Combating Phone Harassment through VoiceAnalysis Filtration of Anonymous Reports

Obonee Kushum
Bangladesh University of Business and Technology
Julkar Nayeen Mahi
Jahangirnagar University,
Milon Biswas (milonbiswas4702@gmail.com)
Bangladesh University of Business and Technology

Method Article

Keywords: Call blocking, Speaker recognition, Spam calls, Harassment, Mobile security, Fraudsters, Mobile phones

Posted Date: August 4th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-52452/v1

Abstract

Given the increasing popularity of smartphones as all-in-one computing devices for corporate work and everyday personal use, it is no wonder that mobile devices have become the most appealing attack surface for today's cyber criminals. In that case obscene or harassing phone calls can be one of the most stressful and frightening invasions of privacy a person experiences. Thus Mobile security has become increasingly important in mobile computing. There exist various applications that block spam calls through the SIM card numbers by establishing a spam database which identities the source of income calls. But unfortunately, their effciency of work is not up to the mark, since its usually pointless to track and block the SIM card number, as the number of spam callers is constantly changed. Considering this point, we are presenting a new concept in which frauds will be recognized through their vocals, even in a noisy environment, with a few seconds of speech, as one can change his number several times but can't change his voice. Here we have used several algorithms and techniques, such as speaker verification, speaker identification, forensic speaker recognition (FSR), spectrogram masking, voice ltering, Mel-Frequency Cepstral Coeffcient (MFCC) and a combination of Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM). Moreover, this system doesn't require any kind of personal information of the users. In this consequence, safety issues also remain in force. Findings of this study will be useful for lawyers, law enforcement agencies, and judges in the courts to recognize their suspects.

Introduction

In our proposed system, to identify the spam callers we have used a voice recognition method that includes both voice identification and voice verification techniques [1]. We also use a simple technique for separating vocals from a noisy environment and a spectrogram masking network [2] to separate the voice of a target speaker from multi-speaker signals from the call recording. Our system only works for unknown numbers, which helps to prevent anyone, from com- plaining against an innocent, for no reason. When an unknown call is arrived in the users phone, the call will be auto recorded through our system. If someone wants to turn off the call recording process for any personal issues or other reasons, the system allows him/her to do so whenever he/she wants. In that case, system wont save any call recordings. But, if the call recording process does'nt turn off by the user, our system will continue its process.

In essence, the proposed system has two scenarios.

First Scenario

It only happens when the unknown caller’s voice doesn’t exist in the database. In this scenario, after receiving the unknown number, if the user doesn’t turn off the automatic call recording process it means he/she may be willing to report this caller.That’s why, once the call is over, the user will be asked whether he/she wants to report that caller or not. If the answer is ”No,” the process will be ”END” and if the answer is ”Yes,” a sorted vocal list will be displayed according to the length or duration of the voice. These voice clips will be extracted from that call recording after eliminating the noise issue.

In this instance, the user basically needs to pick the specific voice from the sorted list that he/she believes was a danger to him/her. If the user gets confused about which vocals, he/she should be picked up, he/she can play these voice clips to ensure. Since the voice will be reported and recorded in the database, it will help to be aware with a notification that it can be a fraud when the threat caller calls again from any number. If a spam caller's voice doesn't already exist in our database, then no notification will be occur in the notification bar. But if he/she thinks it was a threat to him/her, the user can report against this voice.

Second Scenario

It happens when the unknown caller's voice already exists in the database. In this scenario, after receiving the unknown number, a notification will be appeared on the notification bar within a few minutes. It will contain the number of spam reports that have been reported against this specific voice earlier by other victims. That will notify the user to be aware from that spam caller. From this notification the user can get an idea about that person and can take necessary steps to be safe or more observant. The number of spam reports reflects the true depth of the fraud's crime. After the call has ended, if he/she will also willing to report against that person, he/she just needs to pick the specific voice from the sorted list as it was said in the first scenario.

Methods

Proposed Model

The model we proposed for speaker recognition includes both speaker identi- fication and Speaker Verification. These are two major applications of speaker recognition technologies and methodologies [3]. If the speaker claims to be of a certain identity and the voice is used to verify this claim, this is called ver- ification or authentication [4]. On the other hand, identification is the task of determining an unknown speaker's identity [5]. In one way the speaker check is a 1:1 match where the voice of a speaker matches a certain template while the speaker identification is a 1:N match when the voice is compared with several templates. [6].

In figure 1 & 2, our content is explained in detail. Our system only works if there is an incoming unknown call, otherwise it will not yield any results. Also if the unknown caller's voice does not exist in the system's database then no notification will be sent through our system.

If the voice of a spam caller already exists in our database, then two events can take place.

1. A notification will be sent to the user about that spam caller if his voice is previously reported more than 5 times. But

2. If his voice has less than 5 reports, then a notification will be sent only to those users who complained Otherwise, no notification will be sent by our system. So that no one can harass people unjustly. This precaution is taken to save innocent people from defamation.

User can report an incoming call if he/she feels harassed, threatened, tormented, humiliated, embarrassed or otherwise victimised, either he/she receives a spam notification about the unknown caller or not.

When a user wants to report a call, some hidden process will take place that won't be disclosed to the user. In figure 1, phase 1 and 4 will be shown on the phone screen to inform the user about the corresponding actions, but phase 3 and 4 will be hidden from the user. System will perform these tasks to extract the specific voice from a noisy environment and multiple speakers.

The processes shown in figure 1 are almost identical to those indicated in figure 2, but there are 3 additional phases called phase 5, phase 6 and phase 7.Here, phase 7 will appear on the user's screen but phase 5 and 6 are hidden process that sends the user a warning massage.

The following part deals with the working principle of these phases.

Phase 1: Reporting threat call

Immediately after a phone call in which there is a threat of physical harm or violence, the user should report that spam call through our system. In that way, he/she could be freed from this spam caller for the rest of his/her life. Because whenever the user is called by this spam caller from any number, he/she will be notified about this spammer. Also it'll help others to be aware from this spammer.

Phase 2: Extract vocals from noise

In the field of automatic or semi-automatic speaker recognition, background noise is one of the main causes of degradation in performance in various appli- cations of digital speech processing [7]. So we need to reduce background noise, as it helps to improve intelligibility and quality of a speech signal.

Recently, the REpeating Pattern Extraction Technique (REPET) was proposed to separate the repeating background from the non-repeating foreground [8, 9]. The fundamental concept is to define repeating audio components, compare them to repeating the derived models, and extract the repeating patterns through time-frequency masking [10]. While the original REPET (and its extensions) assumes that repetitions happen periodically [11], REPETSIM, a generalization of the method that uses a similarity matrix was further proposed to handle structures where repetitions can also happen intermittently [12].

The only as- sumption is that the repeating background is dense and low-ranked, while the non-repeating foreground is sparse and varied [10].

Repetitions happens in background noise, such as car horn sounds, construc- tion work, crying babies and industrial machinery. All of them have repeated patterns. Considering this point, we have used this algorithm in our proposed system to extract vocals from background noise.

We got the following result by coding this method. Figure 4 shows voices and noise are combined. The vocal element is caused by the wiggly lines above. Our objective is to distinguish them from the instruments we use.

Phase 2: Extract vocals from noise

In the field of automatic or semi-automatic speaker recognition, background noise is one of the main causes of degradation in performance in various appli- cations of digital speech processing [7]. So we need to reduce background noise, as it helps to improve intelligibility and quality of a speech signal.

Recently, the REpeating Pattern Extraction Technique (REPET) was proposed to separate the repeating background from the non-repeating foreground [8, 9]. The fundamental concept is to define repeating audio components, compare them to repeating the derived models, and extract the repeating patterns through time-frequency masking [10]. While the original REPET (and its extensions) assumes that repetitions happen periodically [11], REPETSIM, a generalization of the method that uses a similarity matrix was further proposed to handle structures where repetitions can also happen intermittently [12]. The only as- sumption is that the repeating background is dense and low-ranked, while the non-repeating foreground is sparse and varied [10].

Repetitions happens in background noise, such as car horn sounds, construc- tion work, crying babies and industrial machinery. All of them have repeated patterns. Considering this point, we have used this algorithm in our proposed system to extract vocals from background noise.

We got the following result by coding this method. Figure 4 shows voices and noise are combined. The vocal element is caused by the wiggly lines above. Our objective is to distinguish them from the instruments we use vocals and background noise are separated in two slices.

Phase 3: Separation of spam caller from multiple speaker

The next phase is about to separate the voice of the spam caller from multi- speaker signals by making use of a reference signal from the target speaker. This process is presented in [13]. One way to deal with this issue is to first apply a speech separation system on the noisy audio in order to separate the voices

from different speakers. Therefore, if the noisy signal contains N speakers, this approach would yield N outputs with a potential additional output for the noise [14].

This approach can be easily extended to more than one speaker of interest by repeating the process in turns, for the reference recording of each target speaker [13].

Phase 4: Saving voice in database

In phase 2 and 3, target voice has already been detected. In this phase, system will save this specific voice in database for further actions. Whenever this person will call the user from any number, the system will match his voice with the saving one by some complex method and send a notification to the user, that this man can be harmful or dangerous for him/her.

Phase 5: Voice recognition in database

The identification of a person through speech samples with a forensic quality is challenging.In this phase caller's voice will be checked whether it matches our database. For this purpose we have used a method for forensic speaker recog- nition that has been proposed in [15]. Here each speakers voice is recorded in both clean and noisy environments, through a microphone and a mobile channel though it has shown low equal error rates (EER) with very short test samples. This diversity facilitates its usage in forensic experimentation. The Gaussian mixture model-universal background model is used for speaker modeling and Mel-Frequency Cepstral Coefficients are used to extract features [16].

Phase 6: Creating Spam reports

Whenever a user reports a spam voice in database, system will save that vocal and create a profile for that corresponding spammer, in which the number of spam reports will be stored. If a voice has already been reported by a user that means this voice has an individual profile with it's corresponding spam reports. Hence, if anyone again reports this spam voice, no profile will be created but the number of reports against this person will be increased by our system.

Phase 7: Sending spam notification

Notification will be send when the spam callers voice already exists in the database. In this scenario, after receiving the unknown number, a notification will be appear on the notification bar within a few minutes. It will contain the number of spam reports that have been reported against this specific voice ear- lier by other victims. That will notify the user to be aware from that spam caller. From this notification the user

can get an idea about that person and can take necessary steps to be safe or more observant. After the call has ended, if he/she will also willing to report against that person, he/she just needs to pick the specific voice from the sorted list as it was explained earlier.

Results

Our proposed system has been completed in a few steps. So the result of this research has to be shown in various steps

Extracting vocals from noisy environment

Tables 1, 2, and 3 demonstrate the outcomes for SDR (dB) and OPS, for stereo voice estimates (sim) and stereo noise estimates (noi), for all techniques, respec- tively for subway noise, cafeteria noise and square noise estimates [10]. Here,

- Algorithm 5 is based on a first constrained ICA that estimates the mixing parameters of the target source, followed by a Wiener filtering to enhance the separation results [17]. - Algorithm 8 is based on a first estimation of the noise from the unvoiced segments, followed by DUET [18] and spectral subtraction to refine the results, and a minimum-statisticsbased adaptive procedure to refine the noise estimate [19]. - Baseline is based on a first estimation of the Time Differences Of Arrival (TDOA) of the sources, followed by a maximum likeli- hood target and noise variance estimation under a diffuse noise model, and a multichannelWiener filtering [20]; this is the baseline algorithm proposed by SiSEC.

Table 1. SDR (dB) and OPS results for the subway noises.

		dev*s u*1 *c eA*		dev*s u*1 *c eB*	
		sim	noi	sim	noi
REPET-SIM	SDR	-0.5	15.4	5.2	14.1
	OPS	15.9	31.3	30.7	22.4
Algorithm 5	SDR	0.9	5.7	-2.3	1.8
	OPS	21.7	10.0	33.6	9.7
Algorithm 8	SDR	-7.8	8.1	-0.7	8.2
	OPS	13.4	12.4	32.2	20.1
Baseline	SDR	-5.0	10.9	0.5	9.4
	OPS	20.5	29.9	28.9	18.3

Table 2. SDR (dB) and OPS results for the cafeteria noises.

		dev Ca1 Ce A		dev Ca1 Ce B		dev Ca1 Co B		dev Ca1 Co B	
		sim	noi	sim	noi	sim	noi	sim	noi
REPET-SIM	SDR	5.4	1.3	8.0	3.7	9.2	5.6	9.2	5.6
	OPS	33.6	23.6	23.7	31.0	30.7	26.6	30.7	26.6
Algorithm 5	SDR	4.7	0.8	10.9	2.8	5.1	0.8	5.1	0.8
	OPS	42.9	24.0	35.4	25.3	31.4	17.1	31.4	17.1
Algorithm 8	SDR	3.4	-0.8	6.3	2.1	7.1	3.6	7.1	3.6
	OPS	34.6	18.1	27.5	24.3	31.1	24.4	31.1	24.4
Baseline	SDR	0.3	-3.9	4.7	0.4	-3.5	-7.0	-3.5	-7.0
	OPS	8.9	9.7	33.1	27.8	22.9	8.3	22.9	8.3

Table 3. SDR (dB) and OPS results for the square noises.

		dev Sq1 Ce A		dev Sq1 Ce B		dev Sq1 Co B		dev Sq1 Co B	
		sim	noi	sim	noi	sim	noi	sim	noi
REPET-SIM	SDR	4.4	9.1	5.1	9.5	5.1	10.7	8.6	10.8
	OPS	32.9	27.1	32.1	27.4	34.1	35.8	36.9	31.1
Algorithm 5	SDR	-0.8	0.8	8.7	5.5	-2.8	0.8	10.8	6.5
	OPS	38.4	15.3	26.9	15.8	36.5	17.3	42.6	18.3
Algorithm 8	SDR	1.7	6.5	3.4	7.8	2.2	7.8	6.0	8.3
	OPS	30.3	17.4	33.0	16.4	29.4	14.0	34.4	17.0
Baseline	SDR	-21.1	-16.4	-21.1	-16.7	-17.5	-12.0	-14.4	-12.2
	OPS	23.6	25.9	8.6	17.9	35.0	30.5	14.5	29.9

As we can see the REPET-SIM is nearly always better than that of Algorithm 8 and Baseline and is performing, as well as of Algorithm 5. This makes sense because REPET-SIM models only the noise [10].

Multi-Vocals Separation

In [13], authors have demonstrated the effectiveness of using a discriminatively- trained speaker encoder to condition the speech separation task. Such a system is more applicable to real scenarios because it does not require prior knowledge about the number of speakers and removes the permutation problem. The VoiceFilter model trained on the LibriSpeech data set also shows that the voice recognition WER in two-language scenario decreases from 55.9% to 23.4% and WER in single-speaker situations stay about the same.

Speaker Recognition

In [15], speech of 40 speakers are used to validate the proposed method while recording the mobile channel. The recording of low bandwidth and low-quality devices is not great on mobile channels. The training set includes recording the statements of one paragraph over an average of 30 s and 10 s for testing by using the mobile channel through smooth voice recording; Figure 6 demonstrates the experiments ' DET curves. It can be said from Figure 6 that important efficiency was accomplished using mobile channel recording at a speed of approximately 97.8% with an EER equivalent to 1.98% [15].

Conclusion

This paper emphasized the solution to reduce mobile telephone harassment by filtering anonymous reports using voice recognition and data analysis. We present a new concept whereby the vocals of fraud are recognized with a few seconds of speech, even in a noisy environment, because the SIM card number can be changed a couple of times, but the vocal of a person cannot be changed.It will reduce harassment, threats, torments, humiliation, embarrassment or victimized by phone calls. Our system overcomes the limitation of existing applications and provides more security than other applications, as it doesnt require any personal information of its user.

Declarations

Competing Interests:

The authors declare no competing interests

References

1. Lu, A. B. Brush, B. Priyantha, A. K. Karlson, and J. Liu, “Speakersense: energy efficient unobtrusive speaker identification on mobile phones,” in *International conference on pervasive computing*, pp. 188–205, Springer, 2011.
2. Li and K. C. Sim, “A spectral masking approach to noise-robust speech recognition using deep neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 8, pp. 1296–1305, 2014.
3. A. Reynolds, “Speaker identification and verification using gaussian mixture speaker models,” *Speech communication*, vol. 17, no. 1-2, pp. 91–108, 1995.
4. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.
5. J. Phillips, A. Martin, C. Wilson, and M. Przybocki, “An introduction to evaluating biometric systems,” *Computer*, no. 2, pp. 56–63, 2000.
6. Fine, J. Navratil, and R. A. Gopinath, “A hybrid gmm/svm approach to speaker identification,” in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 1, pp. 417–420, IEEE, 2001.
7. Ozerov and E. Vincent, “Using the fasst source separation toolbox for noise robust speech recognition,” in *International Workshop on Machine Listening in Multisource Environments (CHiME 2011)*, 2011.
8. Rafii and B. Pardo, “A simple music/voice separation method based on the extraction of the repeating musical structure,” in *2011 IEEE International Con- ference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 221–224, IEEE, 2011.
9. Rafii and B. Pardo, “Repeating pattern extraction technique (repet): A simple method for music/voice separation,” *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 1, pp. 73–84, 2012.
10. Rafii and B. Pardo, “Online repet-sim for real-time speech enhancement,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 848–852, IEEE, 2013.
11. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard, “Adaptive filtering for music/voice separation exploiting the repeating musical structure,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 53–56, IEEE, 2012.

12. Rafii and B. Pardo, “Music/voice separation using the similarity matrix.,” in *ISMIR*, pp. 583–588, 2012.

13. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous,
14. J. Weiss, Y. Jia, and I. L. Moreno, “Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking,” *arXiv preprint arXiv:1810.04826*, 2018.
15. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Dis- criminative embeddings for segmentation and separation,” in *2016 IEEE Interna- tional Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 31–35, IEEE, 2016.
16. Algabri, H. Mathkour, M. A. Bencherif, M. Alsulaiman, and M. A. Mekhtiche, “Automatic speaker recognition for mobile forensic applications,” *Mobile Informa- tion Systems*, vol. 2017, 2017.
17. Drygajlo, “From speaker recognition to forensic speaker recognition,” in *Inter- national Workshop on Biometric Authentication*, pp. 93–104, Springer, 2014.
18. Nesta and M. Matassoni, “Robust automatic speech recognition through on-line semi blind source extraction,” in *Machine Listening in Multisource Environments*, 2011.
19. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Transactions on signal processing*, vol. 52, no. 7, pp. 1830–1847, 2004.
20. Rangachari and P. C. Loizou, “A noise-estimation algorithm for highly non-stationary environments,” *Speech communication*, vol. 48, no. 2, pp. 220–231, 2006.
21. Blandin, A. Ozerov, and E. Vincent, “Multi-source tdoa estimation in reverberant audio using angular spectra and clustering,” *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.

Figures

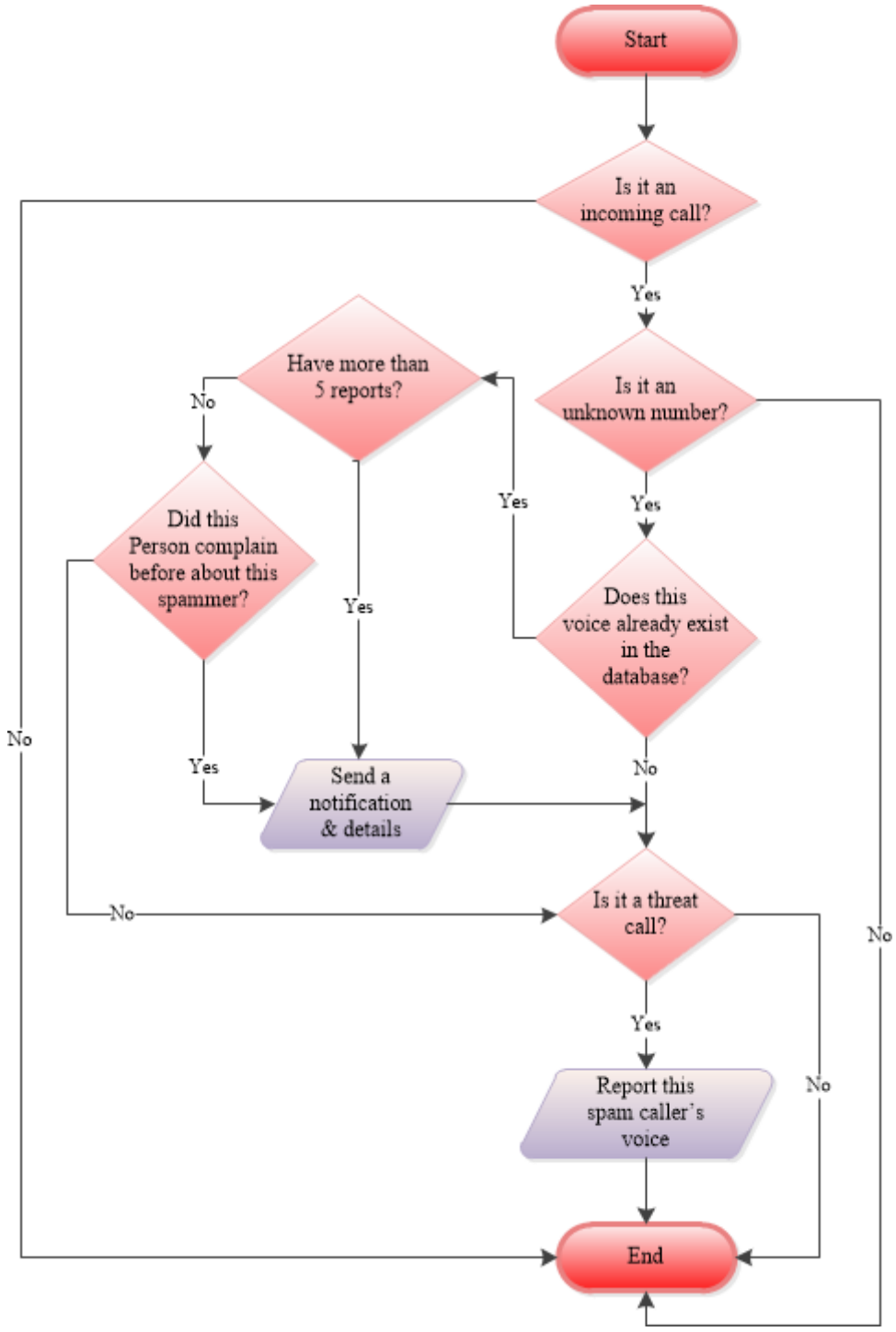

Figure 1

Full system Flow chart

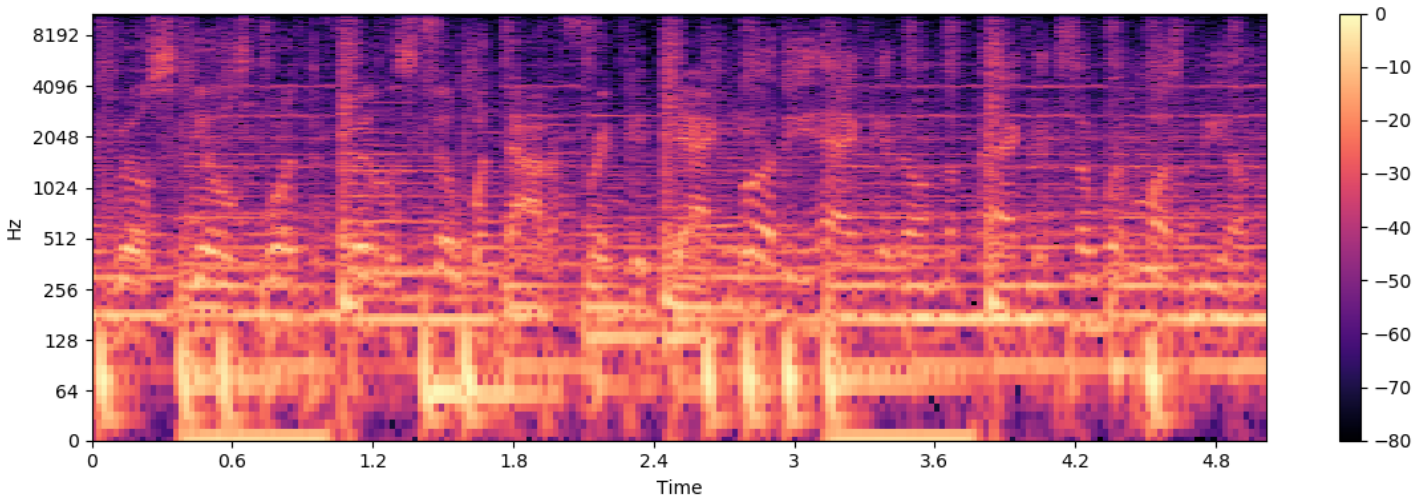

Figure 2

Vocal with noise

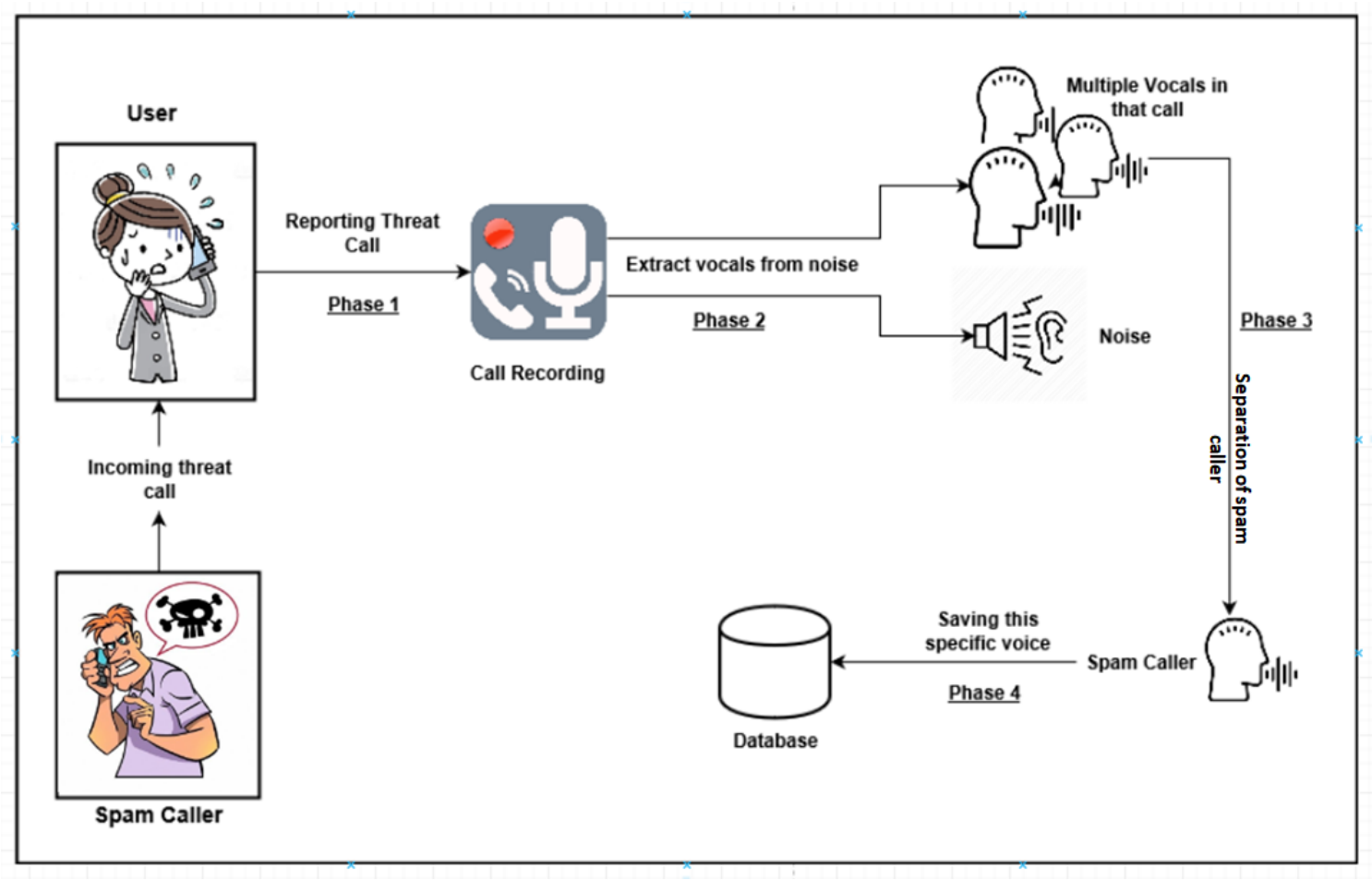

Figure 3

Reporting Spam Voice

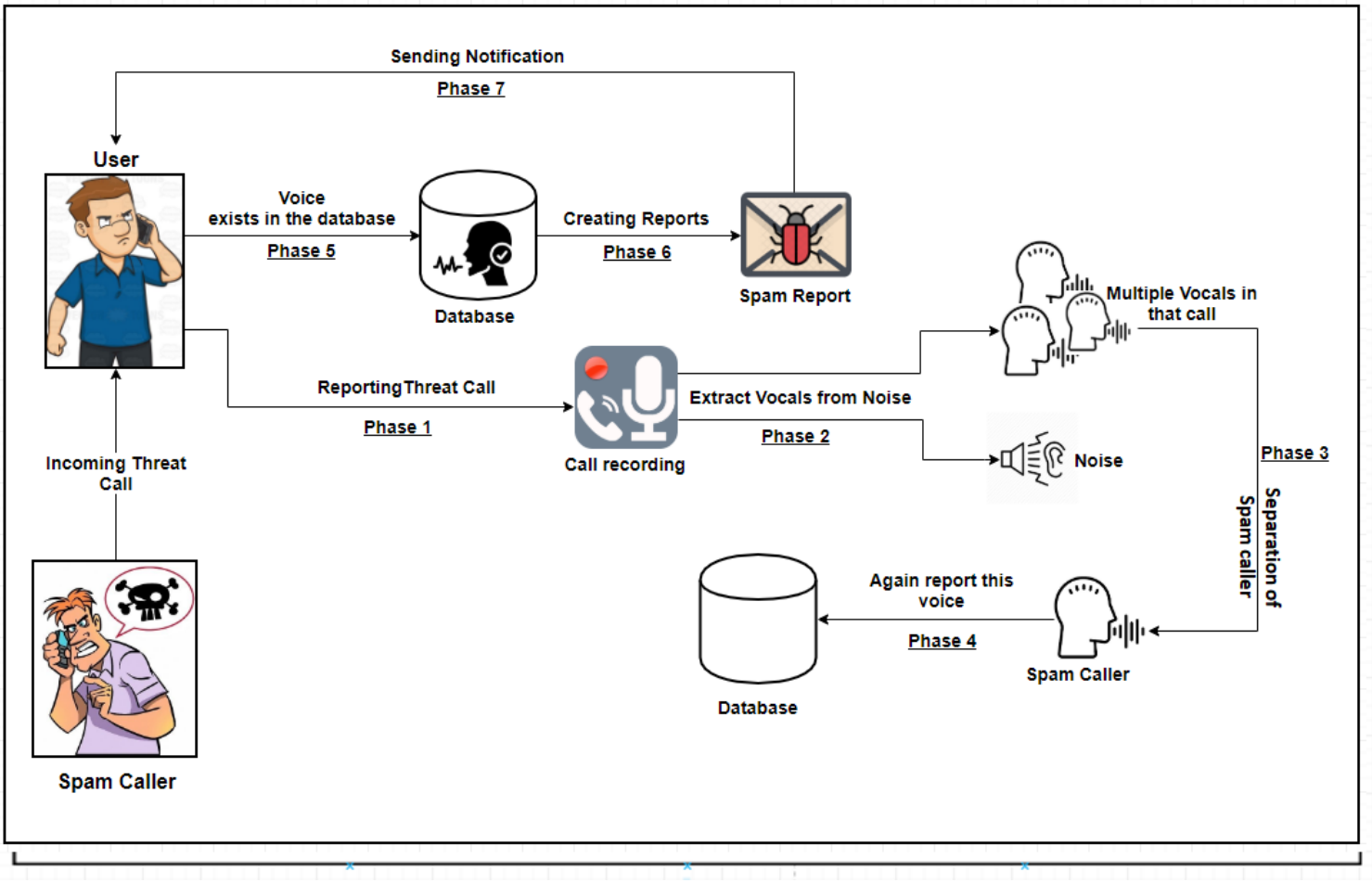

Figure 4

Get notication about spam report

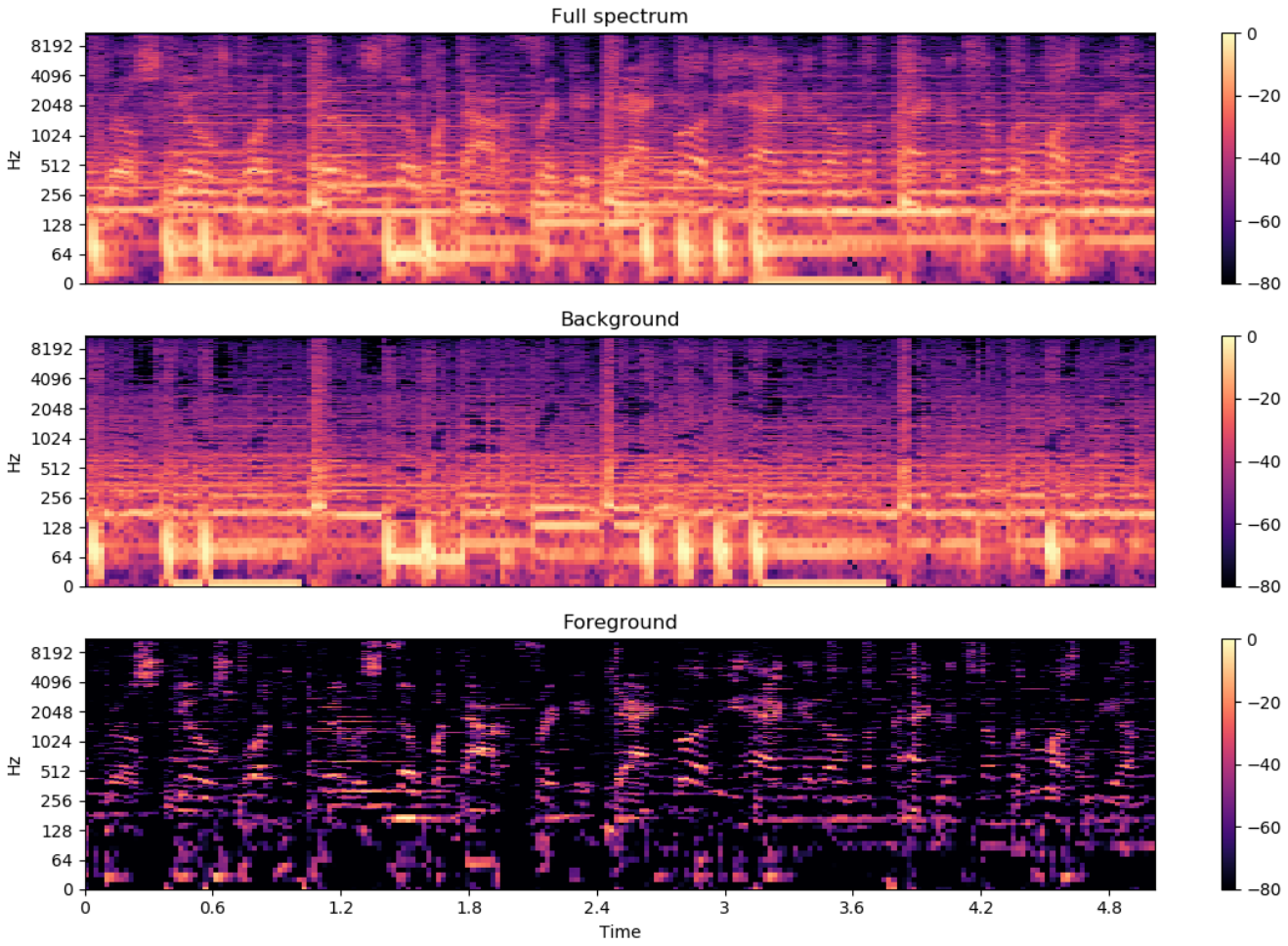

Figure 5

Separating vocal and noise