Spectral Analysis of Audio Signals for the Identification of Spam Over IP Telephony

C. Pörschmann, H. Knospe

Institut f. Nachrichtentechnik, Fachhochschule Köln,Germany, Email: Christoph.Poerschmann@fh-koeln.de

Introduction

With modern computer and telecommunication systems voice calls can be automatically set up and prerecorded speech messages can then be played. As especially in IP-based networks the costs for voice calls are quite low, Spam over IP Telephony (SPIT) can become a serious problem in the near future.

Different approaches have already been carried out developing methods for the identification of SPIT. These systems typically evaluate statistical properties of the voice calls but do not look into the audio data itself. A machine-based analysis the audio signal of a voice call can help to identify SPITs and thus help establishing black-lists of potential spammers. However, it has to be guaranteed that privacy is protected and no content-related data is permanently stored by the system.

In this article, a method adapted from the area of music identification is proposed. This method performs an appropriate analysis of the audio speech signals and identifies if several voice calls with the same audio data are being set up in a telephone network. This sender ID can be added to a black-list of potential spammers who send the same voice message to many receivers. The proposed method can be an effective protection against machine-based SPIT which has the characteristic that the same or a marginally modified message is distributed to many recipients.

Approaches for the identification of SPIT

In modern telephone networks (fixed Telephone networks, mobile networks, IP-based telephony networks) typically all incoming calls are signalized to the receiver. The receiver has to determine the SPIT call and then actively to finish the call. In actual studies SPIT is mainly discussed for IP-Telephony (Voice over IP) [1]. Several activities are currently ongoing in order to identify SPIT and to protect telephone networks from being flooded with SPIT. A number of approaches are considered:

The rejection of voice calls can be based on a black-list of caller IDs and/or a white-list of allowed callers. Furthermore, calls can be filtered on the basis of authenticated caller identities [2] and by analyzing trust or security attributes. These approaches are based on the classification of the callers or a verification of their identity.

Another approach is to use a challenge-response procedure to identify machine-based calling systems [3]. However, this leads to disturbances and to extended call set-up times. Furthermore, there is the risk of false acceptances or false rejections due to intelligent answering machines or incorrect recognitions of the human answer.

Content-based music identification

In this article, a method adapted from the area of music identification is proposed which by an appropriate analysis of the audio speech signals can be used to identify SPIT and to automatically establish black-lists.

Several methods for music identification have been developed; some of them have meanwhile become a commercial success [4, 5]. Commonly, in a first step a so-called “acoustic fingerprint” of the audio track is created. Several parameters are extracted from the audio data: the Spectral Flatness Measure (SFM) and the Spectral Crest Factor (SCF) of a music track are computed (for sequenced time windows and for different frequency bands). Together with title and artist, these spectral parameters are stored in a database [4]. Due to their properties with respect to music identification the SCF and the SFM have been standardized as low level signal features within the MPEG-7 framework. In an alternative approach Wang [5] proposed to use peaks in the spectrogram as features in music identification.

A typical application is a mobile phone capturing an extract of a registered audio track. By comparing its “acoustic fingerprint” to those in the database the captured sequence can be identified. Two characteristics of this method are of great importance: The identified parameters are resistant to influences caused by voice coding (e.g. GSM, AMR), background noise and other modifications. Furthermore, a match is only recognized when exactly the same track is played, thus detection by humming or singing fails.

Spectral based identification of SPIT

The method which is described in the following allows the identification of SPIT calls and an automatic filling of black-lists. The method is based on the idea that SPAM can be identified if a spammer distributes the same message to many different receivers.

A spectral analysis of the audio signal similar to the content-based music identification method is applied. It is assumed that a spammer sends replayed calls to many recipients. Such replayed calls have to be identified and the caller identifier is marked for the black-list. It is then possible to block further calls originating from this caller ID.

To identify SPIT some or all incoming voice calls in a telephone network are analyzed. The spectral features (e.g. SFM and SCF or peaks in the spectrogram) are computed. Replayed calls have very similar characteristics regarding

these features. As already shown by Allamanche et al. [4] these features have the property that they are not significantly influenced by speech coding systems or by other modifications of the audio signal. All these features have in common that it has to be difficult for the caller to modify the audio data automatically in such a way that the identification fails without a significant degradation of the audio quality. The feature vectors of incoming calls and the corresponding caller IDs are stored in a class database. However, the audio data of the voice call is not stored in order to guarantee privacy for the user.

The comparison between the fingerprints of the actual call and the ones in the class database can be performed applying a standard distance metric. If the distance is below a certain limit a similarity between the two sequences is determined and both calls are marked as duplicate (probably SPIT). After a certain number of duplicates, the caller ID is added to a black-list. Further calls from this caller ID are blocked during the signalling phase. The identification still succeeds when there are slight differences between the feature vectors (e.g. caused by noise, speech coding, different order of speech blocks). Figure 1 shows the general set-up of the system.

In order to protect callers which are permitted to send out identical calls (e.g. alarm calls) from being put on a black list, a white-list can be created of those caller IDs. Furthermore, outgoing calls to prerecorded message services (e.g. weather forecast) or messages from answering machines are not affected by this SPIT filter since it only analyzes the audio data of the caller.

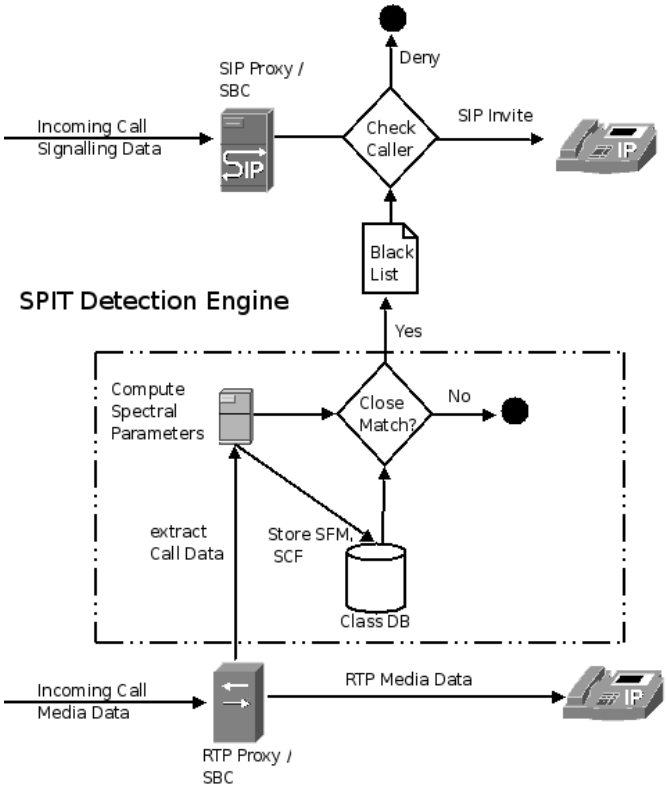

Figure 1: Identification of SPIT based on audio features: The caller ID and feature vectors of incoming calls are stored and compared to those in the database. In case of a high degree of similarity a probable SPIT call is identified. After several calls which are categorized as SPIT the caller ID is added to the black list and further calls from this caller are being blocked.

It should be noted that the identification requires at least two fully established calls and some seconds of incoming audio data for successful replay detection. Furthermore, SPIT calls with varying and spoofed caller IDs could in fact be detected and further analyzed but can not be blocked beforehand during the signalling phase. But most VoIP operators require anyway authentication of callers and trust user identifiers only from selected foreign networks.

Implementation and measurement results

The proposed methods have been implemented in Matlab at Cologne University of Applied Sciences within several diploma and bachelor theses. Demo implementations from the MPEG-7 standardization were used for the determination of the SFM and the SCF parameters. For each voice call, 256 feature vectors with 28 components are stored. Thus 7168 bytes are required to store the acoustic fingerprint of one call. The set of 256 feature vectors is determined from the complete set of vectors by applying a vector quantization method [6].

28 different voice calls (8 kHz sampling rate) with a duration ranging from 20 to 35 seconds have been analyzed and the resulting sets of feature vectors have been stored. Furthermore, since a spammer might slightly modify the audio data, the following modifications were investigated:

- Change of the pitch of the signal (max. 10%)
- Extraction of small sequences (ca. 5 s)
- Amplitude modification (max. 12 dB)
- Add noise with different spectral characteristics
- Linear distortions (high- or low-pass filtering)
- Non-linear distortions (clipping)

The results show that even for the modified audio signals a robust identification can be achieved. Thus most of the described modifications do not hinder the identification of SPIT. However, a significant degradation in the identification can be observed when adding white noise with energy of more than 12 dB below the energy of the speech signal.

In order to enhance the performance for noisy stimuli in a next step a method analyzing peaks in the spectrogram as a fingerprint of the audio signal has been considered. The identification of identical audio sections is performed by a comparison of the positions of the peaks. A similar method has been proposed by Wang [5] for music recognition applications. In our investigations a window length of 128 ms and an overlap of more than 100 ms have been used. Then the signal is divided in several frequency bands and in each frequency band the characteristic maxima are determined. Figure 2 shows the positions of the peaks for a signal without noise. Then different portions of noise were added to the signal and the influence of the noise on the positions of the peaks was analyzed.

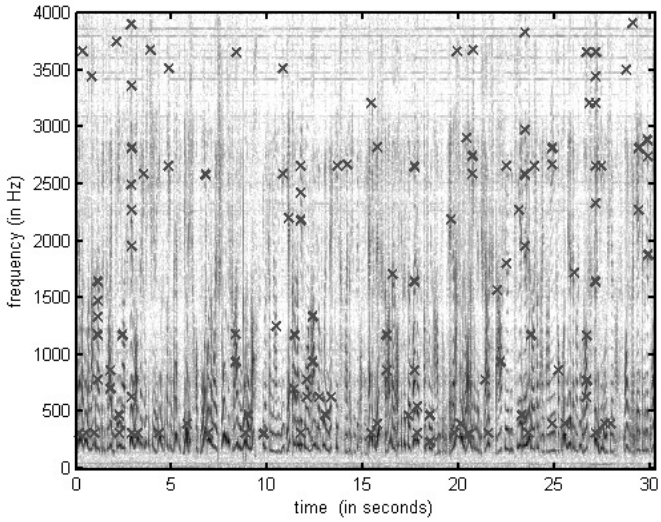

Figure 2: Spectrogram of an unnoised audio signal. In the spectrogram the peaks have been identified and marked.

In Figure 3 the influence of noise on the positions of the peaks is demonstrated. It can be observed that most of the peaks remain unchanged.

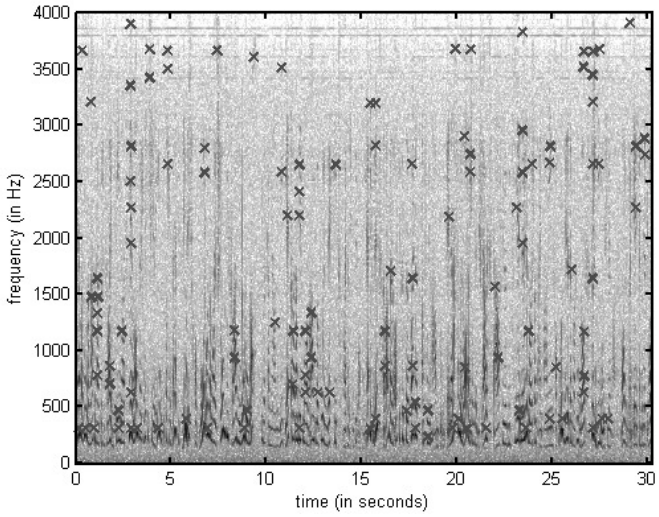

Figure 3: Spectrogram of a noised audio signal (pink noise, 15 dB SNR).

Figure 4 shows the influence of noise on the number of correctly identified peaks. Even for signal to noise ratios of 10 dB at least 40 % of the peaks are correctly identified. This seems to be sufficient for a correct identification of the similar audio signals. In a next step of it will be investigated which number of correctly identified peaks is required in order to correctly identify a SPIT message.

Conclusion

The described method allows the identification of calls with identical or very similar audio data which typically characterize SPIT. The method helps to detect SPIT calls and to generate black-lists of spamming caller IDs. An advantage of the method is that an identification of replayed calls is possible after very few of these calls have been captured by the system. Comparable approaches require a higher number of SPIT calls in order to allow a clear identification. A second advantage is the high reliability of

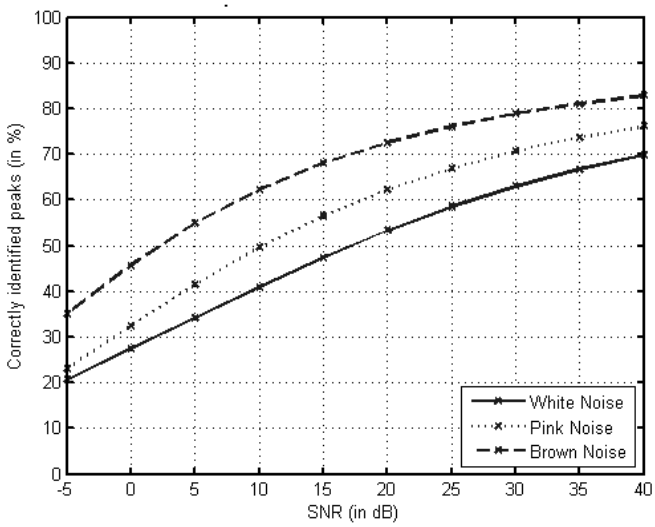

Figure 4: Robustness of the spectral peaks regarding noise. The percentage of unchanged spectral peaks is shown for different types of noise and different SNRs

the feature comparison. Spectral features are determined which have shown their resistance to different modifications (codec, background noise, etc.). Finally privacy is protected as no content-related data (e.g. audio content) is stored.

References

[1] Rosenberg, J., Jennings, C. (2007). The Session Initiation Protocol (SIP) and Spam, IETF, Internet Draft.

[2] Peterson, J., Jennings, C. (2006). Enhancements for Authenticated Identity Management in the Session Initiation Protocol (SIP). IETF, RFC 4474.

[3] Brouwer, S. (2007). Spam protection system for voice calls, Patent EP 1742452.

[4] Allamanche, E. Cremer, M., Fröba, B., Hellmuth, O., Herre, J., Kastner, T. (2001). Content based Identification of Audio Material Using MPEG-7 Low Level Description, *2nd Annual International Symposium on Music Information Retrieval*.

[5] Wang, A. (2006). The Shazam music recognition service. *Communications of the ACM*, 49(8), 44–48.

[6] Linde, Y., Buzo, A. Gray, R.M. (1980). An Algorithm for Vector Quantizer Design, *IEEE Transactions on Communications*, 702-710.

[7] Pörschmann, C., Knospe, H. (2008). Analyse spektraler Parameter des Audiosignals zur Identifikation und Abwehr von Telefon-SPAM, In: Lecture Notes in Informatics, *Proceedings Sicherheit 2008*, Ges. f. Informatik, Volume P-128, 551-555.

[8] Hansen, M., Hansen, M., Möller, J. Rohwer, T., Tolkmit, C., Waack, H. (2006). Developing a Legally Compliant Reachability Management System as a Countermeasure against SPIT, *Proceedings of the 3rd Annual VoIP Security Workshop*, Berlin