International Telecommunications Symposium – ITS2002, Natal, Brazil

Mixed Watermarking-Fingerprinting Approach for Integrity Verification of Audio Recordings

Emilia Gómez[1], Pedro Cano[1], Leandro de C. T. Gomes[2], Eloi Batlle[1], Madeleine Bonnet[2]

[1] Music Technology Group, Pompeu Fabra University, Spain
{pedro.cano, emilia.gomez, eloi.batlle}@iua.upf.es, http://www.iua.upf.es/mtg/

[2] InfoCom-Crip5, Université René Descartes, Paris, France
{tgomes, bonnet}@math-info.univ-paris5.fr, http://www.math-info.univ-paris5.fr/crip5/infocom/

Abstract—We introduce a method for audio-integrity verification based on a combination of watermarking and fingerprinting. The fingerprint is a sequence of symbols ("audio descriptor units") that enables one to identify an audio signal. Integrity verification is performed by embedding the fingerprint into the audio signal itself by means of a watermark. The original fingerprint is reconstructed from the watermark and compared with a new fingerprint extracted from the watermarked signal. If they are identical, the signal has not been modified; if not, the system is able to determine the approximate locations where the signal has been corrupted.

I. Introduction

IN many applications, the integrity of an audio recording must be unquestionably established before the signal can actually be used, i.e. one must be sure that the recording has not been modified. As an example, some countries require oral testimonies to be recorded for later review; it is desirable that the integrity of such recordings be certified before they are used in court.

Integrity verification systems have been proposed as an answer to this need. Two classes of methods are well suited for these applications: *watermarking*, which allows one to embbed data into the signal, and *fingerprinting*, which consists in extracting a "signature" (the fingerprint) from the audio signal.

Integrity verification systems are useful in several scenarios. For speech-related applications, we can mention:

- integrity verification of a previously recorded testimony that is to be used as evidence before a court of law;
- integrity verification of recorded interviews (which could be edited for malicious purposes).

For music and other kinds of recordings, we can mention:

- integrity verification of radio or television commercials;
- integrity verification of music aired by radio stations;
- content preservation of audio signals in general.

After a conceptual description of schemes based solely on fingerprinting and watermarking, we propose a mixed approach that takes advantage of both technologies.

II. Integrity Verification Systems: A Conceptual Review

A. Watermarking-Based Systems

Audio watermarking[1] consists in embedding a mark (the *watermark*) into an audio signal. This mark is also an audio signal and carries data that can be retrieved from the watermarked signal. Ideally, the watermark should not introduce any perceptible degradation in the signal, which means that the original and the watermarked signals should sound exactly the same to the listener. To a certain extent, this can be achieved by using psychoacoustic models such as those found in perceptual coding [2], [3].

In watermarking-based integrity-verification systems, the integrity of a previously watermarked audio signal is determined by checking the integrity of the watermark. We define three classes of integrity-verification systems based on watermarking:

1. Methods based on fragile watermarking, which consist in embedding a fragile watermark into the audio signal (e.g. a low-power watermark). If the watermarked signal is edited, the watermark must no longer be detectable. By "edited", we understand any modification that could corrupt the content of a recording. "Cut-and-paste" manipulations (deletion or insertion of segments of audio), for example, must render the watermark undetectable. In contrast, content-preserving manipulations, such as lossy compression with reasonable compression rates or addition of small amounts of channel noise, should not prevent watermark detection (as long as the content is actually preserved).

Extremely fragile watermarks can also be used to verify if a signal has been manipulated in any way, even without audible distortion. For example, a recording company can watermark the content of its CDs with a very fragile watermark. If songs from this CD are compressed (e.g. in MPEG format), then decompressed and recorded on a new CD, the watermark would not be detected in the new recording, even if the latter sounds exactly as the original one to the listener. A CD player can then check for the presence of this watermark; if no watermark is found, the recording has necessarily undergone illicit manipulations and the CD is refused. The main flaw in this approach is its inflexibility: as the watermark is extremely fragile, there is no margin for the rights owner to define any allowed signal manipulations (except for the exact duplication of the audio signal).

Music Technology Group, Institut Universitari de l'Audiovisual (IUA), Universitat Pompeu Fabra (UPF), Area Estació de França, Passeig de Circumval·lació 8, 08003 Barcelona, Spain, Phone: +34 93 542 2201, Fax: +34 93 542 2202.

InfoCom-Crip5, UFR de Mathématiques et Informatique, Université René Descartes, 45 rue des Saints-Pères, 75270 Paris cedex 06, France, Phone: +33 01 44 55 35 24, Fax: +33 01 44 55 35 35.

[1] For an introductory paper on watermarking and information hiding, see [1].

2. Methods based on semi-fragile watermarking, which are a variation of the previous class of methods. The idea consists in circumventing the excessive fragility of the watermark by increasing its power. This semi-fragile watermark is able to resist slight modifications in the audio signal but becomes undetectable when the signal is more significantly modified. The difficulty in this approach is the determination of an appropriate "robustness threshold" for each application.

3. Methods based on robust watermarking, which consist in embedding a robust watermark into the audio signal. The watermark is supposed to remain detectable in spite of any manipulations the signal may suffer. Integrity is verified by checking whether the information contained in the watermark is corrupted or not.

Watermarking-based integrity-verification systems depend entirely on the reliability of the watermarking method. However, an audio signal often contains short segments that are difficult to watermark due to localized unfavorable characteristics (e.g. very low power or ill-conditioned spectral characteristics); these segments will probably lead to detection errors, particularly after lossy transformations such as resampling or MPEG compression. In integrity-verification applications, this is a serious drawback, since it may not be possible to decide reliably whether unexpected data are a consequence of intentional tampering or "normal" detection errors.

B. Fingerprinting-Based Systems

Audio fingerprinting or **content-based identification** (CBID) methods extract relevant acoustic characteristics from a piece of audio content. The result is a sequence of symbols ("audio descriptor units", ADU), the *fingerprint,* that acts as a kind of signature of the audio signal. If the fingerprints of a set of recordings are stored in a database, each of these recordings can be identified by extracting its fingerprint and searching for it in the database.

In fingerprinting-based integrity-verification systems, the integrity of an audio signal is determined by checking the integrity of its fingerprint. These systems operate in three steps: (1) a fingerprint is extracted from the original audio recording, (2) this fingerprint is stored in a trustworthy database, and (3) the integrity of a recording is verified by extracting its fingerprint and comparing it with the original fingerprint stored in the database.

According to the chosen fingerprinting method, two subclasses of this approach can be defined:

1. Methods sensitive to data modification, based on hashing methods such as MD5 [4]. This class of methods is appropriate when the audio recording is not supposed to be modified at all, since a single bit flip is sufficient for the fingerprint to change. Some robustness to slight signal modifications can be obtained by not taking into account the least-significant bits when applying the hash function.

2. Methods sensitive to content modification, based on fingerprinting methods that are intended to represent the content of an audio recording (such as AudioDNA [5]). This class of methods is appropriate when the integrity check is not supposed to be compromised by operations that preserve audio content (in a perceptual point of view) while modifying binary data, such as lossy compression and resampling.

The main disadvantage of fingerprinting-based methods is the need of additional metadata (the original fingerprint) in the integrity-check phase, thus requiring access to a database.

III. A Combined Watermarking-Fingerprinting System

We propose an integrity-verification approach that combines watermarking and fingerprinting in a single system. The idea consists in extracting the fingerprint of an audio signal and storing it in the signal itself through watermarking, thus avoiding the need of additional metadata during integrity check. Some methods based on this idea have already been described in the literature ([6] for audio, [7], [8] for images and video).

Fig. 1 presents a general scheme of this mixed approach. First, the fingerprint of the original recording is extracted; this fingerprint, viewed as a sequence of bits, is then used as the information to be embedded into the signal through watermarking. As the watermark signal is weak, the watermarked recording should have the same fingerprint as the original recording. Thus, the integrity of this recording can be verified by extracting its fingerprint and comparing it with the original one (reconstructed from the watermark). This procedure will be detailed in the following sections.

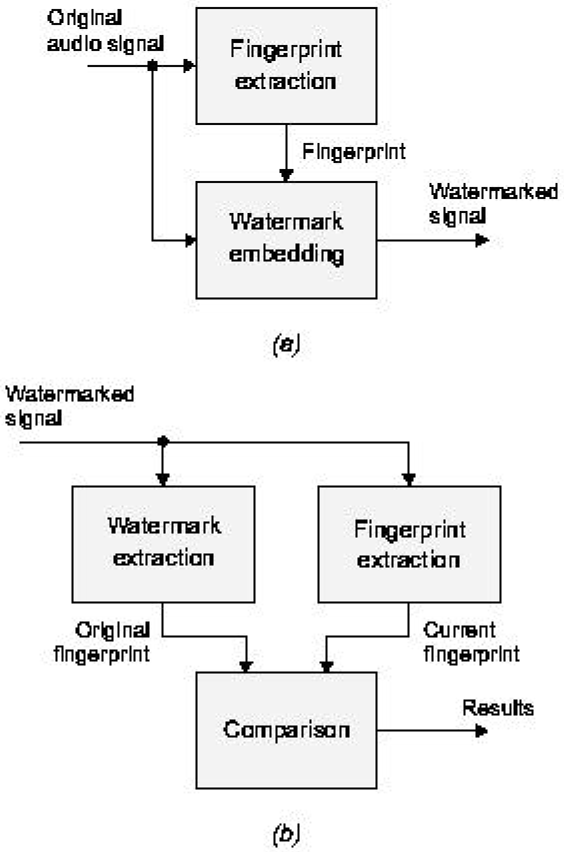

Fig. 1. Block diagram of the mixed approach for audio integrity verification: (a) embedding; (b) detection.

A. *Requirements*

We mention below some of the requirements that are expected to be satisfied by the integrity-verification system and its components:

- the fingerprint should not be modified when transformations that preserve audio content are performed;
- the watermarking scheme must be robust to such transformations;
- the bit rate of the watermarking system must be high enough to code the fingerprint information with strong redundancy;
- the method should be suitable for use with streaming audio, as the total length of the audio file is unknown in applications such as broadcasting [9].

As will be shown in section IV, the first three requirements are fulfilled by the system. The last one is also satisfied, as both the watermark and the fingerprint can be processed “on the fly”.

B. *System features*

This system is particularly well suited for the detection of “cut-and-paste” manipulations, which are exactly the kind of tampering that must be avoided in the case of recorded testimonies or interviews. Distortions that perceptually affect the signal will also be detected; as examples, we can mention:

- time stretching ;
- pitch shifting
- severe distortion through filtering ;
- addition of strong noise.

The system is not only able to detect tampering, but it can also determine the approximate location where the audio signal was corrupted.

C. *Implementation*

C.1 Fingerprint Extraction

The key idea of fingerprinting consists in considering audio as a sequence of *acoustic events*. In the case of speech signals, for example, acoustic events can be directly associated with phonemes. In music modeling, however, this association is not straightforward. The use of musical notes, for instance, would present many disadvantages: several notes may be played simultaneously, and music pieces often contain voice and non-harmonic sounds.

An appropriate approach consists in obtaining the relevant acoustic events — called *Audio Descriptor Units* (ADU) — by means of unsupervised training, i.e. without any previous knowledge of music events. The training process is performed through a modified Baum-Welch algorithm on a corpus of representative music [10].

Shortly, the system works as follows. An alphabet of representative sounds is derived from the corpus of audio signals (constructed according to the kind of signals that the system is supposed to identify). These audio units are modeled by means of Hidden Markov Models (HMM).

The audio signal is processed in a frame-by-frame analysis. A set of relevant-feature vectors is first extracted from the sound. These vectors are then normalized and sent to the decoding block, where they are submitted to statistical analysis by means of the Viterbi algorithm. The output of this chain — the fingerprint — is the most likely ADU sequence for this audio signal. This process is illustrated in Fig. 2.

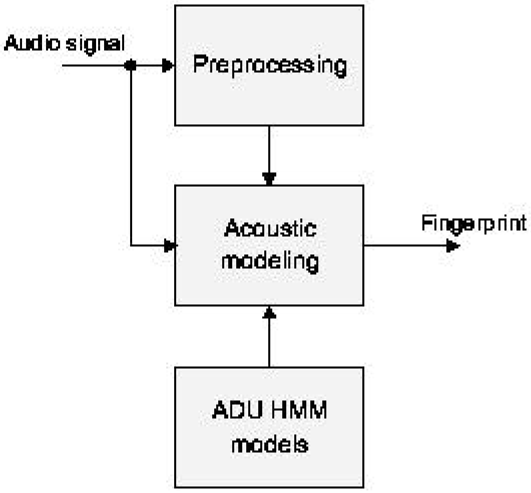

Fig. 2. Fingerprint extraction.

The resulting fingerprint is therefore a sequence of symbols (the ADUs) and time information (start time and duration). The number of different ADUs available to the system can be adjusted, as well as the output rate. The setup used in our experiments corresponds to 16 diferent ADUs (G0, G1, . . . , G15) and an average output rate of 100 ADUs per minute.

C.2 Fingerprint Encoding and Watermark Embedding

Each 8-s segment of the audio signal is treated individually in order to allow for streaming-audio processing. The fingerprint is converted into a binary sequence by associating a unique four-bit pattern to each of the 16 possible ADUs; thus, the average fingerprint bit rate is approximately 7 bits/s. In our experiments, the watermark bit rate is set to 125 bits/s, allowing the fingerprint information to be coded with huge redundancy (which minimizes the probability of error during its extraction). A simple repetition code is employed, with a particular 6-bit pattern (011110) serving as a delimiter between repetitions. To avoid confusion between actual data and delimiters, every group of four or more consecutive bits “1” in the data receives an additional bit “1”, which is suppressed in the detection phase.

The fingerprint data is embedded into the audio signal by means of a watermark. The watermarking system used in our experiments is represented in Fig. 3. The analogy between watermarking and digital communications is emphasized in the figure: watermark synthesis corresponds to transmission (with the watermark as the information-bearing signal), watermark embedding corresponds to channel propagation (with the audio signal as channel noise), and watermark detection corresponds to reception.

The watermark signal is synthesized from the input data by a modulator. In order to obtain a watermark that is spread in frequency (so as to maximize its power and increase its robustness), a codebook containing white, orthogonal Gaussian vec-

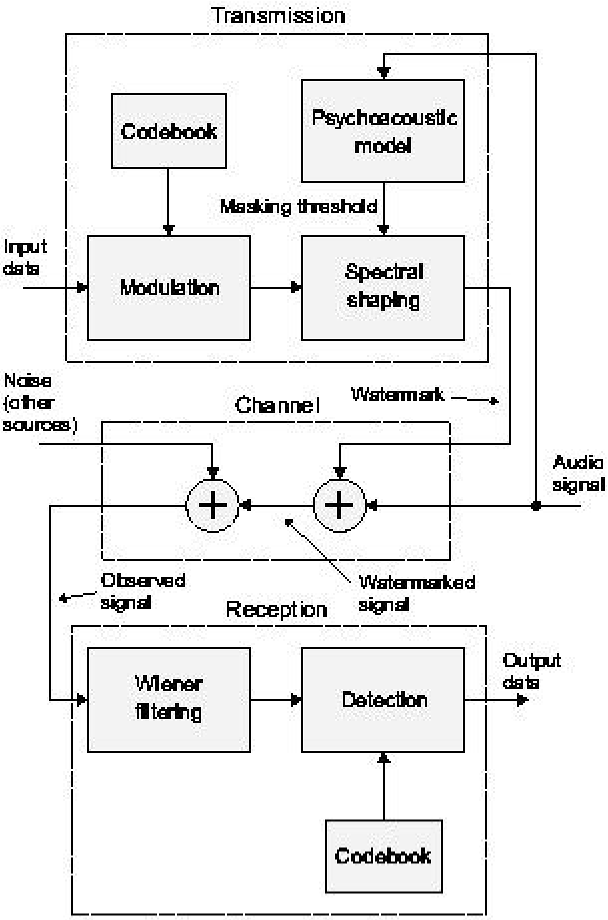

Fig. 3. Watermarking system.

tors is used in the modulator. The number of vectors is a function of the desired bit rate. Each codebook entry is associated with a specific input binary pattern. The modulator output is produced by concatenating codebook vectors according to the input data sequence.

To ensure watermark inaudibility, the modulator output is spectrally shaped through filtering according to a masking threshold (obtained from a psychoacoustic model). This procedure, repeated for each window of the audio signal (≈ 10 ms), produces the watermark. The watermarked signal is obtained by adding together the original audio signal and the watermark.

As transmission and reception must be synchronized, the transmitted data sequence also carries synchronization information. This sequence is structured in such a way that detected data is syntactically correct only when the detection is properly synchronized. If synchronism is lost, it can be retrieved by systematically looking for valid data sequences. This resynchronization scheme, based on the Viterbi algorithm, is detailed in [11] and [12].

C.3 Watermark Detection and Fingerprint Decoding

For each window of the received signal, the watermark signal is strengthened through Wiener-filtering and correlation measures with each codebook entry are calculated. The binary pattern associated with the codebook entry that maximizes the correlation measure is selected as the received data. The syntactic consistency of the data is constantly analyzed to ensure synchronization, as described in the previous section.

The output binary sequence is then converted back into ADUs. For each 8-s audio segment, the corresponding fingerprint data is repeated several times in the watermark (16 times in average). Possible detection errors (including most errors caused by malicious attacks) can then be corrected by a simple majority rule, providing a replica of the original fingerprint of the signal.

C.4 Matching and Report

Finally, the fingerprint of the watermarked signal is extracted (through the same procedure presented in section III-C.1) and compared with the original fingerprint obtained from the watermark. If the two sequences of ADUs match perfectly, the system concludes that the signal has not been modified after watermarking; otherwise, the system determines the instants associated to the non-matching ADUs, which correspond the approximate locations where the signal has been corrupted. Identical ADUs slightly shifted in time are considered to match, since such shifts may occur when the signal is submitted to content-preserving transformations (e.g. MPEG compression).

IV. Simulations

A. *Experimental conditions*

Results of cut-and-paste tests are presented for four 8-s test signals: two songs with voice and instruments (signal "cher", from Cher's "Believe", and signal "estrella_morente", a piece of flamenco music), one song with voice only (signal "svega", Suzanne Vega's "Tom's diner", a cappella version), and one speech signal (signal "the_breakup", Art Garfunkel's "The breakup"). The signals were sampled at 32 kHz and were inaudibly watermarked with a signal to watermark power ratio of 23 dB in average.

B. *Results*

Fig. 4 shows the simulation results for all test signals. For each signal, the two horizontal bars represent the original signal (upper bar) and the watermarked and attacked signal (lower bar). Time is indicated in seconds on top of the graph. The dark-gray zones correspond to attacks: in the upper bar, they represent segments that have been *inserted* in the audio signal, whereas in the lower bar they represent segments that have been *deleted* from the audio signal. Fingerprint information (i.e. the ADUs) is marked over each bar.

For all signals, the original fingerprint was successfully reconstructed from the watermark. Detection errors introduced by the cut-and-paste attacks were eliminated by exploiting the redundancy of the information stored in the watermark.

A visual inspection of the graphs in Fig. 4 shows that the ADUs in the vicinities of the attacked portions of the signal were always modified. These corrupted ADUs allow the system to determine the instant of the attack within a margin of approximately ±1 second.

For the last signal ("the_breakup"), we also observe that the attacks induced two changes in relatively distant ADUs (approximately 2 s after the first attack and 2 s before the second one). This can be considered a false alarm, since the signal was not modified in that zone but

V. Advantages of the Mixed Approach

In this section, we summarize the main advantages of the mixed approach in comparison with other integrity-verification methods:

- No side information is required for the integrity test; all the information needed is contained in the watermark or obtained from the audio signal itself. This is not the case for systems based solely on fingerprinting, since the original fingerprint is necessary during the integrity test. Systems based solely on watermarking may also require side information, as the data embedded into the signal cannot be deduced from the signal itself and must be stored elsewhere;
- Slight content-preserving distortions do not lead the system to "false alarms", since the fingerprint and the watermark are not affected by these transformations. Hashing methods (such as MD5) and fragile watermarks generally do not resist such transformations;
- In general, localized modifications in the audio signal also have a localized effect on the fingerprint, which enables the system to determine the approximate locations where the signal has been corrupted. This is not the case for simple hashing methods, since the effects of a localized modification may be propagated to the entire signal;
- Global signal modifications can also be detected by the system; in this case, the entire fingerprint will be modified and/or the watermark will not be succesfully detected;
- The method is well suited for streaming audio, since all the processing can be done in real time.

VI. Conclusions

In this paper, we have presented a system for integrity verification of audio recordings based on a combination of watermarking and fingerprinting. By exploiting both techniques, our system avoids most drawbacks of traditional integrity-verification systems based solely on fingerprinting or watermarking. Unlike most traditional approaches, no side information is required for integrity verification. Additionally, the effect of localized modifications generally do not spread to the rest of the signal, enabling the system to determine the approximate location of such modifications. Experimental results confirm the effectiveness of the system.

As next steps in this research, we will consider alternatives to further increase overall system reliability, particularly in what concerns false alarms (i.e. signal modifications detected after content-preserving transformations or in zones where the signal was not modified). More efficient coding schemes will also be considered for fingerprint encoding prior to embedding.

Acknowledgments

Leandro de C. T. Gomes would like to thank CNPq for financial support.

References

[1] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, *Information hiding — a survey*, Proceedings of the IEEE, special issue on protection of multimedia content, 87(7):1062-1078, July 1999.

[2] L. Boney, A. Tewfik, and K. Hamdy, *Digital watermarks for audio signals*, International Conference on Multimedia Computing and Systems, Hiroshima, June 1996.

[3] M. Perreau Guimarães, *Optimisation de l'allocation des ressources binaires et modélisation psychoacoustique pour le codage audio*, PhD Thesis, Université Paris V, Paris, 1998.

[4] R. Rivest, *The MD5 Message-Digest Algorithm*, <http://theory.lcs.mit.edu/ rivest/Rivest-MD5.txt>, April 1992.

[5] Recognition and Analysis of Audio, <http://raa.joanneum.at/>.

[6] G. Shaw, *Digital document integrity*, 8th ACM Multimedia Conference, Los Angeles, November 2000.

[7] J. Dittmann, A. Steinmetz, and R. Steinmetz, *Content-based digital signature for motion pictures authentication and content-fragile watermarking*, International Conference on Multimedia Computing and Systems, Florence, June 1999.

[8] J. Dittmann, *Content-fragile watermarking for image authentication*, Proceedings of SPIE, vol. 4314, Bellingham, 2001.

[9] R. Gennaro and P. Rohatgi, *How to sign digital streams*, Advances in Cryptology, CRYPTO'97, 1997.

[10] E. Batlle and P. Cano, *Automatic segmentation for music classification using competitive hidden markov models*, International Symposium on Music Information Retrieval, 2000.

[11] E. Gómez, *Tatouage de signaux de musique (méthodes de synchronisation)*, DEA ATIAM thesis, ENST (Paris Télécom)-IRCAM (Centre Georges Pompidou), 2000, <http://www.iua.upf.es/mtg/>.

[12] L. de C. T. Gomes, E. Gómez, and N. Moreau, *Resynchronization methods for audio watermarking*, 111th AES Convention, New York, November 2001, <http://www.iua.upf.es/mtg/>.

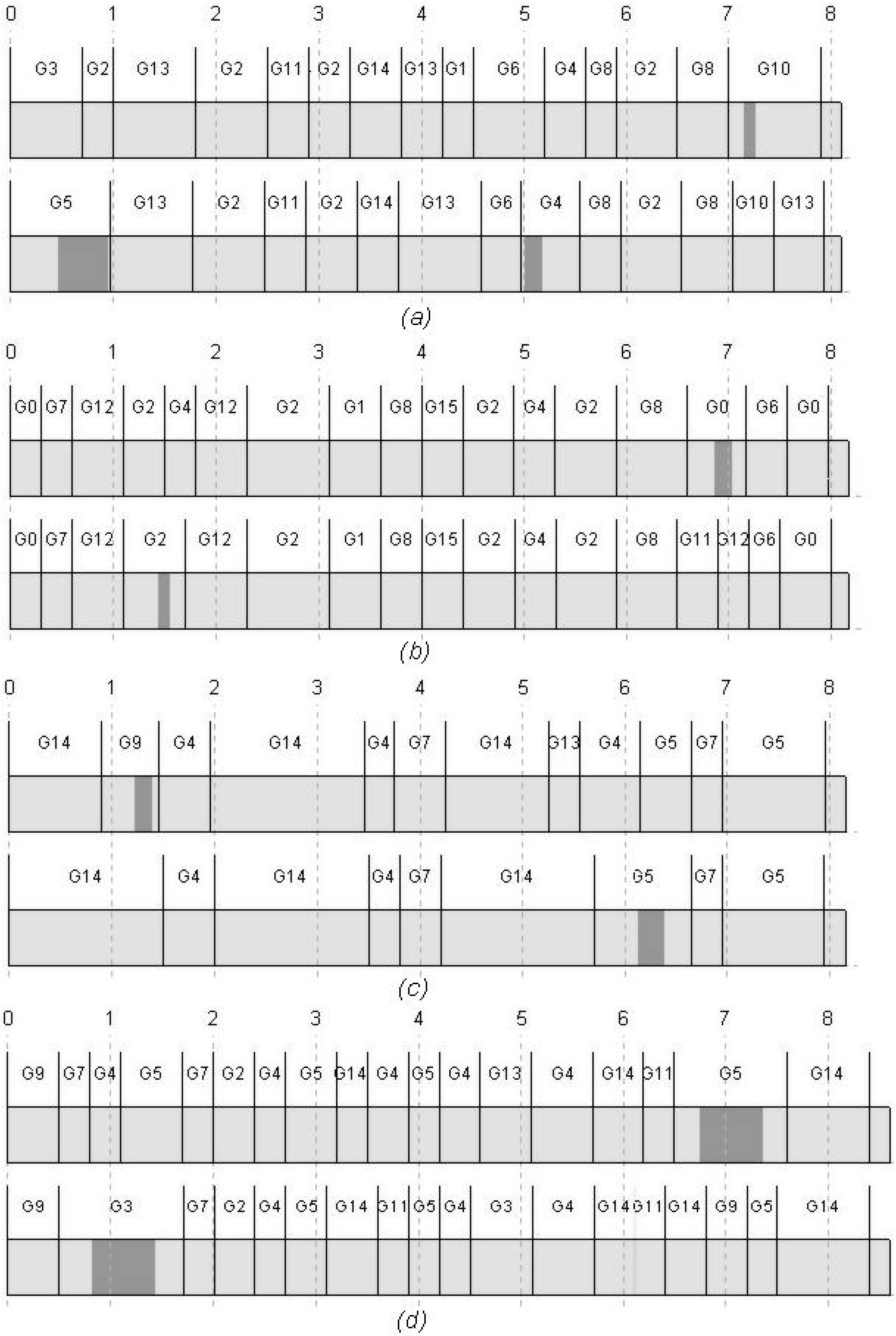

Fig. 4. Simulation results: (a) signal "cher"; (b) signal "estrella_morente"; (c) signal "svega"; (d) signal "the_breakup".