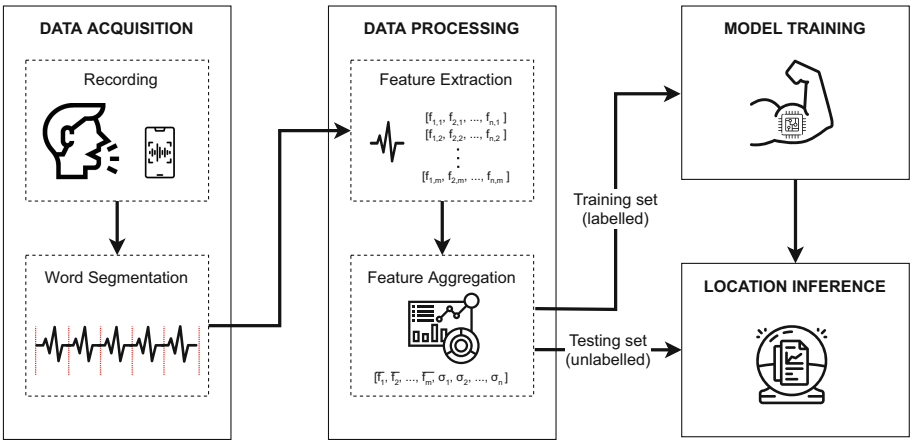


Fig. 3. *ForYourVoiceOnly* attack phases

Data Acquisition. This phase consists of two steps: Recording and Word Segmentation. At the end of the data acquisition phase, the attacker will own two datasets composed of segmented voice messages.

- *Recording*: In this step, the attacker performs two types of data acquisition. The first involves acquiring WhatsApp voice messages recorded by different people (including the victim if allowed by the attack scenario) at some locations or specific positions of interest to build a labeled dataset. The second, for acquiring unlabeled (i.e., both the location or the position are unknown) WhatsApp audio messages of the victim (i.e., test dataset). These two steps do not necessarily have to be consecutive. The attacker can create the labeled dataset even after obtaining the test dataset. The attacker can then choose the locations of interest based on the available information type (e.g., the victim might say she is in one location, but the attacker suspects she is in another known specific location).
- *Word Segmentation*: The attacker segments the recorded voice messages to extract audio fragments related to specific words frequently used in speech [12, 13] (e.g., "*and*", "*of*" and "*the*"). This procedure can be done either manually or by using speech-to-text algorithms[4].

Data Processing. The data processing phase is carried out on both the labeled and the test datasets. This phase consists of two stages: *Feature Extraction* and *Feature Aggregation*.

- *Feature Extraction*: The attacker extracts features that are descriptive of vocal and environmental characteristics: spectral centroid, spectral roll-off, spectral flatness, zero-crossing rate, and Mel-frequency cepstral coefficients [15]. At the end of this step, the attacker has a set of time-frequency features whose dimensionality depends on the duration of the segmented voice message.
- *Feature Aggregation*: Since segmented voice messages may have a variable duration, the attacker needs to process the feature extracted in the previous step to create a feature vector of standardized length. The attacker aggregates the extracted features by calculating the average and the standard deviation as suggested in [7,30]. This procedure allows maintaining information about the magnitude and variability of the data, reducing the total number of features per voice message. At the end of this step, each segmented voice message has a set of 48 associated features.

Model Training. In this phase, the attacker uses only the labeled dataset to train the classification models. The attacker may also decide to train the models using a sub-sample of the dataset based on the owned information. For example, the labeled dataset may contain records from many locations in the acquisition phase, but the attacker has obtained new information about the victim and may discard some of them.

Location Inference. In this phase, the attacker applies the model trained in the *Model Training* phase and predicts the location or the specific location where the victim recorded the message.

[4] https://www.mathworks.com/help/audio/ug/audio-labeler-walkthrough.html.

5 Experimental Setting

In this section, we provide details about the procedure followed during data collection and the characteristics of the obtained dataset. We further provide a comprehensive overview of the machine learning models we used to demonstrate the efficacy of our proposed attack.

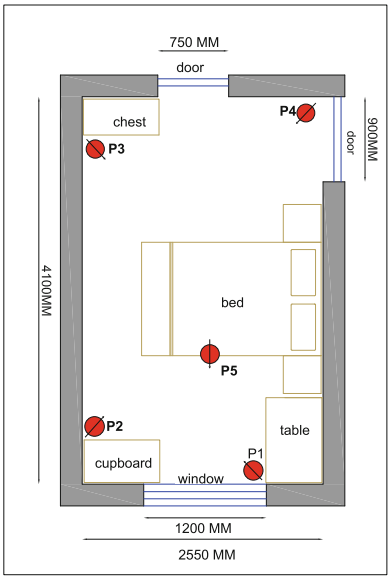
5.1 Data Collection

We performed our data collection at four different real locations. The layouts of these locations are depicted in Fig. 4. In particular, we considered three indoor locations I1 (Fig. 4a), I2 (Fig. 4b), and I3 (Fig. 4c), and one outdoor location O1 (Fig. 4d). Since our goal is to recognize the specific location (or the specific position) from which a voice message is sent for indoor locations, we decided to consider the worst-case where the rooms have a similar layout and furnishings (i.e., bedrooms). Within each of the indoor locations, we further identify five different recording positions: south-east corner (P1), south-west corner (P2), north-west corner (P3), north-east corner (P4), and center (P5). While for O1, we identified a central recording position only (P5).

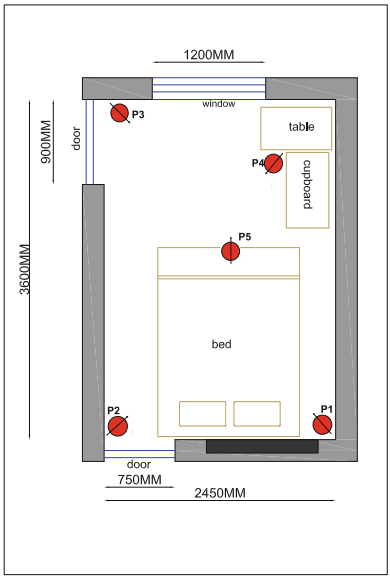
The data collection process involved 15 participants (5 males and 10 females aged 20 to 59 years). In the institution where the experiments were carried out, an IRB approval was not mandatory for this context. All voluntary participants were informed of the actual use of their data and their informed consent was obtained before the recording process. We ensured that the participants held their phones at a distance of about 15 cm from their face at chin level, as shown in Fig. 2. While recording, only the participant was present, and the room doors and windows were closed. To create a more realistic dataset, we asked the participants to use their own smartphone devices[5]. During the collection phase, the participants recorded 30 different voice messages using WhatsApp in all the locations and at each position (see Fig. 2). This results in a total of 150 recordings per indoor location and 30 recordings for the outdoor location. We collected a total of 7200 WhatsApp voice messages, corresponding to 480 recordings per participant.

All the recorded WhatsApp voice messages have a one-second duration (i.e., the minimum duration of a WhatsApp voice message) and contain a single word (i.e., *and*, *of*, or *the*). Specifically, for each position the participants recorded 30 voice messages: 10 pronouncing the word *and*, 10 pronouncing the word *of*, and 10 pronouncing the word *the*. We selected these words based on the OEC, and COCA ranks for most commonly used words during an English conversation [12, 13]. We divided the 30 recordings at a single position into three sequences of 9–12-9. The participant starts the data collection from position P1, recording 9

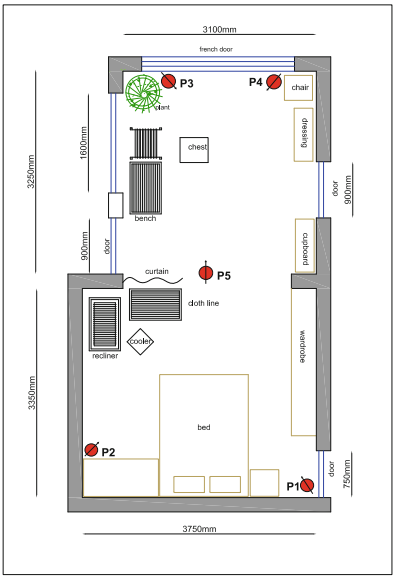
[5] Devices in the data collection: Apple iPhone 7, Apple iPhone X, Apple iPhone 11 pro, Motorola Moto E6, Motorola Moto G3, OnePlus 3, OnePlus 5T, OnePlus 6, OnePlus 6T, OnePLus 6T, OnePlus 8T, OnePlus NORD, Samsung Galaxy A9, Samsung Galaxy A30, and Samsung Galaxy Z Fold 2.



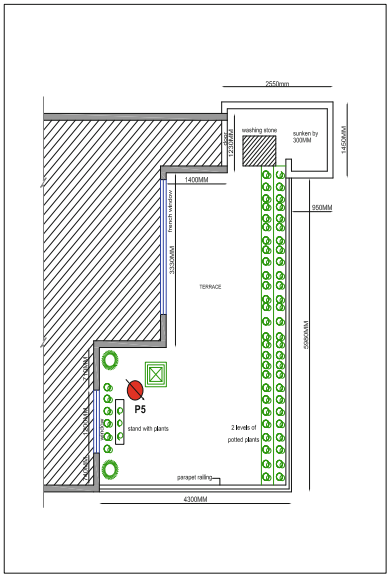
(a) *Indoor location I1 - bedroom.*



(b) *Indoor location I2 - bedroom*



(c) *Indoor location I3 - bedroom*



(d) *Outdoor location O1 - Terrace*

Fig. 4. Location layout and recording positions with orientation considered in the data collection

voice messages at this position (i.e., 3 voice messages per word). Once concluded with this step, the participant moves to P2 in the same location and records 9 voice messages again. After all the five positions are covered in sequence, the participant starts the procedure again from P1, recording 12 voice messages (i.e., 4 voice messages per word). Finally, the participant concludes the data collection with a final set of 9 voice messages per position before moving to the next location. For the O1 location, the participant recorded 30 voice messages from the same position (i.e., P5).

5.2 Feature Extraction

To characterize the location of audio messages, we extracted frame-level features that traditionally were involved in speech recognition and EASR tasks. In particular, for one second of recording (i.e., the minimum duration of a voice message on WhatsApp), we extract 24 features:

- *Zero Crossing Rate (ZCR)*: A temporal feature that indicates the rate at which the signal changes sign [17]. ZCR can also indicate the amount of noise in a signal. A higher ZCR value typically means more noise. ZCR formulation is defined (1)

$$ZCR = \frac{1}{2W_L} \sum_{n=1}^{W_L} |sgn[x_i(n)] - sgn[x_i(n-1)]| \ . \tag{1}$$

  where n is the n-th audio sample and W_L is the length of the considered time window.
- *Spectral Roll-off (SR)*: A spectral feature that measures the bandwidth that contains a certain percentage of the spectral energy [3]. This feature can differentiate harmonic sounds from noisy sounds that usually lie above the roll-off frequency. Further, SR can be used for voiced and unvoiced speech detection [3], and EASR [9]. SR formulation is reported in (2)

$$SR = i \ such \ that \sum_{k=b_1}^{i} |s(k)| = \theta \sum_{k=b_1}^{b_2} s(k) \ . \tag{2}$$

  where s(k) is the power of the k-th frequency bin, θ is the specified frequency threshold, while b_1 and b_2 are the band edges. In this work, we considered a frequency threshold of 85%.
- *Spectral Flatness (SF)*: Also known as Wiener entropy, it is a spectral feature that is used for quantifying how tonal a sound is compared to how noisy it is. SF was applied for singing voice detection [18] and EASR [20] Mathematically this value is calculated as the ratio between the geometric and arithmetic means of a power spectrum. Formally SR can be derived as reported in (3)

$$SF = \frac{(\prod_{k=b_1}^{b_2} s(k))^{\frac{1}{b_2-b_1}}}{\frac{1}{b_2-b_1} \sum_{k=b_1}^{b_2} s(k)} \ . \tag{3}$$

where s(k) is the power of the k-th frequency bin, while while b_1 and b_2 are the band edges.

– *Spectral Centroid (SC)*: Geometrically the centroid represents the arithmetic mean of the positions of the points composing a figure. The spectral centroid is a spectral feature that performs a similar function with respect to a spectrogram. SC is commonly used in music genre classification [1] and it is an indicator of brightness (i.e., upper mid and high frequency content) Mathematically, this value is the weighted mean of the constituent frequencies of a signal, as reported in (4)

$$SC = \frac{\sum_{k=b_1}^{b_2} f(k)s(k)}{\sum_{n=b_1}^{b_2} s(k)} \,. \tag{4}$$

where f(k) is the frequency of the k-th bin, s(k) is the power of the k-th bin, while b_1 and b_2 are the band edges.

– *Mel-Frequency Cepstral Coefficients (MFCCs)*: MFCCs take into account the non-linear behavior of the human auditory system with respect to different frequencies. This is done by converting the spectrum to the mel-scale using a mel filter bank. MFCCs describe the shape of the spectral envelope giving details regarding the timber. MFCCs have been used in the literature for several purposes, such as voice recognition [11] and audio event detection [19]. Furthermore, Gergen *et al.* suggested that MFCCs could be a good descriptor for discerning between anechoic and reverberant signals. In our work, we extracted 20 Mel-frequency cepstral coefficients.

5.3 Machine Learning Models

To identify the location and the specific position in a location of a voice message, we tested four multi-class classifiers: Linear Discriminant Analysis (LDA), Logistic Regression (LR), Ridge Classifier (RC), and Support Vector Machine (SVM). Based on the attack scenario, we applied different strategies to split the data into training, validation, and testing sets:

– **Complete Profiling**: To evaluate the performance of our approach, we apply (for each participant) a nested-cross fold validation. In the outer loop, we use a stratified 5-fold cross-validation on the 480 voice messages recorded by the participant, resulting in 384 recordings in training and 96 in testing per fold. We apply a stratified 3-fold cross-validation in the inner loop on the 384 training recordings, obtaining 256 recordings in training and 128 recordings in validation per fold.

– **Location Profiling**: For this experiment, we consider the entire dataset comprising of 7200 audio recordings, and we apply a nested cross-fold validation. For the outer loop, we apply a user-independent leave-one-out cross-validation, obtaining a testing set containing the recordings of a single participant (i.e., 480). Similarly, in the inner loop, we apply a user-independent

leave-one-out cross-validation on the other 14 participants, obtaining a training set of 13 participants (i.e., 6240 recordings) and a validation set of one participant (i.e., 480 recordings) for each iteration.

– **User Profiling**: In this scenario, we consider the dataset of each participant individually, as for the *Complete Profiling scenario*. Also here, we apply a nested-cross fold validation. Still, different to the *Complete Profiling scenario*, we use a group-k-fold to split the dataset into subsets based on the recording location. We use a group 5-fold cross-validation in the outer loop and a group 4-fold cross-validation for the inner loop. In this way, we split data recorded within the same room into subsets corresponding to each of the 5 recording positions (i.e., P1, P2, P3, P4, and P5). Using this configuration, both the validation and the test sets consist of one subset each, while the training set contains the remaining positions. The recordings from location O1 are excluded from this scenario since they all come from the same position (i.e., P5).

We explored different hyper-parameters by using grid search on all the considered classifiers. In particular, for LDA we vary the solver over [*svd*, *lsqr*, *eigen*]. For LR we vary the solver in [*newton-cg*, *lbfgs*, *liblinear*] and the C value in the range $[10^{-3}, 10^{-2}, \ldots, 10^{1}]$. For RC we vary α from 0.1 to 0.9 with a step size of 0.1, and from 1 to 10 with a step size of 1. Finally, for SVM we tune the values parameter C in the range $[10^{-1}, 10^{0}, \ldots, 10^{3}]$, and γ in the range $[10^{-4}, 10^{-3}, \ldots, 10^{0}]$.

6 Experimental Results

In this section, we report and discuss the results achieved by our approach in the three attack scenarios based on the attack goal: location in Sect. 6.1 or position in Sect. 6.2. Finally, in Sect. 6.3 we prove the applicability of $\mathcal{F}orYourVoiceOnly$ to complex voice messages.

6.1 Location Inference

In Table 1 we show the performance of the classifiers in identifying the location according to the attack scenario, considering the worst case for each scenario (i.e., 4 locations for the *Complete Profiling* and *Location Profiling* scenarios, and 3 locations for the *User Profiling* scenario).

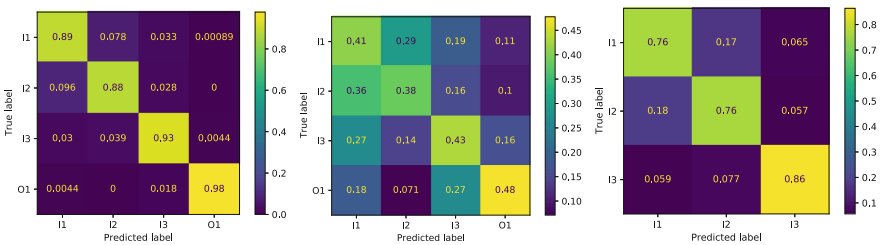
Table 1. Average accuracy of $\mathcal{F}orYourVoiceOnly$ attack for location inference in different attack scenarios

Scenario	LDA	LR	RC	SVM
Complete	0.85 (0.06)	0.85 (0.06)	0.83 (0.06)	0.87 (0.05)
Location	0.41 (0.11)	0.39 (0.10)	0.43 (0.09)	0.35 (0.00)
User	0.80 (0.09)	0.33 (0.04)	0.32 (0.03)	0.33 (0.03)

The scenario where the classifiers perform best is the *Complete Profiling* scenario, where the attacker has the full information available. The results show that in this scenario, all classifiers have accuracy higher than 83%. In particular, the SVM manages to reach an accuracy of 87%. On the contrary, in the *Location Profiling* scenario, there is a consistent drop in performance. In this case, the best classifier is the RC, which reaches an accuracy of 43% (i.e., 18% above the chance level). Lower performance can be attributed to multiple factors:

- Device: the participants used different phones during data collection. The absence of the model in the training set may reduce the accuracy of new test data.
- Training Size: The number of users in training is not enough to ensure sufficient variability in the training features.
- Voice Uniqueness: The distinctiveness of the victim's vocal characteristics cannot be completely replaced, and their lack of training is reflected in performance in testing.

The importance of the victim's voice for the attacker is supported by the results obtained for the *User Profiling* scenario, where the attacker has voice messages from the victim but does not know the specific recording location. In this case, LDA achieves an accuracy of 80% (i.e., only 7% less than in the Complete Profiling scenario), outperforming the others classifiers. In Fig. 5 we show the confusion matrices of the best model per scenario in the location classification. It is interesting to note that the locations I1 and I2 are confused with each other in all three attack scenarios. This is due to the similar layout of the two locations (see Fig. 4). The background noise is instead discriminant for the identification of the external location (i.e., O1). O1 is generally classified better, reaching an accuracy up to 98% in the *Complete Profiling* scenario.



(a) *Complete Profiling scenario.* (b) *Location Profiling scenario* (c) *User Profiling scenario*

Fig. 5. $\mathcal{F}orYourVoiceOnly$ confusion matrices for the best models

Further, we analyzed the influence of the number of locations of interest (i.e., the number of classes to be predicted) on the accuracy of the classification. In the *Complete Profiling* scenario, we obtain an average accuracy of 99% when

we classify an audio message between the outdoor location O1 and one of the indoor locations (i.e., I1, I2, and I3). While when we classify messages between two indoor rooms, we achieve an accuracy ranging from 89% to 95% on this task. Also, in *Location Profiling* scenario, we obtain a higher accuracy if we reduce the location of interest considering O1 and an indoor location. In this case, $\mathcal{F}orY ourV oiceOnly$ correctly predicts the location with an average accuracy of 80%. While for the prediction of internal location pairs, the accuracy remains rather low, ranging from 57% between I1 and I2 to 66% between I1 and I3. Finally, considering the *User Profiling* scenario, reducing the locations of interest to two leads to an average accuracy of 87% in predicting the correct recording location.

Finally, we evaluated $\mathcal{F}orY ourV oiceOnly$ by training the models on a single word, splitting the dataset into three subsets of 2400 audio recordings, each containing the words "*and*", "*of*" and "*the*". Figure 6 depicts the variation of the accuracy of our attack in the *Complete Profiling* scenario between all the locations I1, I2, I3, and O1 using different classifiers and different words. Results show no significant differences between models trained on the specific word and those trained on all words (i.e., combined).

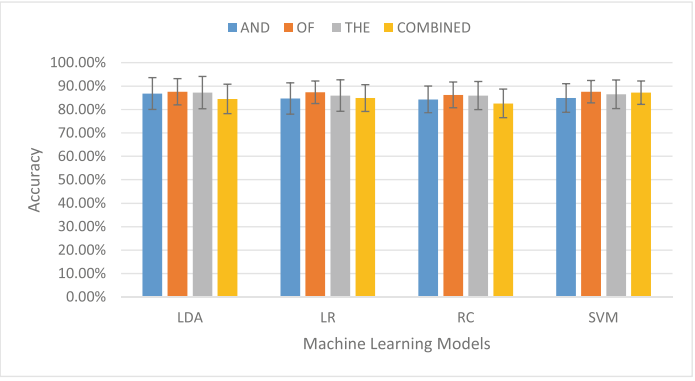


Fig. 6. Performance of machine learning models in classifying the four locations in *Complete Profiling* scenario when trained specifically with one word and all the words (i.e., combined)

6.2 Position Inference

In Table 2 we show the performance of the classifiers in identifying the specific position according to the attack scenario, considering the worst case (i.e., 16 positions - five for each indoor location and one for the outdoor location). Unlike *Location Inference*, here we consider only two attack scenarios (i.e., *Complete Profiling* and *Location Profiling*), since the *User Profiling* scenario assumes that the attacker has no information about the specific position in training. As in Location Inference, even for the position inference, the scenario where the

classifiers perform best is the *Complete Profiling*, and SVM resulted in the best classifier scenario with an accuracy of 61%. Contrarily, in *Location Profiling* scenario models performance is slightly above chance (i.e., 0.0625). The increase in the number of classes to be predicted and the factors already highlighted in Sect. 6.1 (i.e., device, training size, and voice uniqueness) further amplify the performance drop.

Table 2. Average accuracy of $\mathcal{F}or Your Voice Only$ attack for position inference in different attack scenarios

Scenario	LDA	LR	RC	SVM
Complete	0.57 (0.09)	0.55 (0.09)	0.49 (0.08)	0.61 (0.09)
Location	0.13 (0.04)	0.13 (0.04)	0.13 (0.04)	0.07 (0.00)

In Fig. 7 we show the confusion matrix of the best model in the *Complete Profiling* scenario (i.e., SVM). As expected, the model manages to accurately predict O1 (i.e., 98%), demonstrating that this is a trivial task for our attack in this scenario. Regarding the internal locations, Fig. 7 shows a concentration of classification errors in the positions belonging to the true location. In particular, the classification of I3 positions shows less accuracy than I1 and I2. We believe that this can be traced back to the layout of the room. I3 has more than twice the

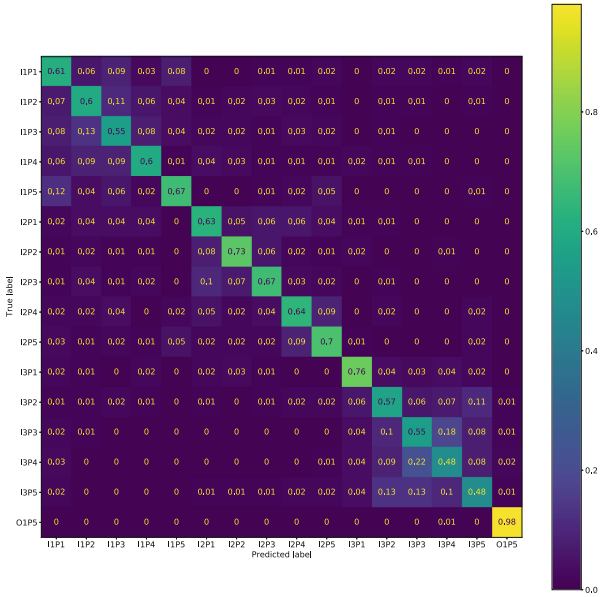


Fig. 7. Confusion matrix for specific position inference for I1, I2, I3 and O1 locations in *Complete Profiling* scenario

surface area of I1 and I2, and the spaces between the recording points and the walls or furniture are much wider. This could lead to a reduction in reverberation and therefore make the recordings more similar. In addition, the best performing position in I3 is P1, which is the recording position with the least open field compared to the other four positions. I1 and I2 generally present better results, but again we can see how room size affects the prediction of the specific location. I2 measures about 2 square meters less than I1 and has a 7% higher average accuracy.

6.3 Extracted Words from Voice Messages

In our experiment, we carried out a data collection on WhatsApp audio messages recording single words pronounced by the participants. However, in a real scenario, voice messages can be of any length. To assess that our approach applied to a real-world context, we carried out a preliminary evaluation on 345 audio samples of words extracted from complex voice messages in the *Complete Profiling* scenario. Also, we reduced the number of rooms in our pool size to 3 (i.e., two indoor bedrooms and one outdoor location-terrace). We noted that $\mathcal{F}orYourVoiceOnly$ reached an average accuracy of 99% in predicting between the outdoor location and any one of the indoor locations. Further, when trying to classify between all the three locations, our attack resulted in an accuracy of 94%. These results demonstrate that $\mathcal{F}orYourVoiceOnly$ can be applied in real-world contexts by extracting single words from a complex voice message.

7 Conclusion

In this paper, we proposed $\mathcal{F}orYourVoiceOnly$, a new attack on voice messages to infer the recording location. $\mathcal{F}orYourVoiceOnly$ leverages attributes such as reverberation and ambient noises, which inadvertently get recorded along with audio messages. We showed the effectiveness of our attack in three realistic attack scenarios: (i) the attacker has previous recordings of the victim in all the selected locations (ii) the attacker has no previous recording of the victim's voice messages (iii) the attacker has previous voice messages of the victim knowing the location they were recorded but does not know the specific position. We demonstrated our attack considering 7,200 voice messages from 15 different users and four environments (i.e., three bedrooms and a terrace). We showed how the possession of audio messages from the victim in known locations dramatically increases the performance of our attack. $\mathcal{F}orYourVoiceOnly$ can infer the user's location among a pool of four known environments with up to 85% accuracy. Moreover, our approach reaches an average accuracy of 93% in discerning between two rooms of similar size and furniture (i.e., two bedrooms) and an accuracy of up to 99% in classifying indoor and outdoor environments.

The results obtained indicate a threat to user privacy. For this purpose, some countermeasures that can be adopted are:

- Adding noise to obscure the leaked information in the audio messages. The noise may also be applied selectively to higher and lower frequencies outside the hearing range so as to not impact the quality of the voice message. This method may prove to act as a countermeasure as we noted variations in the audio signals in the ultrasonic and infrasonic ranges at different locations and positions.
- Shielding the microphone during recording to minimize the environmental noise and to reduce the recorded reverberation.
- Filtering the recorded audio to select only the primary sound source and reducing the information leakage.
- Poisoning the dataset during the training phase (e.g., mislabeling the locations).
- Change the furniture/arrangement of the room.

We believe that the proposed work can be a starting point for developing environment recognition from voice messages that can overcome the limitations of $\mathcal{F}orYourVoiceOnly$. First, the collection of new datasets would allow for more consolidated results and the application of more powerful feature extraction and prediction techniques (e.g., deep learning). The collection of new datasets would also be beneficial for assessing the effect of noisier environments. We made several restrictions during recording, such as having no other member in the rooms during recording, the recordings were done in a relatively quiet and less crowded location. Hence, we expect the behavior to be affected when the noise increases. This can be detrimental or instrumental depending on whether valuable information is obscured or the noise indicates that particular location. Further, it would be helpful to have a more diverse dataset regarding languages, gender, age, nationality, Finally, a new data collection that includes multiple phone holding positions would overcome a limitation of the proposed work.

References

1. Grey, J.M., Gordon, J.W.: Perceptual effects of spectral modifications on musical timbres. J. Acoust. Soc. Am. **63** (5), 1493–1500 (1978)
2. Liu, H., Setiono, R.: Chi2: feature selection and discretization of numeric attributes. In Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence, pp. 388–391. IEEE (1995)
3. Scheirer, E., Slaney, M.: Construction and evaluation of a robust multifeature speech/music discriminator. In 1997 IEEE international conference on acoustics, speech, and signal processing, vol. 2, pp. 1331–1334. IEEE (1997)
4. Kostov, V., Fukuda, S.: Emotion in user interface, voice interaction system. In SMC 2000 conference proceedings. 2000 IEEE International Conference on Systems, Man and Cybernetics. Cybernetics evolving to systems, humans, organizations, and their complex interactions, vol. 2, pp. 798–803. IEEE (2000)
5. Peltonen, V., Tuomi, J., Klapuri, A., Huopaniemi, J., Sorsa, T.: Computational auditory scene recognition. In 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. II-1941. IEEE (2002)

6. Cowling, M., Sitte, R.: Comparison of techniques for environmental sound recognition. Pattern Recogn. Lett. **24**(15), 2895–2907 (2003)
7. Guo, G., Li, S.Z.: Content-based audio classification and retrieval by support vector machines. IEEE Trans. Neural Networks **14**(1), 209–215 (2003)
8. Kim, H.-G., Moreau, N., Sikora, T.: Audio classification based on MPEG-7 spectral basis representations. IEEE Trans. Circuits Syst. Video Technol. **14**(5), 716–725 (2004)
9. Eronen, A.J., et al.: Audio-based context recognition. IEEE Trans. Audio Speech Lang. Process. **14**(1), 321–329 (2005)
10. Chen, L., Gunduz, S., Ozsu, M.T.: Mixed type audio classification with support vector machine. In 2006 IEEE International Conference on Multimedia and Expo, pp. 781–784. IEEE (2006)
11. Bala, A., Kumar, A., Birla, N.: Voice command recognition system based on MFCC and DTW. Int. J. Eng. Sci. Technol. **2**(12), 7335–7342 (2010)
12. Davies, M.: The corpus of contemporary American English as the first reliable monitor corpus of English. Liter. Linguis. Comput. **25**(4), 447–464 (2010)
13. Stevenson, A.: Oxford dictionary of English. Oxford University Press, USA (2010)
14. Hallin, A.E., Fröst, K., Holmberg, E.B., Södersten, M.: Voice and speech range profiles and voice handicap index for males-methodological issues and data. Logoped. Phoniatr. Vocol. **37**(2), 47–61, 2012
15. Okuyucu, Ç., Sert, M., Yazici, A.: Audio feature and classifier analysis for efficient recognition of environmental sounds. In 2013 IEEE International Symposium on Multimedia, pp. 125–132. IEEE (2013)
16. Delgado-Contreras, J.R., Garćıa-Vázquez, J.P., Brena, R.F., Galván-Tejada, C.E., Galván-Tejada, J.I.: Feature selection for place classification through environmental sounds. Procedia Comput. Sci. **37**, 40–47 (2014)
17. Giannakopoulos, T., Pikrakis, A.: Introduction to audio analysis: a MATLAB® approach. Academic Press (2014)
18. Lehner, B., Widmer, G., Sonnleitner, R.: On the reduction of false positives in singing voice detection. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7480–7484. IEEE (2014)
19. Ezgi Küçükbay, S., Sert, M.: Audio-based event detection in office live environments using optimized MFCC-SVM approach. In Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015), pp. 475–480. IEEE (2015)
20. Petetin, Y., Laroche, C., Mayoue, A.: Deep neural networks for audio scene recognition. In 2015 23rd European Signal Processing Conference (EUSIPCO), pp. 125–129. IEEE (2015)
21. Walnycky, D., Baggili, I., Marrington, A., Moore, J., Breitinger, F.: Network and device forensic analysis of android social-messaging applications. Digit. Investig. **14**, S77–S84 (2015)
22. Gomes, E.F., Batista, F., Jorge, A.M.: Using smartphones to classify urban sounds. In Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering, pp. 67–72 (2016)
23. Phan, H., Hertel, L., Maass, M., Mazur, R., Mertins, A.: Learning representations for nonspeech audio events through their similarities to speech patterns. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(4), 807–822 (2016)
24. Eghbal-zadeh, H., Lehner, B., Dorfer, M., Widmer, G.: A hybrid approach with multi-channel i-vectors and convolutional neural networks for acoustic scene classification. In 2017 25th European Signal Processing Conference (EUSIPCO), pp. 2749–2753. IEEE (2017)

25. Khonglah, B.K., Deepak, K.T., Prasanna, S.R.M.: Indoor/outdoor audio classification using foreground speech segmentation. In: INTERSPEECH, pp. 464–468 (2017)
26. Almaadeed, N., Asim, M., Al-Maadeed, S., Bouridane, A., Beghdadi, A.: Automatic detection and classification of audio events for road surveillance applications. Sensors **18**(6), 2018 (1858)
27. Oramas, S., Barbieri, F., Caballero, O.N., Serra, X.: Multimodal deep learning for music genre classification. Trans. Int. Soc. Music Inf. Retr. **1**, 4–21 (2018)
28. Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., Stolcke, A.: The microsoft 2017 conversational speech recognition system. In: 2018 IEEE International Conference on Acoustics, Speech and Signal processing (ICASSP), pp. 5934–5938. IEEE (2018)
29. Chandrakala, S., Jayalakshmi, S.L.: Environmental audio scene and sound event recognition for autonomous surveillance: A survey and comparative studies. ACM Comput. Surv. (CSUR) **52**(3), 1–34 (2019)
30. Nolasco, I., Terenzi, A., Cecchi, S., Orcioni, S., Bear, H.L., Benetos, E.: Audio-based identification of beehive states. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8256–8260. IEEE (2019)
31. Ozkan, Y., Barkana, B.D.: Forensic audio analysis and event recognition for smart surveillance systems. In: 2019 IEEE International Symposium on Technologies for Homeland Security (HST), pp. 1–6. IEEE (2019)
32. Simonetta, F., Ntalampiras, S., Avanzini, F.: Multimodal music information processing and retrieval: survey and future challenges. In: 2019 International Workshop on Multilayer Music Representation and Processing (MMRP), pp. 10–18. IEEE (2019)
33. Faezipour, M., Abuzneid, A.: Smartphone-based self-testing of COVID-19 using breathing sounds. Telemed. e-Health **26**(10), 1202–1205 (2020)
34. Issa, D., Demirci, M.F., Yazici, A.: Speech emotion recognition with deep convolutional neural networks. Biomed. Sig. Process. Control **59**, 101894 (2020)
35. Mushtaq, Z., Shun-Feng, S.: Environmental sound classification using a regularized deep convolutional neural network with data augmentation. Appl. Acoust. **167**, 107389 (2020)
36. Ramírez, J., Flores, M.J.: Machine learning for music genre: multifaceted review and experimentation with audioset. J. Intell. Inf. Syst. **55**(3), 469–499 (2019). https://doi.org/10.1007/s10844-019-00582-9
37. Malik, M., Malik, M.K., Mehmood, K., Makhdoom, I.: Automatic speech recognition: a survey. Multimedia Tools Appl. **80**(6), 9411–9457 (2020). https://doi.org/10.1007/s11042-020-10073-7