A Simple User Interface System for Recovering Patterns Repeating in Time and Frequency in Mixtures of Sounds

Zafar Rafii[1] Antoine Liutkus[2] Bryan Pardo[3]

[1]Gracenote, Media Technology Lab, Emeryville, CA 94608 USA zrafii@gracenote.com
[2]Inria, Villiers-Lès-Nancy, 54600 France antoine.liutkus@inria.fr
[3]Northwestern University, EECS department, Evanston, IL 60208 USA pardo@northwestern.com

Introduction

– Audio editors let users manipulate recordings but still lack tools to allow for the separation of sounds.
– Proposed user interface systems for audio source separation depend on heavy manual annotations.
+ Researchers have demonstrated the importance of repetition for efficient audio source separation.
+ Findings in cognitive psychology also showed that listeners use repetition as a cue to separate sounds.
= We propose to leverage repetition to perform audio source separation for a more intuitive system.

System

Constant Q Transform is first used to transform a recording into a time-frequency representation with a logarithmic frequency resolution.
Normalized 2D Cross-Correlation is then used to identify regions similar to a selected region from which a user wishes to recover a repeating pattern.
Median Filter is finally used to average the similar regions over their repetitions in order to recover the underlying repeating pattern by removing outliers.

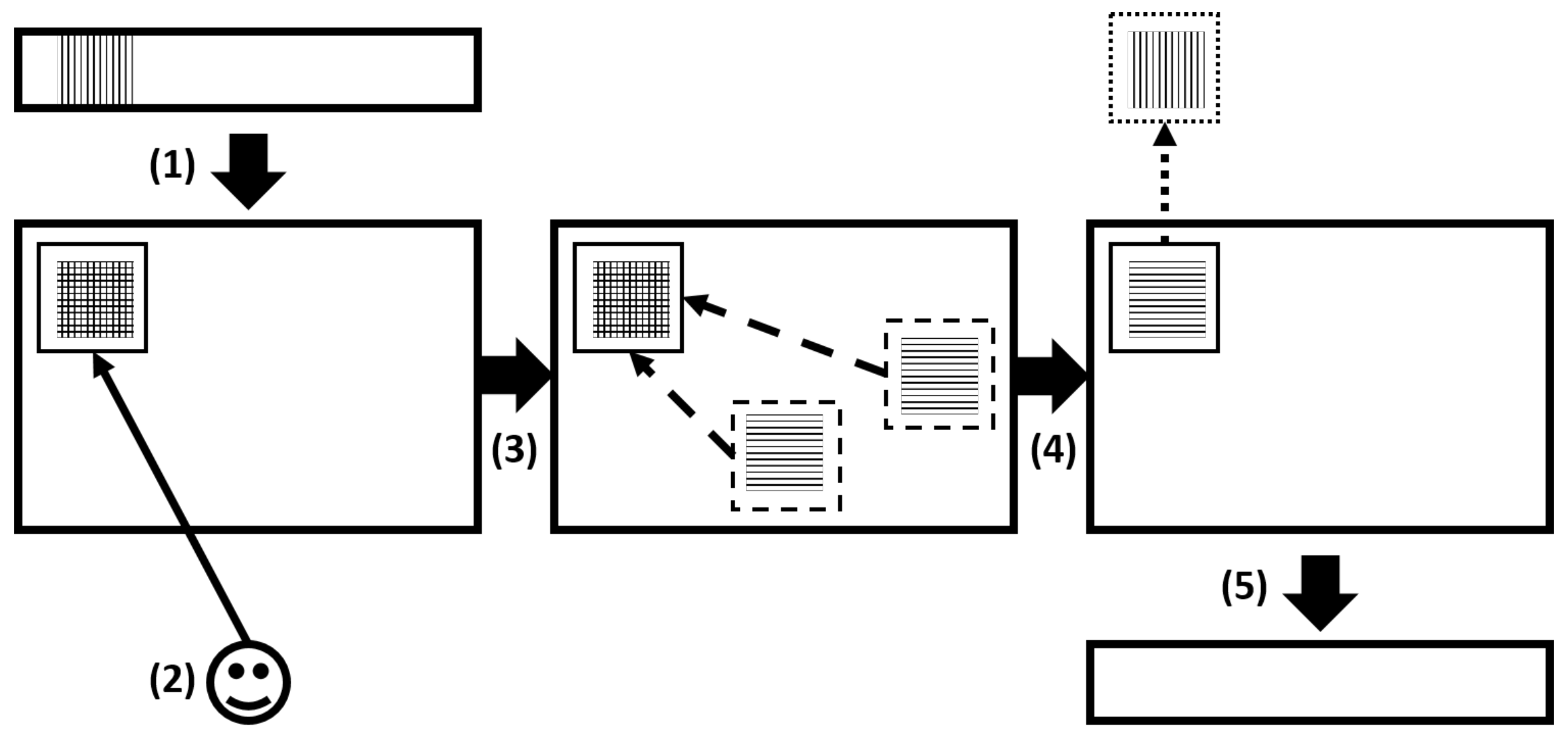

Figure 1: Overview of the system. (1) An audio recording with an undesired element is transformed into a log-frequency spectrogram. (2) The user selects the region of the undesired element in the spectrogram. (3) The selected region is cross-correlated with the spectrogram to identify similar regions where the underlying pattern repeats. (4) The identified regions are averaged over their repetitions and the repeating pattern is recovered. (5) The filtered spectrogram is inverted back to the time-domain with the undesired element removed.

Applications

Recovering a Melody Masked by a Cough

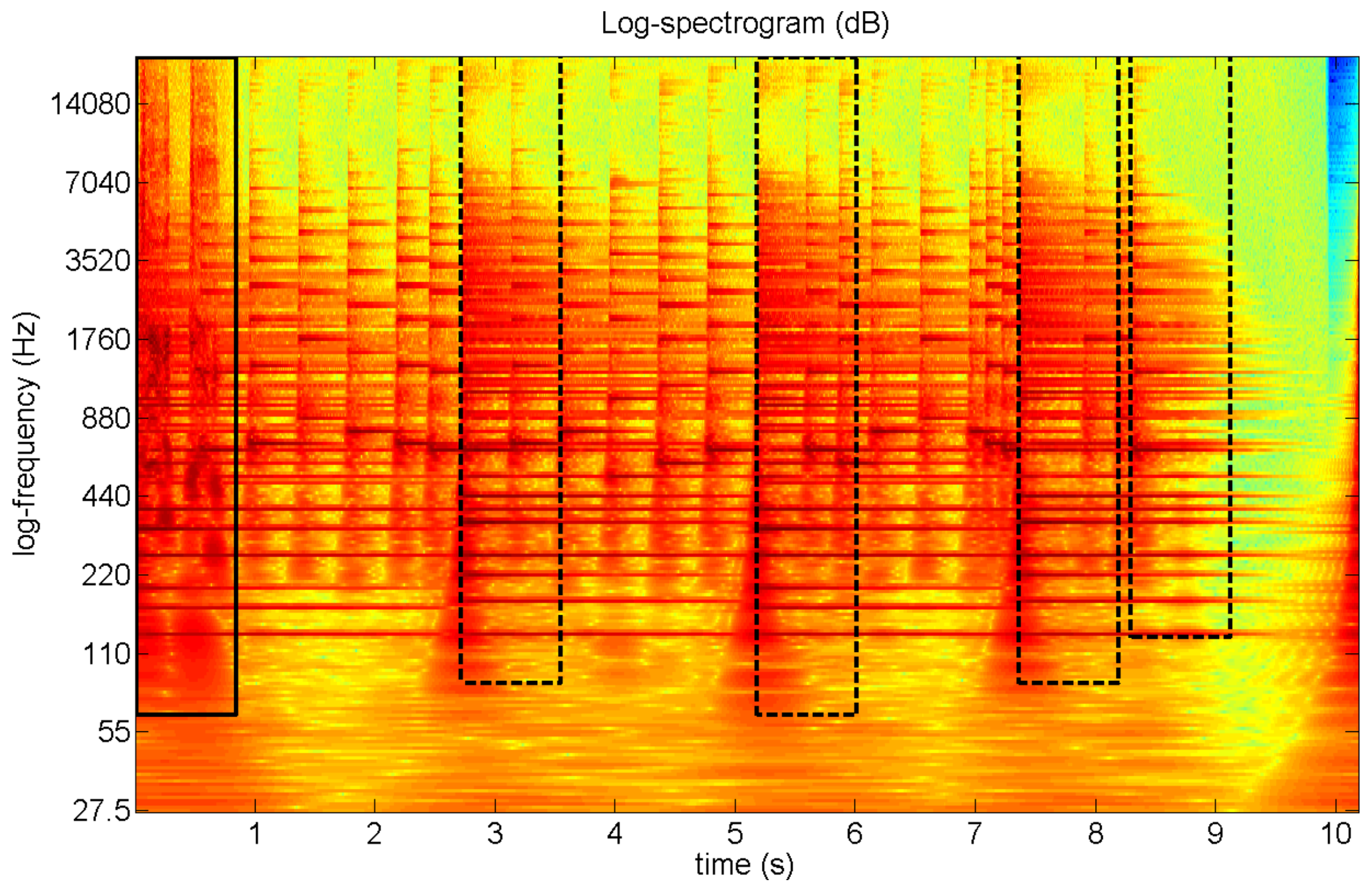

Figure 2: Log-spectrogram of a melody with a cough masking the first note. The user selected the region of the cough (solid line) and the system identified similar regions where the underlying note repeats (dashed lines).

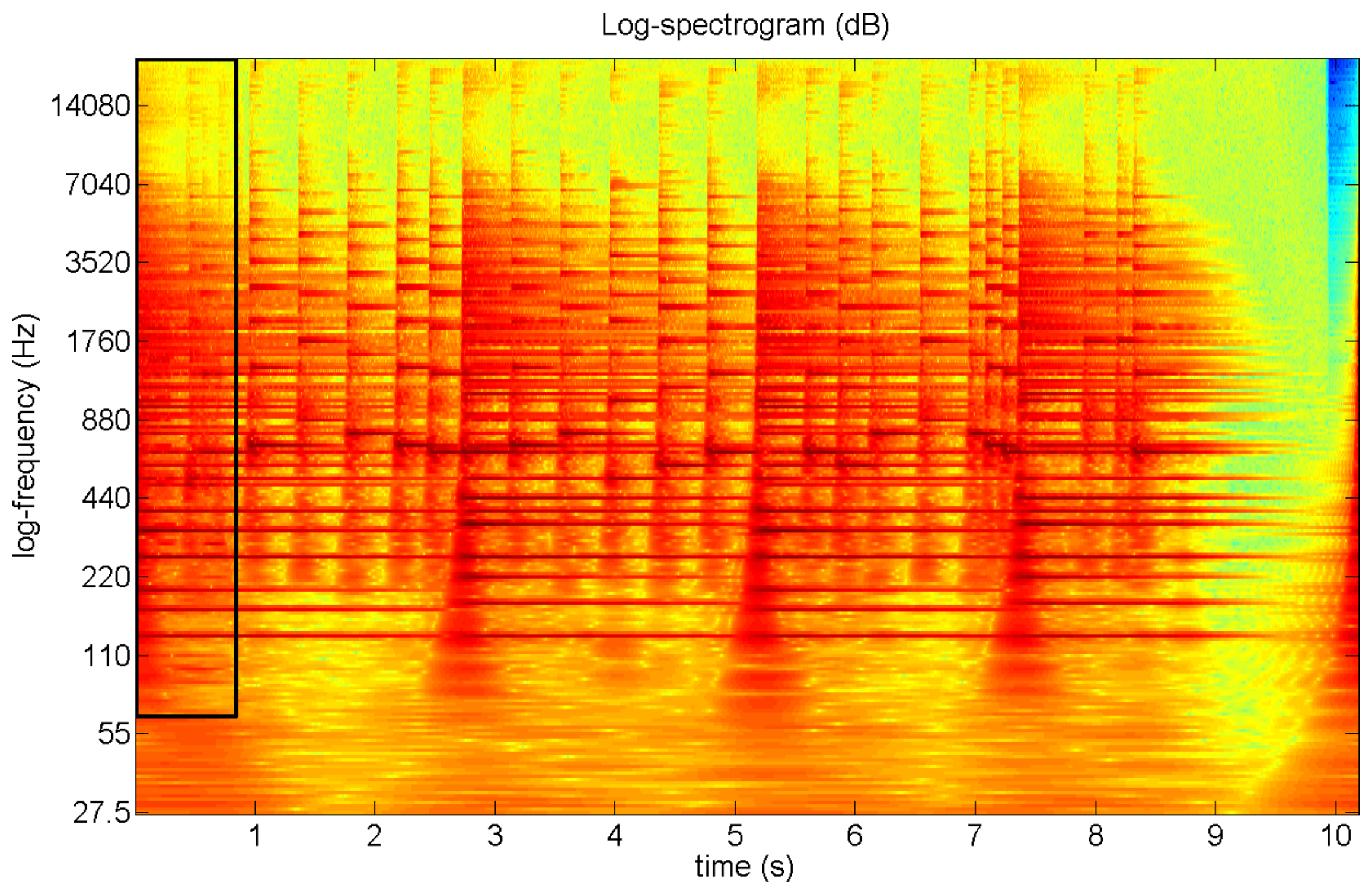

Figure 3: Log-spectrogram of the melody with the first note recovered. The system averaged the identified regions over their repetitions and filtered out the cough from the selected region.

	SDR	SIR	SAR
recovered note	8.70	13.44	13.56
extracted cough	5.91	6.55	11.90

Table 1: Separation performance for the recovered note, and the extracted cough (in dB).

In practice, the whole process only takes a fraction of a second, as the system involves efficient algorithms and simple operations.

Recovering an Accompaniment Masked by Vocals

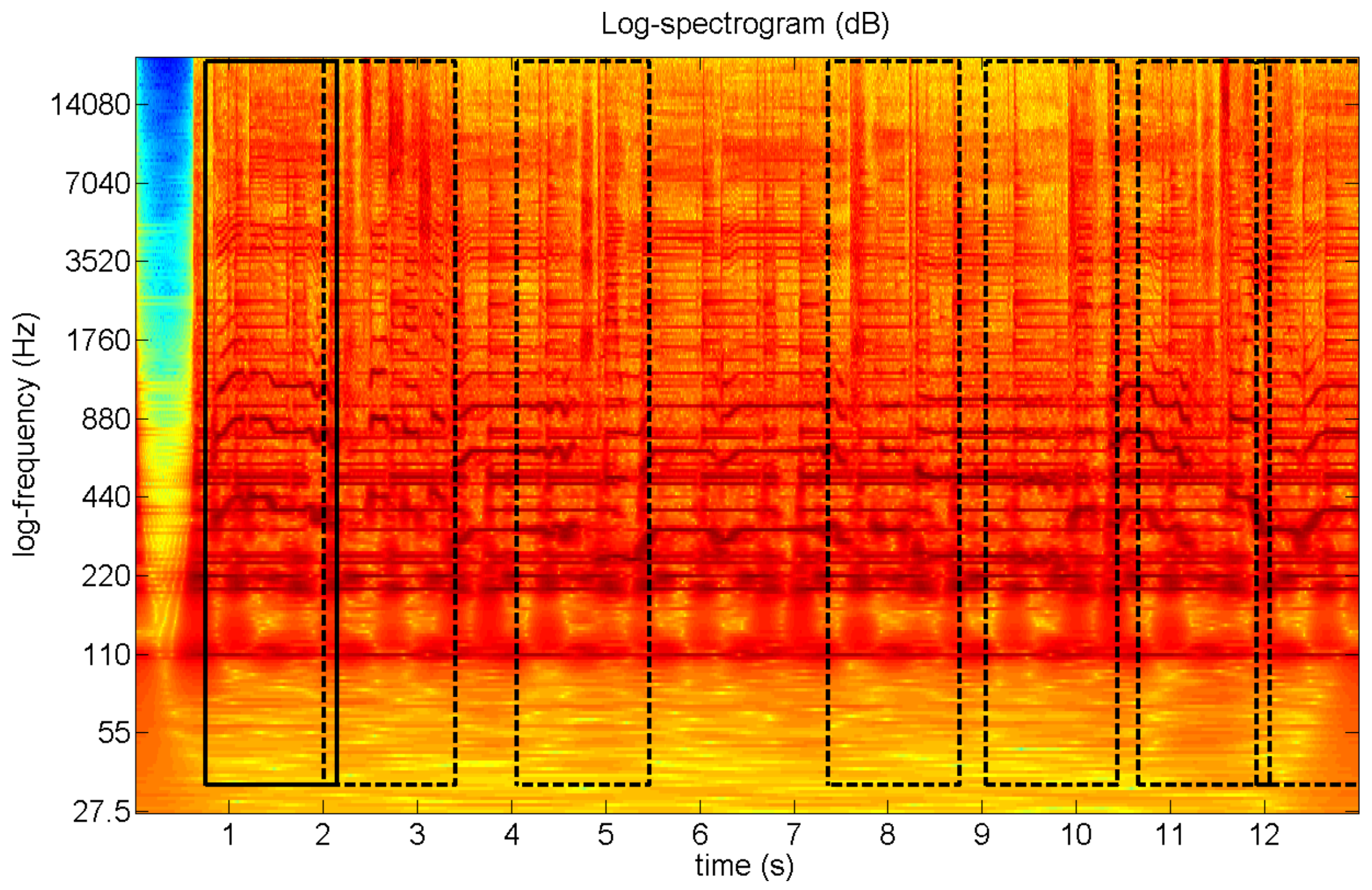

Figure 4: Log-spectrogram of a song with vocals masking an accompaniment. The user selected the region of the first measure (solid line) and the system identified similar regions where the underlying accompaniment repeats (dashed lines).

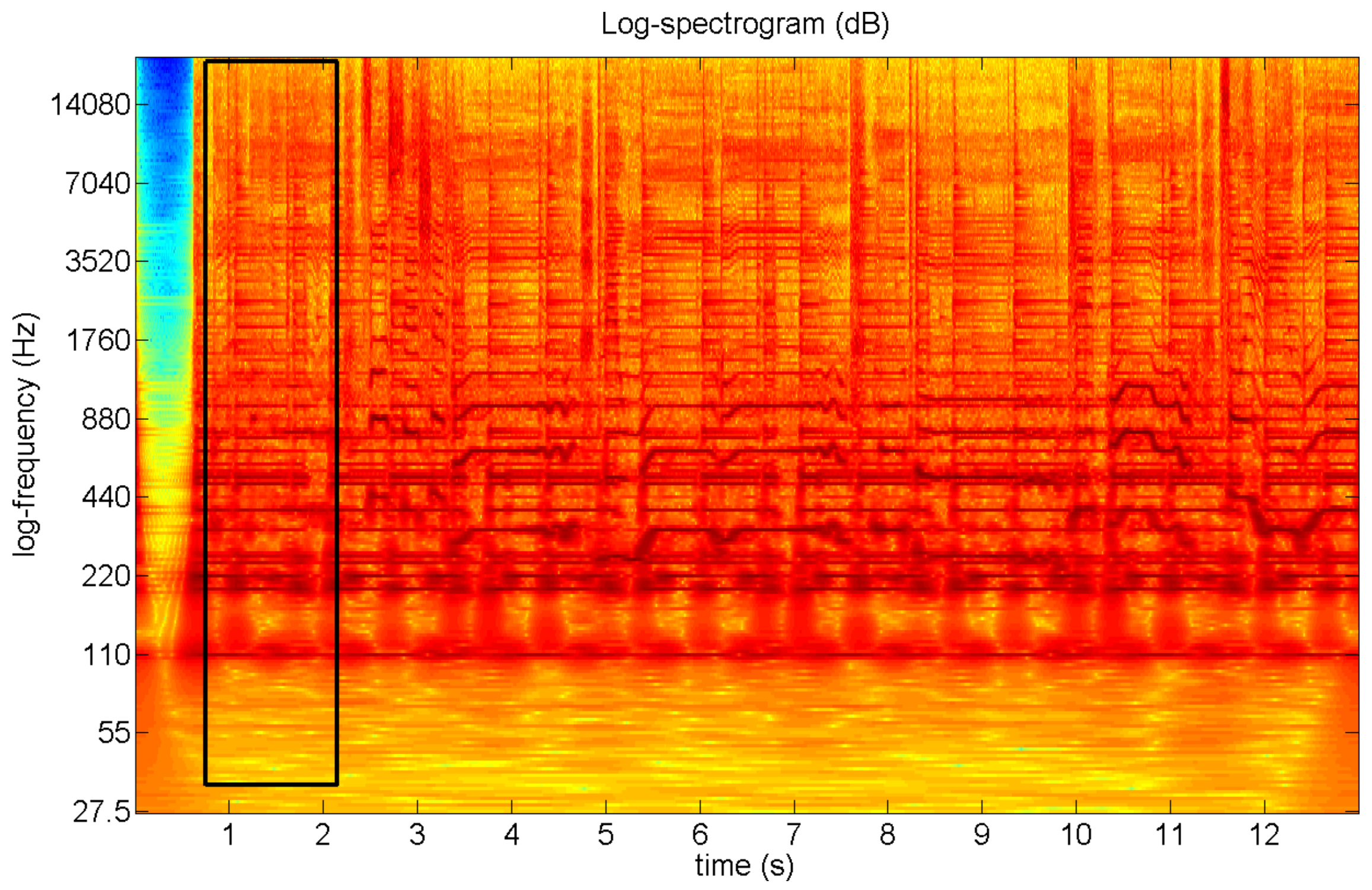

Figure 5: Log-spectrogram of the song with the first measure of the accompaniment recovered. The system averaged the identified regions over their repetitions and filtered out the vocals from the selected region.

	SDR	SIR	SAR
recovered accompaniment	9.01	10.71	14.34
extracted vocals	10.77	24.95	14.32

Table 2: Separation performance for the recovered accompaniment, and the extracted vocals (in dB).

SDR measures the overall separation performance, with SIR measuring the degree of separation between the sources and SAR measuring the quality of the separation of the estimates.

Extracting a Speech Masking a Noise

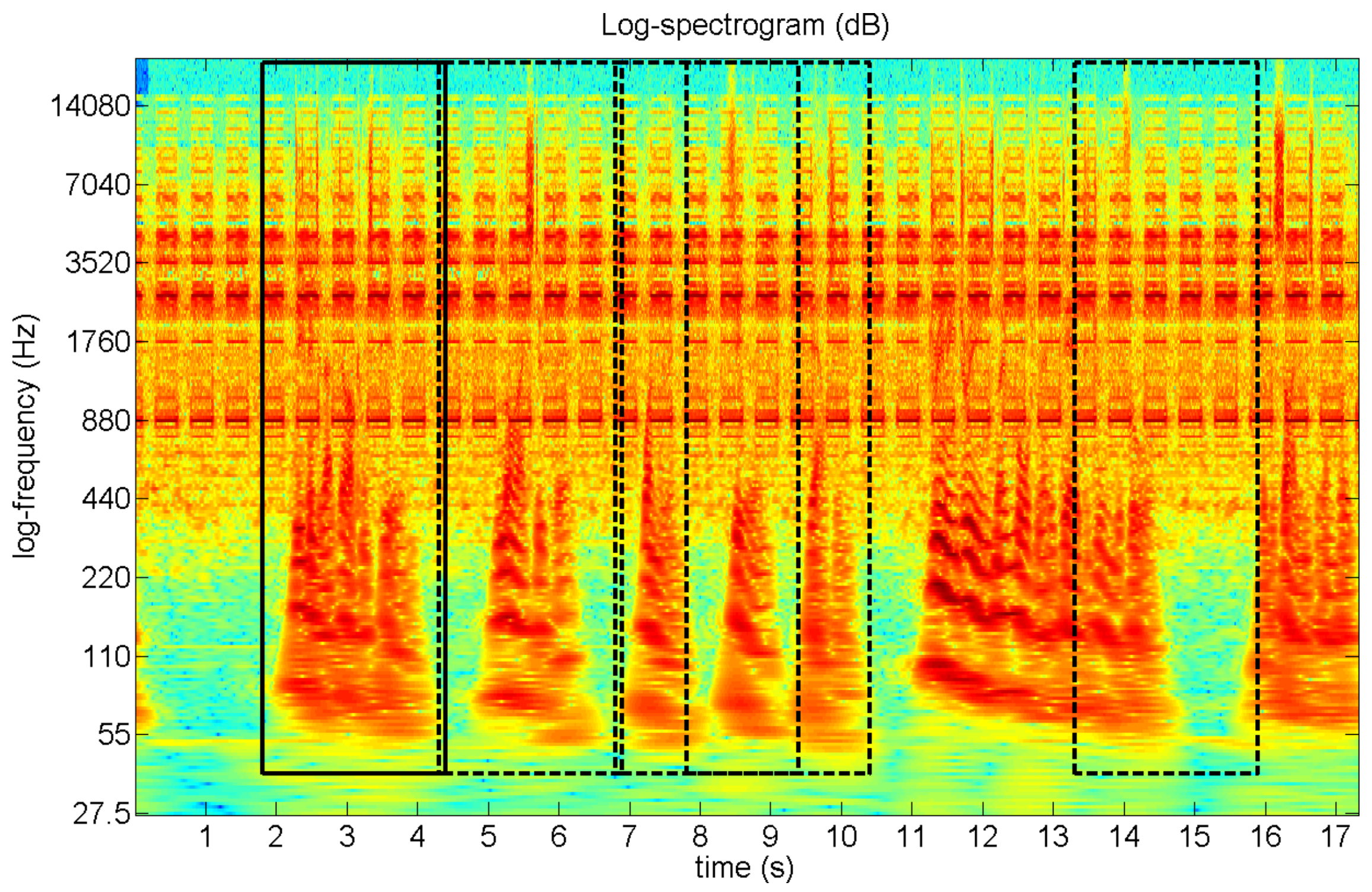

Figure 6: Log-spectrogram of a speech masking a noise. The user selected the region of the first sentence (solid line) and the system identified similar regions where the underlying noise repeats (dashed lines).

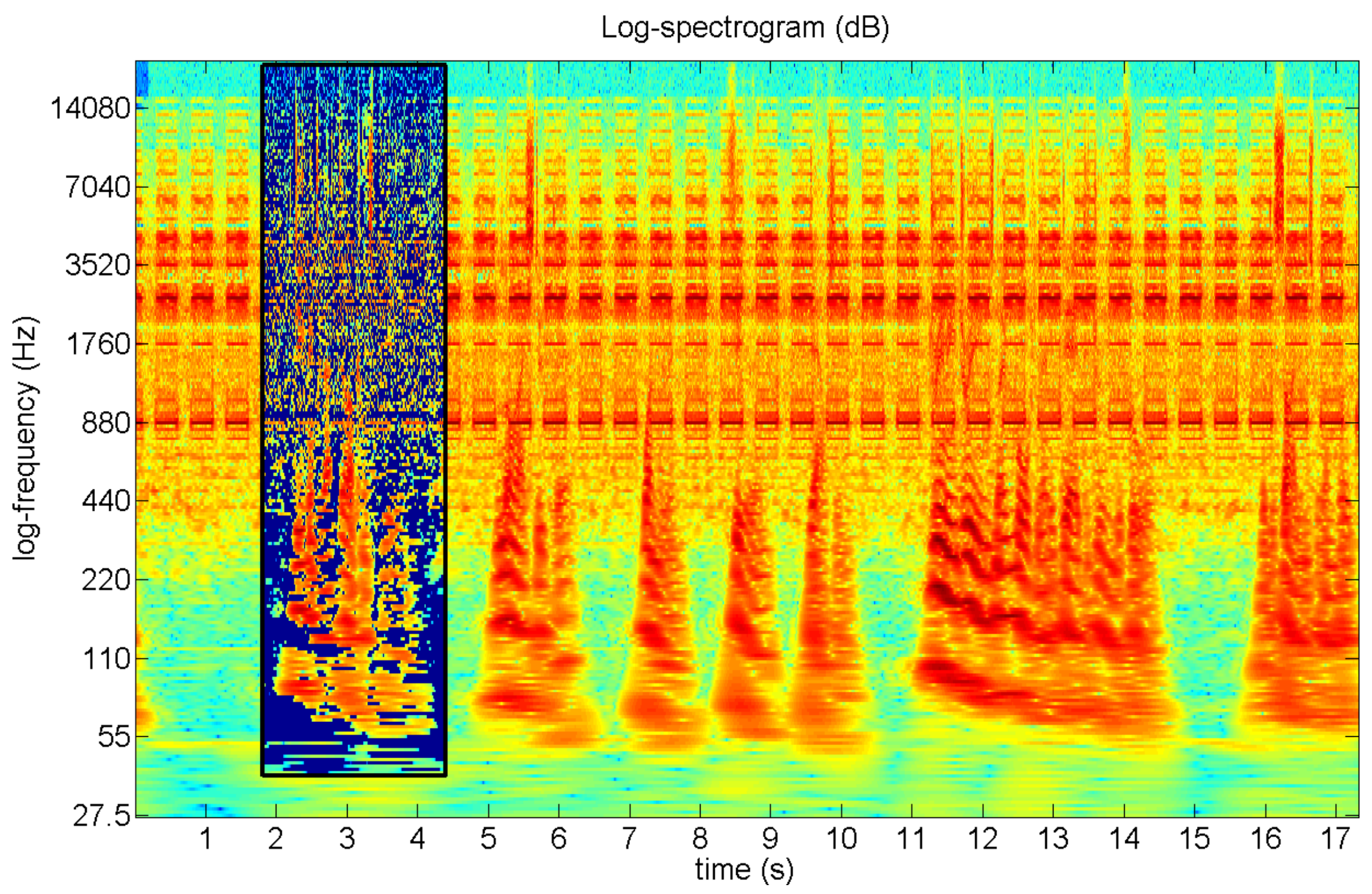

Figure 7: Log-spectrogram of the first sentence of the speech extracted. The system averaged the identified regions over their repetitions and extracted the speech from the selected region.

	SDR	SIR	SAR
extracted speech	6.01	15.64	7.83
filtered noise	9.28	10.31	15.44

Table 3: Separation performance measures for the extracted speech, and the filtered noise (in dB).

Note that, here, we recovered the non-repeating pattern instead of the repeating pattern.
The reader will find the audio examples online (`http://www.zafarrafii.com/repet.html`).
This work was supported in part by National Science Foundation Award 1420971.