ISSN 2007-9737

A Fuzzy Approach to Mute Sensitive Information in Noisy Audio Conversations

Imran Sheikh[1], Balamallikarjuna Garlapati,[2] Srinivas Chalamala,[2] Sunil Kumar Kopparapu[1]

[1] TCS Research and Innovation - Mumbai,
Tata Consultancy Services,
India

[2] TCS Research and Innovation - Hyderabad,
Tata Consultancy Services,
India

{imran.as, balamallikarjuna.g, chalamala.srao, sunilKumar.kopparapu}@tcs.com

Abstract. Audio conversation between a service seeking customer and an agent are common in a voice based call center (VbCC) and are often recorded either for audit purposes or to enable the VbCC to improve their efficiency. These audio recordings invariably contain personal information of the customer, often spoken by the customer to confirm their identity to get personalized services from the agent. This private to a person (P2aP) information is the recordings is a serious concern from the GDPR perspective and can lead to identity theft among other things. In this paper, we propose a robust framework that enables us to reliably spot the P2aP information in the audio and automatically mute it. The main contribution of this paper is the proposal of a novel fuzzy criteria which by design allows for reduced false alarms and at the same time increases the accuracy of the muting process even when the the speech to text conversion process is erroneous. Evaluation on real call center conversations demonstrates the reliability of the proposed approach.

Keywords. Fuzzy-muting, masking audio.

AGT: */how are you today/*

CUS: */i want to know my err my credit card dues/*

AGT: */can i have your name please/*

CUS: */~~ralf ralf edison~~/*

AGT: */mr ~~edison~~. can i confirm your mobile number?/*

CUS: */sure it is ~~zero double one nine nine six two~~/*

AGT: */thank you. please confirm your credit card number/*

CUS: */my credit card number is ~~five three one six nine seven seven four one six nine seven eight two five zero~~/*

Fig. 1. Sample audio conversation with P2aP information to be muted struck off

1 Introduction

Private to a Person (P2aP) Information can easily lead to identity theft if recorded audio conversations, between agents and customers in a call center, get into wrong hands. And more recently, the General Data Protection Regulation (GDPR) [5] makes it strict for an enterprise to refrain from retaining P2aP information. For these reasons, an enterprise, would like to *mute* all confidential (P2aP) information in their audio recordings, while retaining the rest of the conversation for audit or other purposes. Given an audio conversation (see Fig. 1), a muting approach should mute all instances of P2aP information while retaining the rest of the audio as is.

ISSN 2007-9737

A typical P2aP information muting process uses a speech to text (S2T) or a key word spotting (KWS) engine to ”locate” all the instances of P2aP information in the audio and then mute (erase) those portions in the audio. The process of locating P2aP information is dependent on the accuracy of the S2T or KWS engine. While the accuracy of S2T is getting better it is not yet perfectly reliable for conversational natural language speech as is the case during an Agent Customer interaction [9]. Noisy transcriptions, which are inherent in a S2T conversion process lead to false positives (Type I errors — non P2aP information are marked as P2aP), as well as false negatives (Type II errors — actual occurrences of P2aP are missed being identified).

Protection and masking of P2aP information has been widely studied in the fields of data security and data mining [12, 2]. Privacy of human speech in spaces like buildings, offices, hospitals, etc. has also been studied for a long time [4]. The focus, however has been to maintain a degree of privacy between the speaker and the intended listener, independent of the spoken content. Most methods available to mute P2aP information focus on (a) locating P2aP information in the audio conversation and then (b) muting the located regions [8, 11, 3, 7, 6, 17] in the audio. Almost all of them use some kind of S2T or KWS or acoustic pattern matching technique. Some of the more recent techniques focus on pattern matching in erroneous S2T hypothesis [7]. In all cases a binary decision is taken (mute or not-to-mute) based on a threshold associated with the located P2aP information inferred by the S2T engine. This binary decision works well for a S2T process which is accurate, however as mentioned earlier, the S2T process especially for natural spoken conversations is far from perfect.

In this paper, we introduce the concept of fuzzy-muting, a technique which is robust even when the S2T conversion process is not perfect. This is the main contribution of this paper. The rest of the paper is organized as follows Section 2 we dwell on the problems in existing P2aP information muting approaches. We present our approach in Section refsec:approach.

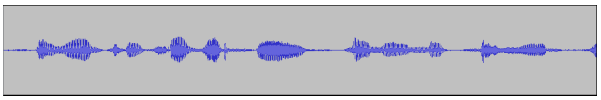

Fig. 2. Audio signal

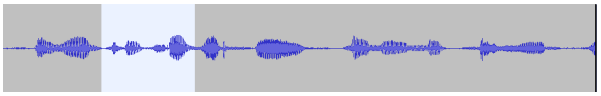

Fig. 3. Portion of the audio signal marked to be muted

Section 4 presents experimental analysis on real audio conversations and we conclude in Section 5.

2 Muting Problems

The basic process of muting P2aP information in audio involves three steps, namely, Step (a) – identify the set of patterns that are representative of the P2aP information, Step (b) – a mechanism to locate these patterns in the actual audio conversation, and then Step (c) – replace the located regions in the audio signal by silence or white noise (as depicted in Fig. 4).

Clearly Step (a) is a preparatory step and typically involves identifying a set of keywords or key-phrases and their possible occurrence patterns for a given conversation scenario and the Step (b) is the audio to text conversion process using a S2T (ASR) or KWS engine. The S2T engine gives a string of recognized words or labels and their corresponding confidence scores (s_c). This confidence score plays a major role in muting P2aP information in existing approaches. There are several issues that crop up. We enumerate them below.

Problem # 1 Speech to text (S2T) process is inherently erroneous for natural language conversational speech, subsequently these lead to false positives (Type I error) and false negatives (Type

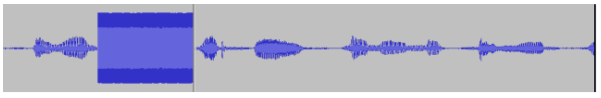

Fig. 4. Muting of the audio signal marked in Fig. 3

II error) in locating P2aP information instances. To overcome the Type I errors existing techniques set a threshold value (τ) on the log likelihood confidence score determined by the S2T engine and mute all the instances of words whose log likelihood score $s_c \geq \tau$ and which corresponding to the P2aP information patterns as determined in Step (a). If t_s denotes the start time of the identified P2aP word or phrase and t_d denotes the duration of the P2aP word and if s_c is the log likelihood score of the P2aP word or phrase as determined by the S2T engine then muting of the signal $x(t)$ is performed as follows:

$$\begin{aligned} \text{if} \quad & (s_c > \tau) \quad x(t_s, t_s + t_d) = 0, \\ \text{else} \quad & (s_c \leq \tau) \quad x(t_s, t_s + t_d) = x(t_s, t_s + t_d). \end{aligned} \tag{1}$$

Note that a high threshold (τ) could lead to false negatives (P2aP information not muted). So the choice of τ becomes very crucial and its choice is dependent on the performance of the S2T engine.

Problem # 2 The S2T conversion process can mis-recognize an audio segment corresponding to P2aP information into a text string or label which does not match the P2aP information patterns at all. For example, */one four two/* could be recognized by the S2T engine as "own for too". Irrespective of the confidence (s_c) of the S2T, there is no way this kind of error can be handled.

Problem # 3 A system based on threshold selection leads to a binary decision, namely, the muting process either renders the identified region muted or leaves it as it is (see (1)). This can lead to non-P2aP regions in an audio recognized by high confidence by the S2T engine to be muted as well as true P2aP regions recognized with lower confidence to be left untouched in the audio conversations.

We now propose an approach which is able to handle all the three problems associated with the conventional approaches to mute P2aP information in an audio conversation.

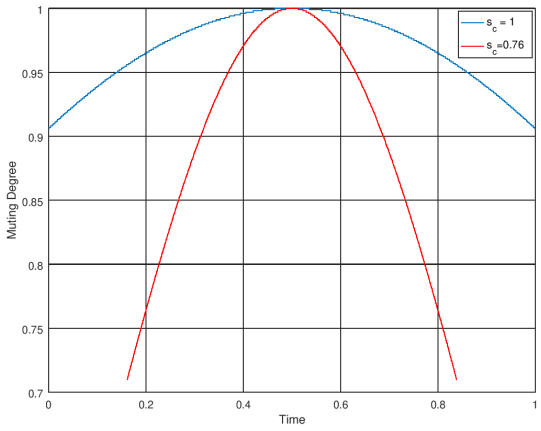

Fig. 5. (a) $\mathcal{F}$ for different confidence scores (s_c) (2)

3 Proposed Approach

Our approach is mainly focused on addressing the problems in the existing systems enumerated earlier and as a result we are able to robustly mute P2aP information even in the presence of S2T conversion errors. To address Problem #1 and Problem #3 we replace the "binary" dependency on choice of the threshold (τ) with a non-binary *fuzzy-muting* using the log-likelihood scores (s_c) determined by the S2T engine. To tackle Problem #2 we use a P2aP *search expansion* technique, which was originally proposed in [1].

3.1 Fuzzy-Muting

Fuzzy-muting, is a non-binary muting operation which allows for decisions that are not hard. The effect of this is that this enables a *usable* robust muting solution even when the S2T conversion process is not perfect. The fuzzy-muting operation is a function of the confidence score s_c and does not explicitly depend on a threshold τ. An example (used in our experiments) of a fuzzy-muting function (muting degree) is a Gaussian window, namely:

$$\mathcal{F}(t, t_d, s_c) = \exp^{-\left\{\frac{\left(t - \frac{t_d}{2}\right)^2}{2{s_c}^2}\right\}}, \tag{2}$$

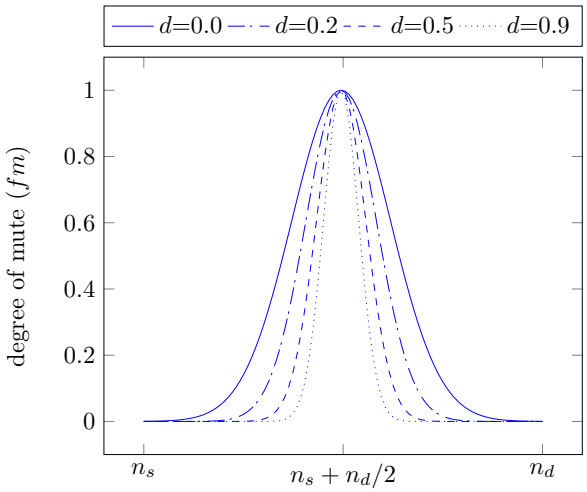

Fig. 6. Fuzzy muting function $\mathcal{F}^d$ for confidence score $s_c = 0.5$ and different values of normalized edit distance d (4)

where, $0 \leq t \leq t_d$ (for purpose of computation). A plot of this muting function is shown in Fig. 5. Note that when the confidence, s_c, of the recognized P2aP information is high then the muting function has a higher degree of mute and will mute most of $(t_s, t_s + t_d)$ interval. Similarly, when the confidence, s_c, is low it takes the form of a peaky Gaussian, meaning only a small portion of the audio around $t_s + t_d/2$ is muted and the rest remains the same. Thus $\mathcal{F}$ does the following:

(a) P2aP information identified with high confidence (true positive) is muted heavily (blue curve, Fig. 5).

(b) P2aP information identified with low confidence (false positive) is muted for only a small portion of the identified duration, so that a non-P2aP information is not completely muted (red curve, Fig. 5).

3.2 Search Expansion

P2aP information regions in the audio signal can be missed due to mis-recognition by the S2T conversion process. In this case the S2T (or KWS) replaces the actual word with other word(s) from the S2T vocabulary (or keyword list) which are phonetically similar. For example, a sequence of numbers */one two six four two/* could be mis-recognized as "one two six photo". Note that "four two" and "photo" are acoustically similar.

Such instances, in our technique are handled in the initial step of constructing P2aP information patterns. In our example, since */four two/* is expected as P2aP information, */photo/* which is acoustically similar is also included as a P2aP information cohort pattern.

When P2aP information cohorts are observed in the S2T output they can be mapped back and considered as confidential information. This mapping can lead to a correct identification of P2aP information or it leads to a new false positive. Either of these cases ultimately undergo fuzzy-muting as discussed earlier and not binary muting as in conventional muting systems.

To incorporate the proposed search expansion and the P2aP information cohort patterns the fuzzy muting function ($\mathcal{F}$) can be updated to:

$$\mathcal{F} \rightarrow \mathcal{F}^d(t, t_d, s_c, d), \tag{3}$$

where, d is the normalized edit distance [13] between the phonemic pronunciation of the expected pattern (*/four two/* in our example) and the observed cohort pattern (*/photo/* in our example). This is incorporated into $\mathcal{F}$ as:

$$\mathcal{F}^d(t, t_d, s_c, d) = \exp^{-\frac{\left(\left(t - \frac{t_d}{2}\right)\left(1 + d^{0.5}\right)\right)^2}{2{s_c}^2}}. \tag{4}$$

In this case, when the edit distance d is zero, i.e. the expected pattern and observed cohort pattern are same, the fuzzy muting function $\mathcal{F}^d$ is same as that in $\mathcal{F}$ (2) and directly dependent on the confidence score s_c. As the edit distance d increases, i.e. the expected pattern and observed cohort pattern become increasingly different, $\mathcal{F}^d$ reduces the degree of mute. This is illustrated in Fig. 6.

Note that as d increases (less acoustic similarity between the actual word and the cohort) the muting becomes narrow, meaning only a small portion of the identified audio segment is muted.

4 Experimental Analysis

4.1 Dataset

For our experiments we use a dataset consisting of real telephone recorded conversations between customers and call center agents of an insurance company operating in the USA. The conversations are in US accented English and the duration of the conversations vary from 1 minute to 25 minutes. For our experiments, we considered the customer's social security number (SSN), as the P2aP information to be muted in these audio conversations. Note that other information, like name of the customer can also be included in the P2aP information to be muted. Of the 50 audio conversations in this dataset, 80% of them contain at least one instance of the P2aP information the rest 20% do not contain any instances of the P2aP information. With these characteristics, our dataset closely resembles a realistic P2aP information muting scenario.

4.2 S2T Setup

Further we use a S2T engine to locate all the instances of P2aP information in the described audio dataset. The public domain Kaldi ASR toolkit [16] is used with the ASpIRE Chain acoustic models [14, 15] trained on conversational telephone speech from the Fisher English corpus. The accompanying pre-trained language model is used without any adaptation to the domain of our dataset. Given the un-adapted acoustic and language models, an average Word Error Rate (WER) of 47.2% is obtained on our dataset [10]. S2T confidence scores (s_c) were generated from the lattice posterior probabilities using a language weight of 1.

4.3 S2T Conversion Error and Confidence Analysis

A total of 3757 P2aP tokens were actually spoken in the conversations. Of these, 65.6% were correctly recognized, 23.4% were substituted by another (cohort) word in the S2T lexicon and 10.8% were deleted (not recognized) by the S2T. One can conclude that there is about 34.2% chance that the expected P2aP information will be in error and will affect a typical P2aP information muting process. The presence of 23.4% of substitution errors in P2aP tokens also indicate that the S2T engine missed these spoken P2aP words and replaced them by some other in-vocabulary word. Similarly, out of the 2860 P2aP tokens hypothesized by the S2T, 86.25% were correctly recognized tokens, 11.22% were substitutions to other words and 2.5% were insertions by the S2T. From this one can derive that about 13.72% of the hypothesized P2aP tokens are false positives and would affect the P2aP information muting process. These observations justify the need to address the problems described in Section 3 and also the need for search expansion.

For a more detailed analysis we present Fig. 7 which shows a distribution of S2T confidence scores for all the P2aP tokens hypothesized by the S2T as well as those originally present in the audio conversations but are substituted by the S2T. Instances where the S2T hypothesizes a P2aP token in place of another word are denoted by Substitutions (I) as they are Type I errors or false positives. Instances where the S2T misses a P2aP token and hypothesizes another word in place of it are denoted by Substitutions (II) as they are Type II errors or false negatives. The first observation from Fig. 7 is that although majority of the correctly hypothesized P2aP tokens have high confidence scores, setting a reasonable threshold like $\tau = 0.8$ would discard at least 7.5% of the correctly hypothesized P2aP tokens which could be part of expected P2aP information. This re-confirms the need to address the Problems #1 and #3 described in Section 3 and the need for a fuzzy muting mechanism.

4.4 Muting Performance

We present an evaluation of the P2aP information muting approaches on our dataset of real call center calls, where SSNs are the P2aP information to be muted. In this evaluation our proposed approach comprises the fuzzy muting function (4). It is compared to a typical P2aP information muting approach which mutes S2T recognized P2aP tokens having a confidence score above a

preset threshold τ (see (1)) to do a binary muting. Additionally we include a third approach in our evaluation, which first finds cohort patterns for P2aP sequences and then applies the typical P2aP information muting approach. In this approach, the confidence score on a P2aP token is modified as:

$$s_c^{'} = \frac{s_c}{(1 + d^{\frac{1}{2}})^{\frac{1}{2}}},$$

using the analogy between (2) and (4).

Our evaluation measure is the average number of SSN digit tokens which are left un-muted. A human listener listens to the muted speech signal and counts the number of SSN digit tokens left un-muted. The listener is allowed to listen to a muted signal multiple times and also to make guesses. The listener was given only those regions of the audio conversation where the user speaks the SSN sequence and not the complete audio conversation because of data confidentiality.

Figure 8 presents an evaluation and comparison of the P2aP information muting approaches. It shows a plot of the average number of P2aP tokens which are left un-muted by a muting approach, against different thresholds (τ) on S2T confidence scores. Approaches based on typical muting, (1), are dependent on the threshold (τ) and as the threshold (τ) is increased more number of P2aP tokens are left un-muted. Smaller thresholds (τ) appear better but they lead to more false alarms (instances of muting when there was no P2aP information) in other parts of the audio. Typical muting with cohorts gives a better performance on smaller thresholds but it degrades as the threshold is increased. Proposed approach comprising the fuzzy muting function (4) does not rely on the confidence threshold (τ) and achieves a significantly better muting performance, where on an average only 1.25 out of 9 SSN digits (P2aP) are left un-muted (see Fig. 8).

5 Discussion and Conclusion

Private to a person (P2aP) information often occur in recorded call center conversations during the process of customer verification which is required for the enterprise to provide personalized services to the customer.

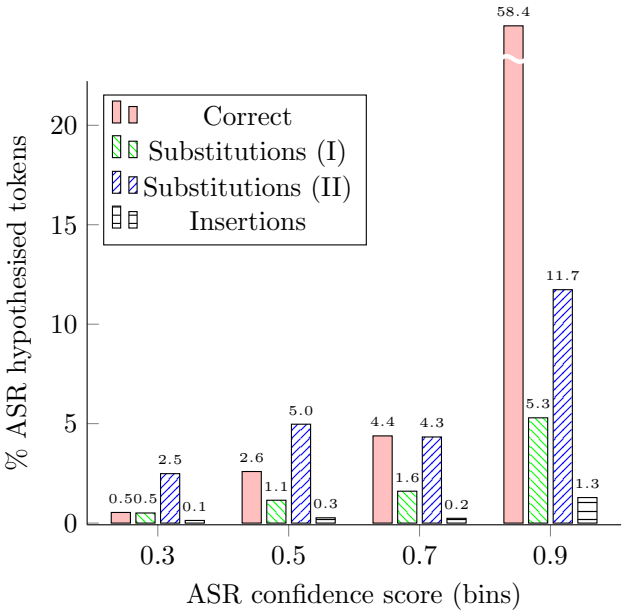

Fig. 7. Distribution of S2T confidence scores for P2aP. (I indicates false positives from S2T and II indicates missed P2aP)

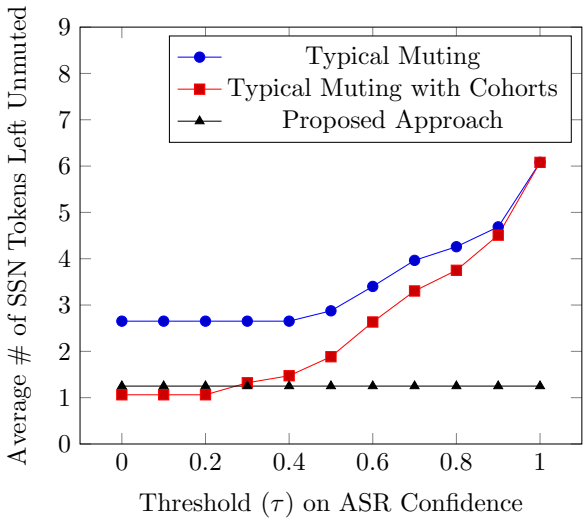

Fig. 8. Comparison of muting approaches. (Proposed Approach comprises fuzzy muting with cohorts)

Enterprises have the need to mute these information for several reasons including government regulations, for example, GDPR. Current muting solutions highly rely on the performance of a S2T engine. While S2T engines perform well in general, they fail to be 100% right when decoding natural spontaneous conversations [9]. In this paper, we have proposed a simple yet effective fuzzy muting function that can work even in the scenario when the S2T is not accurate all the time. The advantage of this approach is that it neither mutes non-P2aP information completely nor does it leave any potential P2aP information (including acoustically similar sounding) untouched, especially when the confidence in recognizing the P2aP information is low. Analysis on real call center conversations justifies the need for a fuzzy approach and the experiments demonstrate the effectiveness of the proposed approach. While in this work we resorted to Gaussian window functions, fuzzy muting could employ more sophisticated muting functions. This includes functions in which degree of mute is dependent on the envelope of the speech signal in the hypothesized region among other functions. Exploration and analysis in this direction is part of our future work.

References

1. **Ahmed, I. & Kopparapu, S. (2013).** Improved method for keyword spotting in audio. *Proc. Acoustics*, New Delhi, India, pp. 1028–1033.
2. **Aldeen, Y. A. A. S., Salleh, M., & Razzaque, M. A. (2015).** A comprehensive review on privacy preserving data mining. *SpringerPlus*, Vol. 4, No. 1, pp. 694.
3. **Bundock, D. S. & Ashton, M. (2008).** System for excluding unwanted data from a voice recording. https://www.google.com/patents/US20080221882. US 20080221882 A1.
4. **Cavanaugh, W. & Hirtle, P. (2006).** Speech privacy in buildings: A review. *The Journal of the Acoustical Society of America*, Vol. 119, No. 5, pp. 3325–3325.
5. **Commission, E. (2018).** 2018 reform of EU data protection rules. https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en.
6. **Doren, G. K. (2013).** Selective security masking within recorded speech. https://www.google.co.in/patents/US8433915. US 8433915 B2.
7. **Faruquie, T. A., Negi, S., & Subramaniam, L. V. (2009).** Protecting sensitive customer information in call center recordings. *IEEE International Conference on Services Computing*, pp. 81–88.
8. **Hillis, W. D., Ferren, B., & Howe, R. (2007).** Method and system for masking speech. https://www.google.com/patents/US7184952. US 7184952 B2.
9. **Kopparapu, S. K. (2014).** *Non-Linguistic Analysis of Call Center Conversations*. Springer Publishing Company.
10. **Kopparapu, S. K. (2014).** Word error rate. https://sites.google.com/site/nlccanalytics/home/wer.
11. **Lee, H., Lutz, S., & Odinak, G. (2007).** Selective security masking within recorded speech utilizing speech recognition techniques. https://www.google.co.in/patents/US20070016419. US 20070016419 A1.
12. **Lodha, S., Patwardhan, N., Roy, A., Sundaram, S., & Thomas, D. (2012).** Data privacy using MASKETEER^{TM}. *Proceedings of the 9th International Conference on Theoretical Aspects of Computing*, ICTAC'12, Springer-Verlag, Berlin, Heidelberg, pp. 151–158.
13. **Marzal, A. & Vidal, E. (1993).** Computation of normalized edit distance and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 15, No. 9, pp. 926–932.
14. **Peddinti, V., Chen, G., Manohar, V., Ko, T., Povey, D., & Khudanpur, S. (2015).** JHU ASpIRE system: Robust LVCSR with TDNNS, iVector adaptation and RNN-LMS. *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 539–546.
15. **Povey, D. (2017).** Kaldi models. http://kaldi-asr.org/models.html.
16. **Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011).** The Kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, IEEE Signal Processing Society, pp. 1–4. IEEE Catalog No.: CFP11SRW-USB.

17. Schachter, J. & Levin, K. D. (2013). System and method for removing sensitive data from a recording. https://www.google.com/patents/US20130266127. US 20130266127 A1.

Article received on 10/01/2019; accepted on 04/03/2019.
Corresponding author is Imran Sheikh.