



(12) **EUROPEAN PATENT APPLICATION**

(43) Date of publication:
21.12.2022 Bulletin 2022/51

(51) International Patent Classification (IPC):
G10H 1/00 ^(2006.01) **G10H 1/40** ^(2006.01)
A63F 13/00 ^(2014.01)

(21) Application number: **21180444.8**

(52) Cooperative Patent Classification (CPC):
G10H 1/0091; G10H 1/42; A63F 13/00;
G10H 2210/041; G10H 2210/051; G10H 2210/071;
G10H 2210/076

(22) Date of filing: **18.06.2021**

(84) Designated Contracting States:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB
GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO
PL PT RO RS SE SI SK SM TR
Designated Extension States:
BA ME
Designated Validation States:
KH MA MD TN

(72) Inventors:

- **KOSTA, Katerina**
London, EC2Y 5EB (GB)
- **NEWTON-REX, Edmund Philip**
London, EC2Y 5EB (GB)
- **SONG, Xuchen**
California, 90066 (US)

(30) Priority: **15.06.2021 GR 20210100391**

(74) Representative: **Dentons UK and Middle East LLP**
One Fleet Place
London EC4M 7WS (GB)

(71) Applicant: **Lemon Inc.**
Grand Cayman, KY1-1205 (KY)

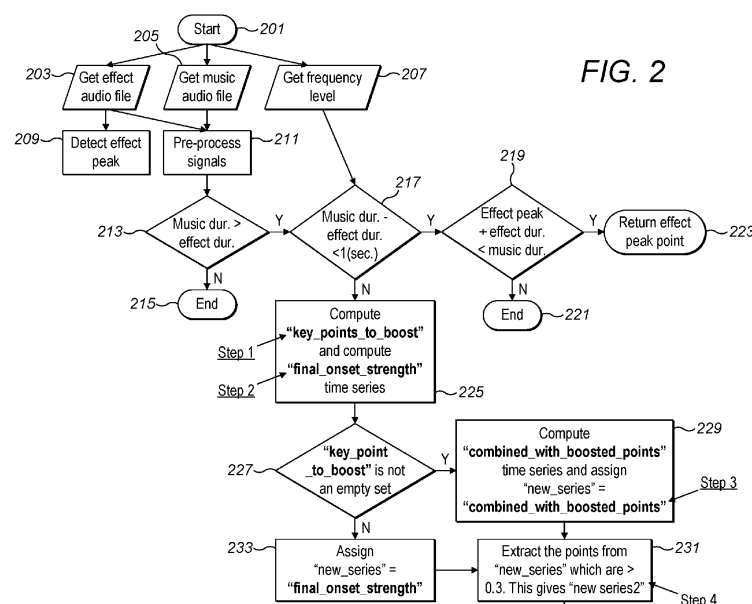
Remarks:

Amended claims in accordance with Rule 137(2) EPC.

(54) **SYSTEM AND METHOD FOR SELECTING POINTS IN A MUSIC AND AUDIO SIGNAL FOR PLACEMENT OF SOUND EFFECT**

(57) A method of selecting points in a music audio signal for placement of a sound effect comprises searching for points in the music audio signal as potential candidate placement points based on one or more criteria such as melspec or CQT. An onset strength time series

for the music audio signal is determined and boosted at points found by the search. Points are then selected from the boosted onset strength time series with a value larger than a predetermined threshold as candidate points for the placement of the sound effect.



Description

[0001] The invention is in the field of mixing audio signals.

5 **BACKGROUND**

[0002] It is known in the art of mixing audio signals to add one or more sound effects to a music audio signal. This might be done in a manual process, for example adding a drum beat or other sound effect in time with the music audio signal. A sound effect may be any audio signal of shorter duration than the music audio signal into which it is to be inserted, such as for example an excerpt from another audio signal.

[0003] Currently, mixing is largely a matter of experimentation or trial and error to produce a new sound mix that is pleasing to the ear of the person making the mix.

[0004] It would be advantageous to provide an automatic method to determine points in time in a piece of music audio for placement of other short audio excerpts ("sound effects"). The points could be chosen to fit with the audio in some way or to have a noticeable impact. With such automation, experimentation with mixing would then be more accessible to those with less experience in this field.

[0005] Some embodiments of the invention described below solve some of these problems. However the invention is not limited to solutions to these problems.

20 **SUMMARY**

[0006] This summary is provided to introduce a selection of concepts in a simplified form that are further described below in the detailed description. This summary is not intended to identify key features or essential features of the claimed subject matter, nor is it intended to be used to determine the scope of the claimed subject matter.

[0007] Some embodiments of the invention provide a method of selecting sound effect placement points in a music audio signal that may be automated and implemented on a computer.

[0008] In one implementation the method comprises determining an onset strength time series for the music audio signal; and selecting points from the onset strength time series with a value larger than a predetermined threshold as candidate points for the placement of the sound effect. Thus, for example, if the method is implemented on a computer, a user may input a sound effect and a music audio signal and a mix of the music and the sound effect may be automatically generated.

[0009] It may be advantageous to include points in addition to those having the highest onset strength. Therefore an optional refinement of the method comprises searching for points in the music audio signal as potential candidate placement points based on one or more criteria; and boosting the onset strength time series at points found by the searching prior to selecting points as candidate points. The criteria may be different from onset strength and may for example be based on the mel-spectrogram or constant-Q transform, both of which will be familiar to those working in music technology. The boosting, which may comprise multiplying the found points by different weights, may yield points that were not identified from onset strength.

[0010] There is also provided here a method of modifying a music audio signal comprising receiving a music audio signal and a sound effect signal, identifying points in the music audio signal for placement of the sound effect according to any of the methods described here, and inserting the sound effect signal into the music audio signal at each of the identified points.

[0011] There is also provided here a computing system comprising at least a memory and a processor, in which the processor is configured, for example by suitable programming, to enable the system to implement any of the methods described here.

[0012] Embodiments of the invention also provide a computer readable medium comprising instructions, for example in the form of an algorithm, which, when implemented in a computing system, cause the system to perform any of the methods described herein.

[0013] Some of the methods and systems to be described in the following enable the selection of points in the music that highlight rhythmical patterns of the music signal or highlight locations with change in the signal spectrum. This is opposed to randomly picking a location, for example in seconds, and placing a sound effect there.

[0014] Music generally has a rhythm, e.g. a strong regular repeated pattern of sound. As noted above a sound effect may be any audio signal of shorter duration than the music. Sound effects may have rhythm but generally do not.

[0015] Features of different aspects and embodiments of the invention may be combined as appropriate, as would be apparent to a skilled person, and may be combined with any of the aspects of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] Embodiments of the invention will be described, by way of example only and with reference to the following drawings, in which:

Figure 1 is a flow chart illustrating a series of operations that may be performed in a method according to some embodiments of the invention;

Figure 2 is a more detailed flow chart showing the operations of figure 1 in more detail;

Figures 3a and 3b show an example of a sound effect signal before and after trimming;

Figure 4 is a graph showing potential candidate points for sound effects overlaid on a music signal;

Figure 5a is a graph showing onset strength determined using a mel-spectrogram;

Figure 5b is a graph showing onset strength determined using a constant-Q transform;

Figure 6 is a graph showing normalised onset strength determined using a mel-spectrogram;

Figure 7 is a graph showing normalised onset strength determined using the constant-Q transform;

Figure 8 is a graph which combines the results of the graphs in figures 6 and 7;

Figure 9 is a normalised version of the graph of figure 8;

Figure 10 is a graph corresponding to figure 9, additionally showing potential candidate points;

Figure 11 is a graph corresponding to figure 10 in which the potential candidate points are boosted;

Figure 12 is a graph in corresponding to figure 11 in which points above a threshold value are extracted;

13a shows an example of a window process for selecting a candidate point;

13b shows an example of final candidate points extracted from a window process; Figure 14a shows a final set of points for placement of a sound effect;

Figure 14b is a graph comparing the potential candidate points of figure 4 with the final set of points of figure 14a.

Figure 15 is a flowchart showing a method of producing a mixed audio signal according to some embodiments of the invention.

[0017] Common reference numerals are used throughout the figures to indicate similar features.

DETAILED DESCRIPTION

[0018] Embodiments of the present invention are described below by way of example only. These examples represent the best ways of putting the invention into practice that are currently known to the applicant although they are not the only ways in which this could be achieved.

[0019] In the methods and systems to be described, a music audio signal and a sound effect are input to the system. The invention is not limited to placement of a single sound effect but will firstly be described with reference to one sound effect for simplicity. Further, the invention is not limited to certain formats for the music audio signal and the sound effect. Suitable formats are .wav and MP3 but others will be familiar to those skilled in the art.

Method Overview

[0020] Figure 1 is an overview of a method of selecting points in a music audio signal according to some embodiments of the invention. The method of figure 1 may be performed in a computing system according to some embodiments of

the invention. The method and system may be implemented in various ways, some examples of which will be described in more detail with reference to figures 2 to 15.

[0021] Referring to figure 1, the music audio signal is analysed in step 1 to search for potential candidate points in time for placement of a sound effect, also referred to here as locations. The search for potential candidate points may be based on at least two different criteria. In the following example the criteria are downbeat and "kick" but other possible criteria may be known to those skilled in this art. Thus the potential candidate points are usually prominent points in the music audio signal due to their amplitude or some other criterion. The outcome of step 1 may be to find one or more points in the music. It is possible that no points will result from this search.

[0022] Any suitable signal processing techniques may be used for the search for points in step 1, for example using a music information retrieval library. In some processes, different libraries may be used for the retrieval of different points, for example according to the criteria being applied for the search for prominent points. Thus for example "Madmom" is a well-known example library for the determination of downbeat positions. Other techniques for downbeat detection are known in the art that are equally suitable. Similarly, automated drum transcripion library "ADTLib" is an example library for the determination of "kick" drum sounds. Both Madmom and ADTLib are available at the Python Package Index <https://pypi.org/>.

[0023] The music audio signal is further analysed in step 2 to determine variations in onset strength. An "onset" is the beginning of a musical note or other sound, both of which may be present in a music audio signal. Techniques for detection of onsets and their strengths are known in the art. The units used to determine onset strength may vary according to the technique that is used. Whatever units are used, the onset strength is a measure of the energy increase in the music audio signal. More than one such technique may be used and their results may be combined. One technique may be performed using a mel spectrogram, also known as a mel-frequency cepstrum. Additionally or alternatively a technique may use an audio signal processing library. For example the "librosa" python package may be used which includes an **onset_strength** attribute that is determined using a mel spectrogram. Another technique may use the constant-Q transform "CQT", in which the data series is transformed into the frequency domain, and the librosa package may also be used for this. The result of each technique may be an onset strength time series with a value for onset strength at each point in time. This is in contrast to step 1 which results in a set of discrete points. The results of each technique may be combined, for example after normalisation, to provide a resulting onset strength time series.

[0024] At step 3, the onset strength time series resulting from step 2 is boosted at any found points from step 1. The boosting comprises multiplying the values of the found points in the time series by different weighting factors. The weighting factors may be determined to depend on a user selectable value that affects the frequency of the identified placement points, in a manner to be described in more detail below. In practice the boosting may be applied to the frame in which each found point is located.

[0025] At step 4, points in the time series resulting from step 3, having an value larger than a predetermined threshold value, are selected as candidate points for the placement of the sound effect. The points selected at step 4 may be a subset of the initial candidate points: downbeat, kick and optionally others, resulting from step 1. However, additional points may be determined from step 4 as a result of the onset strength time series. This can be seen in figure 12 discussed further below where there are more points than in figure 4.

[0026] It may be desirable to reduce the number of points resulting from step 4. In step 5, points that are within a predetermined time window from each other are identified, and for each set of points in the same time window only the point with the largest value in the onset strength time series is selected for the placement of the sound effect.

[0027] It may be desirable to reduce the number of points for placement of the sound effect in addition to or alternatively to the window operation of step 5. Steps 6-8 are aimed at achieving a reasonable number of points for the piece of music.

[0028] In step 6, the selected points are sorted by onset strength. In step 7 an integer number n for the placement of a sound effect is defined, based on the relative durations of the music audio signal and the sound effect to be explained further below. Then at step 8 the n points with the highest onset strength value are selected for the placement of the sound effect.

[0029] At step 9, a peak of the energy of the sound effect signal is aligned with the music audio signal at each of the n points in the music audio signal. In each case the peak may be determined from rms values of the audio determined for each frame. For example the rms function from librosa at <https://librosa.org/doc/main/generated/librosa.feature.rms.html> may be used. in a particular example described further below, where there is more than one peak in the sound effect signal the second peak is used for the alignment. Then at step 10, points for which the sound effect overlaps the beginning or end of the music audio signal by more than a predetermined duration are removed as points for the placement of the sound effect. A different predetermined duration may be used for the beginning and end respectively or the same duration may be used.

[0030] If the result of step 10 is less than a predetermined number of points, one or more of the previous steps may be repeated with different parameters in order to ensure that at least a predetermined number of points is selected. In step 11, step 5 is repeated using a shorter predetermined time window if less than a predetermined number of points has been selected.

[0031] The outcome of the process illustrated in figure 1 is a set of points in the music audio signal for the placement of the sound effect.

[0032] Prior to this invention, the placement of sound effects has not been performed in any organised manner and has been performed by simple experimentation without following any particular set of rules. The method of figure 1 codifies the identification of the points in a repeatable manner so that a person with no skill in the art can mix an audio effect into a music audio signal and will be inspired to experiment with different pieces of music and effects without the laborious experimentation to determine where to place the effects.

Method Detailed Examples

[0033] Some examples of the method of figure 1 will now be described in more detail with reference to figure 2 and figures 3 to 14 which are graphs showing the outcome of the method at successive steps.

[0034] The method of figure 1 or figure 2 may be carried out in any computing system such as a laptop computer, desktop, tablet or smart phone. A system may for example be cloud based, and may for example receive user input via a user device, implement any of the methods described here, and output either a mixed audio signal or a selection of placement points to enable the user device to play back the mixed audio signal.. The invention may be implemented in software using one or more algorithms operating on a suitably configured processor. The steps or operations of the methods may be carried out on a single computer or in a distributed computing system across multiple locations. The software may be client based or web based, e.g. accessible via a server, or the software may be a combination of client and web based software. Those skilled in the art will be familiar with different ways to implement the methods described here in single devices or distributed over multiple devices.

[0035] In the flowchart of figure 2, the method begins with initialisation at operation 201 followed by obtaining or receiving a file for the audio effect at 203 and obtaining or receiving the music audio file at 205. In addition, a frequency value, which may for example be input by the user, is received or obtained at operation 207. The frequency value may be one of a number of predetermined values from which the user may select. In the system to be described further here, the frequency value determines parameters `ratio_weighting`, `boosted_points_strength` and `tolerance_value` used in operations 235, 229, 237. In general the frequency value may be used to determine any one or more of a maximum number of final key points to be extracted, a number of additional points extracted in addition to those with high onset strength and closeness (proximity) of chosen points. It is assumed that the sound effect has an amplitude peak, or point of highest amplitude, and this is detected at 209 and transmitted to the next stage in the method. Both the music audio file and the sound effect audio file may be subject to pre-processing at operation 211 if required.

[0036] The pre-processing may include any one or more of reading the signals with a predetermined sample rate, such as 44100 samples per second, trimming the effect signal for example if there is silence at the beginning and/or end (suitable tools for this are available e.g. from `librosa.effects.trim`) and converting signals to stereo if they are mono by duplicating the mono signal. It should be noted here that the invention is not limited to stereo signals and may equally well be performed to insert mono audio effects into mono musical signals. Figures 3a and 3b show an example of a sound effect signal before and after trimming.

[0037] If the duration of the music is found to be shorter than the duration of the sound effect at 213, the process ends at 215. Ideally the difference between the durations should be more than a predetermined amount, 1 second in the example of figure 2. If the music duration is larger than the effect duration by more than 1 second as determined at 217, the process proceeds to operation 225.

[0038] If the difference between the music duration and the sound effect duration is found at 217 to be less than 1 second or some other predetermined value, it may be decided that a search for placement points is not necessary but nevertheless the method may be used to mix the music and the sound effect. For this purpose, a further test is performed at 219 to determine whether the total of the effect start time to peak time and the effect duration are less than the music duration. If not the process ends at 221. Otherwise at 223 the effect peak point is returned which can then be used as the start location for the effect (see operations 1512 and 1514 in figure 15).

[0039] At operation 225, steps 1 and 2 described in figure 1 are performed to find points for the insertion of sound effects, referred to below as "**key_points_to_boost**" (step 1) and to obtain a **final_onset_strength** time series (step 2).

[0040] In one example, step 1 may comprise:

[0041] For the audio signal input, obtain prominent locations time series:

A. Downbeat positions (`a_1`, `a_2`, ..., `a_n`), where `a_1`, `a_2`, ..., `a_n` locations in seconds detected by `madmom` python library.

B. Positions of 'Kick' drum sounds (`b_1`, `b_2`, ..., `b_n`), where `b_1`, `b_2`, ..., `b_n` locations in seconds detected by `ADTLib` python library.

C. Get the union of points A and B ("**key_points_to_boost**")

[0042] The graph of figure 4 is an example of the result of step 1, with the **key_points_to_boost** overlaid on the music signal.

[0043] In one example, step 2 may comprise the following sub-steps A-F:
For the audio signal input, detect "**final_onset_strength**" time series:

A. **onsets_melspec**: time series in number of frames for onset strength of mel spectrogram feature. Use **librosa.onset.onset_strength** attribute of Librosa python library with the following inputs:

- i. Audio signal converted to mono,
- ii. sampling rate "sr": 44100,
- iii. "feature": librosa.feature.melspectrogram,
- iv. "aggregate": mean,
- v. "fmax": 8000,
- vi. "n_mels": 512

B. **Onsets_CQT**: time series in number of frames for onset strength of CQT feature.

Get the absolute values of the CQT feature from Librosa python library ("C"), then use **librosa.onset.onset_strength** attribute of Librosa python library with the following inputs:

- vii. Sampling rate "sr": 44100,
- viii. "S": amplitude_to_db attribute of Librosa python library with input C and "ref": max.

C. Normalise the values from step 2 A in range [0, 1].

D. Normalise the values from step 2 B in range [0, 1].

E. Sum element-wise the outcome from C and D.

F. Normalise the outcome from E in range [0, 1] (outcome time series: "**final_onset_strength**").

[0044] Figure 5a show the onset strength time series using the mel-spectrogram feature and figure 5b shows the onset strength time series using the CQT feature. Both of figures 5a and 5b are based on the music signal of figure 4, e.g. the outcome of step 2B. Figure 6 shows the normalised onset strength using the mel spectrogram feature, e.g. the outcome of step 2C, and figure 7 shows the normalised onset strength using the CQT feature, e.g. the outcome of step 2D. Figure 8 shows the sum element-wise of the onset strengths of figures 5 and 6, e.g. the outcome of step 2E, and figure 9 shows the **final_onset_strength** time series resulting from step 2F.

[0045] Referring back to figure 2, a check is made at 227 that step 1 has resulted in at least one point. If so the process continues to 229 implementation of step 3 described in general terms above with reference to figure 1.

[0046] At step 3, the locations found in step 1, **key_points_to_boost**, are boosted at the **final_onset_strength** time series with a weight parameter to produce a boosted onset strength time series, referred to below as "**combined_with_boosted_points**" time series.

[0047] In one example, step 3 may be implemented as follows:

[0048] For each key point in **key_points_to_boost**, add the following weight to closest frame in seconds of **final_onset_strength**:

weight: mean of **final_onset_strength** + **boosted_points_strength** * standard deviation of **final_onset_strength**, where **boosted_points_strength** is a selected integer number.

[0049] Then extract the resulted "**combined_with_boosted_points**" time series.

[0050] The results of step 3 for the music audio signal of figure 4 are shown in figures 10 and 11. Figure 10 shows the **final_onset_strength** time series overlaid with the **key_points_to_boost** which are candidate points for the placement of a sound effect. Figure 11 shows the result of the boosting of the key points according to the process of step 3.

[0051] After operation 229, the flow of figure 2 continues to 231 where step 4 is performed: Detect the locations where the **combined_with_boosted_points** values from step 3 are bigger than 0.3. The result of this process is shown in figure 12.

[0052] If step 1 does not result in any points being determined, step 3/operation 229 may be omitted.

[0053] If step 3 is performed, then operation 4 is performed on the boosted onset strength time series. If step 3 is not performed because no points are found in step 1, then operation 3 is performed on the (unboosted) onset strength time

series. The time series on which step 4 is performed is designated in figure 2 "**new_series**".

[0054] At operation 235 step 7 is carried out in which an integer number n of points is defined for the placement of a sound effect based on the relative durations of the music audio signal and the sound effect. This number may be defined as:

Song duration in seconds = sd

Effect duration in seconds = ed

$\text{max_number} = \text{int}(\text{song_dur}/(\text{effect_dur} * \text{ratio_weighting}))$

where **ratio_weighting** is a defined integer number.

If **max_number** is less than zero it is forced to be one.

[0055] At operation 237, step 5 is carried out, for example: recursively extract points detected in step 4 which are close to a neighbour point in time, given a **tolerance_value** in seconds: if they are close ("closeness" defined by the tolerance value), select the one with the biggest **combined_with_boosted_points** value. The **tolerance_value** defines the time window referred to in connection with figure 1. An example initial tolerance value might be 0.3 seconds.

[0056] Figure 13a illustrates an example set of points that may be extracted at step 5 and the selection of a point with the highest value. It may be necessary to repeat this extraction and selection of points using a smaller tolerance value as explained with reference to operation 247. Figure 13b shows a set of extracted points obtained using a tolerance value of 0.3.

[0057] Also at operation 237, following step 5, step 6 is performed in which the onset locations extracted from step 5 are sorted in descending order of their onset strength value.

[0058] Next, in operation 239, step 8 is performed in which the first **max_number** points of the key points from step 6 are obtained. Then, in operation 241, step 10 is carried out in which points are removed that fall into edge categories, for example:

[0059] Remove the key points extracted from step 8 that follow to at least one of the following categories:

- Key point that does not allow the sound effect to play its whole duration (e.g. key point at 500ms before audio signal finishes and duration of sound effect is 800ms with peak from step 9 at 100ms: this means that 200ms of the end of the sound effect would be cut out)
- Key point that does not allow the sound effect to play its beginning (e.g. key point at 500ms of audio signal and the peak extracted from step 9 is aligned to 800ms of sound effect: this means that 300ms of the beginning of sound effect would be cut out).

[0060] At operation 243 a check is made that at least one point remains. If so, then at operation 245 the first **max_number** of points from operation 237, step 6, are obtained and then the sound effect is aligned with the music. Here, the key point may be aligned with the time when the second highest peak value occurs (or the first highest if there is only one peak detected). The points may then be adjusted so that each point defines the start of the sound effect. The result is a set of aligned points at 246, illustrated in figure 14a.

[0061] Figure 14b compares the set of points that result from step 1 shown in figure 4 to the set of aligned points of figure 14a. It can be seen that there are points in figure 14a that are not present in figure 4 and result from the identification of potential candidate points, in this example using the mel-spectrogram and CQT, that have a higher value than 0.3 and are not identified from the onset strength time series. More generally there can be a possibility that points from **final_onset_strength** time series can be higher than the chosen threshold, in this example 0.3, and not part of the "downbeat" and "kick" points. These potential candidates provide a representation sensitive to the frequency bins in octave resolution: essentially the energy is more centralized and a change in pitch affects this type of centralization and this provides some prominent points in the onset strength.

[0062] It will be appreciated from the foregoing that there are songs and other music audio signals where the downbeats are not able to be detected or they are detected wrongly. Using the method described here, it is possible to make sure that there is at least one point where the sound effect can be placed even if downbeat points are not detected. Also it enhances the probability of rhythmically or musically meaningful locations to be selected even if downbeat points are not correct.

[0063] The set of aligned points resulting at 246 in figure 2 that may be used in the flow of figure 15.

[0064] At operation 247, if all points have been removed, step 11 is carried out to ensure that at least one point is obtained for the placement of a sound effect, for example as follows:

- i. Define **max_number**: 1
- ii. Repeat step 5 with **tolerance_value**: 0.015
- iii. While (ii) does not provide a key point which does not fall into the edge cases defined in step 10:
- iv. Increase the **max_number** by 1
- v. do (ii) again.

[0065] At operation 249 an alignment process is carried out similar to that carried out at operation 245. The result is a set of aligned points at 251 that may be used in the flow of figure 15.

[0066] In an optional feature, different frequency levels may be provided for the placement of sound effects according to the density of points found in step 1.

[0067] In a specific example, a system may provide 5 different frequency levels which correspond to levels of how dense is the number of key points detected. Frequency levels may be predetermined, such as: '1', '2', '3', '4', and 'auto', and they may be selectable by the user. Parameter values per frequency level:

```

if frequency == 'auto':
    ratio_weighting = 5
    boosted_key_points_strength = 4
    tolerance_value = 0.015
if frequency == '1':
    ratio_weighting = 7
    boosted_key_points_strength = 5
    tolerance_value = 0.5 if track_duration > 0.5 else track_duration
if frequency == '2':
    ratio_weighting = 5
    boosted_key_points_strength = 4
    tolerance_value = 0.5 if track_duration > 0.5 else track_duration
if frequency == '3':
    ratio_weighting = 4
    boosted_key_points_strength = 4
    tolerance_value = 0.015
if frequency == '4':
    ratio_weighting = 2
    boosted_key_points_strength = 4
    tolerance_value = 0.2

```

[0068] Figure 15 is a flowchart showing how a mixed audio signal, e.g. a music signal mixed with a sound effect, may be produced using placement points obtained by any of the methods described above.

[0069] After initialisation at 1500, a number of inputs are obtained or received in order to perform the mixing. These inputs comprise the points 1501 at which the sound effect is to be placed, for example as determined at 246 or 251 in figure 2, the pre-processed music audio signal 1505 for example resulting from operation 211 of figure 2 and the pre-processed sound effect for example resulting from operation 203.

[0070] Additional optional inputs include "overlap" obtained at 1502. In some implementations of the method, a user may have the option to decide whether sound effects should be permitted to overlap, for example if the duration of the sound effect is longer than the gap between two consecutive placement points. Thus the "overlap" may be a binary choice. If overlap is not to be permitted this can be handled in a number of ways in the mixing process including one or more of shortening one sound effect to end before the next one begins, optionally with a fade out of volume, shortening one sound effect to commence after the previous one has ended, optionally with a fade in. In some implementations the overlap option may be predetermined so that the user has no control over this.

[0071] A further optional additional input is a volume balance 1503 which determines the relative volumes of the music audio signal and the sound effect. Again this may be predetermined or selectable by the user.

[0072] Then at operation 1510 an empty "final_effect_signal" is initialised. This final effect signal may comprise a signal having the duration of the music audio signal into which the sound effects may be placed, to then be mixed with the music audio signal.

[0073] Then, at operation 1511 a check is made as to whether overlap is permitted. If not, then in the illustrated example at 1512 the effect is located to each point such that the point defines the start position of the effect signal and the end of the sound effect signal is trimmed so that the sound effect stops being produced in the presence of another sound

effect point. If overlap is not permitted then at 1514 the effect is located to each point similarly to 1512 but the trimming is not performed. The result of operation 1512 or 1514 is an updated final_effect_signal which is then multiplied at 1516 by a volume level, for example determined by the volume balance at 1503 to produce an effect signal with volume, which is then mixed with the pre-processed music signal at operation 1518 to produce the mixed audio signal output at 1520.

[0074] The methods described in the foregoing may be readily modified to accommodate different sound effects of the same or different durations. For example additional filtering might be included to determine which sound effect(s) could be accommodated at which point(s) and if more than one sound effect could be accommodated at a particular point then one could be selected, for example randomly.

[0075] Some operations or steps of the methods described herein may be performed by software in machine readable form e.g. in the form of a computer program comprising computer program code. Thus some aspects of the invention provide a computer readable medium which when implemented in a computing system cause the system to perform some or all of the steps or operations of any of the methods described herein. The computer readable medium may be in transitory or tangible (or non-transitory) form such as storage media include disks, thumb drives, memory cards etc. The software can be suitable for execution on a parallel processor or a serial processor such that the method steps may be carried out in any suitable order, or simultaneously.

[0076] This application acknowledges that firmware and software can be valuable, separately tradable commodities. It is intended to encompass software, which runs on or controls "dumb" or standard hardware, to carry out the desired functions. It is also intended to encompass software which "describes" or defines the configuration of hardware, such as HDL (hardware description language) software, as is used for designing silicon chips, or for configuring universal programmable chips, to carry out desired functions.

[0077] The embodiments described above are largely automated. In some examples a user or operator of the system may manually instruct some steps of the method to be carried out.

[0078] In the described embodiments of the invention the system may be implemented as any form of a computing and/or electronic system as noted elsewhere herein. Such a device may comprise one or more processors which may be microprocessors, controllers or any other suitable type of processors for processing computer executable instructions to control the operation of the device in order to gather and record routing information. In some examples, for example where a system on a chip architecture is used, the processors may include one or more fixed function blocks (also referred to as accelerators) which implement a part of the method in hardware (rather than software or firmware). Platform software comprising an operating system or any other suitable platform software may be provided at the computing-based device to enable application software to be executed on the device.

[0079] The term "computing system" is used herein to refer to any device with processing capability such that it can execute instructions. Those skilled in the art will realise that such processing capabilities may be incorporated into many different devices and therefore the term "computing system" includes PCs, servers, smart mobile telephones, personal digital assistants and many other devices.

[0080] It will be understood that the benefits and advantages described above may relate to one embodiment or may relate to several embodiments. The embodiments are not limited to those that solve any or all of the stated problems or those that have any or all of the stated benefits and advantages.

[0081] The term "comprising" is used herein to mean including the method steps or elements identified, but that such steps or elements do not comprise an exclusive list and a method or apparatus may contain additional steps or elements.

[0082] The figures illustrate exemplary methods. While the methods are shown and described as being a series of acts that are performed in a particular sequence, it is to be understood and appreciated that the methods are not limited by the order of the sequence. For example, some acts can occur in a different order than what is described herein. In addition, an act can occur concurrently with another act. Further, in some instances, not all acts may be required to implement a method described herein.

[0083] It will be understood that the above description of a preferred embodiment is given by way of example only and that various modifications may be made by those skilled in the art. What has been described above includes examples of one or more embodiments. It is, of course, not possible to describe every conceivable modification and alteration of the above devices or methods for purposes of describing the aforementioned aspects, but one of ordinary skill in the art can recognize that many further modifications and permutations of various aspects are possible. Accordingly, the described aspects are intended to embrace all such alterations, modifications, and variations that fall within the scope of the appended claims.

Claims

1. A computer-implemented method of selecting points in a music audio signal for placement of a sound effect comprising:

determining an onset strength time series for the music audio signal; and
selecting points from the onset strength time series with a value larger than a predetermined threshold as
candidate points for the placement of the sound effect.

- 5 **2.** The method of claim 1 comprising searching for points in the music audio signal as potential candidate placement
points based on one or more criteria; and
boosting the onset strength time series at points found by the searching prior to selecting points as candidate points.
- 10 **3.** The method of claim 2 wherein the boosting comprises multiplying the values of the found points in the time series
by different weights.
- 4.** The method of claim 3 comprising receiving a user selected value for the frequency of placement points, wherein
the weights are calculated based on the user selected value.
- 15 **5.** The method of claim 2 , 3 or 4 wherein the one or more criterial comprise one or both of downbeat and kick.
- 6.** The method of any preceding claim wherein the onset strength time series is determined using a combination of
two or more onset strength determination methods.
- 20 **7.** The method of claim 6 wherein the two or more methods comprise one or both of using a mel spectrogram and
constant-Q transform.
- 8.** The method of any preceding claim comprising identifying in the selected points any which are within a predetermined
time window from another selected point, and for each predetermined time window selecting only the point with the
25 largest weighted value for the placement of a sound effect.
- 9.** The method of claim 8 comprising repeating the selection of only the points with the largest weighted value using
a shorter predetermined time window if less than a predetermined number points is selected.
- 30 **10.** The method of any preceding claim comprising determining a number n of points for the placement of a sound effect,
and reducing the number of selected points to the n points with the highest onset strength value.
- 11.** The method of claim 10 wherein n is based on the relative durations of the music audio signal and the sound effect.
- 35 **12.** The method of claim 9 or claim 10 comprising increasing n if less than a predetermined number of points is selected.
- 13.** The method of any preceding claim comprising temporally aligning a peak of energy from the sound effect signal
with each selected point in the music audio signal.
- 40 **14.** The method of claim 13 comprising removing from the selected points those for which the sound effect overlaps the
beginning of the music audio signal by a predetermined duration or overlaps the end of the music audio signal by
a predetermined duration.
- 45 **15.** A method of modifying a music audio signal comprising:
receiving a music audio signal and a sound effect signal;
identifying points in the music audio signal for placement of the sound effect according to the method of any
preceding claim, and
50 inserting the sound effect signal into the music audio signal at each of the identified points.

Amended claims in accordance with Rule 137(2) EPC.

- 55 **1.** A computer-implemented method of selecting points in a music audio signal for placement of a sound effect com-
prising:
searching for points in the music audio signal as potential candidate placement points based on one or more
criteria;

determining an onset strength time series for the music audio signal; and
boosting the onset strength time series at points found by the searching prior to selecting points as candidate
points; and
selecting points from the boosted onset strength time series with a value larger than a predetermined threshold
as candidate points for the placement of the sound effect.

2. The method of claim 1 wherein the boosting comprises multiplying the values of the found points in the time series by different weights.

3. The method of claim 2 comprising receiving a user selected value for the frequency of placement points, wherein the weights are calculated based on the user selected value.

4. The method of claim 1, 2 or 3 wherein the one or more criteria comprise one or both of downbeat and kick.

5. The method of any preceding claim wherein the onset strength time series is determined using a combination of two or more onset strength determination methods.

6. The method of claim 5 wherein the two or more methods comprise one or both of using a mel spectrogram and constant-Q transform.

7. The method of any preceding claim comprising identifying in the selected points any which are within a predetermined time window from another selected point, and for each predetermined time window selecting only the point with the largest weighted value for the placement of a sound effect.

8. The method of claim 7 comprising repeating the selection of only the points with the largest weighted value using a shorter predetermined time window if less than a predetermined number of points is selected.

9. The method of any preceding claim comprising determining a number n of points for the placement of a sound effect, and reducing the number of selected points to the n points with the highest onset strength value.

10. The method of claim 9 wherein n is based on the relative durations of the music audio signal and the sound effect.

11. The method of claim 8 or claim 9 comprising increasing n if less than a predetermined number of points is selected.

12. The method of any preceding claim comprising temporally aligning a peak of energy from the sound effect signal with each selected point in the music audio signal.

13. The method of claim 12 comprising removing from the selected points those for which the sound effect overlaps the beginning of the music audio signal by a predetermined duration or overlaps the end of the music audio signal by a predetermined duration.

14. A method of modifying a music audio signal comprising:

receiving a music audio signal and a sound effect signal;
identifying points in the music audio signal for placement of the sound effect according to the method of any preceding claim, and
inserting the sound effect signal into the music audio signal at each of the identified points.

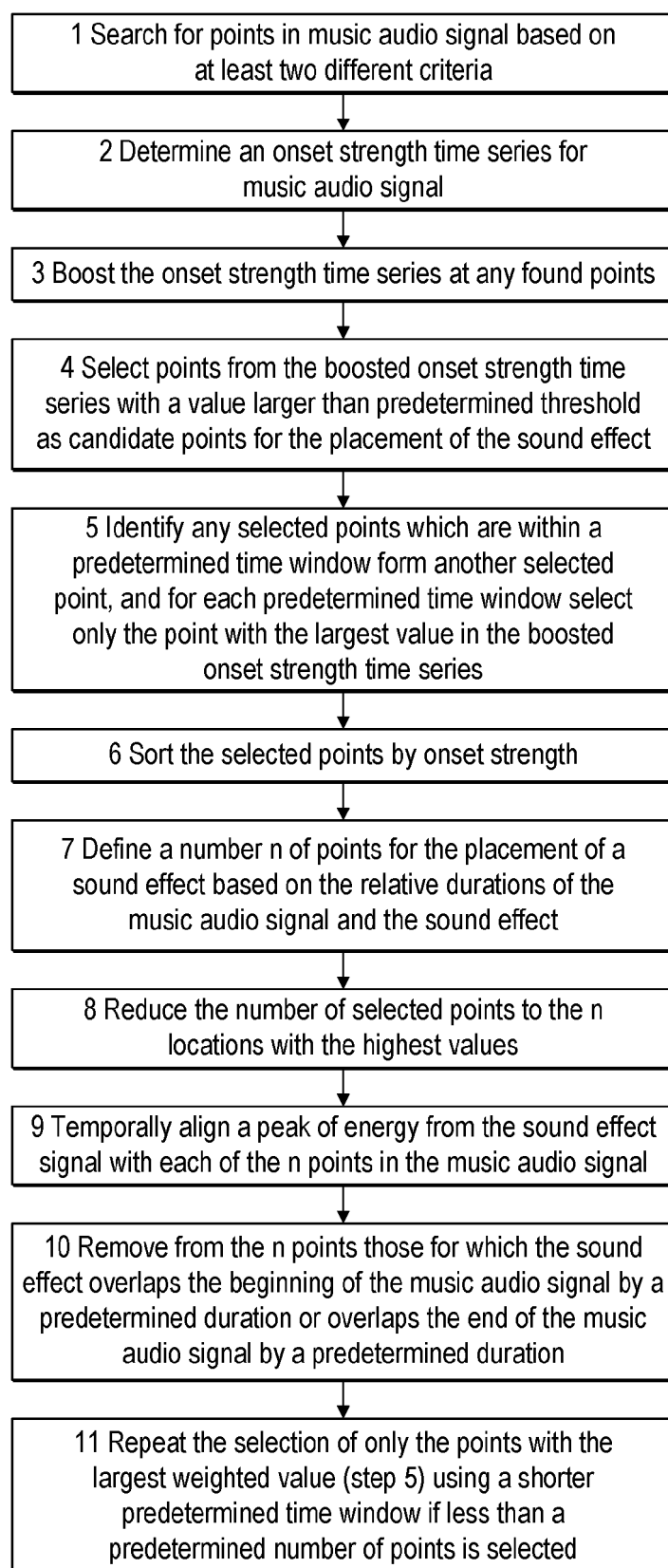
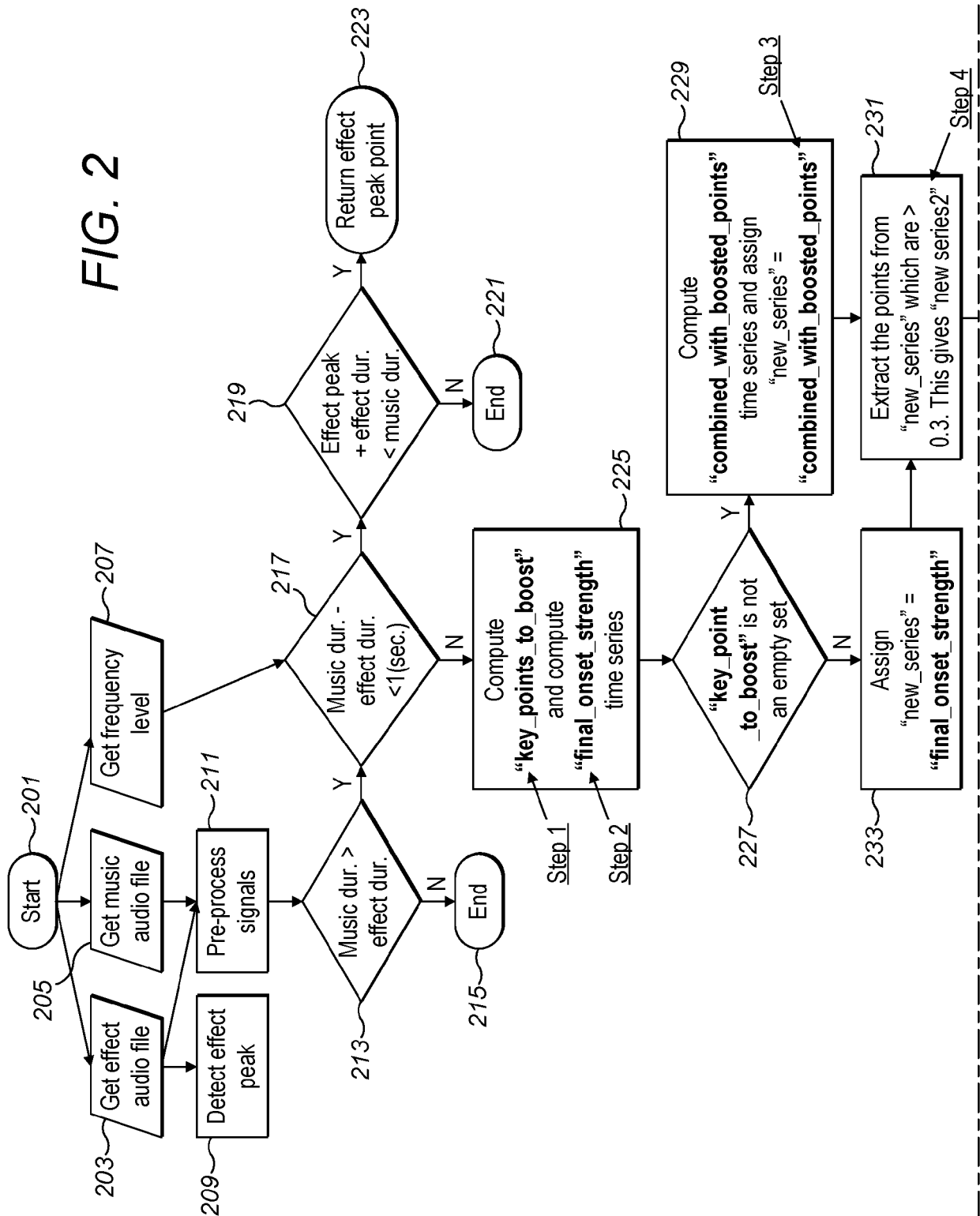
*FIG. 1*

FIG. 2



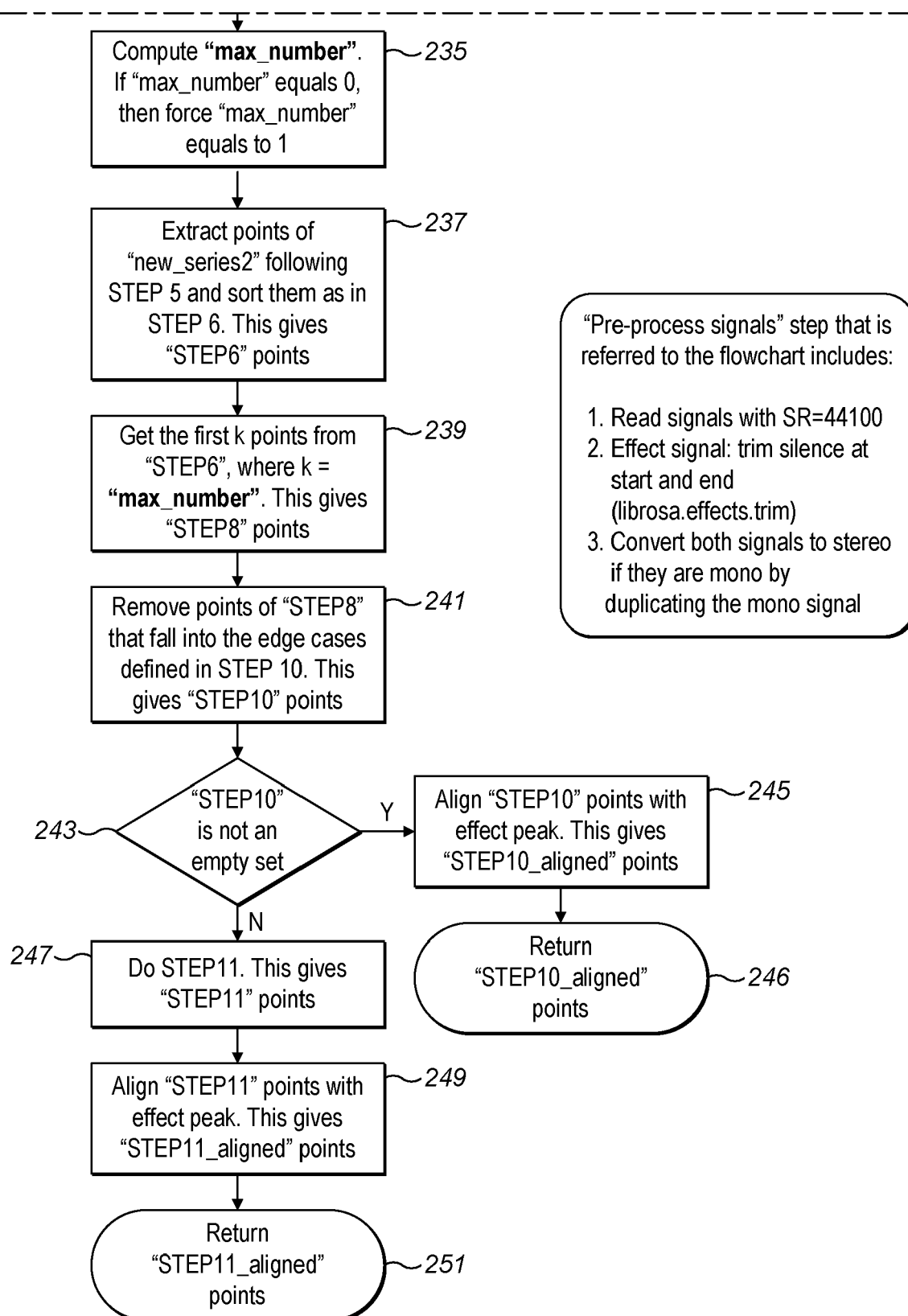


FIG. 2 Cont'd

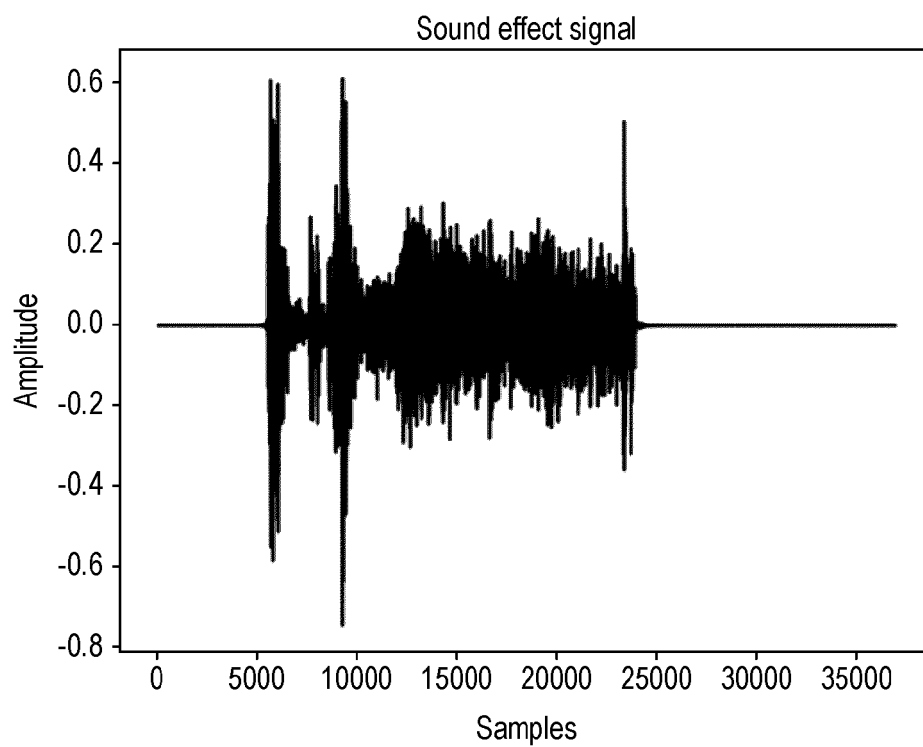


FIG. 3a

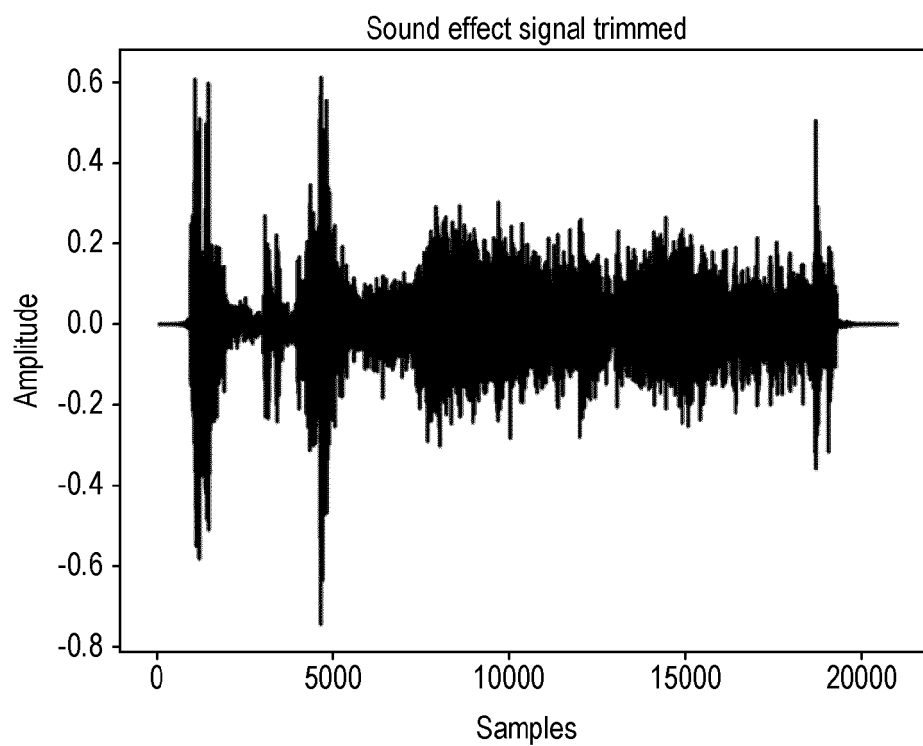


FIG. 3b

Figure 1 step 1 outcome

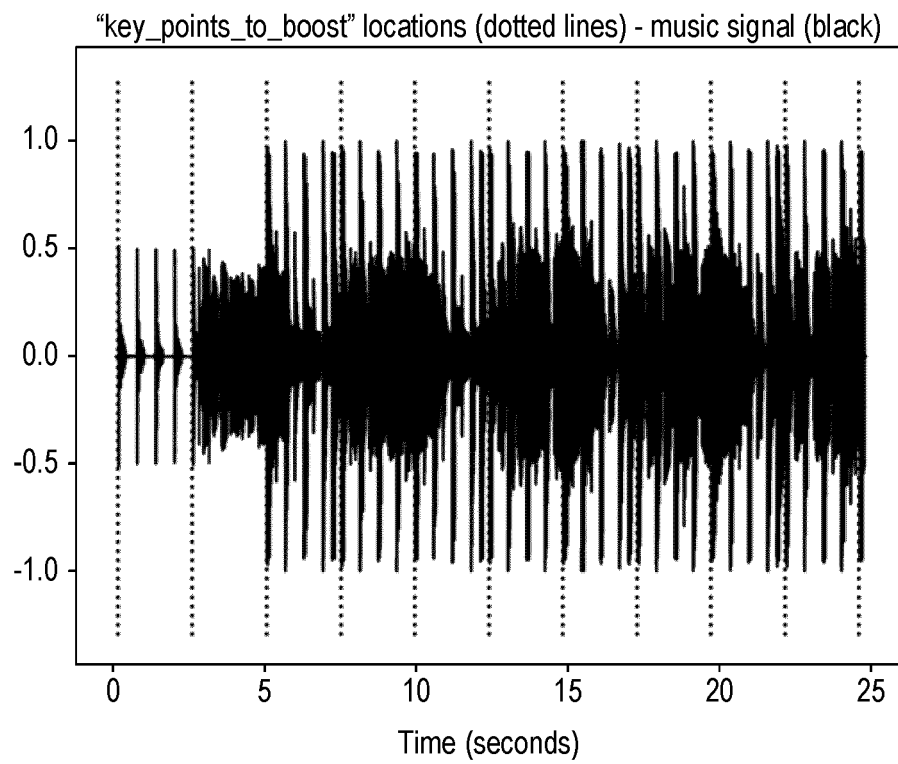


FIG. 4

Figure 1 step 2A outcome

Onset strength using Mel-spectrogram feature

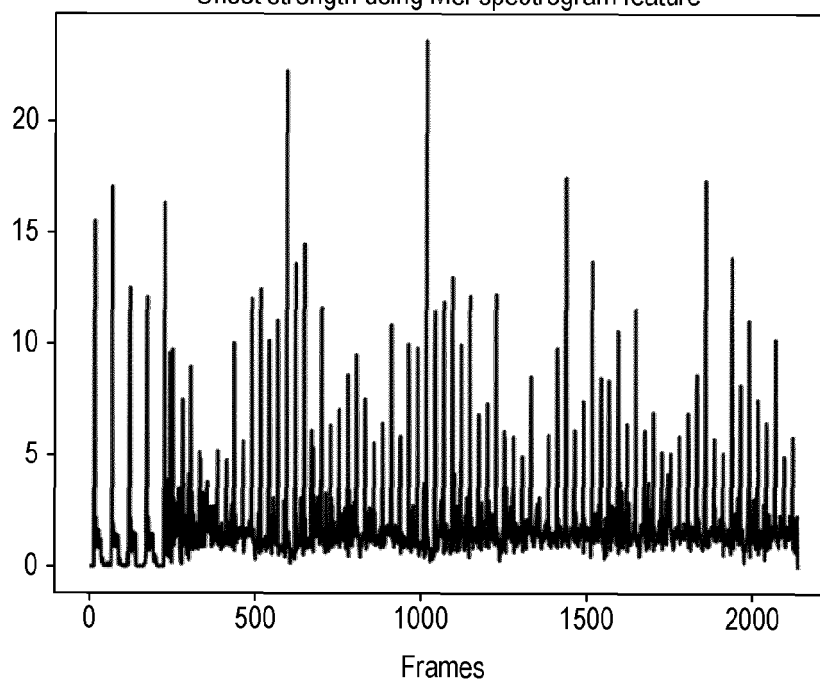


FIG. 5a

Figure 1 step 2B outcome

Onset strength using CQT feature

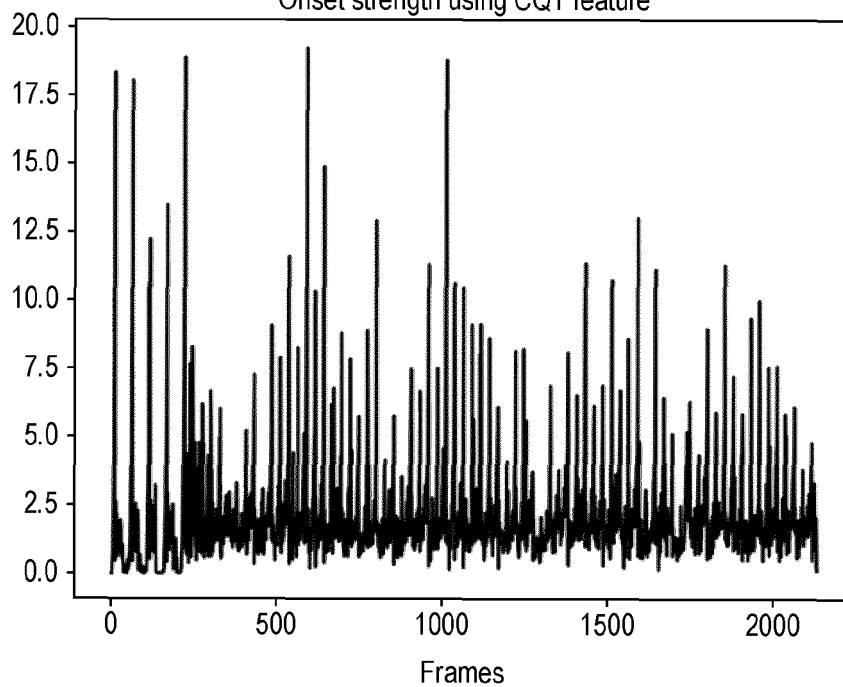


FIG. 5b

Figure 1 step 2C outcome

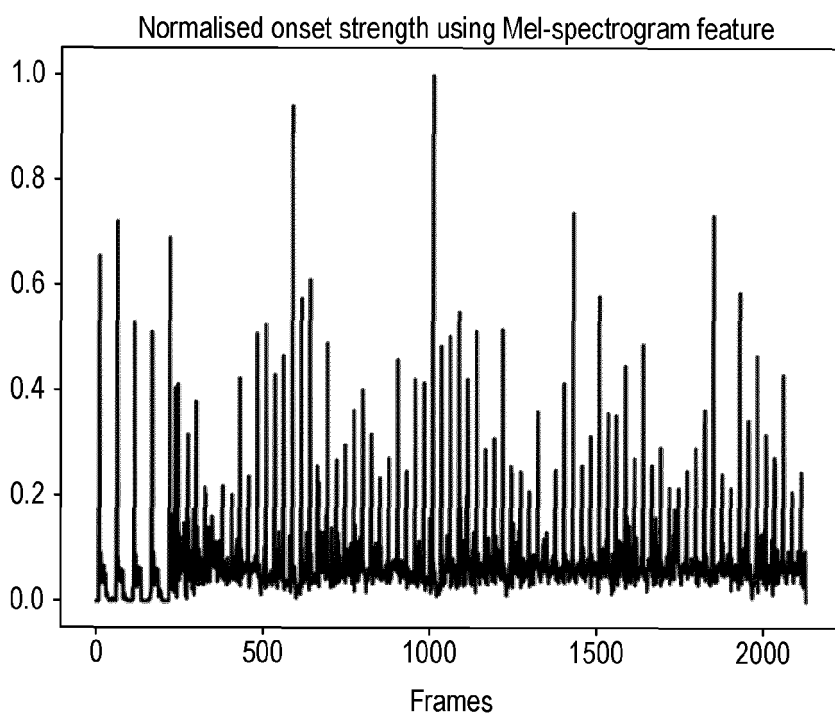


FIG. 6

Figure 1 step 2D outcome

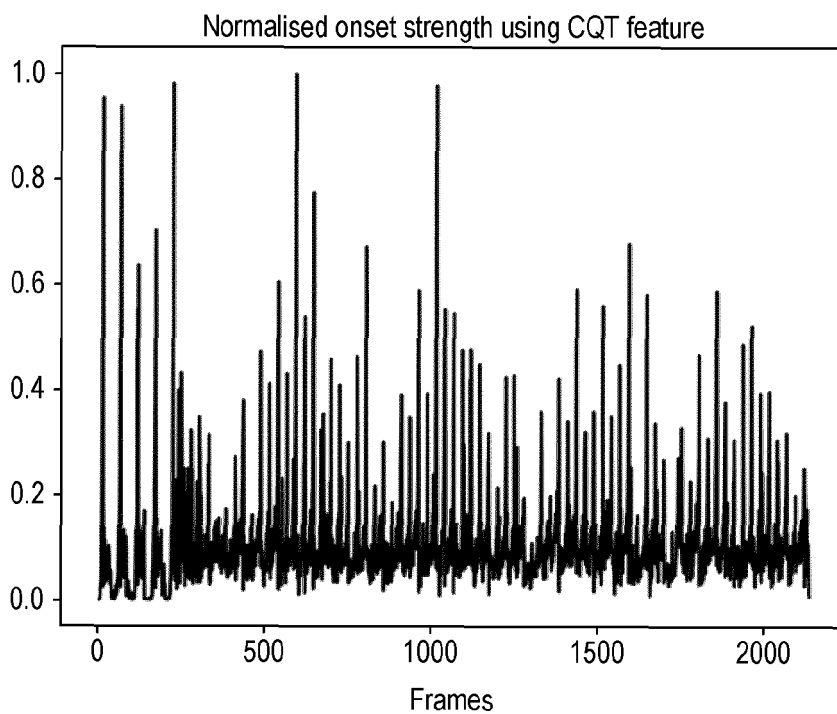


FIG. 7

Figure 1 step 2E outcome

Sum element-wise of normalised mel-spectrogram onset strength
and normalised CQT onset strength

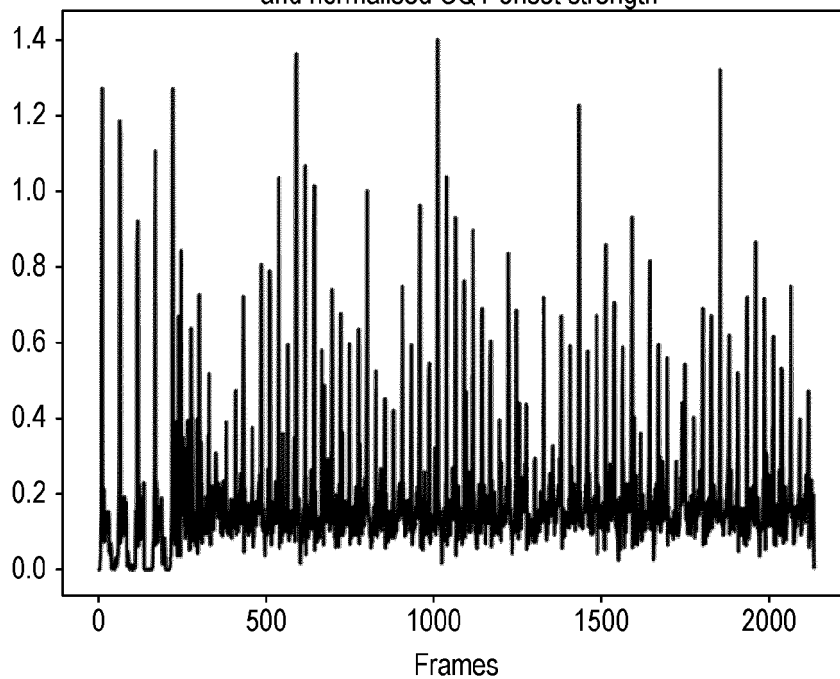


FIG. 8

Figure 1 step 2F outcome

"final_onset_strength" time series

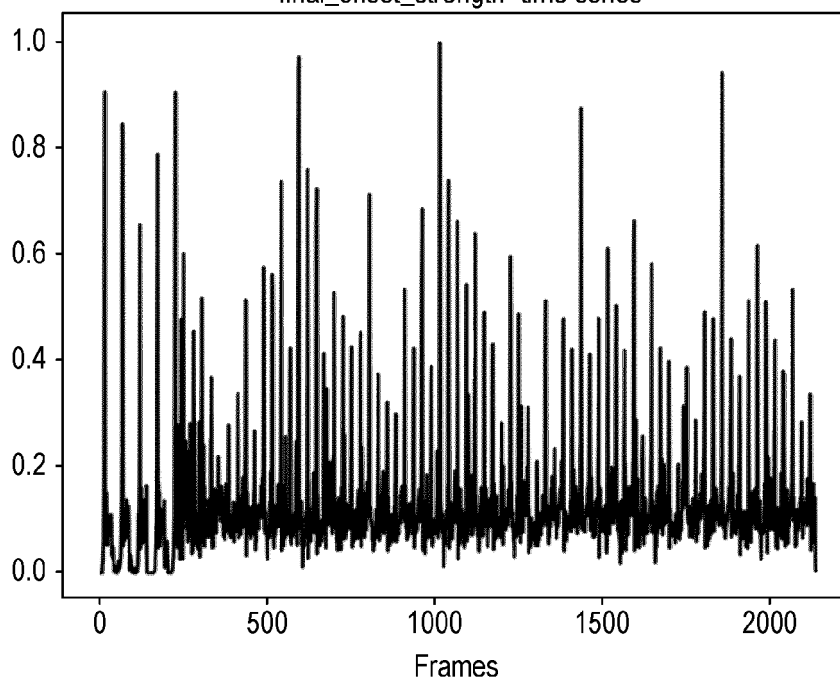


FIG. 9

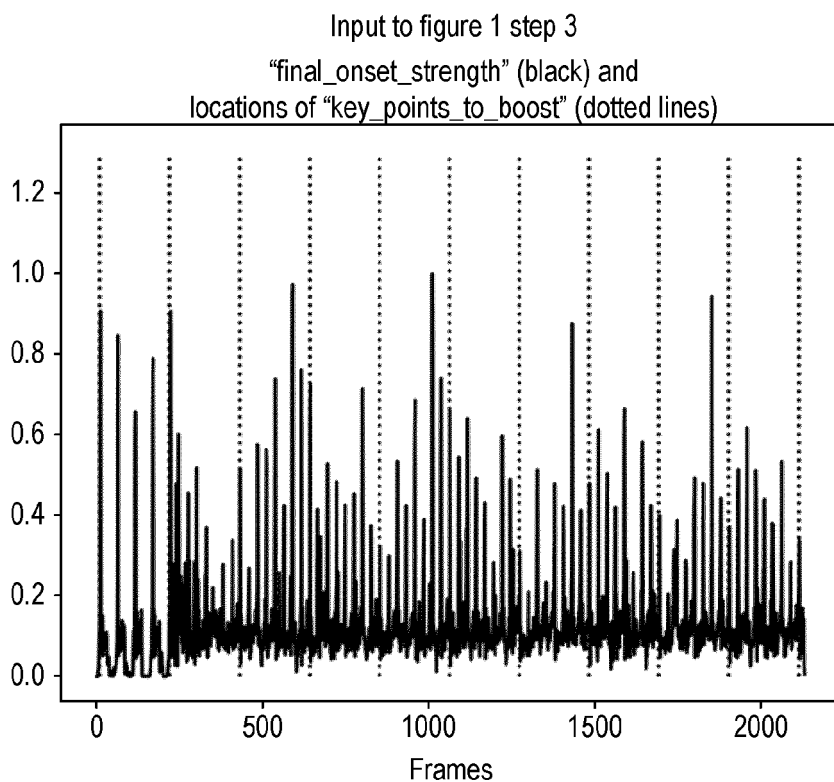


FIG. 10

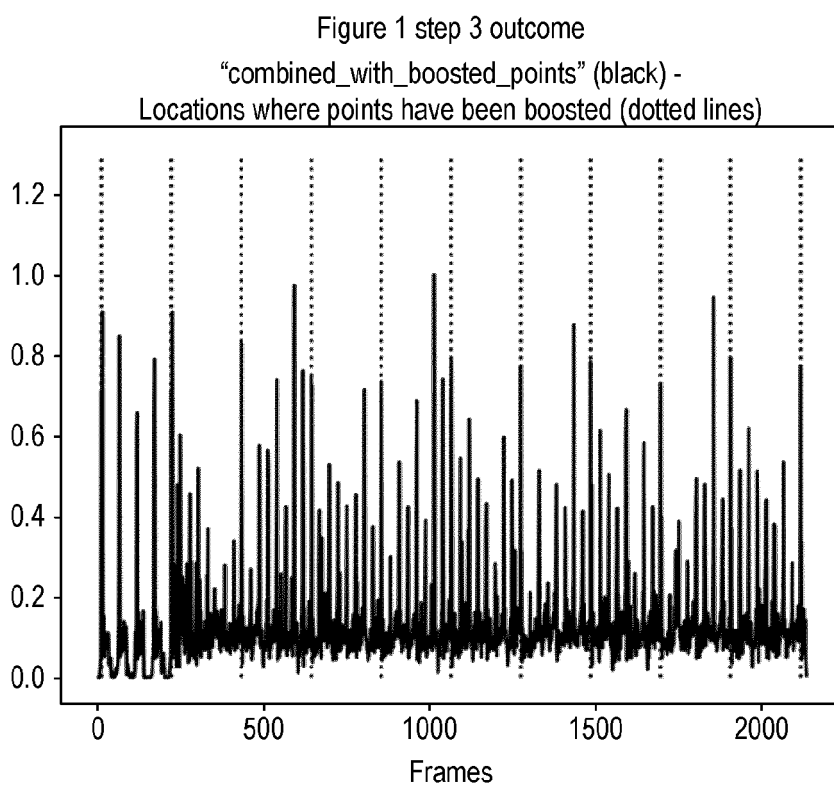


FIG. 11

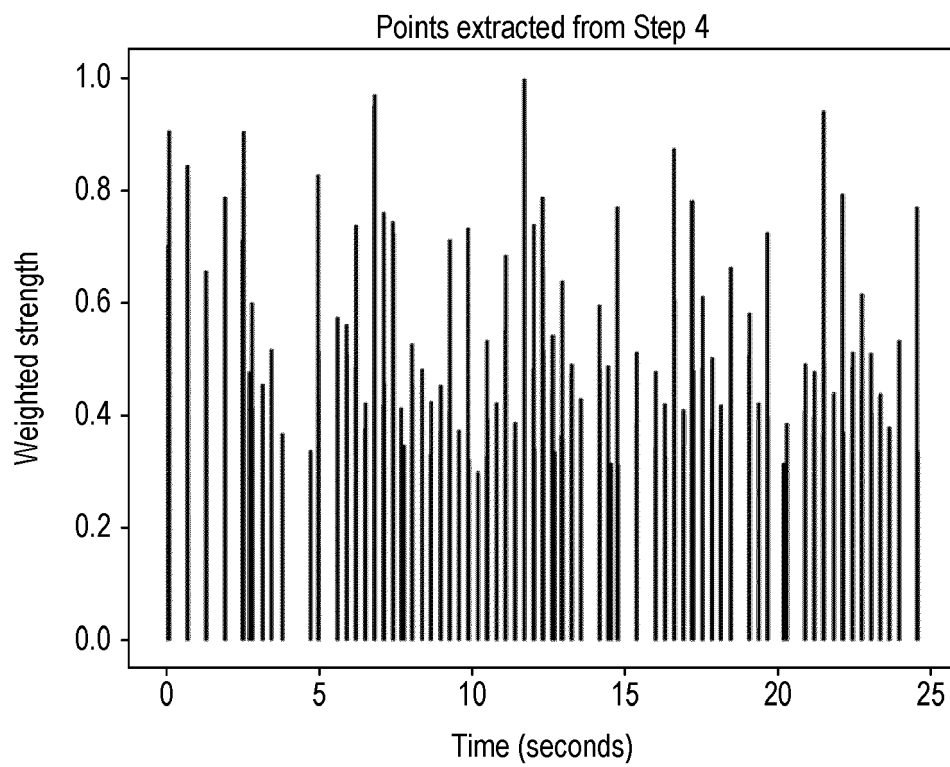


FIG. 12

Example of one iteration of a window process to group candidate points and select the highest one (black line), eliminating the remaining points (dotted lines)

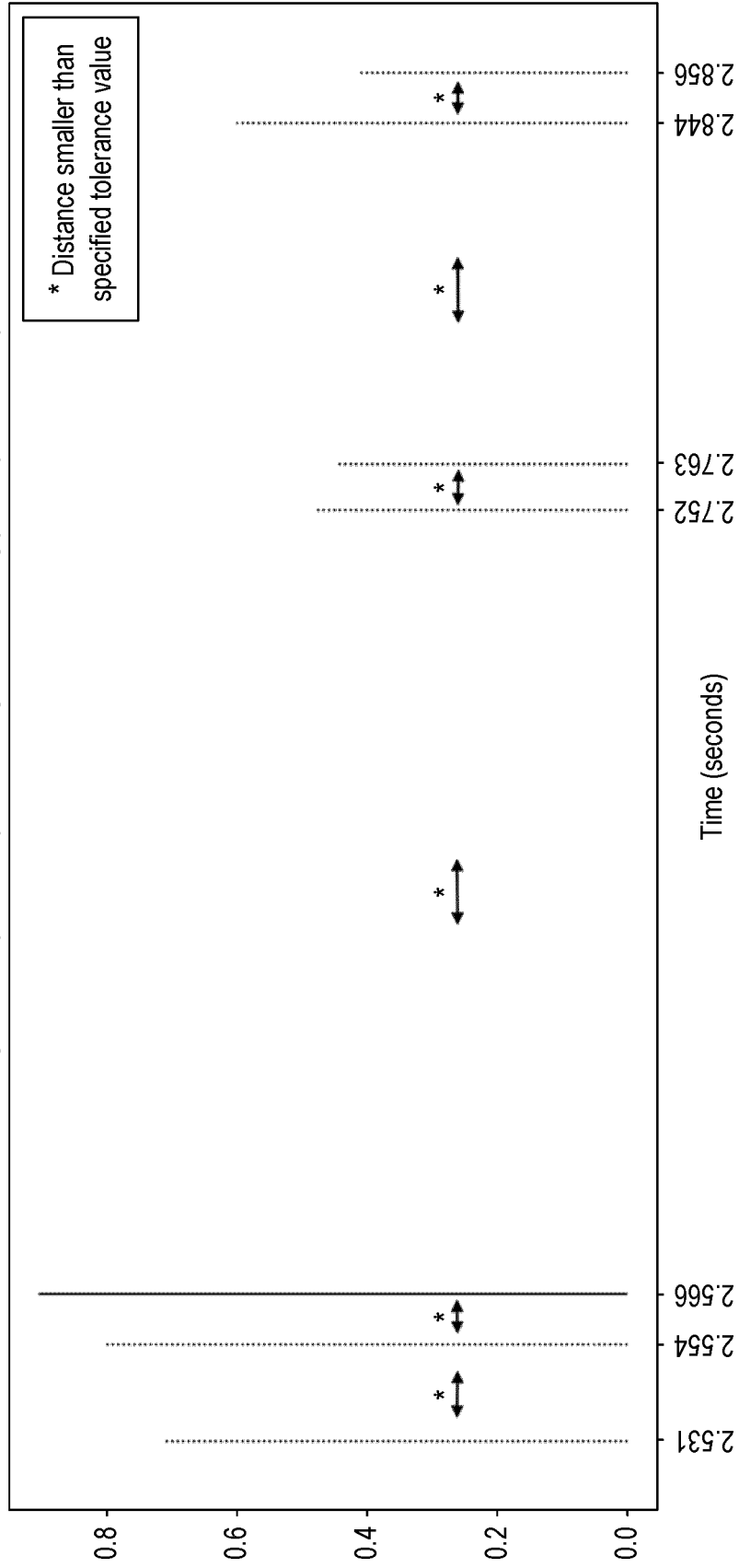


FIG. 13a

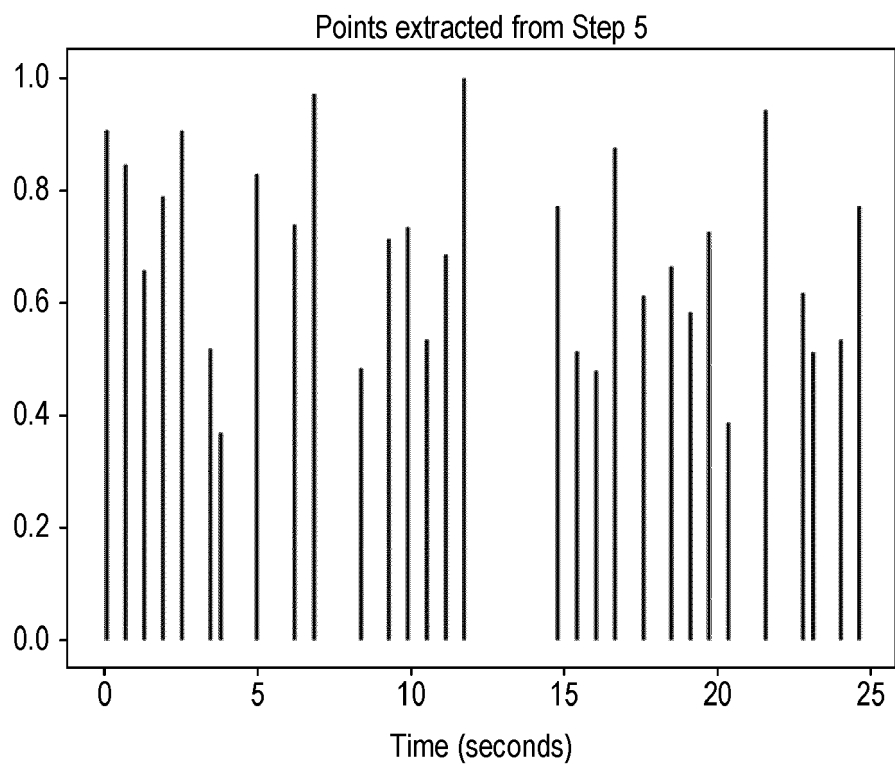


FIG. 13b

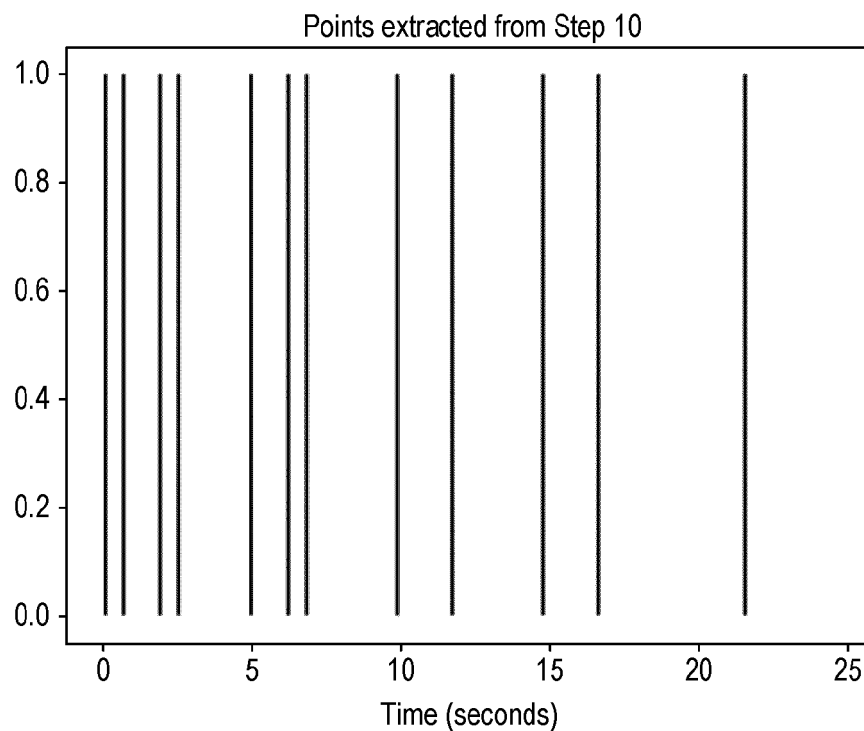


FIG. 14a

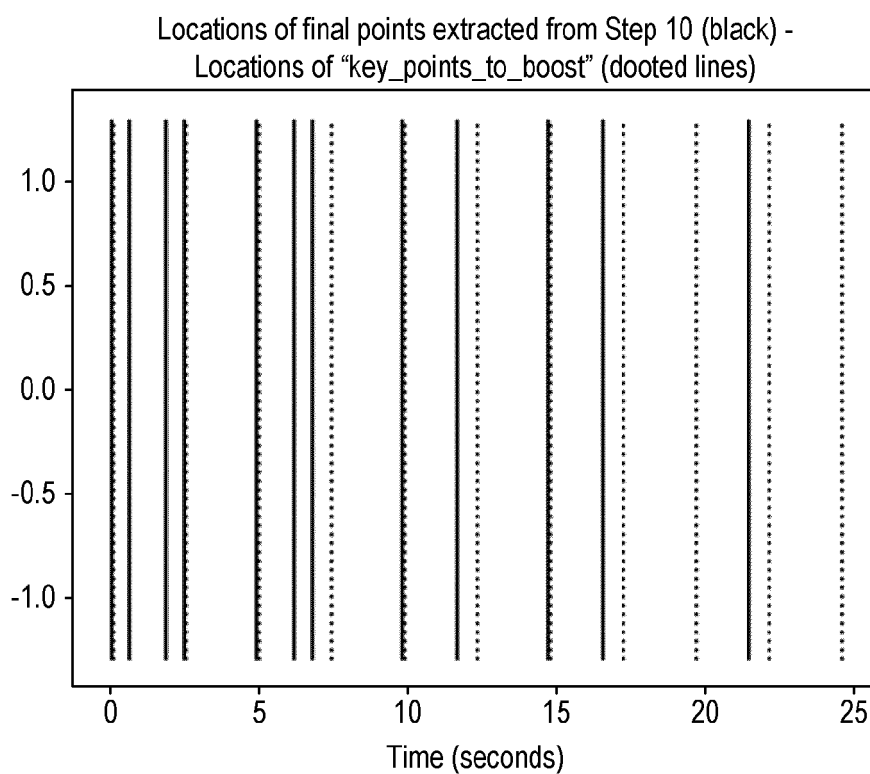


FIG. 14b

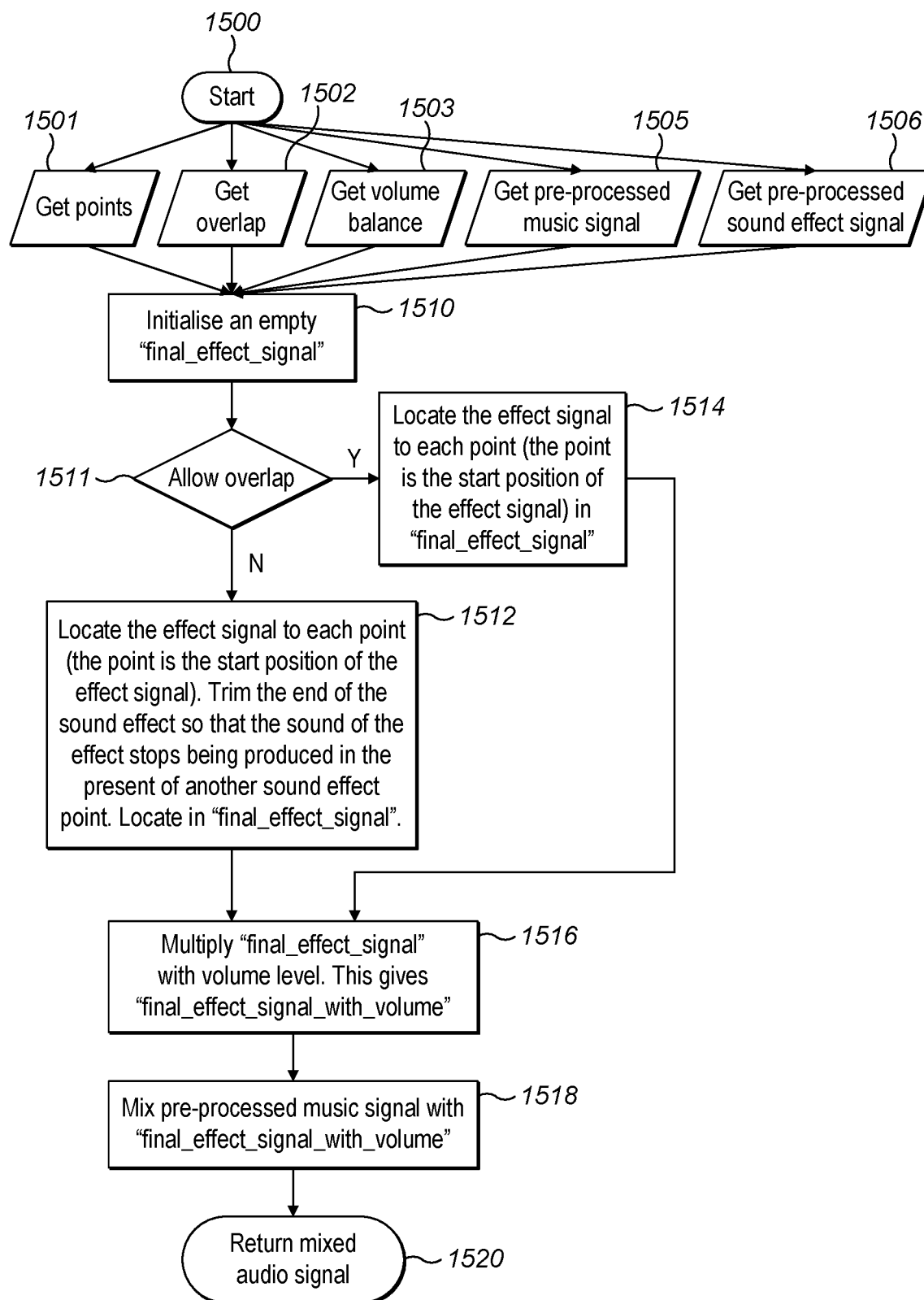


FIG. 15



EUROPEAN SEARCH REPORT

Application Number
EP 21 18 0444

5

10

15

20

25

30

35

40

45

50

55

2

EPO FORM 1503 03.82 (P04C01)

DOCUMENTS CONSIDERED TO BE RELEVANT			
Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
X	SUBRAMANI KRISHNA ET AL: "Energy-Weighted Multi-Band Novelty Functions for Onset Detection in Piano Music", 2018 TWENTY FOURTH NATIONAL CONFERENCE ON COMMUNICATIONS (NCC), IEEE, 25 February 2018 (2018-02-25), pages 1-6, XP033488032, DOI: 10.1109/NCC.2018.8599955 [retrieved on 2019-01-02]	1,2,6, 8-10,12	INV. G10H1/00 ADD. G10H1/40 A63F13/00
Y	* page 3, left-hand column, paragraph IV *	3,7,	
A	* page 4, right-hand column, paragraph IV-D "GROUPING" *	13-15 4,5,11	
Y	MICKAEL ZEHREN ET AL: "Automatic Detection of Cue Points for DJ Mixing", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 16 July 2020 (2020-07-16), XP081721914, * paragraph [04.2]; figure 2a *	3,7	
Y	US 2014/128160 A1 (GILLET OLIVIER [FR] ET AL) 8 May 2014 (2014-05-08) * paragraph [0036] - paragraph [0047]; figures 2,3 * * paragraph [0041] *	13-15	TECHNICAL FIELDS SEARCHED (IPC) G06N G10H G11B A63F
The present search report has been drawn up for all claims			
Place of search Munich		Date of completion of the search 2 December 2021	Examiner Glasser, Jean-Marc
CATEGORY OF CITED DOCUMENTS X : particularly relevant if taken alone Y : particularly relevant if combined with another document of the same category A : technological background O : non-written disclosure P : intermediate document		T : theory or principle underlying the invention E : earlier patent document, but published on, or after the filing date D : document cited in the application L : document cited for other reasons & : member of the same patent family, corresponding document	

02-12-2021

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2014128160 A1	08-05-2014	FR 2974226 A1	19-10-2012
		US 2014128160 A1	08-05-2014
		WO 2012140468 A1	18-10-2012

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82